Des protéines et de leurs interactions aux principes évolutifs des systèmes biologiques

Anne-Ruxandra Carvunis

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **MODELES, METHODES ET ALGORITHMES EN BIOLOGIE, SANTE ET ENVIRONNEMENT**

Arrêté ministériel : 7 août 2006

Présentée par

Anne-Ruxandra CARVUNIS

Thèse dirigée par **Laurent TRILLING** et
codirigée par **Nicolas THIERRY-MIEG** et **Marc VIDAL**

préparée au sein du **Laboratoire du Professeur Marc Vidal** et du **laboratoire Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications de Grenoble**
dans **l'Ecole Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement**

Des protéines et de leurs interactions aux principes évolutifs des systèmes biologiques

Thèse soutenue publiquement le **26 janvier 2011**
devant le jury composé de :

Mr Jacques DEMONGEOT
Professeur, Grenoble, Président
Mr Vincent LOTTEAU
DR INSERM, Lyon, Rapporteur
Mr David SHERMAN
DR INRIA, Bordeaux, Rapporteur
Mr François KEPES
DR CNRS, Evry, Membre

LABORATOIRES DE THESE

Laboratoire TIMC-IMAG, équipe TIMB
Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications de Grenoble
Équipe Traitement de l'Information et Modélisation en Bio-médecine
Unité Mixte de Recherche CNRS Université Joseph Fourier UMR 5525
Domaine de la Merci
38706 La Tronche Cedex
France

Laboratory of Dr. Marc Vidal, Professor
Department of Cancer Biology, Dana-Farber Cancer Institute
Department of Genetics, Harvard Medical School
44 Binney Street, Smith Building 858
Boston MA 02115
USA

Aux Nicolas qui m'ont aidée et inspirée,
À mes parents.

REMERCIEMENTS

Je remercie vivement Marc Vidal pour la confiance dont il a fait preuve en m'accueillant dans son laboratoire et la qualité de son encadrement. L'enthousiasme scientifique de ce visionnaire est communicatif. Il m'a appris l'indépendance et la rigueur, le travail d'équipe et la nécessité de faire des sacrifices à l'autel de la science. Il m'a poussée à aller bien plus loin que ce dont je me croyais capable. Grâce à Marc, le vif intérêt que je portais à la biologie s'est transformé en passion de la recherche.

Je remercie sincèrement Nicolas Thierry-Mieg et Laurent Trilling pour leur soutien sans faille tout au long de ma thèse et leur aide précieuse à la rédaction de ce manuscrit.

Je remercie du fond du cœur les scientifiques talentueux avec qui travailler fut un pur plaisir : Kavitha Venkatesan, Denis Dupuy, Benoît Charloteaux, Jean-François Rual, Nicolas Bertin, Cesar Hildalgo, Murat Tasan, Ilan Wapinski, Muhammed Yildirim, Michael Cusick, Shahid Mukhtar, Matija Dreze, et tout spécialement mon « jeune mentor » Nicolas Simonis.

Je remercie toute ma famille pour leur soutien et leur compréhension, et spécialement ma petite sœur Solange, qui partage ma conviction que l'art et la science sont deux exercices de curiosité et créativité qui émanent du même amour de la vie.

Merci Amélie, Benoit, Elisa, et les autres copines et copains qui non seulement m'ont beaucoup aidée moralement, mais surtout qui m'aiment assez pour lire cette page !

Thanks to my boyfriend Andy, probably the most patient man on this planet.

RESUME

Darwin a révélé au monde que les espèces vivantes ne cessent jamais d'évoluer, mais les mécanismes moléculaires de cette évolution restent le sujet de recherches intenses. La biologie systémique propose que les relations entre génotype, environnement et phénotype soient sous-tendues par un ensemble de réseaux moléculaires dynamiques au sein de la cellule, mais l'organisation de ces réseaux demeure mystérieuse. En combinant des concepts établis en biologie évolutive et systémique avec la cartographie d'interactions protéiques et l'étude des méthodologies d'annotation de génomes, j'ai développé de nouvelles approches bioinformatiques qui ont en partie dévoilé la composition et l'organisation des systèmes cellulaires de trois organismes eucaryotes : la levure de boulanger, le nématode *Caenorhabditis elegans* et la plante *Arabidopsis thaliana*. L'analyse de ces systèmes m'a conduit à proposer des hypothèses sur les principes évolutifs des systèmes biologiques. En premier lieu, je propose une théorie selon laquelle la traduction fortuite de régions intergéniques produirait des peptides sur lesquels la sélection naturelle agirait pour aboutir occasionnellement à la création de protéines *de novo*. De plus, je montre que l'évolution de protéines apparues par duplication de gènes est corrélée avec celle de leurs profils d'interactions. Enfin, j'ai mis en évidence des signatures de la co-évolution ancestrale hôte-pathogène dans l'organisation topologique du réseau d'interactions entre protéines de l'hôte. Mes travaux confortent l'hypothèse que les systèmes moléculaires évoluent, eux aussi, de manière darwinienne.

Mots clés : réseaux d'interactions entre protéines, interactions hôte-pathogène, annotation de génomes, duplications de gènes.

FROM PROTEINS AND THEIR INTERACTIONS TO EVOLUTIONARY PRINCIPLES OF BIOLOGICAL SYSTEMS

ABSTRACT

Darwin exposed to the world that living species continuously evolve. Yet the molecular mechanisms of evolution remain under intense research. Systems biology proposes that dynamic molecular networks underlie relationships between genotype, environment and phenotype, but the organization of these networks is mysterious. Combining established concepts from evolutionary and systems biology with protein interaction mapping and the study of genome annotation methodologies, I have developed new bioinformatics approaches that partially unveiled the composition and organization of cellular systems for three eukaryotic organisms: the baker's yeast, the nematode *Caenorhabditis elegans* and the plant *Arabidopsis thaliana*. My analyses led to insights into the evolution of biological systems. First, I propose that the translation of peptides from intergenic regions could lead to *de novo* birth of new protein-coding genes. Second, I show that the evolution of proteins originating from gene duplications and of their physical interaction repertoires are tightly interrelated. Lastly, I uncover signatures of the ancestral host-pathogen co-evolution in the topology of a host protein interaction network. My PhD work supports the thesis that molecular systems also evolve in a Darwinian fashion.

Key words: protein interaction networks, host-pathogen interactions, genome annotation, gene duplications.

TABLE DES MATIERES

AVANT-PROPOS

J'étais en deuxième année de DEUG, prête à écouter un cours sur la biochimie de la glycolyse, lorsque le professeur annonça avec émotion à l'amphithéâtre que la séquence du génome humain venait d'être publiée. Naturellement, la portée historique de l'évènement m'échappait. Aujourd'hui, je réalise pleinement la chance que je partage avec les autres jeunes scientifiques qui commencent leur carrière au XXI^{ème} siècle. Pour nous, il est parfaitement acquis que 1) le gène est la base de l'hérédité, 2) la cellule est l'unité fonctionnelle et structurale de la vie, 3) la vie est basée sur la chimie et 4) l'évolution par sélection naturelle est caractéristique de la vie. À ces quatre prismes à travers lesquels nous cherchons à comprendre la biologie, et qui chacun correspond à une discipline spécialisée, Sir Paul Nurse propose d'ajouter un cinquième : la vie est un système organisé (Nurse 2003). Des efforts comme le séquençage du génome humain, ou comme ceux auxquels j'ai participé pendant ma thèse, serviront de pierre de base pour construire les concepts qui permettront un jour de comprendre l'organisation des systèmes vivants et leurs propriétés émergentes. La discipline qui entend relever l'immense défi posé par Sir Paul Nurse se nomme « biologie systémique ». Les pages qui suivent, largement inspirées d'une revue parue dans *Médecine/Sciences* dont je suis premier auteur (**Document Joint 1**) intitulée «Biologie systémique : des concepts d'hier aux découvertes de demain » (Carvunis, Gomez et al. 2009), exposent brièvement l'histoire, les techniques et les concepts clés de la biologie systémique. J'introduis ensuite la problématique abordée plus spécifiquement pendant ma thèse, qui concerne l'organisation et l'évolution des systèmes moléculaires formés par les protéines en interaction. Les contributions méthodologiques et conceptuelles principales que mes travaux apportent à cette question sont alors résumées en trois chapitres, chacun accompagné de documents joints, publiés ou non, contenant de plus amples développements et explications ainsi que des références bibliographiques spécifiques. À la fin de ce manuscrit, je discute les limitations et les implications de mes travaux ainsi que des perspectives pour des recherches futures.

INTRODUCTION

La biologie systémique représente un important changement de paradigme scientifique par rapport à la biologie moléculaire des cinquante dernières années.

La biologie moléculaire traditionnelle, de nature « réductionniste », s'est jusqu'ici concentrée principalement sur la caractérisation des composants individuels de la cellule: gènes, protéines, ou encore ARNs non codant, avec pour but de comprendre la Vie à partir de la caractérisation des macromolécules qui la constituent. Toutefois, protéines et ARNs opèrent en interagissant les uns avec les autres, formant ainsi des systèmes dont la complexité peut difficilement être comprise en considérant les molécules une par une. La biologie systémique, de nature « intégrative et holistique », entend comprendre la Vie à partir de ces systèmes. Elle pose les questions biologiques en mettant l'accent sur le « tout » plutôt que sur les « parties ». Intrinsèquement interdisciplinaire, sa méthodologie originale est définie par un aller-retour dynamique entre expérimentation, travail informatique et théorique, formulation de nouvelles hypothèses scientifiques et développement technologique.

La biologie systémique est l'étude des interactions dynamiques entre composants d'un système biologique.

Le terme « Système » vient du grec *sustêma* qui signifie ensemble. Un système est un ensemble d'entités interagissant ou interdépendantes, abstraites ou concrètes, dont l'union forme un tout. En biologie, une certaine confusion règne autour de cette définition car l'échelle d'étude d'un système peut lui valoir ou non l'« appellation système » selon l'opinion de celui qui le considère. En fait, il n'y a pas de limite théorique à la taille d'un système : libre à chacun de définir les bornes le délimitant de son environnement (Borneman, Chambers et al. 2007). Ainsi un facteur de transcription régulant sa propre expression constitue un système, de la même façon que l'ensemble des molécules d'une cellule, l'ensemble des cellules d'un organisme, ou encore l'ensemble des individus d'une population. Différents systèmes peuvent être étudiés à partir des mêmes entités si l'on considère un type d'interaction plutôt qu'un autre. La population d'une ville peut être vue comme un ensemble d'individus partageant des relations économiques, des engagements matrimoniaux ou encore des maladies contagieuses, soit comme trois systèmes distincts. Un système peut aussi intégrer des composants de différentes natures, tel un écosystème comprenant à la fois des proies, des prédateurs et des ressources

naturelles. Enfin, des composants et des interactions de nature différente et traversant plusieurs échelles peuvent former un système unique (**Figure 1A**).

Biologie systémique : des concepts historiques …

Les origines de la biologie systémique remontent aux années cinquante, lorsque C.H. Waddington établit le concept de « paysage épigénétique » (Waddington 1957). Il imagine les cellules passer d'un état de différentiation à l'autre en suivant un trajet dicté par la forme d'un paysage constitué de monts et de vallées (**Figure 1B**, panel supérieur), paysage lui-même généré par les interactions entre gènes (les « piliers » du panel inférieur de la **Figure 1B**). Cette vision de la cellule comme système évoluant d'état en état s'inspire des travaux de M. Delbrück (Delbrück 1949), F. Jacob et J. Monod (Monod and Jacob 1961) qui, après la seconde guerre mondiale, introduisent la notion de système en biochimie (**Figure 1C**), comme un mécanisme susceptible d'expliquer le mystère de la différentiation : comment des cellules au génome identique peuvent-elles adopter des formes et des propriétés aussi différentes qu'un lymphocyte et un neurone? Ils proposent de voir les enzymes et leurs substrats comme les composants de circuits dynamiques, dont les exemples les plus simples sont les boucles de rétroaction négatives et positives. Dans le cas d'une boucle négative, l'augmentation du niveau d'un élément entraîne la diminution de son taux de production ce qui résulte en une stabilisation de sa production et de son abondance, à la manière d'un thermostat. Une boucle positive en revanche a l'effet inverse, et donne lieu à deux scénarios opposés : si la boucle est enclenchée, l'élément encourage sa propre production, sinon il n'est pas produit. Ce type de circuits moléculaires et d'autres plus complexes ont depuis été beaucoup étudiés théoriquement, notamment par R. Thomas et ses collègues (Thomas 1978). La réalité de ces systèmes a été démontrée expérimentalement dès la fin des années cinquante avec l'exemple de l'opéron lactose inductible de A. Novick et M. Weiner chez la bactérie *E. coli* (Novick and Weiner 1957), puis à de multiples reprises au cours des cinquante dernières années, permettant à la biologie systémique de s'affirmer aujourd'hui comme discipline à part entière.

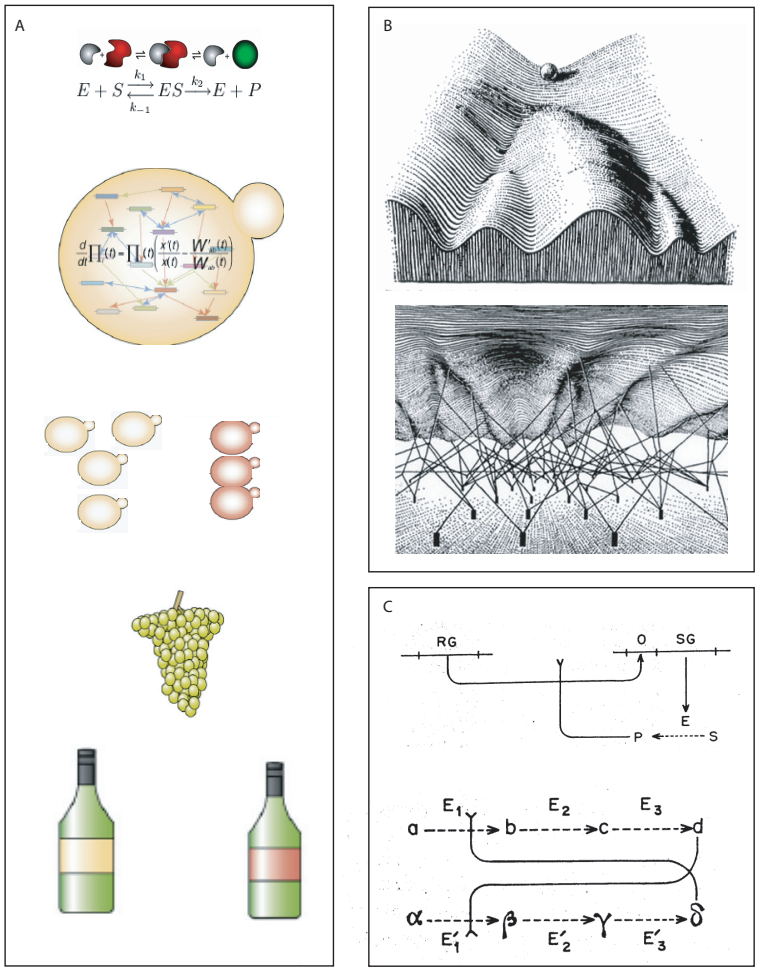


Figure 1 : Illustrations de la notion de système biologique.
A. Schématisation de systèmes biologiques à différentes échelles d'étude. De haut en bas : une enzyme et son substrat, ensemble des voies métaboliques dans une levure, population constituée de plusieurs levures individuelles partageant les mêmes ressources, la grappe de raisin comme écosystème, le vin comme produit d'intégration. Si chacun de ces systèmes peut être étudié séparément, ils peuvent aussi être intégrés dans un unique système mixte. Par exemple, l'institut de recherche pour le vin australien propose d'étudier en détail comment les systèmes métaboliques de différentes levures (ici représentées par une population jaune et une population rouge), en fermentant le même raisin, produisent des produits de dégradation caractéristiques (Borneman, Chambers et al. 2007). Puisque ce sont ces produits qui sont responsables de la variété aromatique des vins (représentée ici par les étiquettes jaunes et rouges), l'institut suggère que ce savoir moléculaire pourra être intégré à la géographie des vignobles et au système économique dynamique de l'offre et de la demande, afin de produire pour les marchés de consommateurs ciblés les vins dont le goût leur plaira.
B. Illustration de la notion de « paysage épigénétique » de C.H. Waddington, d'après (Waddington 1957)
C. Deux systèmes biologiques imaginés par Jacob et Monod, d'après (Monod and Jacob 1961). RG : « gène régulateur » (répresseur de transcription) ; SG : « gène structural » (codant pour une enzyme) ; E(')(1,2,3) : enzymes ; S : substrat ; P : produit ; a, b, c, d, α, β, γ, δ, métabolites. Dans l'exemple du haut, le produit d'une réaction enzymatique inhibe (>-) la répression de l'expression du gène codant pour l'enzyme. Il s'agit d'une version de la boucle de rétroaction positive. Dans l'exemple du bas, le produit final d'une voie métabolique inhibe la première réaction d'une autre voie métabolique, et réciproquement. Il en résulte un système bistable où une seule de ces deux voies métaboliques peut être active à la fois.

… remis à jour grâce au développement de technologies expérimentales et de nouveaux outils d'analyse

Biologie de synthèse : reconstruire un système pour mieux le comprendre

À la frontière de la science-fiction, cette discipline reconstitue des circuits moléculaires *in vivo*, ou bien en invente de toutes pièces en fusionnant des domaines d'ADN provenant de multiples espèces. Parmi ses travaux fondateurs, l'« interrupteur » (Hasty, McMillen et al. 2002) est un système artificiel bistable composé de deux promoteurs et deux répresseurs de transcription croisés chez *E. coli*. Le résultat est une bactérie qui produit une protéine recombinante si et seulement si elle s'est trouvée en contact avec l'inducteur de l'un des promoteurs dans le passé, sans que son patrimoine génétique n'ait été modifié d'aucune manière. En d'autres termes, la bactérie se « différencie » à travers un processus de « mémorisation » d'un changement purement environnemental. La preuve qu'une boucle de rétroaction positive peut être à l'origine de la différentiation cellulaire a donc été faite.

Génomique fonctionnelle : expérimentation à l'échelle de la cellule entière

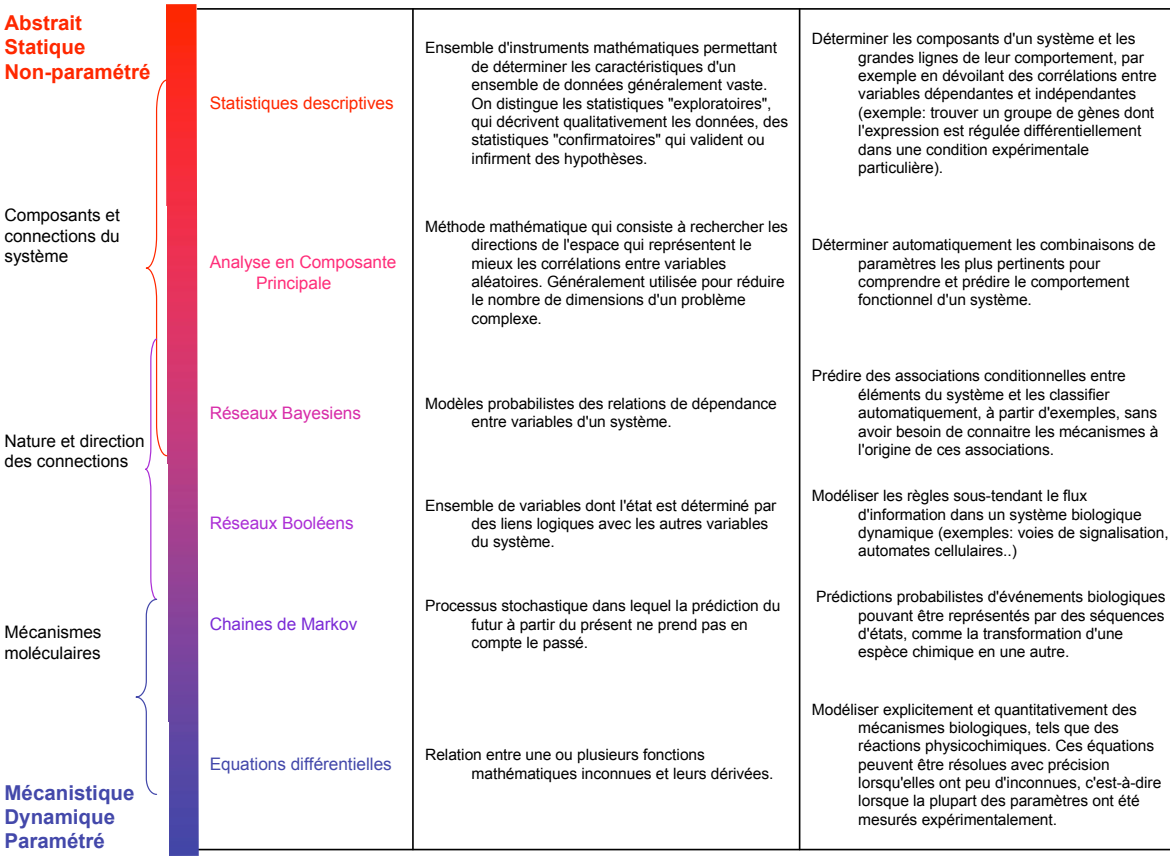
Aussi fondamentaux qu'ils soient, les travaux de la biologie de synthèse restent encore loin d'atteindre la complexité du système « cellule vivante ». C'est à cette échelle que les défis de la biologie systémique se posent aujourd'hui. Comment ces petits circuits moléculaires sont-ils liés les uns aux autres ? Comment communiquent-ils pour répondre aux besoins du « tout » cellulaire ? Pour parvenir à une compréhension globale de la cellule, il faut pouvoir observer simultanément *tous* les gènes, *toutes* les protéines, *tous* les ARNs, ainsi que *toutes* leurs interactions. À cette fin se développent de nombreuses techniques systématiques à haut débit, souvent par miniaturisation et robotisation de techniques préexistantes (**Tableau I**). L'ensemble de ces efforts constitue la génomique fonctionnelle, parfois confondue avec la biologie systémique tant elle en est une sous-discipline essentielle. Comme exemples emblématiques, on peut citer la puce à ADN (une seule puce mesure quantitativement l'expression des 6000 gènes de la levure à la fois), la cartographie du réseau des interactions moléculaires à l'échelle du protéome, ou encore l'établissement de réseaux génétiques. On comprendra aisément que l'analyse et l'intégration (Ge, Walhout et al. 2003) de telles quantités d'information s'accompagnent nécessairement de l'introduction de nouvelles méthodes mathématiques et informatiques.

Exemples de techniques identifiant les interactions entre composants cellulaires	
ADN/ADN	5C (capture de la conformation des chromosomes-copie carbone)
protéine/ADN	immuno-précipitation de chromatine suivie d'analyse de puce à ADN
protéine/protéine	système double-hybride, co-immuno-précipitation suivie de spectrométrie de masse
Exemples de techniques mesurant les états induits par ces interactions	
variations de l'expression des ARN messagers	puces à ADN, SAGE (analyse sérielle de l'expression des gènes)
abondance, localisation et modifications post-traductionnelles des protéines	spectrométrie de masse, marquage suivi de microscopie en fluorescence, puces à protéines

Tableau I: Exemples de techniques expérimentales utilisées à haut débit (liste non-exhaustive). D'après (Ideker and Lauffenburger 2003).

Gestion des données et modélisation : transfert de savoir des « sciences dures » et apparition de nouvelles problématiques

Prenons l'exemple d'une simple boucle de rétroaction positive où l'élément encourage sa propre production. À quelle vitesse ? Quelle quantité d'inducteur permet-elle de générer une quantité souhaitée de produit ? En fait, indépendamment de l'échelle d'étude, la compréhension quantitative de tout système a besoin d'analyse numérique. Répondre aux problèmes spécifiques à la biologie s'est révélé un défi fascinant pour les sciences exactes qui ont non seulement adapté leurs concepts, mais ont aussi trouvé dans les systèmes biologiques une source de nouvelles recherches théoriques. Réciproquement, l'apport de nouvelles méthodes d'analyse a permis aux biologistes de s'ouvrir à de nouvelles problématiques. Ainsi le contrôle qualité, l'estimation de paramètres à partir d'un nombre limité de mesures ou le contrôle de la stochasticité, problèmes classiques pour les ingénieurs, sont maintenant maniés par les biologistes pour parvenir à des analyses plus précises de leurs données. Au cœur de la biologie systémique se trouvent l'identification et la modélisation des réseaux à travers lesquels gènes et protéines interagissent pour effectuer les opérations cellulaires. Les outils mathématiques utilisés pour aborder ces questions correspondent à différents niveaux d'abstraction (**Figure 2**). De façon générale, les analyses à haut niveau d'abstraction décrivent les propriétés qualitatives des systèmes, tandis que les modèles à bas niveau d'abstraction produisent des prédictions quantitatives (Ideker and Lauffenburger 2003). En effet, ces travaux dépendent étroitement de l'acquisition systématique de données biologiques chiffrées.



Méthodologies	Définition et enjeux généraux	Exemples d'applications en biologie des systèmes
Statistiques descriptives	Ensemble d'instruments mathématiques permettant de déterminer les caractéristiques d'un ensemble de données généralement vaste. On distingue les statistiques "exploratoires", qui décrivent qualitativement les données, des statistiques "confirmatoires" qui valident ou infirment des hypothèses.	Déterminer les composants d'un système et les grandes lignes de leur comportement, par exemple en dévoilant des corrélations entre variables dépendantes et indépendantes (exemple: trouver un groupe de gènes dont l'expression est régulée différentiellement dans une condition expérimentale particulière).
Analyse en Composante Principale	Méthode mathématique qui consiste à rechercher les directions de l'espace qui représentent le mieux les corrélations entre variables aléatoires. Généralement utilisée pour réduire le nombre de dimensions d'un problème complexe.	Déterminer automatiquement les combinaisons de paramètres les plus pertinents pour comprendre et prédire le comportement fonctionnel d'un système.
Réseaux Bayesiens	Modèles probabilistes des relations de dépendance entre variables d'un système.	Prédire des associations conditionnelles entre éléments du système et les classifier automatiquement, à partir d'exemples, sans avoir besoin de connaitre les mécanismes à l'origine de ces associations.
Réseaux Booléens	Ensemble de variables dont l'état est déterminé par des liens logiques avec les autres variables du système.	Modéliser les règles sous-tendant le flux d'information dans un système biologique dynamique (exemples: voies de signalisation, automates cellulaires..)
Chaines de Markov	Processus stochastique dans lequel la prédiction du futur à partir du présent ne prend pas en compte le passé.	Prédictions probabilistes d'événements biologiques pouvant être représentés par des séquences d'états, comme la transformation d'une espèce chimique en une autre.
Equations différentielles	Relation entre une ou plusieurs fonctions mathématiques inconnues et leurs dérivées.	Modéliser explicitement et quantitativement des mécanismes biologiques, tels que des réactions physicochimiques. Ces équations peuvent être résolues avec précision lorsqu'elles ont peu d'inconnues, c'est-à-dire lorsque la plupart des paramètres ont été mesurés expérimentalement.

Figure 2 : Exemples de méthodologies mathématiques et computationnelles pour l'analyse des systèmes biologiques, des plus abstraites aux plus mécanistiques (liste non-exhaustive). D'après (Ideker and Lauffenburger 2003).

La science des réseaux, une branche prometteuse de la biologie systémique.

La recherche sur l'organisation des systèmes vivants est fortement motivée par l'espoir de parvenir à maîtriser leur perturbation en cas de maladie. Certains groupes de recherche dont le nôtre cartographient les interactions entre protéines de pathogènes et protéines hôtes (Dyer, Neff et al. ; Uetz, Dong et al. 2006; Calderwood, Venkatesan et al. 2007; de Chassey, Navratil et al. 2008), et nous avons entrepris de mesurer quelles interactions protéiques sont perdues ou retenues par les protéines lorsqu'on leur impose des mutations connues pour être associées à des maladies génétiques (Zhong, Simonis et al. 2009). Ce type d'approche s'inscrit dans le contexte d'une nouvelle discipline, émergeant des sciences théoriques traditionnelles : la science des réseaux. Sa plus grande découverte est que les propriétés structurelles des réseaux informent sur la nature des tâches qu'ils peuvent accomplir. Ainsi, la connectivité des réseaux réels (internet, chaînes alimentaires, transports publics, interactions protéiques…) les distinguent des modèles classiques de réseaux aléatoires (Barabasi and Albert 1999). Par exemple, la distribution des degrés (nombre de connections par composant) dans les réseaux réels tend à suivre une loi de puissance, tandis que celle de réseaux aléatoires suit généralement une loi en forme de cloche. Cela signifie que dans les réseaux réels un petit nombre de composants établissent de nombreuses connections (on les appelle les *hubs*) tandis que la plupart des composants font peu de connections. Cette propriété rend les réseaux réels très robustes en cas d'« erreur » ou mal-fonctionnement de leurs composants, mais en revanche très sensibles aux « attaques » dirigées contre les *hubs (Albert, Jeong et al. 2000)*. En biologie (**Figure 3**), cette observation soulève des questions fondamentales : l'évolution favorise-t-elle ces structures pour protéger les réseaux cellulaires contre les erreurs induites par les mutations ? Les pathogènes attaquent-ils préférentiellement les *hubs* ?

Parmi les réseaux réels, des différences topologiques révèlent des classes universelles. Ainsi, les motifs de trois ou quatre composants surreprésentés dans des systèmes de transfert d'énergie (chaînes alimentaires) ne sont pas les mêmes que ceux surreprésentés dans des systèmes de traitement d'information (réseaux de transcription, réseaux neuronaux, puces électroniques) (Milo, Shen-Orr et al. 2002). Placer dans un même cadre théorique des systèmes de nature différente promet donc de révéler encore bien des lois propres à tous les systèmes complexes. Quelles seront les retombées de la biologie des réseaux sur la santé publique ? Nombreux axes de recherche comme ceux que nous venons de citer sont déjà en cours d'étude, et il est clair que l'avenir nous réserve des découvertes d'amplitude.

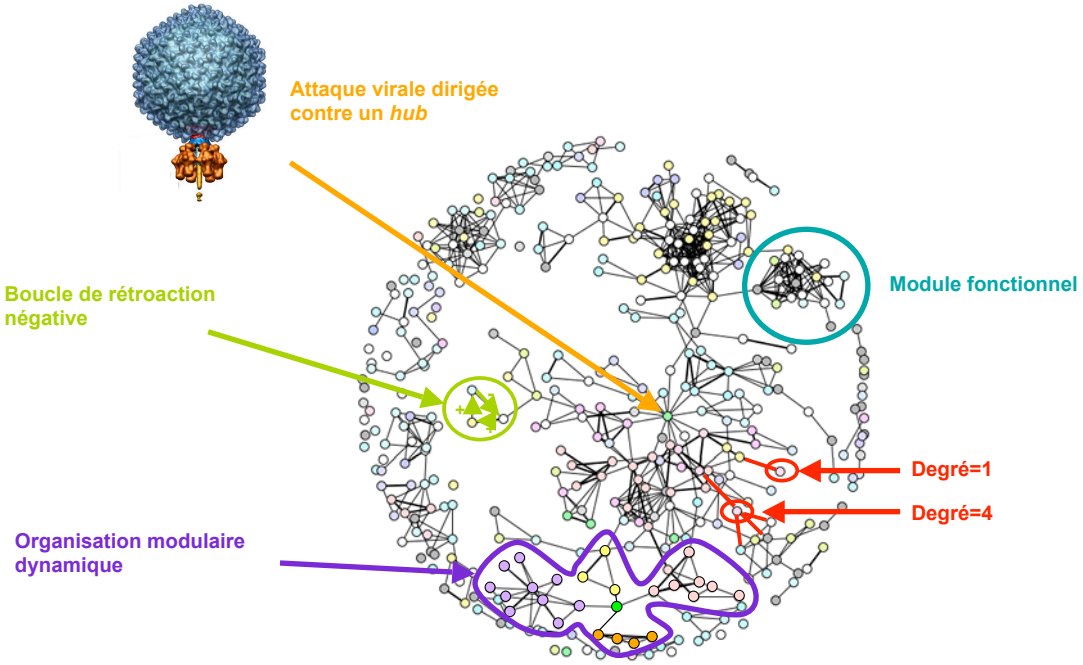


Figure 3 : Représentation théorique d'un réseau d'interactions entre protéines. Dans ce réseau, chaque nœud représente une protéine et chaque lien entre deux nœuds une interaction biophysique ou biochimique. Les différentes couleurs attribuées aux nœuds représentent des classes fonctionnelles (exemples : enzyme, facteur de transcription, localisation membranaire, co-expression…). Les différentes épaisseurs des liens illustrent que les interactions peuvent être plus ou moins stables, ou bien démontrées avec plus ou moins de certitude. La notion de degré est exemplifiée en rouge : une protéine n'interagissant qu'avec un seul partenaire est de degré 1, alors qu'une autre avec quatre partenaires est de degré 4. Lorsqu'un groupe de protéines partagent plus d'interactions entre elles que la moyenne, on peut supposer qu'elles forment un module fonctionnel (Gunsalus, Ge et al. 2005) (en bleu). Si ce module contient des protéines de fonction inconnue, on peut prédire qu'elles partagent certaines des fonctions des autres membres. Bien que la plus grande partie de ce que nous connaissons sur le réseau d'interactions physiques entre protéines soit statique, ce savoir peut servir de plate-forme sur laquelle surimposer des informations dynamiques, comme la co-expression. Au sein de la forme violette, les nœuds de la même couleur sont des protéines dont l'expression est co-régulée. Les nœuds roses, jaunes, violets et oranges sont regroupés par couleur, tandis que le nœud vert lie ces différents groupes entre eux. Cela suggère que le nœud vert a probablement une fonction d'intégration et de coordination dynamique des modules unicolores (Ge, Walhout et al. 2003). De la même façon, des données cinétiques peuvent se superposer aux interactions physiques statiques et placer dans un contexte plus général de petits systèmes dynamiques (en vert). La flèche orange illustre que les cibles préférées des virus seraient les protéines les plus connectées, ou *hubs*.

Mesurer expérimentalement les réseaux d'interactions protéiques

Dans ce cadre scientifique, il apparaît indispensable de mesurer expérimentalement l'ensemble des interactions entre bio-molécules, qui constitue le réseau « interactome » de chaque cellule. Cette tâche est rendue difficile par la taille et la complexité de ce réseau. En effet, la cellule contient de multiples catégories de bio-molécules (protéines, ADN, ARN, lipides, sucres,…) formant des interactions de nature variée. Par exemple, une interaction protéine-ADN peut déclencher la transcription de gènes cibles s'il s'agit d'un facteur de transcription se liant à un promoteur, structurer la chromatine dans le cas des histones, ou encore participer à la réplication comme dans le cas des polymérases. De la même façon, les interactions physiques entre protéines peuvent être de force et de dynamique très différentes. Lorsque des protéines interagissent de façon solide et permanente, elles forment de véritables machines moléculaires, appelées complexes, comme l'ATP-synthase. D'autres interactions fortes mais non permanentes peuvent être contrôlées par des modifications chimiques, des changements de localisation cellulaire ou de conformation, comme l'association de la sous-unité alpha des protéines G avec les sous-unités bêta et gamma, qui dépend de l'hydrolyse du GTP. Enfin, de nombreuses interactions entre protéines sont faibles et temporaires, par exemple celles entre récepteurs membranaires et protéines de la matrice extracellulaire qui assistent la mobilité cellulaire (Levy and Pereira-Leal 2008). On considère aussi que certaines interactions entre protéines sont directionnelles, par exemple dans la signalisation cellulaire lorsqu'une kinase phosphoryle son substrat. La compréhension et la modélisation des systèmes cellulaires requièrent que tous ces types d'interactions soient pris en compte, avec leur force, leur direction, et leurs conditions d'application. Cependant, il faut aujourd'hui choisir entre détail et quantité, car les biotechnologies ne permettent pas encore de mesurer ces paramètres à grande échelle.

Dans le cas des interactions entre protéines, deux solutions expérimentales partielles et complémentaires, sont utilisées pour résoudre ce problème d'échelle (Yu, Braun et al. 2008). L'une consiste à purifier *in vivo* des complexes protéiques entiers à partir de protéines « appâts », ce qu'il est aujourd'hui possible de réaliser à haut débit grâce aux avancées en spectrométrie de masse (Nilsson, Mann et al. ; Rigaut, Shevchenko et al. 1999). L'autre consiste à interroger systématiquement toutes les paires de protéines envisageables pour découvrir toutes les interactions binaires potentielles, c'est-à-dire générer des cartes d'interactions « physiquement possibles ». Ces cartes peuvent aussi servir à extraire un nombre limité d'interactions candidates auxquelles attribuer poids, direction et fonction via des expériences supplémentaires. La technique expérimentale la plus répandue permettant de

construire de telles cartes d'interactions binaires est le double hybride en levure (Figure 4), semi automatisé depuis le début du XXIe siècle pour être applicable à grande échelle (Fields and Song 1989; Walhout and Vidal 2001). La constante amélioration de ses protocoles permet aujourd'hui d'interroger jusqu'à ~80,000,000 paires de protéines en quelques mois, ce qu'une équipe de recherche dont je fais partie a effectué en 2009. La construction et l'analyse de telles cartes nécessitent des travaux de recherche et des réalisations importantes en bioinformatique, qui ont constitué la majeure partie de mes travaux de thèse.

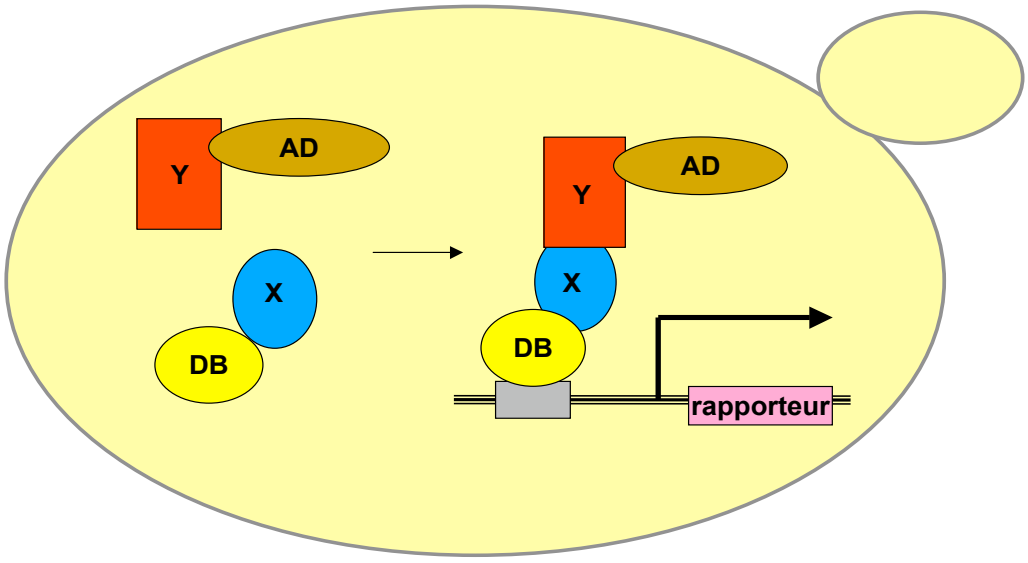


Figure 4 : Principe du double hybride en levure. La protéine X (en bleu) est fusionnée au domaine de liaison à l'ADN (« DB », en jaune) du facteur de transcription Gal4, tandis que la protéine Y (en rouge) est fusionnée à son domaine activateur («AD », en marron). Une interaction physique entre les deux protéines reconstitue le facteur de transcription, ce qui déclenche la transcription d'un gène rapporteur (en rose). Pour une description plus détaillée du double hybride en levure, y compris des souches utilisées, des contrôles et des spécificités de son utilisation à grande échelle, se référer aux articles sur l'interactome d'un nématode et d'une plante associés à ce manuscrit (**Documents Joints 5 et 6** respectivement).

Problématique

Une hypothèse centrale de la biologie systémique propose que les relations entre génotype et phénotype soient sous-tendues par un ensemble de réseaux moléculaires dynamiques au sein de la cellule. Mes travaux de thèse testent et étoffent cette hypothèse à la lumière de la célèbre formule du généticien T. Dobzhansky : *rien n'a de sens en biologie sauf à la lumière de l'évolution.*

L'immense diversité des formes de vie qui peuplent notre planète, des bactéries aux primates en passant par les plantes et les insectes, résulte de l'évolution incessante des espèces selon le principe de sélection naturelle décrit dès le XIXe siècle par C. Darwin. Les travaux quasi-concomitants de G. Mendel, puis de H. de Vries et T. H. Morgan, en posant les bases de la génétique moderne, ont mis en évidence que l'évolution des espèces est sous-tendue par celle de leurs gènes. L'accumulation de mutations dans l'ADN crée de la variabilité héréditaire parmi les génotypes d'une population. Cette variabilité représente le substrat de la sélection naturelle et de la dérive génétique qui, ensemble, façonnent les fréquences alléliques (Hartl and Clark 1997). De plus, certains mécanismes épigénétiques comme l'organisation chromatinienne ou la conformation tridimensionnelle des protéines prions jouent aussi un rôle évolutif (Halfmann, Alberti et al.). S'il est bien établi que les changements écologiques influencent l'évolution des espèces, l'impact des dynamiques évolutives sur les traits écologiques et la vitesse de l'évolution demeurent sujets de recherches intenses (Schoener).

L'hypothèse de l'horloge moléculaire, proposée dans les années 1960 par E. Zuckerkandl et L. B. Pauling, suppose que les mutations ont lieu de façon aléatoire et à vitesse constante. Cette notion intuitive bien que controversée (Kumar 2005) est intimement liée à la théorie neutraliste de l'évolution (Kimura 1968), et la plupart des algorithmes spécialisés dans l'analyse de séquence reposent sur ses principes (Durbin, Eddy et al. 1998). Dans le cas des protéines, le taux de mutations non synonymes apparaît en général largement inférieur à celui de mutations synonymes, indiquant selon l'hypothèse de l'horloge moléculaire une sélection dite « purifiante », qui stabilise leur fonction. Tout changement dans le taux de mutations non synonymes d'une protéine reflète donc potentiellement une modification des pressions de sélection s'exerçant sur cette protéine. Le même raisonnement s'applique aux variations du nombre de copies d'un locus encodant une protéine donnée. Retracer l'évolution des protéines ouvre donc des portes vers l'identification des processus moléculaires impliqués dans l'adaptation de l'organisme à son environnement, et donc indirectement vers une meilleure compréhension de l'organisation cellulaire.

La biologie systémique propose que les fonctions des protéines reposent en

grande partie sur les réseaux d'interactions physiques qu'elles effectuent entre elles et avec les autres macromolécules cellulaires. Si cette hypothèse centrale de la biologie systémique est vraie, alors on devrait pouvoir observer que ces réseaux évoluent eux aussi, comme les génotypes et comme les phénotypes, de façon darwinienne. C'est à la recherche de signatures d'une telle évolution des systèmes biologiques que ma thèse est consacrée.

Je me suis concentrée sur les systèmes formés par les protéines et leurs interactions physiques binaires, que j'appellerai « réseaux interactomes » dans la suite de ce manuscrit. À ce jour, il n'existe aucun organisme dont le réseau interactome entier soit connu, et ce pour deux raisons principales. D'une part, connaître la liste de toutes les protéines encodées par un génome représente un défi technique et conceptuel considérable. D'autre part, il n'existe aucun moyen expérimental permettant de déterminer toutes les interactions physiques binaires possibles entre un groupe de protéines défini. Je me suis donc attachée à améliorer les méthodes visant à caractériser les nœuds (protéines) et les liens (interactions physiques binaires) des réseaux interactomes chez plusieurs organismes eukaryotes. Ces travaux m'ont permis de dégager certains principes organisationnels et évolutifs des systèmes biologiques exposés dans ce manuscrit.

Présentation du plan

Le premier chapitre de ma thèse révèle que les racines profondes de la difficulté à annoter les gènes encodant les protéines ont un rapport direct avec l'évolution des génomes et la définition même de ce que sont les protéines. Les conclusions de cette étude, publiées dans le journal Genome Research en juillet 2008 (Li, Carvunis et al. 2008), m'ont amenée à concevoir une théorie sur l'origine évolutive des protéines. Cette théorie est basée sur l'observation que des mutations dans les régions intergéniques transcrites peuvent aboutir à la formation de phases ouvertes de lectures. Elle est également exposée dans le premier chapitre de ce manuscrit.

Le second chapitre est consacré à la caractérisation des liens des réseaux interactome, les interactions physiques binaires entre protéines. J'y présente mes propositions conceptuelles et mes réalisations bioinformatiques optimisant la construction expérimentale des réseaux interactome de grande taille par double hybride en levure, et permettant de mesurer la qualité technique et la pertinence biologique de ces réseaux.

Le troisième et dernier chapitre de ma thèse est consacré à l'étude des principes évolutifs de l'un de ces réseaux. J'y présente mes observations et interprétations concernant la relation entre la structure du réseau interactome de la plante *Arabidopsis thaliana* et deux types de perturbations naturelles intimement liées à son évolution: les duplications de gènes, et l'attaque par deux phyto-pathogènes.

DOCUMENT JOINT 1

Titre : Biologie systémique : des concepts d'hier aux découvertes de demain.

Auteurs : Anne-Ruxandra Carvunis, Elisa Gomez, Nicolas Thierry-Mieg, Laurent Trilling, Marc Vidal

Description : texte de vulgarisation sur la biologie systémique et ses applications médicales paru été 2009 dans le magazine *Médecine/Sciences*.

Contribution : Marc Vidal, invité par le magazine à écrire un texte pour une issue spéciale, m'a proposé d'en être l'auteur principal. Il m'a aidée à rassembler la bibliographie nécessaire et à rédiger le texte. Elisa Gomez, Nicolas Thierry-Mieg et Laurent Trilling m'ont aidée à rédiger le texte.

Médecine Sciences

Médecine/Science | Numéro Double Juin Juillet 2009 | Volume 25 | n° 6

Biologie systémique : des concepts d'hier aux découvertes de demain

Systems biology: from yesterday's concepts to tomorrow's discoveries

L'idée selon laquelle les gènes et leurs produits sont les unités fondamentales de la biologie a profondément marqué la pensée scientifique de la seconde moitié du xxe siècle. Aujourd'hui, cette approche réductionniste est remise en cause par la renaissance de la biologie systémique, qui a pour objets d'étude les systèmes formés par les produits de gènes en interaction. Le développement de cette discipline est intimement lié aux développements biotechnologiques qui permettent d'interroger ces interactions systématiquement, ainsi qu'aux avancées théoriques qui rendent possible leur modélisation. En recherche fondamentale comme biomédicale, la biologie systémique s'affirme comme un paradigme de choix pour comprendre les propriétés émergentes des systèmes biologiques complexes. <

Auteur principal :
Anne-Ruxandra Carvunis
Adresse : M. Vidal : Center for cancer systems biology (CCSB) and Department of cancer biology, Dana-Farber Cancer Institute, 1, Jimmy Fund Way, Boston, Massachusetts 02115, États-Unis. Department of Genetics, Harvard Medical School, 77, avenue Louis Pasteur, Boston, Massachusetts 02115, États-Unis. N. Thierry-Mieg, L. Trilling : TIMC-IMAG, CNRS UMR5525, Faculté de médecine, 38706 La Tronche Cedex, France. A.R. Carvunis : Center for cancer systems biology (CCSB) and Department of cancer biology, Dana-Farber Cancer Institute, 1, Jimmy Fund Way, Boston, Massachusetts 02115, États-Unis. Department of Genetics, Harvard Medical School, 77, avenue Louis Pasteur, Boston, Massachusetts 02115, États-Unis. TIMC-IMAG, CNRS UMR5525, Faculté de médecine, 38706 La Tronche Cedex, France. E. Gomez : Inserm U833, Collège de France, Chaire de médecine expérimentale, 75005 Paris, France.
email : anne-ruxandra_carvunis@dfci.harvard.edu

Co-auteur(s) :
Elisa Gomez | |

Nicolas Thierry-Mieg

Laurent Trilling

Marc Vidal | marc_vidal@dfci.harvard.edu

The idea that genes and their products are the fundamental units of biology has profoundly influenced our scientific thinking during the second half of the past century. Today, this reductionism is challenged by a renaissance of a systems understanding of biology, focusing on the systems formed by interacting gene products rather than on individual gene products. This discipline, based on a complementary and more holistic approach, keeps expanding its scope thanks to biotechnological innovations as well as theoretical modeling. This review aims at showing how and why, since the beginning of the 21st century, in fundamental as well as biomedical research, systems biology is proving a promising paradigm for understanding emerging properties of complex biological systems.

La biologie systémique représente un important changement de paradigme scientifique par rapport à la biologie moléculaire des cinquante dernières années.

La biologie moléculaire traditionnelle, de nature « réductionniste », s'est jusqu'ici concentrée principalement sur la caractérisation des composants individuels de la cellule, gènes, protéines, ou encore ARN non codants, avec pour but de comprendre la vie à partir de la caractérisation des macromolécules qui la constituent. Toutefois, protéines et ARN opèrent en interagissant les uns avec les autres, formant ainsi des systèmes dont la complexité peut difficilement être comprise une molécule à la fois. La biologie systémique, de nature « intégrative et holistique », entend comprendre la vie à partir de ces systèmes. Elle pose les questions biologiques en mettant l'accent sur le « tout » plutôt que sur les « parties ». Intrinsèquement interdisciplinaire, sa méthodologie originale est définie par un aller-retour dynamique entre expérimentation, travail informatique et théorique, formulation de nouvelles hypothèses scientifiques et développement technologique.

La biologie systémique est l'étude des interactions entre les composants d'un système biologique

Le terme « système » vient du grec *sustêma* qui signifie ensemble. Un système est un ensemble d'entités interagissant ou interdépendantes, abstraites ou concrètes, dont l'union forme un tout. En biologie, une certaine confusion règne autour de cette définition car l'échelle d'étude d'un système peut lui valoir ou non l'appellation « système » selon l'opinion de celui qui le considère. En fait, il n'y a pas de limite théorique à la taille d'un système : libre à chacun de définir les bornes le délimitant de son environnement[1]. Ainsi, un facteur de transcription régulant sa propre expression constitue un système, de la même façon que l'ensemble des molécules d'une cellule, l'ensemble des cellules d'un organisme, ou encore l'ensemble des individus d'une population. Différents systèmes peuvent être étudiés à partir des mêmes entités si l'on considère un type d'interaction plutôt qu'un autre. La population d'une ville peut être vue soit comme un ensemble d'individus partageant des relations économiques, des engagements matrimoniaux ou encore des maladies contagieuses, soit comme trois systèmes distincts. Un système peut aussi intégrer des composants hétérogènes, tel un écosystème comprenant à la fois des proies, des prédateurs et des ressources naturelles. Enfin, des composants et des interactions de natures différentes et traversant plusieurs échelles peuvent former un système unique *(Figure 1A)*.

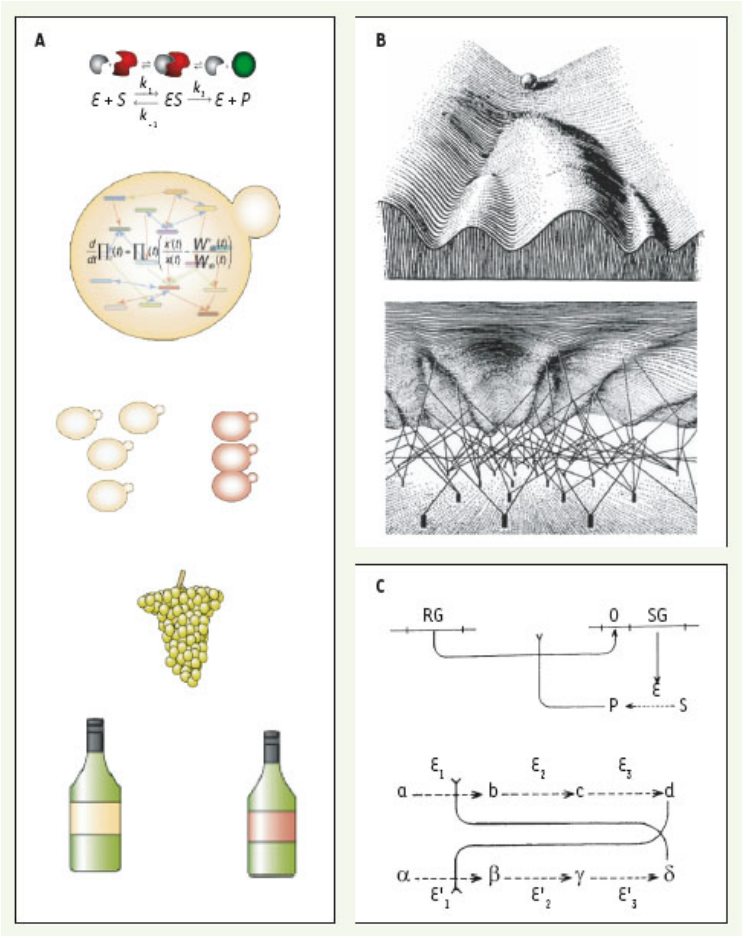


Figure 1. Illustrations de la notion de système biologique. A. Schématisation de systèmes biologiques à différentes échelles d'étude. De haut en bas : une enzyme et son substrat, ensemble des voies métaboliques dans une levure, population constituée de plusieurs levures individuelles partageant les mêmes ressources, la grappe de raisin comme écosystème, le vin comme produit d'intégration. Si chacun de ces systèmes peut être étudié séparément, ils peuvent aussi être intégrés dans un unique système mixte. Par exemple, l'Institut de recherche pour le vin australien propose d'étudier en détail comment les systèmes métaboliques de différentes levures (ici représentées par une population jaune et une population rouge), en causant la fermentation du même raisin, produisent des produits de dégradation caractéristiques [1]. Puisque ce sont ces produits qui sont responsables de la variété aromatique des vins (représentée ici par les étiquettes jaunes et rouges), l'institut suggère que ce savoir moléculaire pourra être intégré à la géographie des vignobles et au système économique

dynamique de l'offre et de la demande, afin de produire pour les marchés de consommateurs ciblés les vins dont le goût leur plaira. ***B.*** Illustration de la notion de « paysage épigénétique » de C.H. Waddington (d'après [2]).***C.*** Deux systèmes biologiques imaginés par F. Jacob et J. Monod (d'après [4]). RG : « gène régulateur » (répresseur de transcription) ; SG : « gène structural » (codant pour une enzyme) ; $E^{(')}_{(1,2,3)}$: enzymes ; S : substrat ; P : produit ; a, b, c, d, α, β, γ, δ métabolites. Dans l'exemple du haut, le produit d'une réaction enzymatique inhibe (>-) la répression de l'expression du gène codant pour l'enzyme. Il s'agit d'une version de la boucle de rétroaction positive. Dans l'exemple du bas, le produit final d'une voie métabolique inhibe la première réaction d'une *autre* voie métabolique, et réciproquement. Il en résulte un système bistable où une seule de ces deux voies métaboliques peut être active à la fois.

Biologie systémique : des concepts historiques…

Les origines de la biologie systémique remontent aux années 1950, lorsque C. H. Waddington établit le concept de « paysage épigénétique » [2]. Il imagine les cellules passer d'un état de différenciation à l'autre en suivant un trajet dicté par la forme d'un paysage constitué de monts et de vallées *(Figure 1B, panneau supérieur)*, paysage lui-même créé par les interactions entre gènes *(les « piliers » du panneau inférieur de la Figure 1B)*. Cette vision de la cellule comme système évoluant d'état en état s'inspire des travaux de M. Delbrück [3], F. Jacob et J. Monod [4] qui, après la Seconde Guerre mondiale, introduisent la notion de système en biochimie *(Figure 1C)*, comme un mécanisme susceptible d'expliquer le mystère de la différenciation : comment des cellules au génome identique peuvent-elles exprimer des formes et des propriétés aussi différentes que celles d'un lymphocyte et d'un myocarde ? Ils proposent de voir les enzymes et leurs substrats comme les composants de circuits dynamiques, dont les exemples les plus simples sont les boucles de rétroaction négatives et positives. Dans le cas d'une boucle négative, l'augmentation du niveau d'un élément entraîne la diminution de son taux de production, ce qui a pour résultat une stabilisation de sa production et de son abondance, à la manière d'un thermostat. Une boucle positive en revanche a l'effet inverse, et donne lieu à deux scénarios opposés : si la boucle est enclenchée, l'élément encourage sa propre production, sinon il n'est pas produit. Ce type de circuits moléculaires et d'autres plus complexes ont depuis été très étudiés théoriquement, notamment par R. Thomas et ses collègues [5]. La réalité de ces systèmes a été démontrée expérimentalement dès la fin des années 1950 avec l'exemple de l'opéron lactose inductible de A. Novick et M. Weiner chez la bactérie *E. coli* [6], puis à de multiples reprises au cours des cinquante dernières années, ce qui a permis à la biologie systémique de s'affirmer aujourd'hui comme discipline à part entière.

…remis à jour grâce au développement de technologies expérimentales et de nouveaux outils d'analyse

Biologie synthétique : reconstruire un système pour mieux le comprendre

À la frontière de la science-fiction, cette discipline reconstitue des circuits moléculaires *in vivo*, ou bien en invente de toutes pièces en fusionnant des domaines d'ADN provenant de multiples espèces. Parmi ses travaux fondateurs [7], l'« interrupteur » est un système artificiel bistable composé de deux promoteurs et deux répresseurs de transcription croisés chez *E. coli*. Le résultat est une bactérie qui produit une protéine recombinante si et seulement si elle a été en contact avec l'inducteur de l'un des promoteurs dans le passé, sans que son patrimoine génétique n'ait été modifié d'aucune manière. En d'autres termes, la bactérie se « différencie » à travers un processus de « mémorisation » d'un changement purement environnemental. La preuve qu'une boucle de rétroaction positive peut être à l'origine de la différenciation cellulaire a donc été faite.

Génomique fonctionnelle : expérimentation à l'échelle de la cellule entière

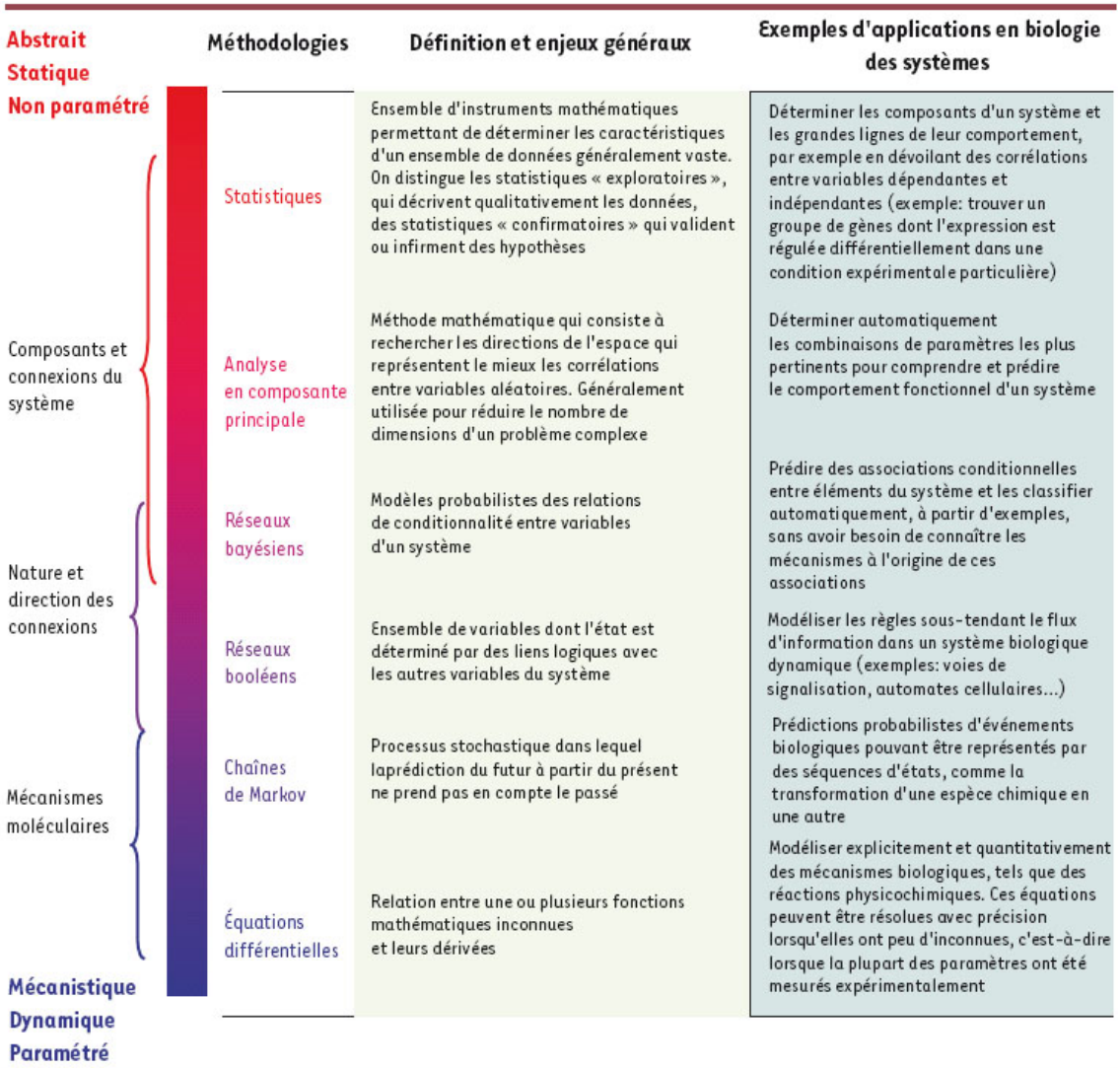
Aussi fondamentaux soient-ils, les travaux de la biologie synthétique sont encore loin d'atteindre la complexité du système « cellule vivante ». C'est à cette échelle que les défis de la biologie systémique se posent aujourd'hui. Comment ces petits circuits moléculaires sont-ils liés les uns aux autres ? Comment communiquent-ils pour répondre aux besoins du « tout » cellulaire ? Pour parvenir à une compréhension globale de la cellule, il faut pouvoir observer simultanément *tous* les gènes, *toutes* les protéines, *tous* les ARN, ainsi que *toutes* leurs interactions. À cette fin se développent de nombreuses techniques systématiques à haut débit, souvent par miniaturisation et robotisation de techniques préexistantes *(Tableau I)*. Parmi les exemples emblématiques de ces efforts, on peut citer la puce à ADN (une seule puce mesure quantitativement l'expression des 6 000 gènes de la levure à la fois), la cartographie du réseau « interactome » des interactions moléculaires à l'échelle du protéome, ou encore l'établissement de réseaux génétiques. On comprendra aisément que l'analyse et l'intégration [8] de telles quantités d'information s'accompagnent nécessairement de l'introduction de nouvelles méthodes mathématiques et informatiques.

Exemples de techniques identifiant les interactions entre composants cellulaires	
ADN/ADN	5C (capture de la conformation des chromosomes-copie carbone)
protéine/ADN	immuno-précipitation de chromatine suivie d'analyse de puce à ADN
protéine/protéine	système double-hybride, co-immuno-précipitation suivie de spectrométrie de masse
Exemples de techniques mesurant les états induits par ces interactions	
variations de l'expression des ARN messagers	puces à ADN, SAGE (analyse sérielle de l'expression des gènes)
abondance, localisation et modifications post-traductionnelles des protéines	spectrométrie de masse, marquage suivi de microscopie en fluorescence, puces à protéines

Tableau I. Exemples de techniques expérimentales utilisées à haut débit (liste non exhaustive) (d'après [9]).

Gestion des données et modélisation : transfert de savoir des « sciences dures » et apparition de nouvelles problématiques

Prenons l'exemple d'une simple boucle de rétroaction positive où l'élément encourage sa propre production. À quelle vitesse ? Quelle quantité d'inducteur permet de générer une quantité souhaitée de produit ? En fait, indépendamment de l'échelle d'étude, la compréhension quantitative de tout système nécessite une analyse numérique. Répondre aux problèmes spécifiques à la biologie s'est révélé un défi fascinant pour les sciences exactes qui ont non seulement adapté leurs concepts, mais ont aussi trouvé dans les systèmes biologiques une source de nouvelles recherches théoriques. Réciproquement, l'apport de nouvelles méthodes d'analyse a permis aux biologistes de s'ouvrir à de nouvelles problématiques. Ainsi, les biologistes manient à présent le contrôle de qualité, l'estimation de paramètres à partir d'un nombre limité de mesures ou le contrôle de la stochasticité, problèmes classiques pour les ingénieurs, afin de parvenir à des analyses plus précises de leurs données. Au cœur de la biologie systémique se trouvent l'identification et la modélisation des réseaux à travers lesquels gènes et protéines interagissent pour effectuer les opérations cellulaires. Les outils mathématiques utilisés pour aborder ces questions correspondent à différents niveaux d'abstraction *(Tableau II)*. De façon générale, les analyses à haut niveau d'abstraction décrivent les propriétés qualitatives des systèmes, tandis que les modèles à bas niveau d'abstraction produisent des prédictions quantitatives [9]. En effet, ces efforts dépendent étroitement de l'acquisition systématique de données biologiques chiffrées.



Méthodologies	Définition et enjeux généraux	Exemples d'applications en biologie des systèmes
Statistiques	Ensemble d'instruments mathématiques permettant de déterminer les caractéristiques d'un ensemble de données généralement vaste. On distingue les statistiques « exploratoires », qui décrivent qualitativement les données, des statistiques « confirmatoires » qui valident ou infirment des hypothèses	Déterminer les composants d'un système et les grandes lignes de leur comportement, par exemple en dévoilant des corrélations entre variables dépendantes et indépendantes (exemple: trouver un groupe de gènes dont l'expression est régulée différentiellement dans une condition expérimentale particulière)
Analyse en composante principale	Méthode mathématique qui consiste à rechercher les directions de l'espace qui représentent le mieux les corrélations entre variables aléatoires. Généralement utilisée pour réduire le nombre de dimensions d'un problème complexe	Déterminer automatiquement les combinaisons de paramètres les plus pertinents pour comprendre et prédire le comportement fonctionnel d'un système
Réseaux bayésiens	Modèles probabilistes des relations de conditionnalité entre variables d'un système	Prédire des associations conditionnelles entre éléments du système et les classifier automatiquement, à partir d'exemples, sans avoir besoin de connaître les mécanismes à l'origine de ces associations
Réseaux booléens	Ensemble de variables dont l'état est déterminé par des liens logiques avec les autres variables du système	Modéliser les règles sous-tendant le flux d'information dans un système biologique dynamique (exemples: voies de signalisation, automates cellulaires...)
Chaînes de Markov	Processus stochastique dans lequel laprédiction du futur à partir du présent ne prend pas en compte le passé	Prédictions probabilistes d'événements biologiques pouvant être représentés par des séquences d'états, comme la transformation d'une espèce chimique en une autre
Équations différentielles	Relation entre une ou plusieurs fonctions mathématiques inconnues et leurs dérivées	Modéliser explicitement et quantitativement des mécanismes biologiques, tels que des réactions physicochimiques. Ces équations peuvent être résolues avec précision lorsqu'elles ont peu d'inconnues, c'est-à-dire lorsque la plupart des paramètres ont été mesurés expérimentalement

Tableau II. Exemples de méthodologies mathématiques et computationnelles pour l'analyse des systèmes biologiques, des plus abstraites aux plus mécanistiques (liste non exhaustive) (d'après [9]).

La biologie systémique, porteuse d'espoir en recherche pharmaceutique

Classiquement, l'industrie pharmaceutique procède de façon réductionniste. Après identification d'une cible thérapeutique potentielle, des substances chimiques interagissant spécifiquement avec cette cible sont développées, puis leurs effets sont testés dans des modèles *in vitro* et animaux. Ce n'est que lors des essais cliniques que les médicaments se trouvent enfin dans le contexte du patient, et c'est souvent à ce moment que des effets indésirables comme la toxicité ou le manque d'efficacité sont révélés. Ainsi, la moitié des essais cliniques aux États-Unis est un échec, la mise sur le marché d'un médicament prend en moyenne entre sept et douze ans, et environ 250 milliards de dollars publics et privés sont investis chaque année pour seulement une soixantaine de nouveaux traitements, dont la plupart sont de nouvelles versions de médicaments existants, seulement une quinzaine sont vraiment nouveaux. Comme en recherche fondamentale, la biologie systémique est perçue comme l'approche qui améliorera à l'avenir le rendement des découvertes pharmaceutiques en considérant le « contexte biologique » plus tôt dans le processus.

Ses succès sont déjà visibles…

L'application de la biologie systémique par l'industrie pharmaceutique se caractérise par l'introduction de modélisations à l'échelle systémique, de techniques de mesures associées, et par l'utilisation de bases de données. Le simple enregistrement dans ces bases de données de résultats disponibles dans la littérature clinique et les archives des hôpitaux et entreprises représente déjà un progrès immense qui permet de prédire l'effet clinique d'une espèce chimique en analysant les résultats d'autres espèces de structure similaire. Ce sont les *start-up*, comme Genestruct, BioSeek ou Merrimack Pharmaceuticals qui les premières ont choisi la biologie systémique, soit en proposant aux grandes entreprises de tester expérimentalement leurs molécules candidates par les techniques expérimentales de la biologie systémique, soit en utilisant la modélisation pour révéler de nouvelles cibles thérapeutiques potentielles. Les effets bénéfiques sont déjà visibles sur l'efficacité et la sécurité des nouveaux traitements, et on prédit à long terme une diminution générale du coût et du temps nécessaires à la production.

…mais ses limites se profilent aussi

Les nouveaux médicaments ainsi rationnellement conçus ont tendance à être ultra-spécifiques. Ainsi certains traitements contre le cancer, comme Tarceva[1], sont extrêmement efficaces lorsque la tumeur porte une mutation précise, mais généralement inutiles dans le reste des cas. Malheureusement, comme les technologies de séquençage disponibles aujourd'hui ne permettent pas d'identifier le profil génétique de chaque patient, les hôpitaux doivent trop souvent soumettre leurs patients à des chimiothérapies douloureuses et coûteuses sans savoir si leurs tumeurs portent ou non ces mutations. C'est pourquoi la recherche de « bio-marqueurs », groupes de gènes ou protéines faciles à mesurer et assez indicatifs de l'état du système et donc susceptibles de permettre d'établir un diagnostic ou un pronostic, suscite beaucoup d'enthousiasme. L'un de ces succès concerne le cancer du sein : les profils de transcription d'une soixantaine de gènes sont suffisamment discernables pour constituer les signatures de différents types de cancer, et donc d'identifier les patientes candidates à une chimiothérapie sans besoin de séquençage [10]. Cependant, l'approche par les « bio-marqueurs » reste purement corrélative : savoir qu'un profil est statistiquement lié à un état du système n'informe pas directement sur les événements responsables des phénotypes observés. Nos efforts pour améliorer la vision systémique des maladies ne pourront donc vraiment porter leurs fruits que lorsque la biologie des systèmes sains sera elle-même mieux élucidée.

La science des réseaux, une branche prometteuse de la biologie systémique

Bien que quelques *start-up* s'y consacrent, la recherche sur les mécanismes par lesquels les maladies perturbent les systèmes reste majoritairement fondamentale. Certains groupes de recherche, dont le nôtre, cartographient les interactions entre protéines virales et protéines hôtes [12, 13], et nous avons entrepris de mesurer quelles interactions protéiques sont perdues ou retenues par les protéines lorsqu'on leur impose des mutations mendéliennes connues pour être associées au cancer. Ce type d'approche s'inscrit dans le contexte d'une nouvelle discipline qui émerge des sciences théoriques traditionnelles : la science des réseaux. Sa plus grande découverte est que les propriétés structurelles des réseaux informent sur la nature des tâches qu'ils peuvent accomplir. Ainsi, la connectivité des réseaux réels (Internet, chaînes alimentaires, transports publics, interactions protéiques, etc.) les distingue de réseaux aléatoires. Par exemple, la distribution des degrés (nombre de connexions par composant) dans les réseaux réels tend à suivre une loi de puissance, tandis que celle de réseaux purement aléatoires suit typiquement une loi en forme de cloche. Cela signifie que dans les réseaux réels, un petit nombre de composants font beaucoup de connexions (on les appelle les *hubs*), tandis que la plupart des composants font peu de connexions. Cette propriété rend les réseaux réels très robustes en cas d'« erreur » ou de dysfonctionnement de leurs composants, mais en revanche très sensibles aux « attaques » dirigées contre les *hubs* [14]. En biologie *(Figure 2)*, cette observation soulève des questions fondamentales : l'évolution favorise-t-elle ces structures pour protéger les réseaux cellulaires contre les erreurs induites par les mutations ? Les pathogènes attaquent-ils de préférence les *hubs* ?

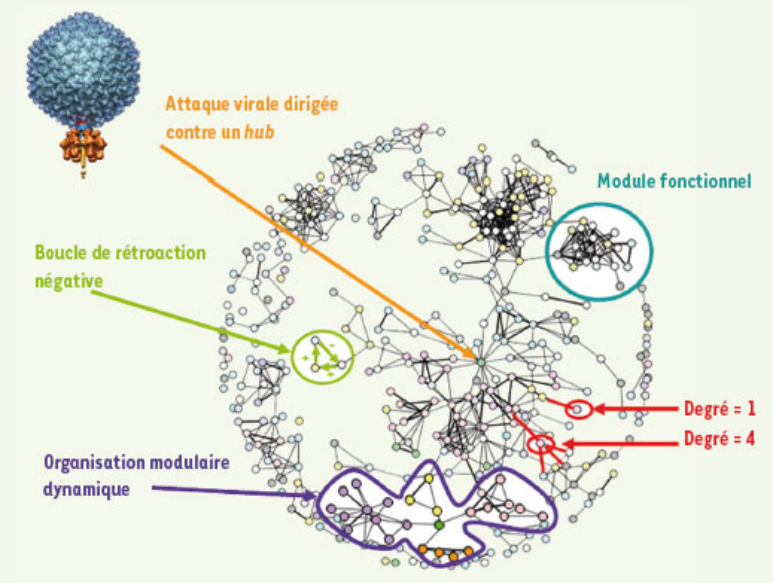


Figure 2. Représentation théorique d'un réseau d'interactions entre protéines. Dans ce réseau, chaque nœud représente une protéine et chaque lien entre deux nœuds une interaction biophysique ou biochimique. Les différentes couleurs attribuées aux nœuds représentent des classes fonctionnelles (exemples : enzyme, facteur de transcription, localisation membranaire, coexpression, etc.). Les différentes épaisseurs des liens illustrent que les interactions peuvent être plus ou moins stables, ou bien démontrées avec plus ou moins de certitude. La notion de degré est représentée en rouge : une protéine n'interagissant qu'avec un seul partenaire est de degré 1, alors qu'une autre avec quatre partenaires est de degré 4. Lorsque dans un groupe les protéines partagent plus d'interactions entre elles que la moyenne, on peut supposer qu'elles forment un module fonctionnel [11] (en bleu). Si ce module contient des protéines de fonction inconnue, on peut prédire qu'elles partagent certaines des fonctions des autres membres. Bien que la plus grande partie de ce que nous connaissons sur le réseau d'interactions physiques entre protéines soit statique, ce savoir peut servir de plate-forme sur laquelle surimposer des informations dynamiques, comme la coexpression. Au sein de la forme violette, les nœuds de la même couleur sont des protéines dont l'expression est corégulée. Les nœuds roses, jaunes, violets et oranges sont regroupés par couleur, tandis que le nœud vert lie ces différents groupes entre eux. Cela suggère que le nœud vert a probablement une fonction d'intégration et de coordination dynamique des modules unicolores [8]. De la même façon, des données cinétiques peuvent se superposer aux interactions physiques statiques et placer dans un contexte plus général de petits systèmes dynamiques [8] (en vert). La flèche orange illustre que les cibles préférées des virus sont les protéines les plus connectées, ou *hubs*.

Parmi les réseaux réels, des différences topologiques révèlent des classes universelles. Ainsi, les motifs de trois ou quatre composants surreprésentés dans des systèmes de transfert d'énergie (chaînes alimentaires) ne sont pas les mêmes que ceux surreprésentés dans des systèmes de traitement d'information (réseaux de transcription, réseaux neuronaux, puces électroniques) [15]. Placer dans un même cadre théorique des systèmes de natures diverses promet donc de révéler encore bien des lois propres à tous les systèmes complexes. Quelles seront les retombées médicales de la biologie des réseaux ? De nombreux axes de recherche comme ceux que nous venons de citer sont déjà en cours d'étude, mais il est clair que l'avenir nous réserve de grandes découvertes.

L'union fait la force

Aujourd'hui la biologie systémique réunit informaticiens, statisticiens, mathématiciens, physiciens, ingénieurs, chimistes, médecins et biologistes autour des mêmes questions scientifiques. Économistes et politiciens sont prêts à s'engager demain dans la même direction pour traiter les défis de santé publique de manière systémique, qu'il s'agisse de comprendre la dynamique de propagation de maladies contagieuses à travers les structures humaines ou de mesurer l'impact des programmes sociaux sur l'obésité [16]. Ainsi, de l'identification de cibles thérapeutiques à l'invention de nouveaux outils diagnostiques en passant par l'établissement de politiques de santé publique, la biologie systémique est présente à toutes les étapes du processus médical. En recherche fondamentale, son succès entraîne un véritable changement de paradigme, remplaçant la notion de gène comme unité centrale de la biologie par une vision holistique. La biologie systémique porte donc beaucoup d'espoir pour mieux comprendre le vivant et les maladies. Sa force provient sans doute de son inhérente

interdisciplinarité. Nul champ scientifique n'est plus à même qu'un autre de trouver les réponses aux grandes questions posées par la nature, mais c'est l'échange entretenu entre les disciplines qui peut mener aux grandes découvertes scientifiques.

REMERCIEMENTS

Nous remercions les docteurs Nicolas Simonis et Thomas Jubault pour leur aide à la rédaction de cette synthèse. M.V. est chercheur qualifié honoraire au Fonds de la recherche scientifique (FRS-FNRS, Communauté française de Belgique).

CONFLITS D'INTÉRÊT

Les auteurs déclarent n'avoir aucun conflit d'intérêt.

RÉFÉRENCES

1. Borneman AR, Chambers PJ, Pretorius IS. Yeast systems biology: modeling the winemaker's art. *Trends Biotechnol* 2007 ; 25 : 349-55.

2. Waddington C. H. *The strategy of the genes*. London : George Allen and Unwin, 1957.

3. Delbrück, M. Discussion. In : *Unités biologiques douées de continuité génétique*. Paris : Éditions du CNRS, 1949 : 33-5.

4. Monod J, Jacob F. Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harb Symp Quant Biol* 1961 ; 26 : 389-401.

5. Thomas R, D'Ari R. *Biological feedback*. Boca Raton, Florida : CRC Press, 1990 : 316 p.

6. Novick A, Weiner M. Enzyme induction as an all-or-none phenomenon. *Proc Natl Acad Sci USA* 1957 ; 43 : 553-66.

7. Hasty J, McMillen D, Collins JJ. Engineered gene circuits. *Nature* 2002 ; 420 : 224-30.

8. Ge H, Walhout AJ, Vidal M. Integrating "omic" information: a bridge between genomics and systems biology. *Trends Genet* 2003 ; 19 : 551-60.

9. Ideker T, Lauffenburger D. Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends Biotechnol* 2003 ; 21 : 255-62.

10. Van't Veer LJ, Dai H, van de Vijver MJ, *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002 ; 415 : 484-5.

11. Gunsalus KC, Ge H, Schetter AJ, *et al*. Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* 2005 ; 436 : 861-5.

12. De Chassey B, Navratil V, Tafforeau L, *et al*. Hepatitis C virus infection protein network. *Mol Syst Biol* 2008 ; 4 : 230.

13. Calderwood MA, Venkatesan K, Xing L, *et al* Epstein-Barr virus and virus human protein interaction maps. *Proc Natl Acad Sci USA* 2007 ; 104 : 7606-11.

14. Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. *Nature* 2000 ; 406 : 378-82.

15. Milo R, Shen-Orr S, Itzkovitz S, *et al*. Network motifs: simple building blocks of complex networks. *Science* 2002 ; 298 : 824-7.

16. Newell B, Proust K, Dyball R, *et al*. Seeing obesity as a systems problem. *NSW Public Health Bulletin* 2007 ; 18 : 214-8.

1. Tarceva (erlotinib, Laboratoires Roche) est une petite molécule qui cible la voie de signalisation du récepteur du facteur de croissance épidermique humain (EGFR).

CHAPITRE 1 : A LA RECHERCHE DES *NŒUDS* DE L'INTERACTOME

Des difficultés de l'annotation de génomes…

Bien que le génome de la levure du boulanger, *Saccharomyces cerevisiae*, soit complètement séquencé depuis 1996 (Goffeau, Barrell et al. 1996), on ignore toujours le nombre de protéines qu'il encode. Lors de sa parution, toutes les phases ouvertes de lecture (« open reading frames » ou ORFs) possibles contenant au moins 100 codons furent considérées comme des gènes codants pour des protéines, soit un total de ~5900. Depuis, environ 400 ORFs contenant moins de 100 codons ont été ajoutées à ce catalogue, prouvant que la longueur d'une ORF n'est pas nécessairement indicative de son potentiel codant. Il est pourtant nécessaire de maintenir ce critère de longueur car le génome de la levure contient plus de 160000 ORFs de plus de 10 codons, y compris 15000 entre 50 et 99 codons dont seulement ~2% encoderaient vraiment des protéines (Fisk, Ball et al. 2006). Les ORFs courtes sont donc considérées comme non fonctionnelles jusqu'à ce que le contraire soit prouvé. Quant aux ORFs de plus de 100 codons, leur longueur ne garantit pas qu'elles encodent des protéines fonctionnelles. Plus de 1000 longues ORFs n'ont pas de rôle biologique connu, et sont classées comme « prédites mais non caractérisées » par le consortium "The *Saccharomyces* Genome Database Project" (SGD), qui recense tous les changements du génome de la levure et de ses annotations depuis sa parution (Pena-Castillo and Hughes 2007). Comment savoir si ces ORFs correspondent à des erreurs de prédiction, ou à des gènes codant pour des protéines dont la fonction n'a tout simplement pas encore été découverte ?

Une solution partielle à ce problème est apparue grâce à la naissance de la génomique comparative eucaryote en 2003, lorsque les génomes d'autres levures furent séquencés : ne garder que les ORFs conservées dans plusieurs espèces (Brachat, Dietrich et al. 2003; Cliften, Sudarsanam et al. 2003; Kellis, Patterson et al. 2003). Ainsi, 10% des ORFs de *S. cerevisiae* furent déclassées par SGD et qualifiées de « douteuses ». Les espèces de levures nouvellement séquencées ont été annotées de telle sorte qu'elles ne possèdent aucune ORF propre, mais seulement des ORFs partagées avec au moins une autre espèce. Depuis, la même approche a été appliquée à plusieurs autres organismes eucaryotes, dont l'humain (Stein, Bao et al. 2003; Clamp, Fry et al. 2007; Clark, Eisen et al. 2007).

N'était-il pas imprudent d'éliminer ainsi ces ORFs, avant d'explorer la possibilité qu'elles participent à la spécificité de chaque espèce? C'est la question que mes co-auteurs et moi-même avons posée dans l'article « Revisiting the

Saccharomyces cerevisiae predicted ORFeome », qui est paru en juillet 2008 dans le journal *Genome Research* et pour lequel j'ai conduit et réalisé la quasi-totalité des analyses bioinformatiques (Li, Carvunis et al. 2008) (**Document Joint 2**). Nous avons regroupé les ORFs éliminées par les études citées ci-dessus sous le terme « ORFs orphelines ». En ré-analysant les résultats de 13 études de génomique fonctionnelle, nous avons montré que 80% de ces orphelines produisaient un ARN ou un peptide détectable *in vivo* (**Figure 5**). De plus, nous avons développé et évalué un algorithme d'apprentissage basé sur une approche bayésienne pour intégrer ces 13 sources de données expérimentales. Cet algorithme a montré que la plupart des ORFs orphelines ont une forte probabilité d'être effectivement des gènes codant pour des protéines. Une indication supplémentaire confortant l'hypothèse de validité de ces ORFs est venue du fait que nous avons recherché et constaté la conservation de leur séquence entre différentes souches de levure au sein de la même espèce, *S. cerevisiae*. Nous avons conclu que l'évaluation d'une prédiction d'ORF devrait se baser sur de multiples niveaux d'analyse, principalement expérimentaux, et non sur leur seule conservation entre plusieurs espèces.

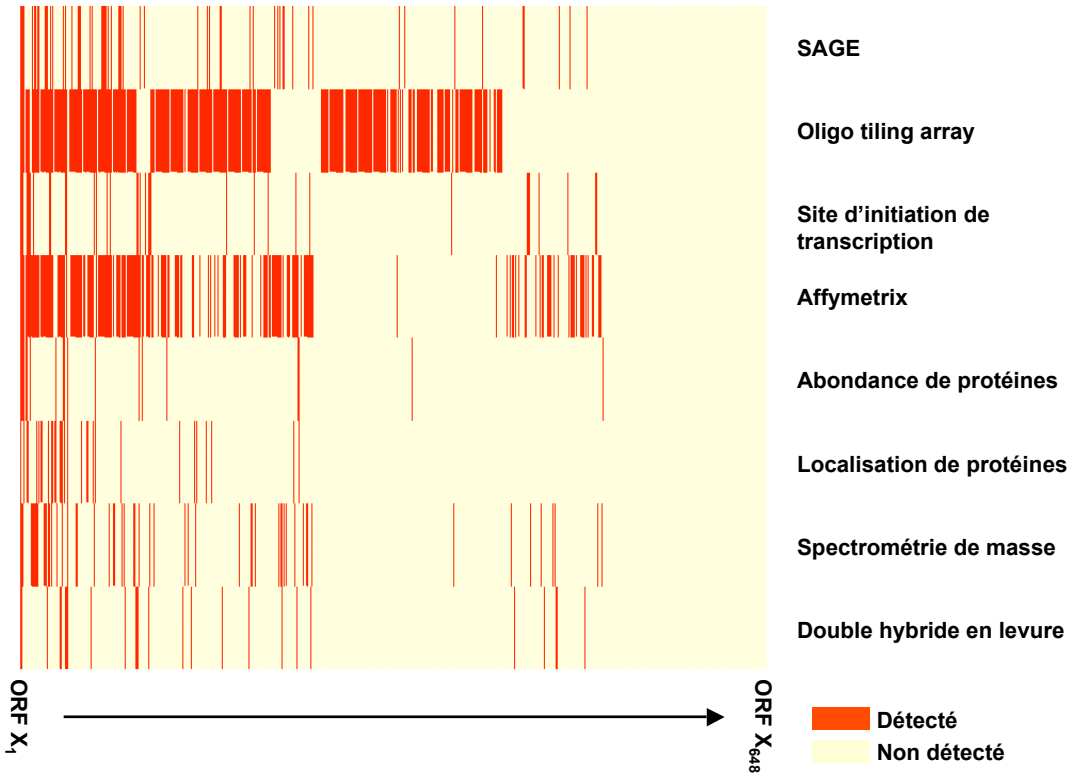


Figure 5 : Support expérimental pour les ORFs orphelines. Chaque colonne représente une ORF. Les colonnes sont ordonnées de l'ORF avec le plus de support expérimental (ORF X1, à gauche) à celle avec le moins de support expérimental (ORF X648; à droite). Les sources de données expérimentales sont regroupées par type d'approche : transcriptionnelles en haut, traductionnelles en bas. En tout, 477 ORFS orphelines semblent transcrites, 180 traduites, et 145 les deux.

… Vers une théorie sur la naissance des gènes

Que des gènes encodent des protéines tout en n'existant que dans la levure du boulanger suggère qu'ils sont apparus récemment au cours de l'évolution. Or, lors de la publication de nos résultats, il n'existait aucun mécanisme biologique connu expliquant comment une telle genèse serait possible. Les gènes codants pour des protéines étaient supposés dériver simplement d'autres gènes codants pour des protéines similaires, principalement par duplication, fusion-fission ou transfert latéral (Long, Betran et al. 2003). J'ai donc entrepris, en collaboration avec Ilan Wapinski qui était alors étudiant en thèse dans le laboratoire d'Aviv Regev (Broad Institute, Boston), de comprendre les mécanismes à l'origine de la naissance de nouveaux gènes (**Document Joint 3**).

Dans ce dessein, Ilan Wapinski et moi-même avons assigné à chacune des ORFs de *S. cerevisiae* recensées par SGD un « âge », correspondant à une estimation de la période d'apparition de leur ancêtre dans l'arbre phylogénétique des levures. Cela m'a permis d'explorer de nombreux paramètres biologiques et de constater que les ORFs les plus jeunes sont plus courtes et moins exprimées que les ORFs plus anciennes. De plus, elles se trouvent souvent dans les régions télomériques, ou bien partiellement à cheval avec d'autres ORFs sur l'autre brin d'ADN. Enfin, les pressions de sélection semblent contraindre leurs séquences moins sévèrement que celles des ORFs ancestrales. L'ensemble de ces constats partiellemet inter-dépendants m'a suggéré un mécanisme simple pour la naissance de nouveaux gènes au cours de l'évolution.

J'émets l'hypothèse que les milliers d'ORFs courtes qui compliquent l'annotation des génomes (voir ci-dessus) constituent un réservoir génétique de potentiels nouveaux gènes (**Figure 6**). En effet, si ces ORFs courtes se trouvent dans l'une des nombreuses régions intergéniques transcrites (Nagalakshmi, Wang et al. 2008), la machinerie de traduction peut y avoir accès et les traduire, produisant ainsi des peptides dont le devenir dépend alors de la sélection naturelle. Cela expliquerait pourquoi les ORFs annotées que nous avons identifiées comme récemment apparues sont courtes, faiblement exprimées, sous faibles pressions de sélection et localisées dans les régions télomériques ou bien partiellement à cheval avec d'autres ORFs sur l'autre brin d'ADN, où la formation d'ORFs aléatoires est favorisée. Pendant que je travaillais sur cette question, plusieurs gènes apparus selon ce mécanisme ont été découverts, et des études bioinformatiques réalisées par d'autres groupes de recherche ont exploré le même sujet (Kaessmann). Pourtant, ces publications se concentrent sur les ORFs annotées par SGD, et aucune ne mentionne que ce mécanisme pourrait agir sur les milliers d'ORFs courtes des régions considérées

intergéniques par SGD. Or, si c'était le cas, nombres de petits peptides pourraient venir s'ajouter aux protéines, ARNs non-codants, lipides et autres composants connus du système cellulaire.

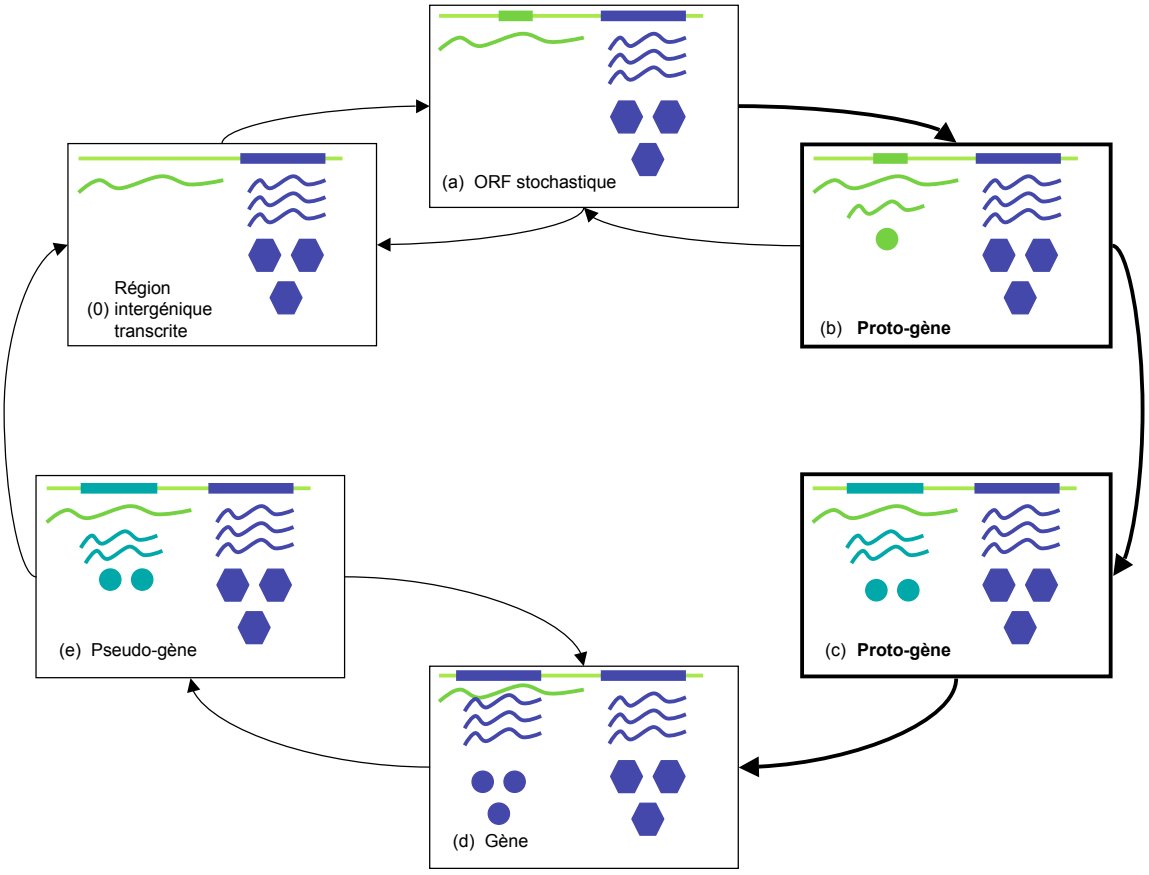


Figure 6 : Boucle conceptuelle où des régions intergéniques donnent naissance à des proto-gènes, qui peuvent devenir gènes, puis pseudo-gènes, et enfin redevenir régions intergéniques. Les flèches représentent des transitions entre les différentiels « états codants » (encadrés) d'une région du génome au cours de l'évolution ; les transitions représentées par des flèches épaisses représentent mes propositions tandis que les plus fines sont reconnues. Dans les états codants, les lignes épaisses indiquent la présence d'ORFs et les lignes fines leur absence; les lignes courbes représentent des molécules d'ARN, et les hexagones et cercles les produits de traduction. Dans une région intergénique transcrite située près d'un gène codant pour une protéine (0), une ORF apparaît par chance (a). Si cette ORF est traduite en un peptide, elle devient un proto-gène (b). Selon les pressions de sélection à l'oeuvre, des mutations peuvent supprimer cette traduction (a), ou modifier graduellement le proto-gène, dont la composition et le niveau d'expression approchent progressivement ceux d'un gène tandis que le produit de traduction développe de nouvelles propriétés (c). Petit à petit, le proto-gène devient un « véritable » gène, évolutivement stable (d). Ce nouveau gène peut, de manière réversible (Gerstein and Zheng 2006), devenir un pseudo-gène (Jacq, Miller et al. 1977) (e), et même retourner à l'état de région intergénique dépourvue d'ORF (0). Je représente pseudo-gènes et proto-gènes de la même façon, car leur composition nucléotidique est intermédiaire entre celles de gènes codants pour des protéines et celle de régions intergéniques, et car leurs produits de traduction ne sont probablement pas fonctionnels. Pseudo-gènes et proto-gènes diffèrent cependant fondamentalement, en ce que les pseudo-gènes ont des homologues mais les proto-gènes sont uniques.

L'opportunité de tester cette hypothèse s'est présentée lorsque les résultats de la première expérience de « profilage ribosomal » à grande échelle pour la levure ont été publiés (Ingolia, Ghaemmaghami et al. 2009). Cette expérience utilise une technique de séquençage nouvelle génération pour savoir quels fragments d'ARN sont protégés par les ribosomes, et donc *a priori* en cours de traduction. Avec l'aide de Muhammed Yildirim, étudiant post-doctoral dans le laboratoire de David Bartel (Whitehead Institute, Boston), j'ai ré-analysé les données brutes de séquençage issues de cette expérience. En appliquant des critères stricts pour limiter les faux positifs, j'ai découvert que plus de 500 ORFs courtes provenant des régions supposées intergéniques étaient protégées par les ribosomes et donc très probablement traduites. Non seulement ce résultat conforte-t-il l'hypothèse selon laquelle les protéines peuvent apparaître *de novo*, mais il pourrait avoir des implications en biologie moléculaire indépendamment de l'aspect évolutif.

Si mes résultats montrent que des ORFs courtes provenant des régions intergéniques sont vraisemblablement traduites par les ribosomes, il est néanmoins possible que les peptides correspondants soient rapidement dégradés. Si certains de ces peptides font partie du système cellulaire, comment acquièrent-ils une fonction ? Deviennent-ils des protéines à part entière, et si oui comment? Pour l'instant je ne peux répondre à ces questions que par des conjectures, inspirées de la littérature sur l'évolution *in vitro* (Keefe and Szostak 2001) ou sur les liens entre régulation de la traduction et « l'évolvabilité » (Halfmann, Alberti et al.), qui montrent que les séquences aléatoires possèdent souvent des propriétés biochimiques intéressantes exploitables en cas de stress environnemental. Quel que soit le rôle de ces peptides, même s'ils ne font qu'occuper des ribosomes pour réguler la traduction d'autres protéines, je pense qu'il est important de les prendre en compte pour modéliser le fonctionnement du système cellulaire.

DOCUMENT JOINT 2

Titre : Revisiting the *Saccharomyces cerevisiae* predicted ORFeome.

Auteurs : Li QR*, Carvunis AR*, Yu H*, Han JD*, Zhong Q, Simonis N, Tam S, Hao T, Klitgord NJ, Dupuy D, Mou D, Wapinski I, Regev A, Hill DE, Cusick ME, Vidal M

Description : article de recherche original paru dans le magazine *Genome Research* en 2008.

Contribution : Co-premier auteur de cet article, j'en ai initié la plupart et effectué la totalité des analyses bioinformatiques. QR Li avait commencé ce projet plusieurs années avant que je ne commence ma thèse et avait déjà des résultats expérimentaux. H Yu a proposé d'utiliser l'apprentissage en machine pour ce projet et me l'a enseigné. QR Li et moi avons rédigé l'article, aidées surtout de ME Cusick et M Vidal qui a supervisé le projet. Les autres auteurs ont apporté diverses contributions intellectuelles et techniques.

Letter

Revisiting the *Saccharomyces cerevisiae* predicted ORFeome

Qian-Ru Li,[1,6] Anne-Ruxandra Carvunis,[1,2,6] Haiyuan Yu,[1,6] Jing-Dong J. Han,[1,6,7] Quan Zhong,[1] Nicolas Simonis,[1] Stanley Tam,[1] Tong Hao,[1] Niels J. Klitgord,[1] Denis Dupuy,[1] Danny Mou,[1] Ilan Wapinski,[3,4] Aviv Regev,[3,5] David E. Hill,[1] Michael E. Cusick,[1] and Marc Vidal[1,8]

[1]*Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA;* [2]*TIMC-IMAG, CNRS UMR5525, Faculté de Médecine, 38706 La Tronche Cedex, France;* [3]*Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA;* [4]*School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA;* [5]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

Accurately defining the coding potential of an organism, i.e., all protein-encoding open reading frames (ORFs) or "ORFeome," is a prerequisite to fully understand its biology. ORFeome annotation involves iterative computational predictions from genome sequences combined with experimental verifications. Here we reexamine a set of *Saccharomyces cerevisiae* "orphan" ORFs recently removed from the original ORFeome annotation due to lack of conservation across evolutionarily related yeast species. We show that many orphan ORFs produce detectable transcripts and/or translated products in various functional genomics and proteomics experiments. By combining a naïve Bayes model that predicts the likelihood of an ORF to encode a functional product with experimental verification of strand-specific transcripts, we argue that orphan ORFs should still remain candidates for functional ORFs. In support of this model, interstrain intraspecies genome sequence variation is lower across orphan ORFs than in intergenic regions, indicating that orphan ORFs endure functional constraints and resist deleterious mutations. We conclude that ORFs should be evaluated based on multiple levels of evidence and not be removed from ORFeome annotation solely based on low sequence conservation in other species. Rather, such ORFs might be important for micro-evolutionary divergence between species.

[Supplemental material is available online at www.genome.org.]

Comparative genomics, involving homology searching of genome sequences between evolutionarily related species, is a powerful tool for predicting functional regions in a genome sequence without prior biological knowledge. To date, complete genome sequences are available for more than 500 different organisms across all three domains of life (Liolios et al. 2006). Comparative genomics of bacteria, yeast, worm, fly, and human have led to extensive revision of complete sets of predicted protein-encoding open reading frames (ORFs), or "ORFeomes" (McClelland et al. 2000; Brachat et al. 2003; Cliften et al. 2003; Kellis et al. 2003; Stein et al. 2003; Clamp et al. 2007; Clark et al. 2007). Removal from earlier versions of predicted ORFeomes of ORFs that are poorly or not conserved in other species ("orphan ORFs") is a critical revision proposed by these comparative genomic studies. The principle underlying removal of orphan ORFs is that selective constraints on functional DNA sequences should prevent deleterious mutations from occurring (Hardison 2003).

However, lack of evolutionary conservation does not guarantee lack of functional significance. It may be imprudent to eliminate putative ORFs from predicted ORFeomes solely based on lack of cross-species conservation. Different species, no matter how evolutionarily close, might express distinct ORF products. In support of this possibility, the pilot Encyclopedia of DNA Elements (ENCODE) project on 1% of the human genome has revealed that experimentally identified functional elements are not necessarily evolutionary constrained (Birney et al. 2007). In addition, although evolutionary conservation implies functionality for the product of a predicted ORF, experimental validation is required to demonstrate its biological significance. Therefore, cautious experimental reinvestigation of the functionality of predicted ORFs is needed to improve the accuracy of genome annotation.

To this end we set out to examine potential functionality of orphan ORFs in *Saccharomyces cerevisiae* based on available experimental evidence. Three independent comparative genomic analyses (Brachat et al. 2003; Cliften et al. 2003; Kellis et al. 2003) have predicted 648 annotated ORFs as "spurious" or "false," representing 10% of originally annotated ORFs. Notably, 10 out of these 648 orphan ORFs have since been validated as functional by small-scale experiments. For example, although YDR504C lacks clear orthologs in other yeast species, its deletion causes lethality upon exposure to high temperature while in stationary phase (Martinez et al. 2004). Given the time-consuming efforts of traditional "one-gene-at-a-time" inquiries, many predicted ORFs have not been individually characterized. However, as the first sequenced eukaryotic organism, *S. cerevisiae* has been used inten-

[6]**These authors contributed equally to this work.**
[7]**Present address: Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China.**
[8]**Corresponding author.**
E-mail marc_vidal@dfci.harvard.edu; fax (617) 632-5739.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.076661.108.

sively for functional genomics and proteomics studies, providing valuable functional evidence that allow further evaluation of coding potential of the orphan ORFs.

Using currently available functional genomics and proteomics data sets, we collate functional evidence for a significant portion of *S. cerevisiae* orphan ORFs, finding that many orphan ORFs produce detectable transcripts and/or translated products. Using a naïve Bayes model, we predict the likelihood that any *S. cerevisiae* ORF encodes a functional product and show that the number of orphan ORFs with potential functional significance is higher than expected by chance. Notably, we provide experimental verification for strand-specific transcription of many orphan ORFs. Finally, we report that interstrain intraspecies genome sequence variation is lower across orphan ORFs than in intergenic regions. Altogether our results demonstrate that orphan ORFs should not be excluded from current ORFeome annotation simply because they fail to show interspecies sequence conservation. We suggest that orphan ORFs should be included in future genome-wide experimental studies to reveal their bona fide identity either as functional ORFs or as randomly occurring misannotated ORFs.

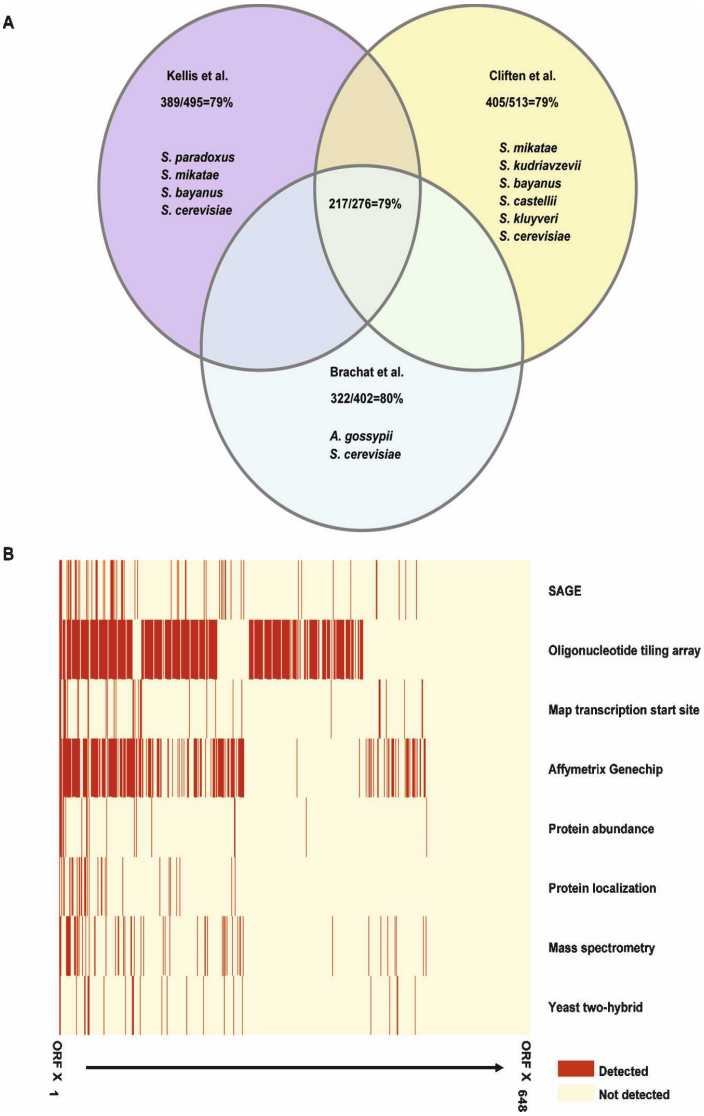


Figure 1. Experimental evidence for *S. cerevisiae* orphan ORFs. (*A*) Percentages indicate proportions of orphan ORFs detected at least in one of 13 functional genomics and proteomics data sets (Table 1). Note that ORFs rejected by all three comparative genomic studies analyzed here (Brachat et al. 2003; Cliften et al. 2003; Kellis et al. 2003) show similar percentages. (*B*) Supporting experimental evidence for each of 648 ORFs observed as orphan by three comparative genomic studies (Brachat et al. 2003; Cliften et al. 2003; Kellis et al. 2003). Complete lists of ORFs and supporting experimental evidence are in Supplemental Table 2. Columns are ordered from the ORF with most evidence (ORF X_1; *left*) to the one with the least evidence (ORF X_{648}; *right*). Data sets were grouped together by type of experimental approach, transcriptional on *top* and translational at the *bottom*. In total, there are 477 orphan ORFs with transcriptional evidence, 180 with translational evidence, and 145 with both transcriptional and translational evidence.

Results

Evidence for biological significance of *S. cerevisiae* orphan ORFs

The genome annotation of *S. cerevisiae* has undergone continuous modification through computational and experimental efforts since the original release in 1996 (Goffeau et al. 1996; Fisk et al. 2006). Three independent comparative genomic analyses compared the conservation of DNA or predicted protein sequences among several ascomycete species (Brachat et al. 2003; Cliften et al. 2003; Kellis et al. 2003), recommending that 402, 513, and 495 ORFs, respectively, be removed from the *S. cerevisiae* predicted ORFeome because their putative counterparts in other yeast species accumulate stop codons and frame-shift mutations (Fig. 1A). The union of these three comparative analyses is a set of 648 orphan ORFs called "spurious" or "false" in these studies (Fig. 1A).

High-throughput functional genomics and proteomics approaches have recently accelerated functional characterization of predicted ORFs. Several of these genome-wide approaches, such as gene-expression profiling or in vivo characterization of protein complexes, have detected transcripts or translated products of orphan ORFs. For example, in a proteome-wide purification of yeast protein complexes (Krogan et al. 2006), 85 proteins identified by mass spectrometry were encoded by orphan ORFs.

To provide a systematic reanalysis of *S. cerevisiae* orphan

ORFs, we collected 13 large-scale studies (Table 1) informing on either transcription or translation of orphan ORFs. The transcriptome studies included tiling arrays (David et al. 2006), high-density Affymetrix chip analysis (Holstege et al. 1998), SAGE analysis (Velculescu et al. 1997), and cDNA sequencing (Miura et al. 2006). Because many (69%) of the orphan ORFs overlap with another annotated ORF, we only included transcriptome studies able to detect strand-specific transcripts. Protein–protein interaction studies included proteome-scale yeast two-hybrid screens (Uetz et al. 2000; Ito et al. 2001) and affinity pull-downs of tagged proteins followed by mass spectrometry (Gavin et al. 2002, 2006; Ho et al. 2002; Krogan et al. 2006). For yeast two-hybrid studies, we considered an ORF being translated only if its product was involved in a protein–protein interaction as a prey. Protein expression studies included global surveys of protein abundance (Ghaemmaghami et al. 2003) and subcellular localization (Kumar et al. 2002; Huh et al. 2003).

Out of the 648 orphan ORFs, most (79%) have been detected in at least one of these data sets. The proportion of orphan ORFs detected was nearly the same for ORFs rejected by each of the three comparative genomics analyses independently (80% for Brachat, 79% for Cliften, and 79% for Kellis) and for the 276 orphan ORFs discarded by all three (79%) (Fig. 1A). Among the 648 orphan ORFs, many were detected by more than one approach. In total, 145 orphan ORFs (22%) were both detected as transcripts and translated products (Fig. 1B). A similar distribution of functional evidence was observed for the orphan ORFs rejected by all three comparative genomic analyses (Supplemental Fig. 1).

Evaluating biological significance of *S. cerevisiae* ORFs by a naïve Bayes approach

High-throughput approaches have inherently limited coverage (not all ORFs are detectable) and precision (detection of some ORFs might be artifactual). Therefore information from large-scale data sets needs to be accepted cautiously. We chose a naïve Bayes model to quantify the observations reported above, because this approach can integrate dissimilar types of data sets into a common probabilistic framework with maximal coverage and precision (Jansen et al. 2003; Yu et al. 2004). By use of such an integration scheme, evidence (i.e., features) from several data types can be accumulated to estimate with increasing confidence the likelihood that an ORF encodes a functional product.

As with any machine learning algorithm, naïve Bayes models need a training set of gold standard positives (GSPs) and negatives (GSNs). The *Saccharomyces* Genome Database (SGD), the arbiter of genome annotation for budding yeasts, has categorized all *S. cerevisiae* ORFs into three major groups based on conservation across species and on available experimental characterization: "verified" (4449 ORFs), "uncharacterized" (1333 ORFs), and "dubious" (823 ORFs) (Fisk et al. 2006). Both verified ORFs and uncharacterized ORFs are conserved across species. Verified ORFs have clear small-scale experimental evidence for the existence of functional ORF products, but uncharacterized ORFs do not. Dubious ORFs are thought not to encode a functional product due to (1) lack of conservation across species, and/or (2) absence of any small-scale experiment demonstrating detectable mRNA or protein production or phenotypic effects. We used all 4449 verified ORFs as the GSPs and all 823 dubious ORFs as the GSNs. Although ideally the GSNs should be depleted of functional ORFs, this cannot exactly be true for the dubious set. However, the dubious set is likely enriched with nonfunctional ORFs. It is common practice to use an "enriched" set of negatives in training data sets (Miller et al. 2005; Xia et al. 2006).

We calculated the ratio of the fraction of GSPs present in each of the 13 functional genomics and proteomics data sets divided by the fraction of GSNs present in each data set, which measures the confidence levels (Supplemental Table 1). The product of these ratios of the 13 data sets for each ORF is defined as the likelihood ratio (*LR*) of an ORF, i.e., the likelihood of each ORF to encode a functional product (see Methods). We used the base 10 logarithmic form of *LR* (*LLR*) as final prediction scores (Supplemental Table 2). Out of the large-scale studies integrated, several did measure similar biological features of ORFs and ORF products. However, we treated all 13 data sets as independent features, due to the low correlation between them (Supplemental Tables 3, 4).

To evaluate the performance of the naïve Bayes model, we used threefold cross-validation (see Methods). After randomly dividing both the GSPs and GSNs into three separate equal sets, we used two of the three sets as the training set to calculate *LLR*s and the remaining set as the test set to identify positives and negatives. The true-positive rate (TP rate: fraction of GSPs that are predicted to be functional) and the false-positive rate (FP rate: fraction of GSNs that are predicted to be functional) were calculated at different *LLR* cutoffs. The resulting couplets (TP rate–FP rate) were used to plot a receiver operating characteristic (ROC) curve. We ran this process three times so that each of the three sets was a test set and the remaining two constituted the training set. Each ROC curve looked similar (Supplemental Fig. 2), which

Table 1. Thirteen functional genomics and proteomics data sets integrated in our analysis

Functional genomics and proteomics data sets	Evidence detected	Approach category
Velculescu et al. 1997: Transcriptome characterized by SAGE	mRNA transcript	SAGE
David et al. 2006: Transcriptome characterized by oligonucleotide tiling array	mRNA transcript	Oligonucleotide tiling array
Miura et al. 2006: Full-length cDNA analysis	mRNA transcript	Map transcription start site
Holstege et al. 1998: Measurement of the transcripts abundance	mRNA transcript	Affymetrix GeneChip
Ghaemmaghami et al. 2003: Expression of TAP-tagged proteins	Protein expression	Protein abundance
Huh et al. 2003: Localization of GFP-tagged proteins	Protein localization	Protein localization
Kumar et al. 2002: Subcellular localization of transposon-tagged proteins	Protein localization	Protein localization
Gavin et al. 2002: Protein complexes characterization	Peptide sequence	Mass spectrometry
Ho et al. 2002: Protein complexes characterization	Peptide sequence	Mass spectrometry
Gavin et al. 2006: Protein complexes characterization	Peptide sequence	Mass spectrometry
Krogan et al. 2006: Protein complexes characterization	Peptide sequence	Mass spectrometry
Ito et al. 2001: Protein–protein interaction mapping by yeast two-hybrid	Protein physical interaction	Yeast two-hybrid
Uetz et al. 2000: Protein–protein interaction mapping by yeast two-hybrid	Protein physical interaction	Yeast two-hybrid

validated the overall quality of our training set. A final ROC curve was plotted by using potential *LLR* cutoffs from all three training subsets and their associated TP rate and FP rate based on the predictions from the complete training set (Fig. 2A). The significant deviation of the final ROC curve from the 45° random ROC line indicates that our model has substantial predictive value (area under ROC curve = 0.982). To assess the contribution of each data set to the final prediction scores, we successively omitted one data set and repeated the training and cross-validation procedures. We plotted ROC curves for all procedures (Supple-

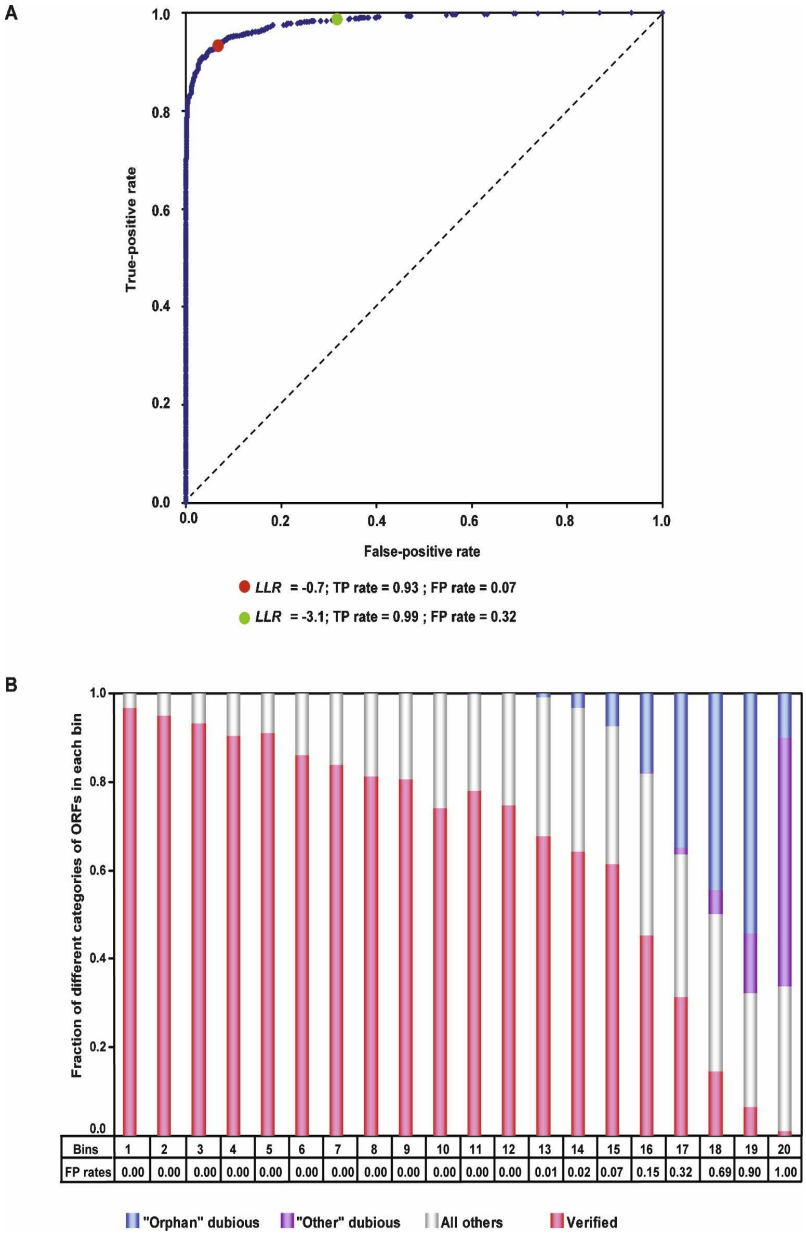


Figure 2. Evaluating functionality of *S. cerevisiae* ORFs. (*A*) ROC curve (blue) for naïve Bayes predictions based on 13 functional genomics and proteomics data sets. The diagonal (black dotted line) is the expected ROC curve for random, where the TP rate equals the FP rate. The two *LLR* cutoffs highlighted on the curve were used later as thresholds for categorizing orphan ORFs. (*B*) All 6718 *S. cerevisiae* ORFs were divided into 20 bins by decreasing *LLR*. Each bin has similar numbers of ORFs. The false-positive rates associated with the minimum *LLR* in each bin are listed. Distributions of verified ORFs, orphan dubious ORFs, "other" dubious ORFs, and all other ORFs in each bin are shown. Orphan dubious ORFs tend to have a higher *LLR* than ORFs classified as dubious for other reasons.

mental Fig. 3) and observed little difference when excluding any single data set. Thus it seems that no single data set dominates the prediction.

We divided all 6718 *S. cerevisiae* ORFs into 20 bins ranked by decreasing *LLR*, with each bin containing similar numbers of ORFs. Verified ORFs localized mostly in the higher *LLR* bins (92.5% of all verified ORFs distributed between bin 1 and bin 15), while dubious ORFs localized in lower *LLR* bins (only 4.98% of dubious ORFs distributed between bin 1 and bin 15) (Fig. 2B). Such segregation between verified ORFs and dubious ORFs was expected, given that the ORFs used in the training as GSPs (verified ORFs) are bound to have a higher *LLR* than the ones used in the training as GSNs (dubious ORFs). An unanticipated result of the naïve Bayes predictions is that orphan dubious ORFs have overall higher *LLR* ($P < 10^{-15}$ by Mann-Whitney *U* test) (Fig. 2B) than ORFs classified as dubious for reasons other than strict lack of interspecies sequence conservation (e.g., a mutant phenotype described for the ORF could be ascribed to mutation of an overlapping well-characterized ORF) (Fisk et al. 2006). This suggests that orphan dubious ORFs might be more likely to encode functional products than "other" dubious ORFs.

For an ORF to be considered "most-likely" functional in our naïve Bayes predictions, its posterior odds (the product of the prior odds and the likelihood ratio) has to be larger than 1 (see Methods). We can estimate that the prior odds for any given ORF to be most-likely functional is ~5.4 (4449 GSPs divided by 823 GSNs). Hence, we used $LLR = \log_{10}1/5.4 = -0.7$ (FP rate = 0.07) as the cutoff for an ORF to be most-likely functional (bins 1–15). Among the 648 orphan ORFs, 54 ORFs with $LLR \geq -0.7$ were thus assigned to a set of most-likely functional orphan ORFs. Although the percentage of verified ORFs decreased significantly from bin 16 to bin 20 compared with the first 15 bins (Fig. 2B), there were still 3.4% and 2.5% of verified ORFs (152 and 111 ORFs) in bins 16 and 17, respectively. We classified the 199 orphan ORFs in bins 16 and 17, with an *LLR* between -0.7 (FP rate = 0.07) and -3.1 (FP rate = 0.32), as "moderately-likely" to encode a functional product. The remaining 395 orphan ORFs distributed between bins 18 and 20 were called "least-likely" functional ORFs. Detectability limitations in the large-scale data sets integrated in our predictions may have biased against these least-likely ORFs. Integration of new lines of experimental evidence in the future could still potentially identify promising functional ORF candidates among the least-likely ORFs.

Experimental evidence for expression of *S. cerevisiae* orphan ORFs

We next experimentally measured mRNA expression for orphan ORFs using reverse transcription–polymerase chain reaction (RT-PCR) (Fig. 3A). Strand specificity was needed to ensure that the transcripts detected were transcribed from the predicted DNA strand and to exclude artifacts caused from read-through transcription on the opposite strand (Craggs et al. 2001).

We tested strand specificity on two verified *S. cerevisiae* ORFs that both contain introns: YER133W (*GLC7*) and YBR078W (*ECM33*) (see Methods). Given the presence of introns in these ORFs, the sense-strand transcripts should be appreciably shorter in length than the antisense-strand transcripts. Spliced transcripts of the expected sizes were obtained in reactions where strand-specific primer was added for cDNA synthesis (Fig. 3B). No RT-PCR products were obtained in reactions without RT, demonstrating absence of contaminating genomic DNA in the poly(A) mRNA template preparation. No RT-PCR products were observed in the absence of cDNA primer for first-strand cDNA synthesis, demonstrating that the second step of standard PCR amplification contained no active reverse transcriptase for the synthesis of incorrect strand cDNA from antisense strand–specific primer. The identities of RT-PCR products were confirmed by sequencing.

Thereafter we applied our strand-specific RT-PCR to 201 orphan ORFs that do not overlap any other annotated ORF. The requirement for nonoverlap further reduces the false-positive rate, because it is less likely that there would be any transcription from the incorrect strand. Among 201 nonoverlapping orphan ORFs tested under conditions of growth on rich media, RT-PCR products of expected size were obtained for 105 ORFs (Supplemental Table 2). Although the available supporting experimental evidence for these 105 ORFs is not strikingly different from the ORFs whose transcripts were not detected by strand-specific RT-PCR (Supplemental Fig. 4), the detected ORFs have a significantly higher average *LLR* (-3.4 ± 0.2) than the ones undetected by RT-PCR (-3.8 ± 0.2, $P = 0.03$ by Mann-Whitney *U* test) (Fig. 3C), demonstrating the validity and robustness of our predictions for positives. In particular, YJL199C, a dubious ORF, has the highest *LLR* among 201 tested ORFs and was detected by RT-PCR. YJL199C was recently predicted to encode a metabolic protein based on large-scale protein–protein interaction studies (Samanta and Liang 2003).

Notably, out of 49 orphan ORFs tested that had not been detected by any of the 13 data sets (Table 1), 29 were expressed (Supplemental Table 2), among which YPR096C was recently found to encode a ribosome-interacting protein (Fleischer et al. 2006) and YOR235W was shown through a genome-wide phenotypic analysis to be involved in DNA recombination events (Alvaro et al. 2007). Therefore, we suggest that more experimentation is needed before rejecting ORFs from the *S. cerevisiae* ORFeome annotation.

Interstrain intraspecies sequence conservation for *S. cerevisiae* orphan ORFs

The available experimental evidence from large-scale data sets, combined with our experimental support for many orphan ORFs, implies that lack of interspecies conservation does not necessarily dispel the bona fide functionality of an ORF. Functional orphan ORFs may have a relaxed selective constraint due to their dispensable roles in other species and may therefore rapidly lose sequence similarity even in closely related species (Schmid and Aquadro 2001). However, select species–specific functions may stringently constrain sequence divergence of functional orphan ORFs within species (Domazet-Loso and Tautz 2003). Therefore, we examined the intraspecies conservation of orphan ORFs in *S. cerevisiae,* using single nucleotide polymorphism (SNP) information from genome resequencing of multiple strains of *S. cerevisiae* by the *Saccharomyces* Genome Resequencing Project (SGRP) (http://www.sanger.ac.uk/Teams/Team71/durbin/sgrp/index.shtml). Among the 37 currently available strain sequences, four (SK1, W303, Y55, and DBVPG6765) have been sequenced at twofold coverage or higher. We used the SNP data from these four genomes to assess nucleotide variation in different genomic regions across *S. cerevisiae* strains. We compared nucleotide divergence among three genomic features: orphan ORFs, nonorphan ORFs, and intergenic regions, considering only the regions that do not overlap with any other annotated ORF (see Methods).

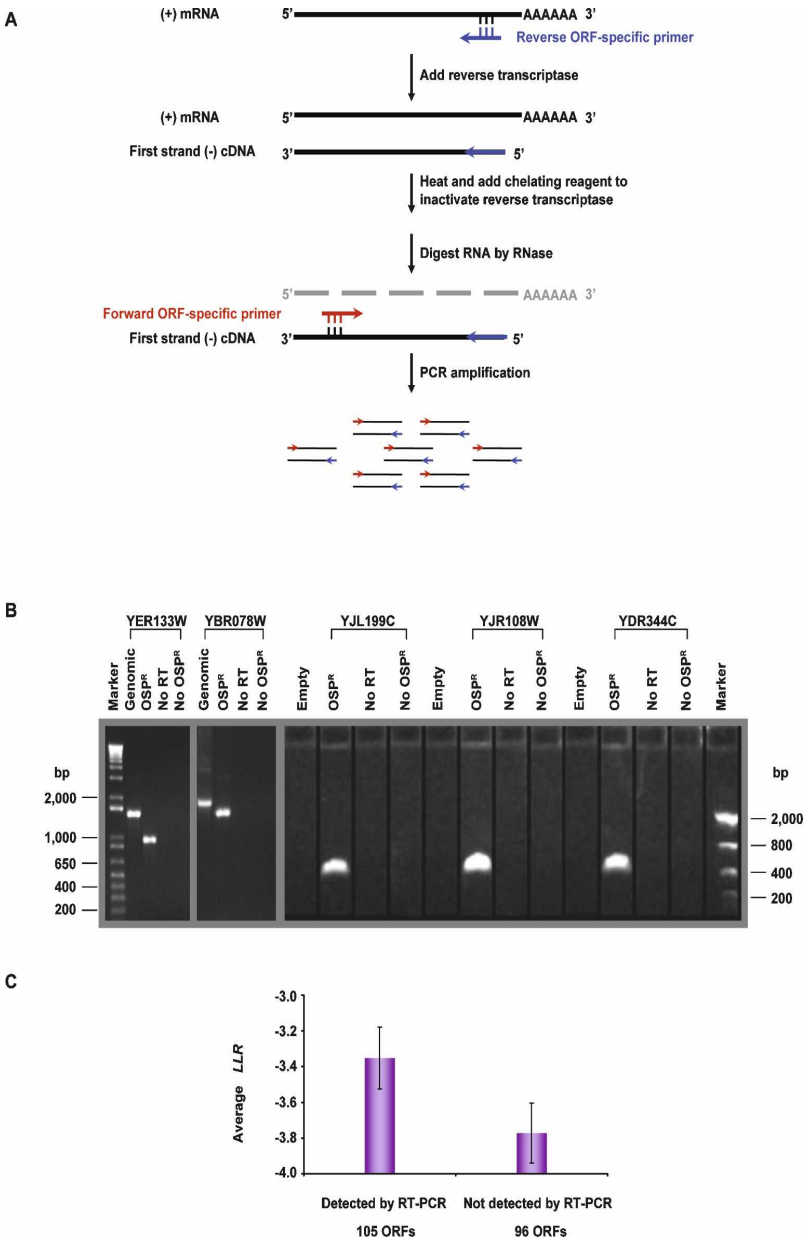


Figure 3. Two-step strand-specific RT-PCR. (*A*) Schematic diagram of the strand-specific RT-PCR procedure. (*B*) Electrophoretic analysis of strand-specific RT-PCR products. Reverse ORF-specific primers (OSP^R), with sequences complementary to the ORF-coding strand, were used for first-strand cDNA synthesis. Second-step PCR amplifications used a pair of forward (OSP^F) and reverse ORF-specific primers (OSP^R). As controls, the first step of RT-PCR was performed without reverse transcriptase for detecting contamination by genomic DNA, or without the OSP^R primer for detecting residual reverse transcriptase activity in second-step PCR reactions. Two intron-containing verified ORFs, YER133W (genomic DNA length: 1464 bp; coding sequence length: 939 bp) and YBR078W (genomic DNA length: 1737 bp; coding sequence length: 1407 bp), were used to test the strand specificity. An extra control for these two verified ORFs was a standard PCR action using yeast genomic DNA as template and the same pair of ORF-specific primers. The observed difference in the length of PCR products amplified from genomic DNA versus poly(*A*) mRNA manifested the strand specificity. Strand-specific RT-PCR results of 201 nonoverlapping orphan ORFs were analyzed on 1% agarose E-gel (Invitrogen). Of the reactions 53% (105 ORFs) gave rise to visible RT-PCR products of the expected sizes. Three orphan ORFs, YJL199C (327 bp), YJR108W (372 bp), and YDR344C (444 bp), are shown as examples of successful RT-PCR reactions. (*C*) Comparison of the average *LLR* between nonoverlapping ORFs detected and undetected by strand-specific RT-PCR. Error bars, SEM.

Across the four strains analyzed, orphan ORFs showed higher nucleotide divergence (7.0 ± 0.4 SNPs per kb) than did nonorphan ORFs (3.7 ± 0.1 SNPs per kb, $P < 10^{-5}$ by Mann-Whitney *U* test), but less than intergenic regions (15.5 ± 0.2 SNPs per kb, $P < 10^{-15}$ by Mann-Whitney *U* test) (Fig. 4A). Such intermediate nucleotide divergence for orphan ORFs suggests that at least a portion of them are subject to significant intraspecies evolutionary constraints. Such "interstrain intraspecies" conservation of orphan ORFs indicates potential functionality of an ORF in addition to experimental evidence.

Among the 648 orphan ORFs, the most-likely functional ones displayed a significantly lower nucleotide divergence (3.8 ± 0.7 SNPs per kb) than both moderately-likely (6.4 ± 0.8 SNPs per kb, $P = 0.016$ by Mann-Whitney *U* test) and least-likely ORFs (7.7 ± 0.6 SNPs per kb, $P = 0.005$ by Mann-Whitney *U* test) (Fig. 4B). Although the moderately-likely category does have a lower nucleotide divergence than least-likely category, the difference is not significant ($P > 0.05$ by Mann-Whitney *U* test). Because different types of SNPs, such as synonymous or nonsynonymous substitutions, could have distinct effects on an ORF product, we applied another test to compare sequence conservation among the three groups, measuring the percentage of ORFs with preserved reading frames (absence of stop codons or frameshift mutations) across all four *S. cerevisiae* strains. A decreasing trend was observed from most-likely to least-likely ORFs (Fig. 4C), with significant differences among the three categories

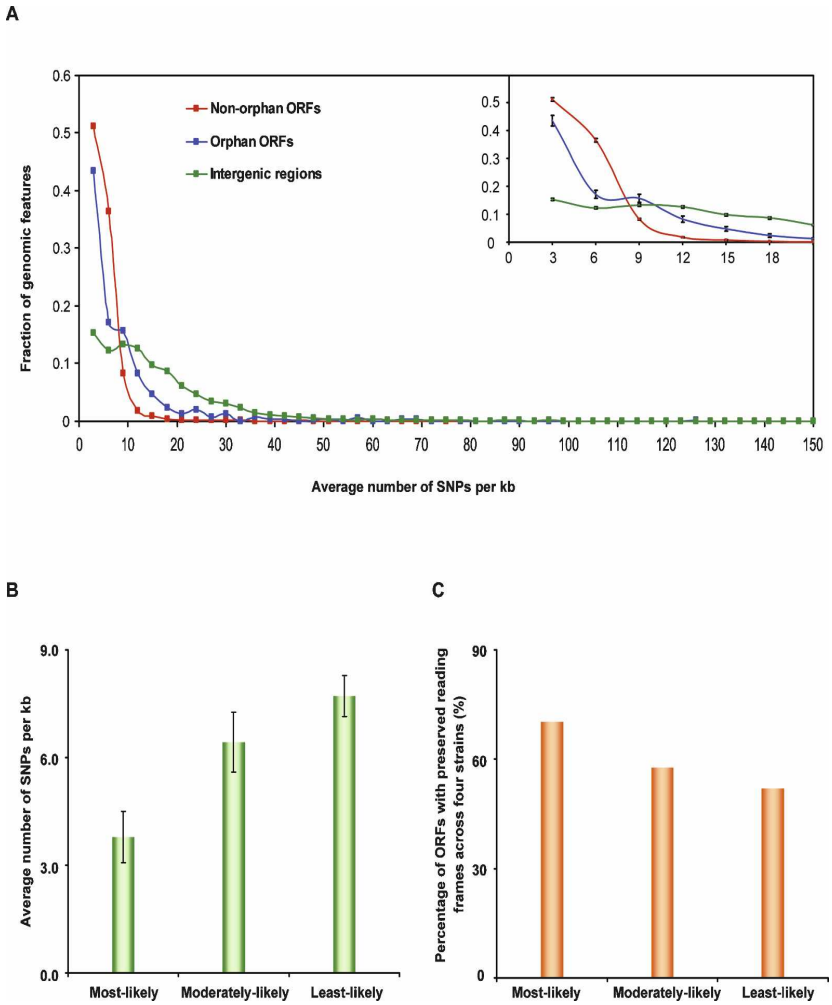


Figure 4. Interstrain intraspecies sequence conservation for orphan ORFs. (*A*) Distribution of nucleotide divergence in different genomic features. We binned three types of genomic features, (1) non-orphan ORFs (red curve), (2) orphan ORFs predicted by three comparative genomic analyses (blue curve) (Brachat et al. 2003; Cliften et al. 2003; Kellis et al. 2003), and (3) intergenic regions (green curve), using a window of an average three SNPs per kb across four *S. cerevisiae* strains. Each dot represents the fraction of genomic features in each bin. Numbers on the *X*-axis represent the maximum number of SNPs per kb in each bin. For instance the first bin collects the genomic regions that have between zero and three SNPs per kb in four strains. The inset zooms in on the 0–21 SNPs per kb range with SEM displayed. (*B*) Comparison of nucleotide divergence among three predicted categories of orphan ORFs based on their *LLR*s. Error bars, SEM in each category. (*C*) Comparison of the percentage of ORFs among the three predicted categories of orphan ORFs that have reading frames preserved across four *S. cerevisiae* strains.

($P = 0.03$ by χ^2 test). The coexistence of high interstrain intraspecies conservation with high likelihood of functionality demonstrates that some orphan ORFs face functional constraints that protect them from deleterious intraspecies mutations.

In summary, analysis of nucleotide variation in multiple *S. cerevisiae* strains, combined with multiple lines of experimental evidence, suggest that reevaluation of the functionality of all ORFs, especially orphan ORFs, is warranted.

Discussion

We report here that many interspecies nonconserved ORFs or orphan ORFs predicted by comparative genomic analyses in *S. cerevisiae* show evidence of transcription or translation, as reported in various functional genomics or proteomics data sets. We used a naïve Bayes probabilistic integration of a heterogeneous set of large-scale data sets to predict the likelihood that a predicted ORF encodes a functional product. Threefold cross-validation demonstrated high performance for this approach, which revealed that orphan ORFs are more likely functional than are ORFs classified as dubious for reasons other than strict lack of sequence conservation across species. Independent strand-specific RT-PCR confirmed that many orphan ORFs are indeed expressed. Although presence of transcripts is not sufficient by itself to conclude that an ORF encodes a functional product, the correspondence between our RT-PCR results and naïve Bayes prediction scores demonstrated both the potential functionality of orphan ORFs and the robustness of our prediction method. Confirming that orphan ORFs could be functional, many show signs of interstrain intraspecies negative selection, such as lower nucleotide divergence than intergenic regions and retaining an intact reading frame in multiple *S. cerevisiae* strains.

Collectively our findings argue that the likelihood that an ORF encodes a functional product is best evaluated by combining multiple lines of experimental and evolutionary evidence (Snyder and Gerstein 2003). The potential functionality of orphan ORFs in *S. cerevisiae* suggests that experimentally verified functional sequences are not always conserved across species. Such nonconserved functional sequences might be responsible for species-specific phenotypic differences, making *S. cerevisiae* "cerevisiae" and not some other species in the *Saccharomyces* genus. An alternative explanation is that there are some functional elements evolving neutrally and conferring no specific benefit to the organism (Birney et al. 2007). Either way, experimental investigation has an irreplaceable role in determining biologically relevant DNA sequences. Comparative genomics has demonstrated analytic power in predicting functional regions before availability of any experimental information (Hardison 2003). When experimental information does become available (mainly from high-throughput functional genomics and proteomics analyses), then its integration should revise the genome annotation accordingly. The naïve Bayes model implemented here can be readily applied to all organisms.

Although we provide confidence scores about the likelihood of a predicted ORF to encode a functional product, comprehensive functional characterization of an ORF needs more concrete evidence from genetics, cell biology, and biochemistry than simple evidence of transcription or translation. The functional genomics or proteomics data sets used in our naïve Bayes predictions only investigated a few growth conditions, generally growth on rich media, limiting investigation of functions unique to the development and physiology of *S. cerevisiae*. Given the limited functional information obtained so far under laboratory conditions about uncharacterized ORFs (Pena-Castillo and Hughes 2007), perhaps what is needed are studies of yeast cells outside the laboratory. Upon such a shift, data sets generated under diverse conditions will become available, and our approach will then be available to aid precise and powerful annotation of genomes.

Methods

Large-scale data sets analysis

We collected 13 published functional genomics and proteomics data sets of *S. cerevisiae*, summarized in Table 1 with references to the data sources. Only ORFs identified by the same primary SGDID in the publication and in the January 2007 version of SGD annotation were included. We assigned "presence" or "absence" of transcript or translated product of every ORF in each data set. For protein complexes characterization data sets (Gavin et al. 2002, 2006; Ho et al. 2002; Krogan et al. 2006) all proteins that were identified as peptides were considered "present," independent of further filtration by the investigators. For high-throughput yeast two-hybrid (Uetz et al. 2000; Ito et al. 2001), only proteins identified as preys were considered present. Only protein–protein interactions classified as "core" by Ito et al. (2001) were included. Transcripts identified by SAGE (Velculescu et al. 1997) and assigned to "class 1" by the investigators were considered present; all others, absent. We divided the Affymetrix Genechip data (Holstege et al. 1998) into two groups: intensity of expression strictly positive but less than or equal to 1, and intensity strictly more than 1. These two groups were treated separately in the naïve Bayes model. The normalized intensity of expression per probe (David et al. 2006) was averaged, and the percentage of probes whose intensity was higher than this average was considered as the intensity of expression of each ORF. We then extracted four groups (undetected, intensity strictly positive but less than 0.4, intensity strictly more than or equal to 0.4 but less than 0.8, and intensity strictly more than or equal to 0.8) that were treated separately in the naïve Bayes model. The remaining data sets were not reprocessed.

The naïve Bayes model

If the numbers of positives are known among the total number of ORFs, the "prior" odds of finding a positive are

$$O_{prior} = \frac{P(pos)}{P(neg)} = \frac{P(pos)}{1 - P(pos)}.$$

The "posterior" odds are the odds of finding a positive after considering N different feature data sets with values $f_1 \ldots f_N$:

$$O_{post} = \frac{P(pos|f_1 \ldots f_N)}{P(neg|f_1 \ldots f_N)}.$$

The likelihood ratio LR is defined as

$$LR(f_1 \ldots f_N) = \frac{P(f_1 \ldots f_N|pos)}{P(f_1 \ldots f_N|neg)}.$$

According to Bayes rule, the posterior odds can be expressed as

$$O_{post} = LR(f_1 \ldots f_N)O_{prior}.$$

If the N features are conditionally independent, LR can be simplified to

$$LR(f_1 \ldots f_N) = \prod_{i=1}^{N} L(f_i) = \prod_{i=1}^{N} \frac{P(f_i|pos)}{P(f_i|neg)}.$$

LR can be computed from contingency tables relating positive and negative examples with the *N* features (we binned the feature values $f_1 \ldots f_N$ into discrete intervals). Since O_{prior} is a fixed value, O_{post} is determined by *LR*. We used log-likelihood ratio ($\log_{10}$ *LR* or *LLR*) as the final prediction score. The higher the *LLR* of a certain ORF, the more likely it is a positive, i.e., a functional ORF.

Threefold cross-validation

We divided the whole training set into three subsets randomly. We then trained the model with two subsets and tested its performance on the third subset. We repeated this step three times so that each subset was used once to test the performance. We calculated the ROC curve with the predictions for the whole training set by combining the results from the three repeated tests.

Strand-specific RT-PCR

S. cerevisiae strain S288C was grown in yeast extract-peptone-dextrose (YPD) medium at 30°C to mid-exponential phase. Yeast cells were then harvested and used for total RNA isolation with an RNeasy kit (Qiagen). Poly(A) RNA was subsequently enriched by Oligotex mRNA kit (Qiagen). Before RT-PCR experiments, Poly(A) RNA was subjected to DNA-free DNase treatment (Ambion) to eliminate genomic DNA contamination. Genomic DNA was extracted from yeast culture by the DNeasy blood and tissue kit (Qiagen). We modified a strand-specific RT-PCR method previously described (Craggs et al. 2001), using the GeneAmp thermostable rTth reverse transcriptase RNA PCR kit (Applied Biosystems). DNase-treated poly(A) RNA sample (25 ng) was denatured for 5 min at 70°C with 2 µL of 10× rTth reverse transcriptase buffer and 1 µL of 10 µM reverse ORF-specific primer complementary to the ORF-coding strand (OSP^R). While the template and the primer were still incubating at 70°C, a preheated reaction mixture was added consisting of 2 µL of 10 mM $MnCl_2$ solution, 1.6 µL of 10 mM dNTP mix, and 2.5U of rTth polymerase. The temperature was lowered for 2 min to 55°C for annealing and then raised for 30 min to 70°C for the first-strand cDNA synthesis. After the cDNA synthesis, 20 µL of prewarmed 1× chelating buffer was added to chelate Mn^{2+} followed by heating the mixture for 30 min at 98°C to inactivate the reverse transcriptase activity of rTth. Second-step PCR reactions were performed in a 50-µL reaction volume using one-tenth of the synthesized first-strand cDNA as template, forward ORF-specific primer (OSP^F) and OSP^R as primers, and one unit of High Fidelity Platinum Taq polymerase (Invitrogen). The OSP^R complementary to the ORF-coding strand was used in both first-strand cDNA synthesis and second-step PCR amplification. The OSP^F complementary to the opposite strand was used only in the second-step PCR amplification. Both OSP^R and OSP^F were designed using the OSP Program (Hillier and Green 1991). The OSP^R starts from the last nucleotide of the termination codon, while the OSP^F starts from A of the ATG initiation codon. Primers used for RT-PCR of 201 nonoverlapping orphan ORFs are listed in Supplemental Table 5.

Interstrain intraspecies conservation analysis

SNP information from the four strains SK1, Y55, DBVPG6765, and W303 were extracted from the website of the Sanger Institute *Saccharomyces* Genome Resequencing Project (http://www.sanger.ac.uk/Teams/Team71/durbin/) on September 18, 2007 (R. Durbin and E. Louis, pers. comm.). The preassembly SNPs were taken into account only when their quality was "confirmed." They were mapped to the ORFeome of the reference strain S288C as annotated by SGD on January 2007, as well as to intergenic regions that are annotated as "not feature" (ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/intergenic/NotFeature.fasta.gz). The nucleotide divergence of each ORF was then computed by averaging the number of SNPs per kb found in each of the four strains, counting insertions and deletions as one event independently of their length. For overlapping ORFs, only the regions unique to the ORFs themselves were considered for counting SNPs. To be considered as a preserved reading frame in our analysis, the ORF had to show neither stop codons nor frame-shift mutations in any of the four strains. The reading frame of an ORF was not considered preserved if the ORF had an insertion or deletion (indel) longer or equal to 20 bp, no matter whether the indel caused a frame-shift or not.

Acknowledgments

We thank R. Durbin and E. Louis for providing SNP information and F. Roth (HMS) for helpful discussions. We thank the members of the Vidal Lab and the Center for Cancer Systems Biology (CCSB) for their scientific and technical support, especially M. Boxem, K. Venkatesan, M. Yildirim, K. Salehi-Ashtiani, M. Dreze, S. Milstein, and C. Fraughton. This work was supported by an Ellison Foundation grant awarded to M.V. and by Institute Sponsored Research funds from the Dana-Farber Cancer Institute Strategic Initiative awarded to M.V. and CCSB.

References

Alvaro, D., Lisby, M., and Rothstein, R. 2007. Genome-wide analysis of Rad52 foci reveals diverse mechanisms impacting recombination. *PLoS Genet.* **3:** e228. doi: 10.1371/journal.pgen.0030228.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Brachat, S., Dietrich, F., Voegeli, S., Zhang, Z., Stuart, L., Lerch, A., Gates, K., Gaffney, T., and Philippsen, P. 2003. Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. *Genome Biol.* **4:** R45. doi: 10.1186/gb-2003-4-7-r45.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M.F., Kellis, M., Lindblad-Toh, K., and Lander, E.S. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci.* **104:** 19428–19433.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:** 203–218.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301:** 71–76.

Craggs, J.K., Ball, J.K., Thomson, B.J., Irving, W.L., and Grabowska, A.M. 2001. Development of a strand-specific RT-PCR based assay to detect the replicative form of hepatitis C virus RNA. *J. Virol. Methods* **94:** 111–120.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. 2006. A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci.* **103:** 5320–5325.

Domazet-Loso, T. and Tautz, D. 2003. An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res.* **13:** 2213–2219.

Fisk, D.G., Ball, C.A., Dolinski, K., Engel, S.R., Hong, E.L., Issel-Tarver, L., Schwartz, K., Sethuraman, A., Botstein, D., Cherry, J.M., et al. 2006. *Saccharomyces cerevisiae* S288C genome annotation: A working hypothesis. *Yeast* **23:** 857–865.

Fleischer, T.C., Weaver, C.M., McAfee, K.J., Jennings, J.L., and Link, A.J. 2006. Systematic identification and functional screens of

uncharacterized proteins associated with eukaryotic ribosomal complexes. *Genes & Dev*. **20:** 1294–1307.
Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.-M., Cruciat, C.-M., et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415:** 141–147.
Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., et al. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440:** 631–636.
Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. 2003. Global analysis of protein expression in yeast. *Nature* **425:** 737–741.
Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. 1996. Life with 6000 genes. *Science* **274:** 546–567.
Hardison, R.C. 2003. Comparative genomics. *PLoS Biol.* **1:** e58. doi: 10.1371/journal.pbio.0000058.
Hillier, L. and Green, P. 1991. OSP: A computer program for choosing PCR and DNA sequencing primers. *PCR Methods Appl*. **1:** 124–128.
Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415:** 180–183.
Holstege, F.C.P., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95:** 717–728.
Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. 2003. Global analysis of protein localization in budding yeast. *Nature* **425:** 686–691.
Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci*. **98:** 4569–4574.
Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302:** 449–453.
Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.
Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., et al. 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440:** 637–643.
Kumar, A., Agarwal, S., Heyman, J.A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., et al. 2002. Subcellular localization of the yeast proteome. *Genes & Dev*. **16:** 707–719.
Liolios, K., Tavernarakis, N., Hugenholtz, P., and Kyrpides, N.C. 2006. The Genomes On Line Database (GOLD) v.2: A monitor of genome projects worldwide. *Nucleic Acids Res*. **34:** D332–D334.
Martinez, M.J., Roy, S., Archuletta, A.B., Wentzell, P.D., Anna-Arriola, S.S., Rodriguez, A.L., Aragon, A.D., Quinones, G.A., Allen, C., and Werner-Washburne, M. 2004. Genomic analysis of stationary-phase and exit in *Saccharomyces cerevisiae*: Gene expression and identification of novel essential genes. *Mol. Biol. Cell* **15:** 5295–5305.
McClelland, M., Florea, L., Sanderson, K., Clifton, S.W., Parkhill, J., Churcher, C., Dougan, G., Wilson, R.K., and Miller, W. 2000. Comparison of the *Escherichia coli* K-12 genome with sampled genomes of a *Klebsiella pneumoniae* and three *Salmonella enterica* serovars, Typhimurium, Typhi and Paratyphi. *Nucleic Acids Res*. **28:** 4974–4986.
Miller, J.P., Lo, R.S., Ben-Hur, A., Desmarais, C., Stagljar, I., Noble, W.S., and Fields, S. 2005. Large-scale identification of yeast integral membrane protein interactions. *Proc. Natl. Acad. Sci*. **102:** 12123–12128.
Miura, F., Kawaguchi, N., Sese, J., Toyoda, A., Hattori, M., Morishita, S., and Ito, T. 2006. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc. Natl. Acad. Sci*. **103:** 17846–17851.
Pena-Castillo, L. and Hughes, T.R. 2007. Why are there still over 1000 uncharacterized yeast genes? *Genetics* **176:** 7–14.
Samanta, M.P. and Liang, S. 2003. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc. Natl. Acad. Sci*. **100:** 12579–12583.
Schmid, K.J. and Aquadro, C.F. 2001. The evolutionary analysis of "orphans" from the *Drosophila* genome identifies rapidly diverging and incorrectly annotated genes. *Genetics* **159:** 589–598.
Snyder, M. and Gerstein, M. 2003. Defining genes in the genomics era. *Science* **300:** 258–260.
Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. 2003. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol*. **1:** e45. doi: 10.1371/journal.pbio.0000045.
Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403:** 623–627.
Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E., Hieter, P., Vogelstein, B., and Kinzler, K.W. 1997. Characterization of the yeast transcriptome. *Cell* **88:** 243–251.
Xia, Y., Lu, L.J., and Gerstein, M. 2006. Integrated prediction of the helical membrane protein interactome in yeast. *J. Mol. Biol*. **357:** 339–349.
Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.-D.J., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. 2004. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res*. **14:** 1107–1118.

Received January 29, 2008; accepted in revised form May 5, 2008.

DOCUMENT JOINT 3

Titre : Proto-genes.

Auteurs : Anne-Ruxandra Carvunis, Ilan Wapinski, Muhammed A. Yildirim, Nicolas Simonis, Benoit Charloteaux, Cesar Hidalgo, Laurent Trilling, Nicolas Thierry-Mieg, Michael E. Cusick, Marc Vidal

Description : Le brouillon le plus récent d'article décrivant mon hypothèse selon laquelle la traduction de peptides des régions intergéniques peut aboutir à la naissance de nouveaux gènes chez la levure. Ce texte est écrit pour le format *Brevia* du magazine *Science*. Il n'est pas terminé, mais les parties manquantes sont indiquées.

Contribution : Je suis à l'origine des idées, des analyses, de l'écriture et de la plupart des calculs présentés dans ce document. Michael Cusick, Nicolas Thierry-Mieg et Marc Vidal m'ont aidée à développer scientifiquement ce projet depuis le début (2008). Ilan Wapinski m'aide principalement pour les analyses de séquences et Muhammed Yildirim pour la cartographie informatique des régions protégées par les ribosomes. Les autres auteurs, dont la liste et l'ordre restent à déterminer, ont apporté diverses contributions intellectuelles et techniques.

Proto-genes

Anne-Ruxandra Carvunis,[1,2] Ilan Wapinski,[3] Muhammed A. Yildirim,[4] Nicolas Simonis,[1] Benoit Charloteaux,[1] Cesar Hidalgo,[5] Nicolas Thierry-Mieg,[2] Michael E. Cusick,[1] Marc Vidal[1]

[1]Center for Cancer Systems Biology (CCSB) Department of Cancer Biology of Dana-Farber Cancer Institute and Department of Genetics of Harvard Medical School, Boston, MA 02115, USA. [2]Computational and Mathematical Biology Group, TIMC-IMAG, CNRS UMR5525 and Université de Grenoble, Faculté de Médecine, 38706 La Tronche cedex, France. [3]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. [4]Center for International Development and Harvard University, Cambridge, MA 02138, USA. [5]The Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02142, USA

Where do protein-coding genes come from? Most proposed mechanisms for the evolution of new protein-coding genes involve the remodeling of ancestral ones, often after gene duplication or lateral gene transfer (*1*). Recent findings indicate that truly novel protein-coding genes can also rarely arise *de novo*, if random mutations in a transcribed DNA sequence lead to a start and a stop codon separated by triplets of nucleotides, creating an open reading frame (ORF), and if this ORF accesses the translation machinery (*2*). We reasoned that such *de novo* protein-coding gene birth must require "noisy" translation of non-coding RNA. Here, focusing on the yeast *S. cerevisiae* as a model organism, we reveal widespread translation of intergenic regions, and link this phenomenon to a surprisingly frequent *de novo* emergence of new protein-coding genes. The corresponding polypeptides, intermediary entities in a continuum from intergenic to genic DNA, appear to evolve under natural selection towards either fixation of a functional protein-coding gene, or extinction. We name these ephemeral entities "proto-genes" (Figure 1).

To detect *de novo* gene birth in *S. cerevisiae*, we analyzed the ~6,000 ORFs predicted to encode proteins (hereafter pORFs) contained in its genome (Figure S1). We estimated the time at which the ancestor sequence of each pORF most likely appeared, using comparative genomics among 14 yeast species (Figure S2A). Consistent with previous studies (*3*), ~2% of pORFs were not conserved at all, a third showed intermediary conservation, and the majority appeared ancient (Figure S2B). To verify that we did not mistake fast-evolving ancient pORFs for recent pORFs, we focused on non-conserved pORFs that overlapped conserved ones. The paralogs of these conserved pORFs overwhelmingly did not overlap any pORF. Although possibly the non-conserved pORFs recently and independently lost their paralogs, as well as all of their homologs in 14 other yeast species, a more likely scenario is that they appeared *de novo* after the duplication events (Figure S3). Average RNA expression rate, chromosomal location and sequence composition of non-conserved pORFs also supported their *de novo* origination (Figures S4-6). We observed signatures of intra-species purifying selection in 15% of the pORFs that likely appeared after *S. cerevisiae* split from its sister species *S. paradoxus*. Relative to estimations of gene fixation rate after small-scale duplications (*4*), *de novo* gene birth appears 3-15 times more frequent.

In addition to ~6,000 pORFs, the *S. cerevisiae* genome contains over 100,000 short ORFs thought to appear and disappear stochastically through random mutations (hereafter sORFs) (Figure S1). We hypothesized that sORFs may constitute the reservoir from which new protein-coding genes arise *de novo*. Indeed, the average length of pORFs correlates with their evolutionary age (Figure S7, (*3*)) suggesting that pORFs may derive from sORFs through evolutionary lengthening, occurring maybe by duplication and fusion with other ORFs (*3*), or by loss of stop or shift of start codon (*5*). To test our hypothesis, we re-analyzed a published genome-wide ribosome profiling experiment (*6*), revealing translation of ~500 sORFs, including ~30 exhibiting signatures of purifying selection

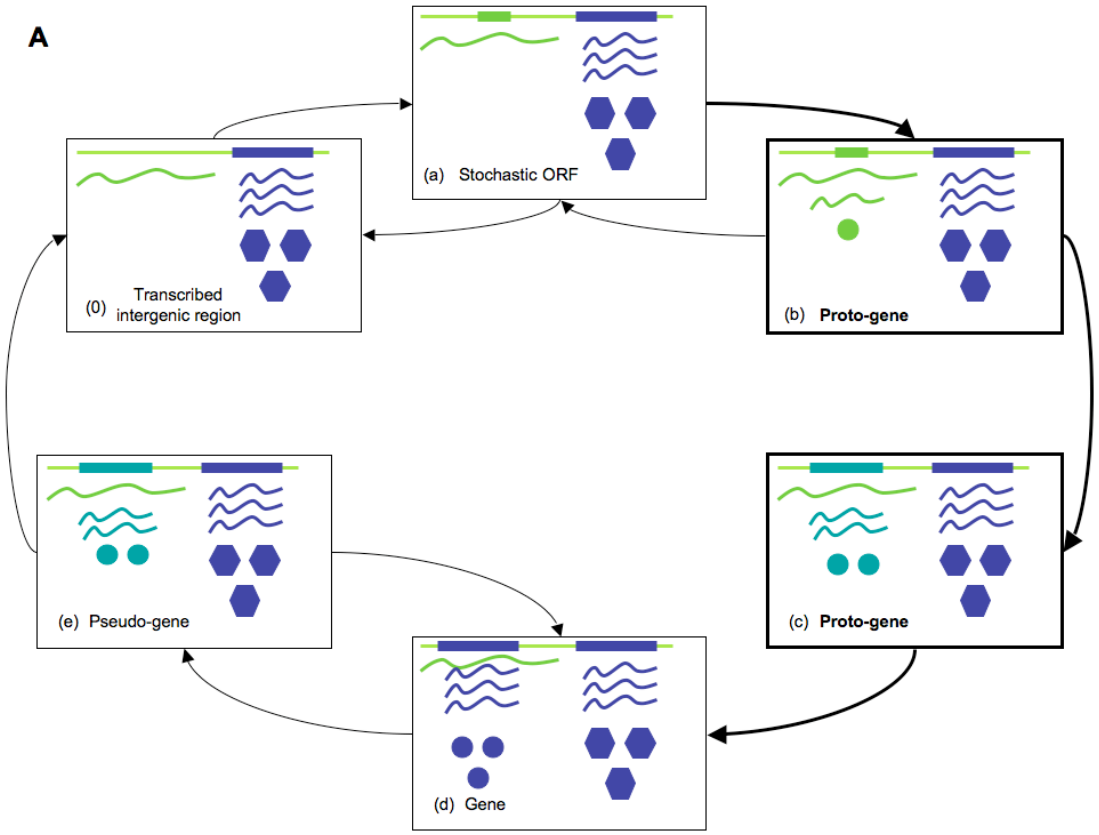
among *S. cerevisiae* strains. The translational landscape of the yeast intergenic regions thus appears larger than previously thought. While the cellular role of these polypeptides remains to be explored, their expression provides a mechanism for *de novo* birth of new protein-coding genes. That translational noise could present evolutionary advantages is in line with hypotheses regarding the way prions speed evolution of new traits in yeasts (*7*). We propose that non-conserved pORFs and sORFs constitute a new class of macromolecules that we name "proto-genes" (Figure 1), evolutionarily unstable entities bridging a continuum between intergenic and genic DNA.

Figure Legend

A. Conceptual evolutionary loop from intergenic regions to proto-genes, to genes, to pseudo-genes and back to intergenic regions
Arrows represent transitions between coding states (in frames) of a single DNA locus across evolutionary time; transitions indicated by thick arrows are proposed in the present article, while those depicted by thin ones are already recognized. Within the coding states frames, thick straight lines represent ORFs, thin ones ORF-free DNA; curved lines designate RNA, and filled hexagons and circles indicate translated products. In a transcribed intergenic region (0) close to a protein-coding gene, an ORF appears stochastically (a). If it is translated, the ORF becomes a proto-gene (b). Depending on selective pressures, mutations can either suppress this translation event (a), or gradually change the sequence composition of the proto-gene, which becomes closer to that of a regular gene (turquoise), increasing in length and expression level, and developing new protein properties (c). Gradually, the proto-gene reaches the status of a stable protein-coding gene (d). This newly formed gene could reversibly (*8*) become a pseudo-gene (*9*) (e), and eventually return to the ORF-free intergenic state (0). Pseudo-genes and proto-genes are depicted in the same way because they both represent intermediary coding states. However, a major difference between them is that pseudo-genes have homologs, while proto-genes do not.
B. YOL038C-A, a pORF specific to *S. cerevisiae*. Alignment of syntenic DNA sequences corresponding to the YOL038C-A locus in *S. cerevisiae* and three related species. Start and stop codons indicated by arrows.
C. sORF expression. Browser view of a sORF that is transcribed and translated in rich media according to raw data from (*6*). Footprint_rich, Footprint_starved: regions detected in ribosomes in rich and starved media, respectively. mRA_rich, mRNA_starved: regions detected as transcribed in rich and starved media, respectively.

References

1. M. Long, E. Betran, K. Thornton, W. Wang, *Nat Rev Genet* **4**, 865 (Nov, 2003).
2. H. Kaessmann, *Genome Res* **20**, 1313 (Oct).
3. D. Ekman, A. Elofsson, *J Mol Biol* **396**, 396 (Feb 19).
4. L. Z. Gao, H. Innan, *Science* **306**, 1367 (Nov 19, 2004).
5. M. G. Giacomelli, A. S. Hancock, J. Masel, *Mol Biol Evol* **24**, 457 (Feb, 2007).
6. N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, J. S. Weissman, *Science* **324**, 218 (Apr 10, 2009).
7. R. Halfmann, S. Alberti, S. Lindquist, *Trends Cell Biol*, (Jan 12).
8. M. Gerstein, D. Zheng, *Sci Am* **295**, 48 (Aug, 2006).
9. C. Jacq, J. R. Miller, G. G. Brownlee, *Cell* **12**, 109 (Sep, 1977).


A
(a) Stochastic ORF
(b) Proto-gene
(c) Proto-gene
(d) Gene
(e) Pseudo-gene
(0) Transcribed intergenic region

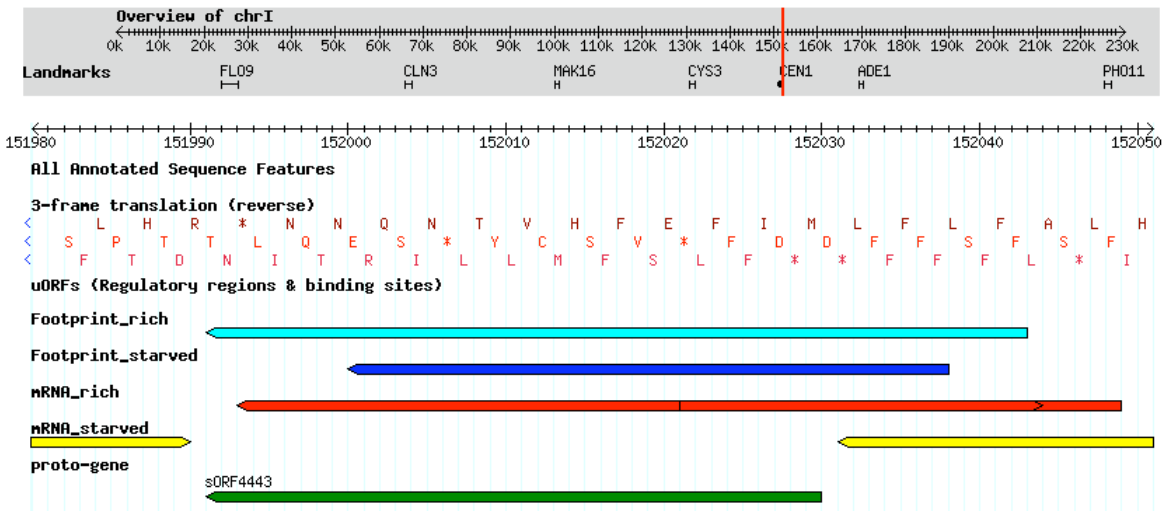
B

```
S. cerevisiae   ATT--GATGGGATATTCAACTTATATGAT-TTGA---TGAAATACA--TG 327
S. paradoxus    CTT--GATGGGCTATATTGCTTATATAAT-GTGA---TAGAGAATA--TG 324
S. mikatae      TTC--GATAGACCCAGGCTCTTCTATATCAGTAA---TAAAAAAAAAATG 331
S. bayanus      TCTAGGATCTATGCTTGCCCTGGCTCGTGCTTGAAAGCAAGGAATATATA 350
                   ***          **         * *        * *  *

S. cerevisiae   GGGTCCTTCCTGAGGAAAGCGGCTACAACAAATTTATTCAATAGCATAAA 377
S. paradoxus    GGATCTTTCCTGAGGTG-GCGGCTATTACAGTTTTCGCCCATAAGCATGA 373
S. mikatae      GGGTCTCTTGTGAGGAA-GCGGCTAGCACACTTTTTGCGCATGAGTTGAG 380
S. bayanus      GGATA-TTACTGAGAAG--CGGCTATAGCAGTTTCTGTCCTCGAGTGAAA 397
                ** *   *  ****     ******   **  **

S. cerevisiae   AAAAAGGAAGGTACAAAACAGAGCCATGTCATAGAGCAAAGAAAGAGCAC 427
S. paradoxus    AAAATGGAACATACAAAACAGAAGTGTAACATACAGCAAAGAAAAAGCAC 423
S. mikatae      AA----GAAGGTACATAACAGAGGCATATTATACATTAAAGAAAAAGGTT 426
S. bayanus      AATA--CGACATACAATTCCGAGGCATGAAATACAGC------------- 432
                **       *  ****   * **    *   *** *
```

C



Proto-genes

Supplementary online material

Anne-Ruxandra Carvunis,[1,2,3] Ilan Wapinski,[4] Muhammed A. Yildirim,[5] Nicolas Simonis,[1,2] Benoit Charloteaux,[1,2] Cesar Hidalgo,[6] Nicolas Thierry-Mieg,[3] Michael E. Cusick,[1,2] Marc Vidal[1,2]

[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, MA 02215, USA. [2]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [3]Computational and Mathematical Biology Group, TIMC-IMAG, CNRS UMR5525 and Université de Grenoble, Faculté de Médecine, 38706 La Tronche cedex, France. [4]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. [5]Center for International Development and Harvard University, Cambridge, MA 02138, USA. [6]The Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02142, USA.

Methods

Sequence analysis

We compared the sequence of the reference strain of *S. cerevisiae* with i) the sequences of other yeast species and ii) the sequences of other strains of *S. cerevisiae*. The sequence and annotations, downloaded in October 2007 from the *Saccharomyces* Genome Database (SGD) (*1*) by the SGRP group (*2*), were used as a reference for the sequence analyses. Paralogous relationships between *S. cerevisiae* pORFs were downloaded from the Ensembl Compara website (*3*). We extracted 107,723 sORFs, defined as any sequence between canonical start and stop codons that was i) longer than 30 nucleotides, ii) not contained in a larger ORF on the same frame iii) not overlapping a pORF, a rRNA, a tRNA, a ncRNA, a snoRNA or a uORF (*4*) on the same strand.

1 Comparative analysis between yeast species

To evaluate the time at which the ancestor sequence of each pORF most likely appeared, we compared its nucleotide and amino-acid sequences with sequences from 14 other yeast species (*S. paradoxus, S. mikatae, S. bayanus, C. glabrata, S. castelii, K. lactis, A. gossypii, K. waltii, D. hansenii, Y. lipolytica, C. albicans, A. nidulans, N. crassa,* and *S. pombe*), organized in an evolutionary tree (Figure S2A). Sequences and annotations for these species were downloaded from websites listed at http://www.broad.mit.edu/regev/orthogroups/ (*5*) in September 2007.

1.1 Evolutionary tree generation

To infer the phylogeny of the 14 yeast species, we first aligned all one-to-one orthologs across these species, using the MUSCLE alignment software (*6*), and we concatenated these alignments. We then created approximately 1,000 hypothetical peptide alignments by sampling each time 10,000 positions from this concatenated alignment, and the phylogeny was reconstructed using the PHYLIP software package (*7*). This sampling process was repeated multiple times, generating the same phylogeny each time, with the exception of the branches that indicated the speciation events separating *S. castellii*, *C. glabrata*, and the *Saccharaomyces sensu stricto* species. Reconstructions sometimes indicated that *S. castellii* was monophyletic with the *Saccharomyces sensus stricto* group, and sometimes that *S. castellii* was more closely related to *C. glabrata*. We decided to consider the *Saccharamyces* group more closely related to *C. glabrata* than to *S. castellii*, according to genome reconstructions performed by Scannell and coworkers (*8*).

1.2 “Age” calculations

The “age” of a pORF was defined as the number of divergent branches in the phylogenetic tree that separates *S. cerevisiae* from the most distant species in which even a slight sequence similarity with this pORF, or any of its paralogs, could be detected. To find such similarities, we proceeded as follows for each pORF: 1) the ORF-specific sequence that did not overlap another annotated pORF was extracted; 2) this sequence was blasted against genomic and protein sequences from the other yeasts, using three distinct blast programs (*9*) (blastp, tblastx, tblastn), and blast hits with an e-value of less than 0.01 were considered positive hits; 3) the number of divergent branches in the phylogenetic tree between *S. cerevisiae* and the species showing a positive hit for the pORF was counted. The largest of these numbers was considered the pORF’s “age”, except when the pORF had a paralog of higher “age”. Note that most results presented here are robust to slight modifications of this method (different blast programs, different e-value cutoffs). Only 5 *S. cerevisiae* pORFs had homologs in *C. glabrata* but not in *S. castellii*, which can be explained by the small size of the *C. glabrata* genome, or the short evolutionary distance between *C. glabrata* and *S. castellii*. To avoid counting 5 proteins as an entire category, we grouped them with other proteins of age 4.

2 Comparative analysis of *S. cerevisiae* strains

To gain insight into the micro-evolution of *S. cerevisiae* pORFs, we compared their sequences in the reference strain to their sequences in seven other strains (SK1, Y55, DBVPG6765, DBVPG1373, DBVPG6044, YPS606 and W303). SNPs from these strains were extracted from the website of the Sanger Institute *Saccharomyces* Genome Resequencing Project (http://www.sanger.ac.uk/Teams/Team71/durbin/) (*2*) in March 2010. Sequenced SNPs were taken into account only when their quality was superior to 55.

The protein evolution rate of each pORF and sORF was evaluated in two steps: first, homologous nucleotide sequences in each strain were aligned using MUSCLE (*6*), and next, the PAML software package was used to estimate the rates of synonymous (dS) and non-synonymous (dN) substitutions (*10*). Method for p-value estimation is missing.

External datasets

1 mRNA abundance

The expression level of each pORF was estimated from two published genome-wide datasets, chosen because they represent two different strand-specific experimental techniques: oligo-tiling array (*11*), and RNA-seq (*12*). In the oligo-tiling array dataset, the normalized intensity of expression per probe was averaged, and the percentage of probes within a pORF with intensity higher than this average was defined as the expression level of this pORF. Estimations of expression level by Nagalaksmi and coworkers (*12*) were not reprocessed.

2 Ribosome foot-printing

Raw reads from the ribosome footprinting experiments (*4*) were downloaded from Short Read Archive maintained by NCBI with accession number SRA008252. Initial analysis by the authors discarded the reads mapping to dubious ORFs and other overlapping features, which are important for our analyses. Therefore, all the Illumina sequencing reads were re-mapped to the reference genome (CYGD) with the Bowtie short read mapping program (*13*) using the first 25 nucleotides as the ‘seed’ region. Reads with multiple matches to the reference genome were discarded. In order to maximize the matched reads, reads that had failed to map with no mismatches in the seed region were re-fed into Bowtie allowing

one mismatch in the seed region and those still failed to align were mapped with two mismatches in the seed region. In total 85-95% of the reads were aligned to the reference genome. Short read runs were subsequently grouped into four categories: Ribosome footprinting in rich conditions (8 runs), ribosome footprinting in amino-acid starved conditions (6 runs), and corresponding mRNA-seq runs (rich, 4 runs and starved, 3 runs). All the runs in each group were merged to create a single mapping file. Genomic regions with less than five reads were filtered out from each mapping file to increase the quality of the resulting alignments. We considered ORFs to be showing signs of translation when they 80% of their length was covered by mRNA and ribosome footprint reads in either of the two conditions. These stringent criteria detect 3134 pORFs and 547 sORFs.

Supplementary Figure Legends

Figure S1: Abundance of sORFs in the yeast intergenic regions. Distribution of ORF length in base pairs observed in the *S. cerevisiae* genome: sORFs (green), pORFs (blue); the vertical black line represents a prediction threshold of 300 base pairs (*14*).

Figure S2: Estimating the evolutionary age of *S. cerevisiae* genes.
A. Principle of evolutionary age estimation. The sequence of every pORF in *S. cerevisiae* (indicated here by the pORFs A, B, B', C and D, where the pORFs B and B' are paralogs) is compared to the sequence of every coding and intergenic region of the 14 yeast species organized in the phylogenetic tree represented on the left (see method). The 11 divergent branches in this tree form 10 age categories (see method). Positive sequence similarity hits are represented by red dots, negative ones by crosses. pORF A shares sequence similarity with every species in the tree, and is thus assigned the maximum age. pORF B only shares sequence similarity with species up to *S. castellii* and is assigned age 6. pORF B' shares similarity with species up to *C. glabrata*, which corresponds to age 4; however, it is assigned age 5 because it is a paralog of pORF B. pORF C only shares similarity with *S. pombe*, and is assigned the age 10, under the assumption that it is actually ancient, but that it may have been lost from all the other yeast species in the tree except *S. cerevisiae*. pORF D does not share sequence similarity with any species in the tree: it is specific to *S. cerevisiae*, and is assigned age 1.
B. Number of *S. cerevisiae* pORFs assigned to each age category, from 1 to 10.

Figure S3: Distinguishing *de novo* gene birth and rapid sequence divergence. Three evolutionary scenarios leading to the same observation, namely that a gene with a species-specific sequence (A, red rectangle) overlaps a gene of higher evolutionary age (B, blue rectangles) that has paralogs (set B', striped blue rectangles). On the left: the species-specific gene truly appeared *de novo* and the genes in set B' do not overlap other genes. On the right: since the ancestor of B overlapped a gene (purple rectangle) before duplicating, members of set B' overlap set A'. Recent sequence divergence makes gene A appear species-specific, although it is not of recent *de novo* origin. In the middle, the ancestor of B overlapped a gene before duplicating, but all of the paralogs of this gene (set A') were independently lost; if all the orthologs and paralogs of this gene were also lost independently in the other yeast species, its sequence would appear species-specific, although it is truly of ancient origins. Considering set A as all pORFs of age 1, only 3 of 138 genes in set B' also overlapped another ORF.

Figure S4: Correlation between RNA expression level and evolutionary age. Median mRNA expression levels for each age category. Estimation of mRNA abundance from: A)

Oligo-tiling array (*11*), B) RNA-seq (*12*). Error bars represent standard error of the median.

Figure S5: Recent pORFs tend to partially overlap other pORFs. This plot shows the proportion of partially overlapping pORFs as a function of their evolutionary age. Fully overlapping pORFs are excluded from the dataset. Error bars represent standard error of the proportion. Since the non-coding strand of genes is markedly depleted in stop codons (*15*), and is pervasively transcribed via bidirectional promoters (*16*), it theoretically presents a higher probability that a random ORF will appear within this strand, while also offering easier access to the transcription machinery. That recent pORFs often partially overlap other pORFs is therefore consistent with a *de novo* origination.

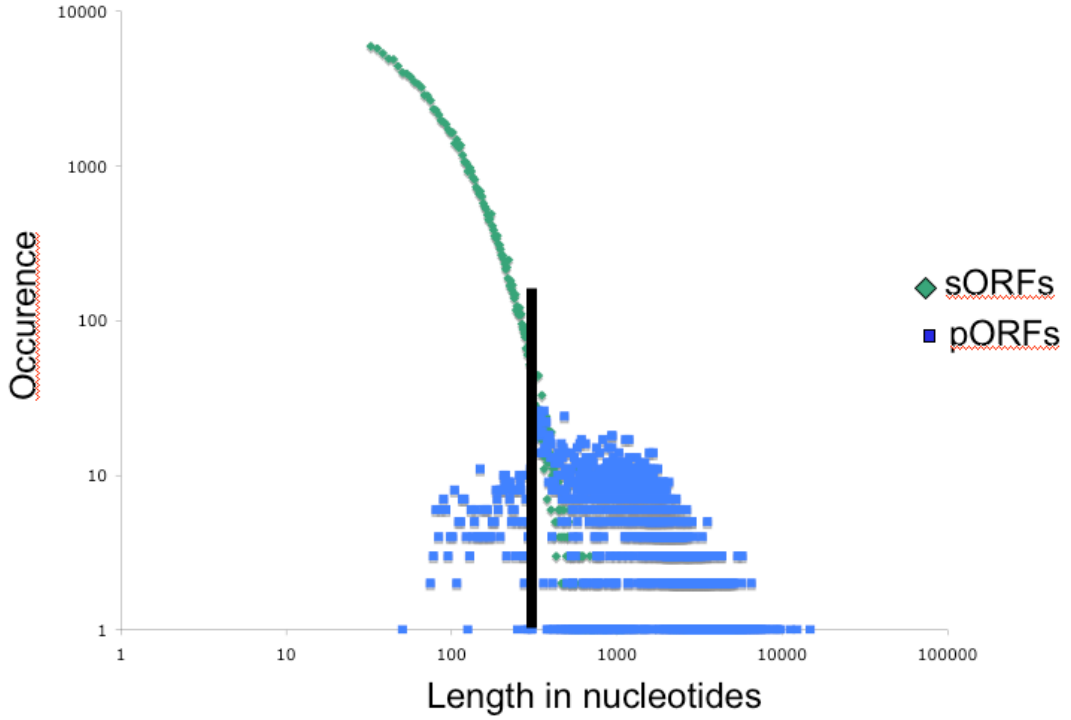
Figure S6: The sequence composition of pORFs of low evolutionary age is intermediary between that of sORFs and of pORFs of higher evolutionary age. For each amino acid (X axis), the Y axis represents the ratio between its frequency in recent pORFs (1-4) and i) ancient pORFs (5-10) (orange); or ii) sORFs (green).

Figure S7: Correlation between pORF length and evolutionary age. Error bars are standard errors of the mean.

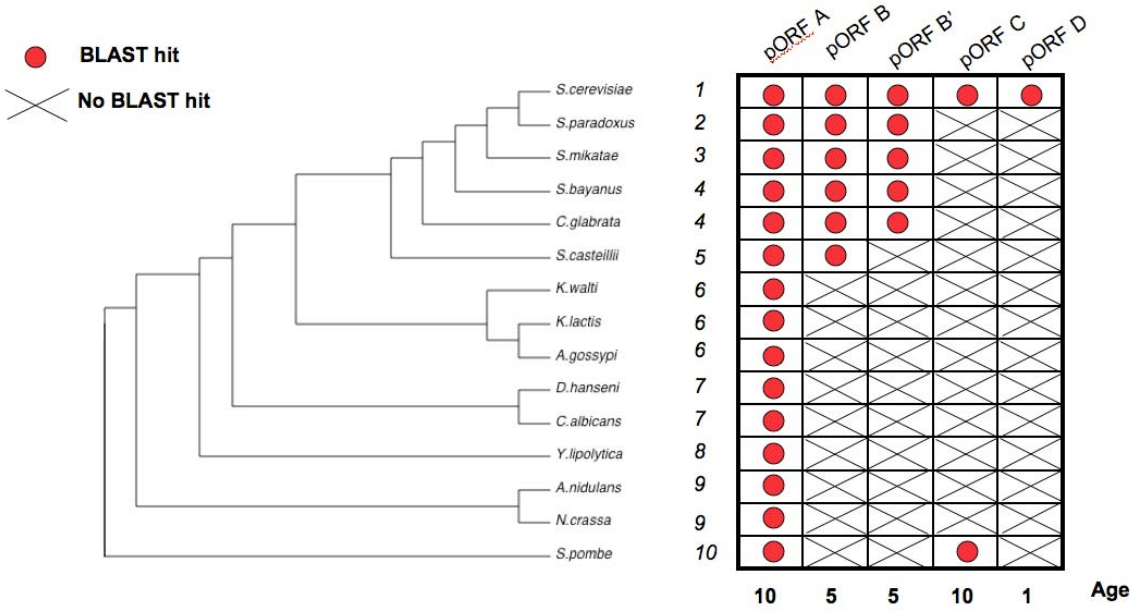
Supplementary References

1. D. G. Fisk *et al.*, *Yeast* **23**, 857 (Sep, 2006).
2. G. Liti *et al.*, *Nature* **458**, 337 (Mar 19, 2009).
3. A. J. Vilella *et al.*, *Genome Res* **19**, 327 (Feb, 2009).
4. N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, J. S. Weissman, *Science* **324**, 218 (Apr 10, 2009).
5. I. Wapinski, A. Pfeffer, N. Friedman, A. Regev, *Nature* **449**, 54 (Sep 6, 2007).
6. R. C. Edgar, *Nucleic Acids Res* **32**, 1792 (2004).
7. J. Felsenstein. (Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, 2005).
8. D. R. Scannell, K. P. Byrne, J. L. Gordon, S. Wong, K. H. Wolfe, *Nature* **440**, 341 (Mar 16, 2006).
9. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389 (Sep 1, 1997).
10. Z. Yang, *Mol Biol Evol* **15**, 568 (May, 1998).
11. L. David *et al.*, *Proc Natl Acad Sci U S A* **103**, 5320 (Apr 4, 2006).
12. U. Nagalakshmi *et al.*, *Science* **320**, 1344 (Jun 6, 2008).
13. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, *Genome Biol* **10**, R25 (2009).
14. S. G. Oliver *et al.*, *Nature* **357**, 38 (May 7, 1992).
15. A. Tats, T. Tenson, M. Remm, *BMC Genomics* **9**, 463 (2008).
16. Z. Xu *et al.*, *Nature* **457**, 1033 (Feb 19, 2009).

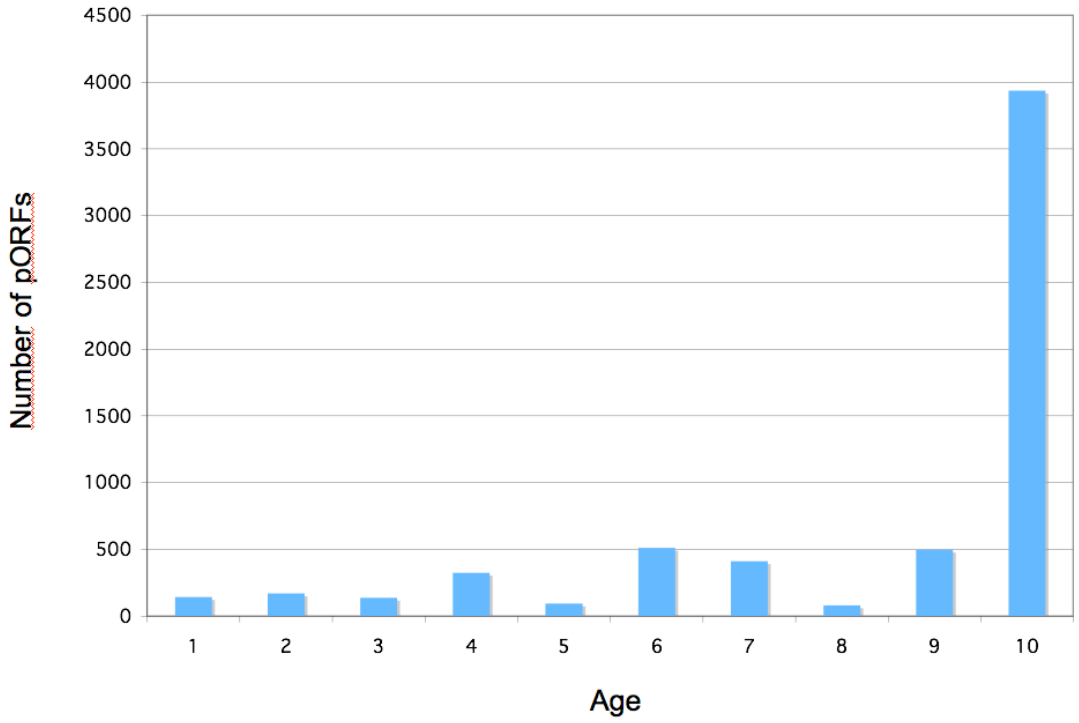
S1


10000
1000
100
10
1
Occurence
1
10
100
1000
10000
100000
Length in nucleotides
sORFs
pORFs

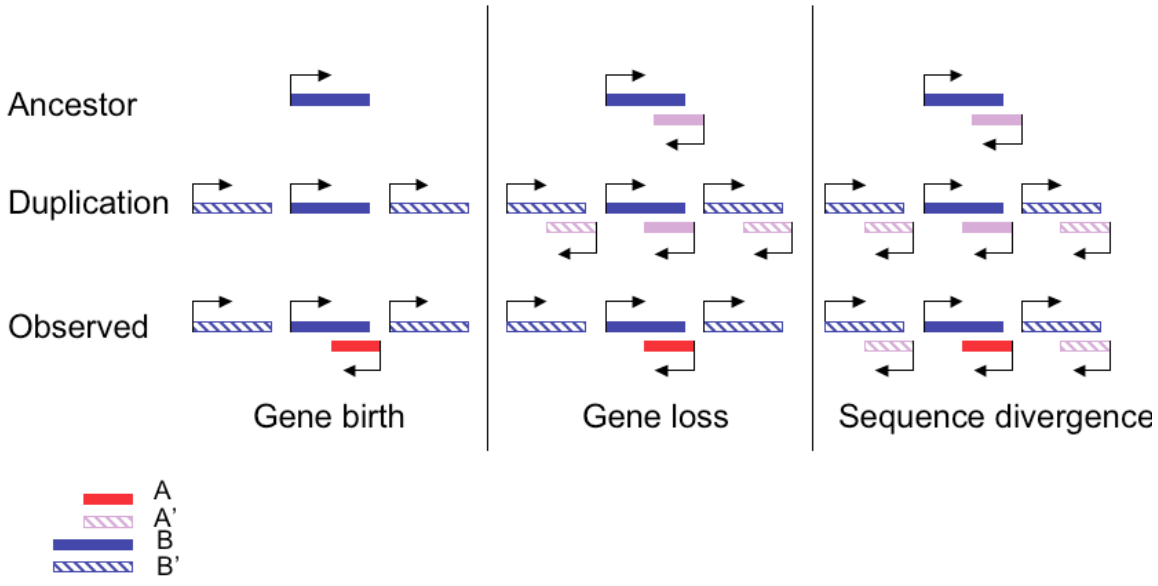
S2A


BLAST hit
No BLAST hit
S.cerevisiae
S.paradoxus
S.mikatae
S.bayanus
C.glabrata
S.casteillii
K.walti
K.lactis
A.gossypi
D.hanseni
C.albicans
Y.lipolytica
A.nidulans
N.crassa
S.pombe
pORF A
pORF B
pORF B'
pORF C
pORF D
1
2
3
4
4
5
6
6
6
7
7
8
9
9
10
10
5
5
10
1
Age

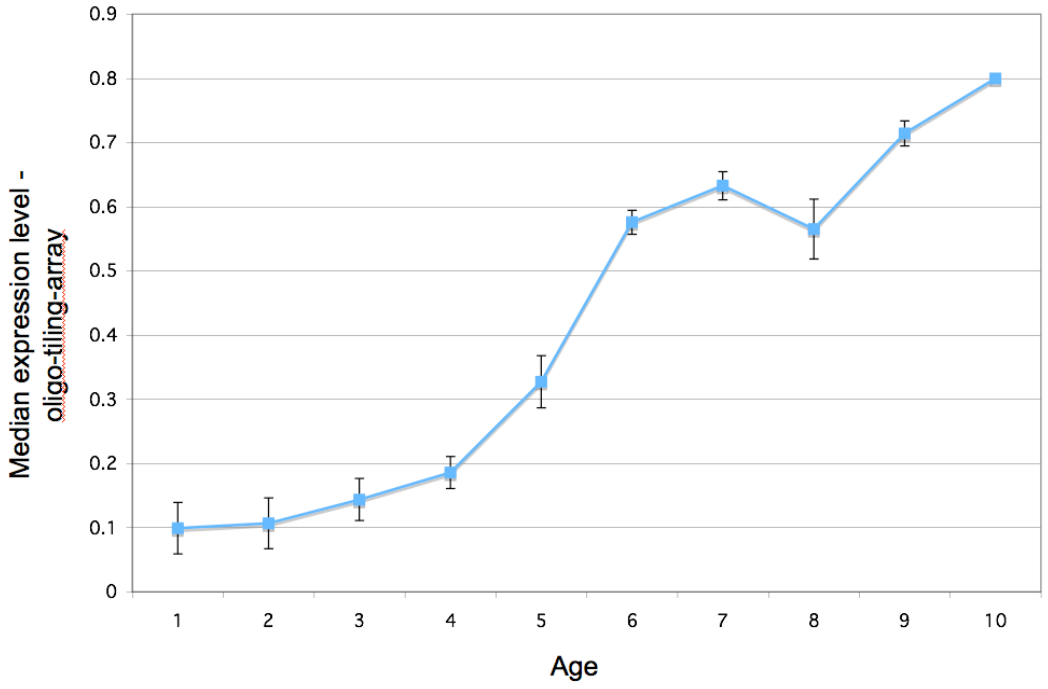
S2B


4500
4000
3500
3000
2500
2000
1500
1000
500
0
Number of pORFs
1
2
3
4
5
6
7
8
9
10
Age

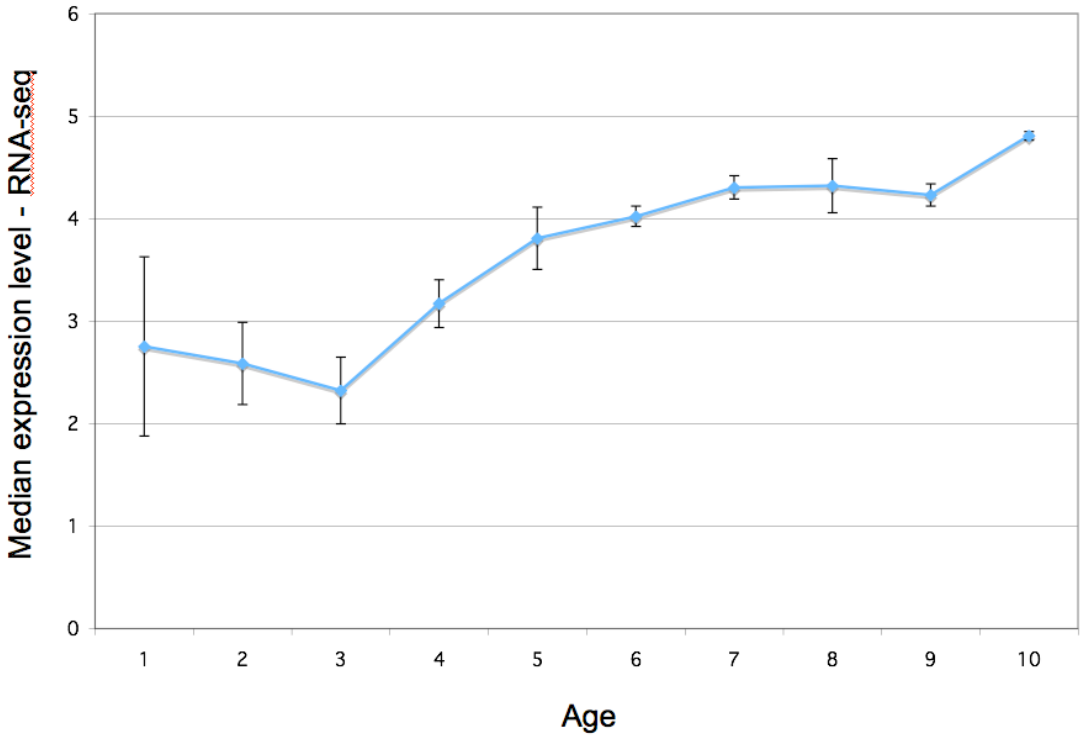
S3


Ancestor
Duplication
Observed
Gene birth
Gene loss
Sequence divergence
A
A'
B
B'

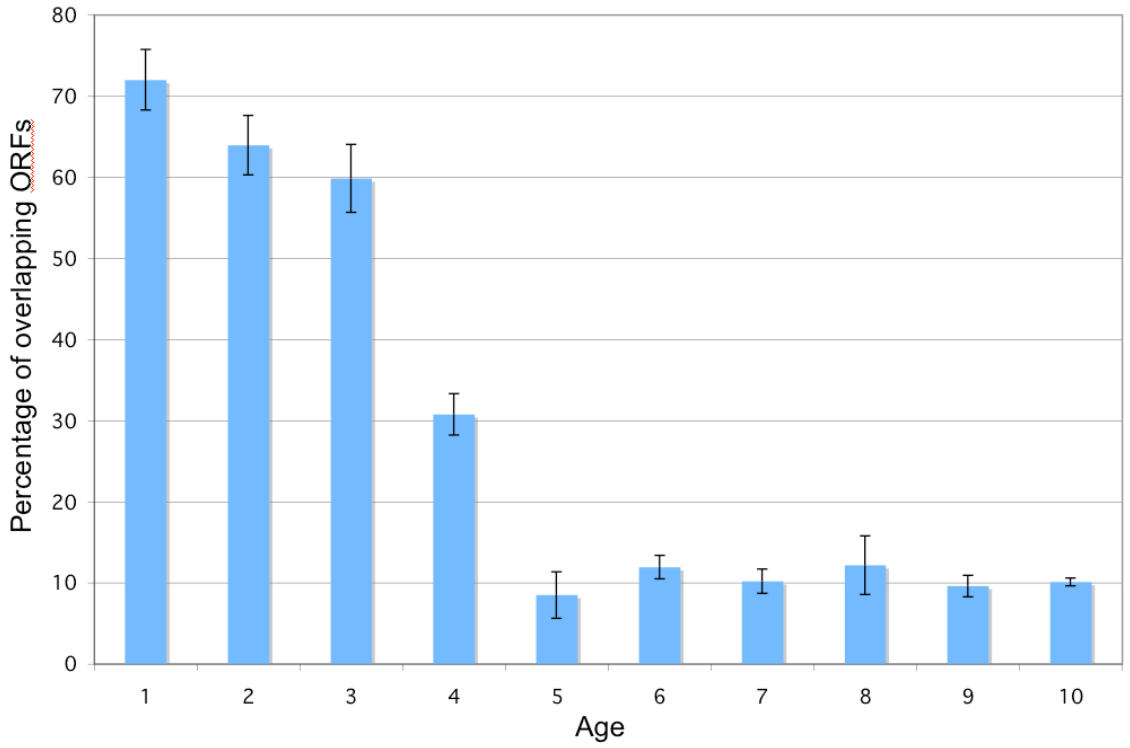
S4A


Median expression level - oligo-tiling-array
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
9
10
Age

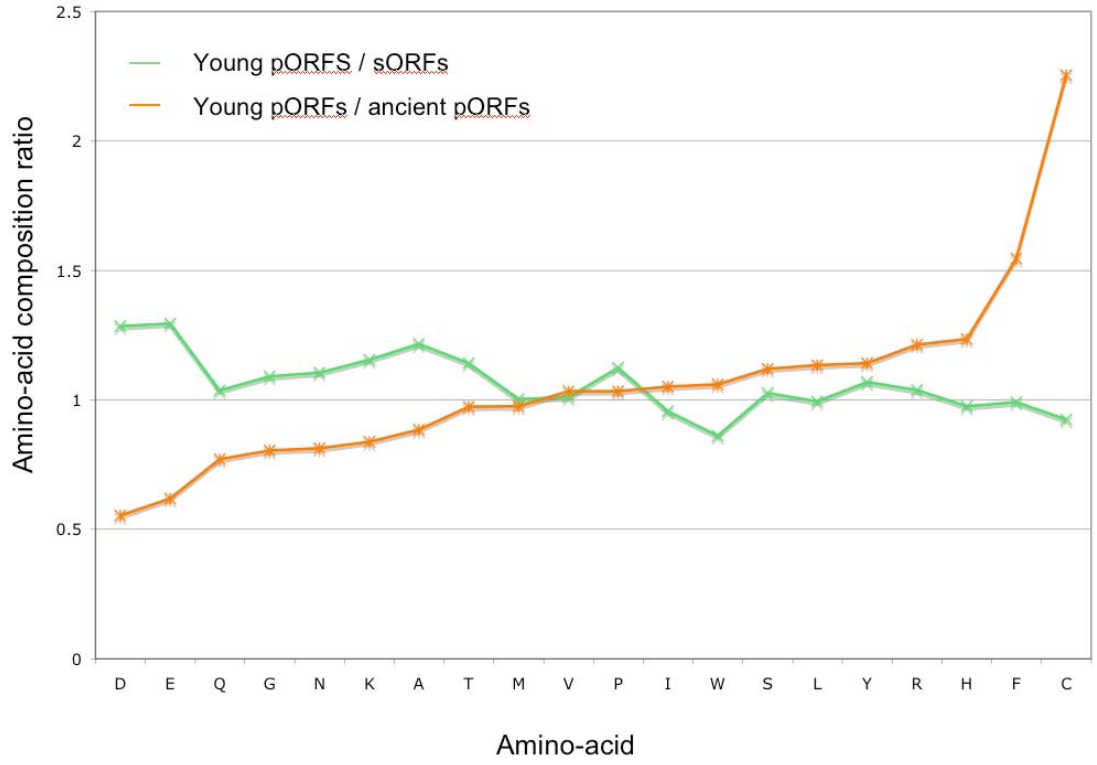
S4B


Median expression level - RNA-seq
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
Age

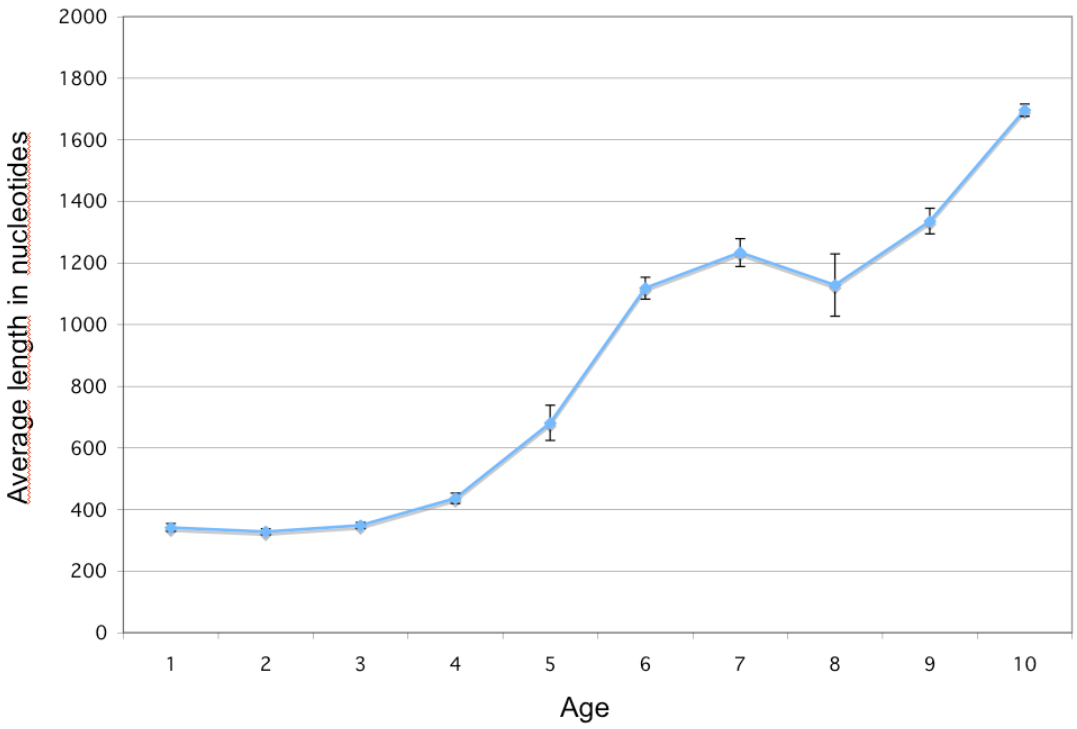
S5


Percentage of overlapping ORFs
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
Age

S6


Young pORFS / sORFs
Young pORFs / ancient pORFs
Amino-acid composition ratio
2.5
2
1.5
1
0.5
0
D E Q G N K A T M V P I W S L Y R H F C
Amino-acid

S7


Average length in nucleotides
2000
1800
1600
1400
1200
1000
800
600
400
200
0
1 2 3 4 5 6 7 8 9 10
Age

CHAPITRE 2 : A LA RECHERCHE DES *LIENS* DE L'INTERACTOME : ASPECTS INFORMATIQUES DE LA DETECTION EXPERIMENTALE D'INTERACTIONS PHYSIQUES ENTRE PROTEINES

Les systèmes cellulaires reposent sur les interactions physiques entre biomolécules, en particulier entre protéines. Mesurer expérimentalement toutes les interactions physiques entre protéines représente un défi encore insurmonté aujourd'hui. Pour cette raison, de nombreuses méthodes bioinformatiques ont été développées pour les prédire (Plewczynski and Ginalski 2009). Bien que ces prédictions puissent être utiles, elles ne remplacent pas les mesures expérimentales. Plusieurs bases de données publiques ont pour objectif la curation de la littérature scientifique afin de recenser toutes les interactions physiques entre protéines qui ont été expérimentalement détectées au fil des années. Puisque leur contenu provient d'expériences focalisées et publiées, ces bases de données sont souvent considérées comme le standard en matière d'interactions entre protéines. Cependant, mon collègue Michael Cusick (CCSB, Boston) a montré avec mon aide dans l'article « Literature-Curated Interactions » publié dans le journal *Nature Methods* en janvier 2009 (**Document Joint 4**) que l'activité de curation est sujette à différentes sources d'erreurs qui rendent ces bases de données moins fiables que communément présumé, bien qu'elles restent une ressource précieuse (Cusick, Yu et al. 2009). Même si elle était de qualité irréprochable, l'information recensée dans ces bases de données ne saurait en aucun cas être utilisée telle quelle pour modéliser le réseau interactome dans sa globalité, en raison du biais dû à l'intérêt scientifique ayant motivé chaque expérimentation. Par exemple, les protéines impliquées dans le cancer, nettement plus étudiées que d'autres protéines de fonction inconnue, apparaîtront artificiellement avoir beaucoup d'interactions. Ne pas avoir conscience de ce biais peut conduire à de mauvaises interprétations (Yu, Braun et al. 2008). Les approches de cartographie systématique expérimentale, comme le double hybride en levure (« yeast two-hybrid » en anglais, Y2H), ne souffrent pas de ce type de biais d'investigation et sont donc plus adaptées à l'analyse de réseau.

Mesures quantitatives des limites d'un réseau interactome expérimental: application au nématode

Malgré leur caractère systématique, les réseaux interactomes générés par Y2H ne sont pas sans limitations, notamment en raison d'artefacts techniques qu'il faut

identifier et éviter (Walhout and Vidal 2001). En particulier, et contrairement à ce que suggèrent les titres de certains articles (Uetz, Giot et al. 2000; Ito, Chiba et al. 2001), ces réseaux ne sont que des échantillons imparfaits et incomplets de l'interactome réel. Une avancée conceptuelle considérable dans le domaine de la cartographie expérimentale de réseaux interactomes a été développée par le groupe de Marc Vidal pendant les deux premières années de ma thèse: la *mesure quantitative* de ces limitations. Kavitha Venkatesan a défini dans l'article « An empirical framework for binary interactome mapping » (Venkatesan, Rual et al. 2009) quatre paramètres qui caractérisent les limitations d'un réseau interactome : la complétion, la sensibilité liée à la technique, la sensibilité liée à l'échantillonnage, et la précision (**Figure 7,** définitions dans la légende).

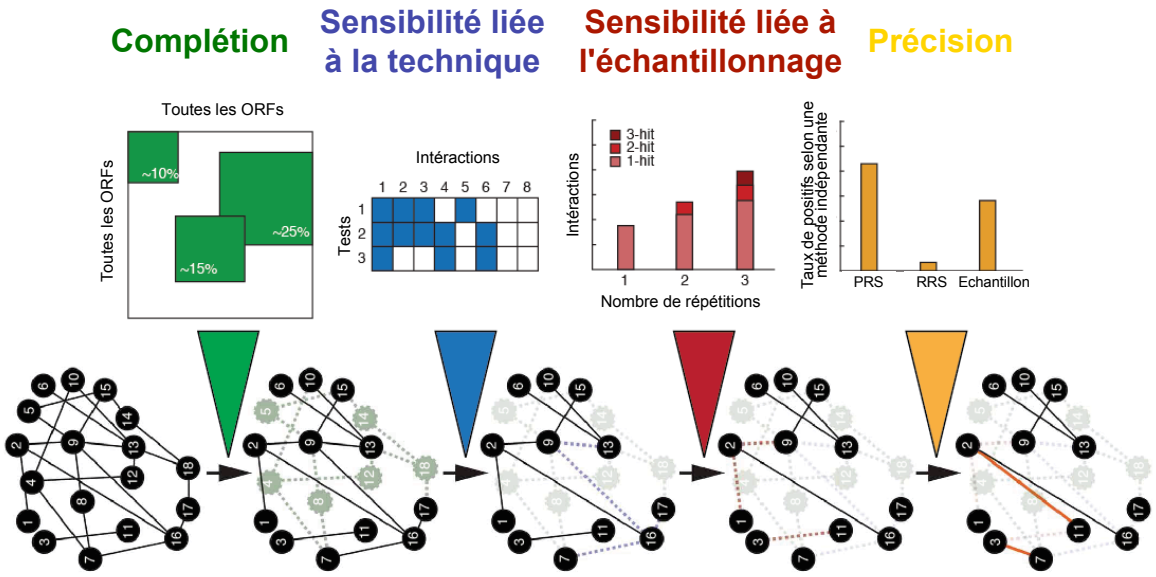


Figure 7 : Stratégie pour la mesure quantitative des limitations d'un interactome expérimental, d'après (Venkatesan, Rual et al. 2009). Illustration des concepts de :
- complétion : fraction de toutes les paires de protéines possibles testées dans l'expérience
- sensibilité liée à la technique : fraction de toutes les interactions physiques confirmées entre protéines détectables avec cette technique
- sensibilité liée à l'échantillonnage : fraction de toutes les interactions détectables avec cette technique identifiées au niveau de saturation de l'expérience
- précision : fraction des paires de protéines identifiées par l'expérience qui correspondent à des vraies interactions physiques et non des artefacts

Ces quatre paramètres peuvent être calculés indépendamment (voir partie supérieure de cette figure) et combinés pour définir les limitations d'un interactome, et estimer la taille de l'interactome complet d'un organisme. Pour plus d'explications sur la démarche à suivre pour calculer ces paramètres et les erreurs associées voir (Simonis, Rual et al. 2009), et l'article sur l'interactome d'*Arabidopsis* joints à ce manuscrit. Les représentations schématiques de réseaux (partie inférieure de cette figure) illustrent l'influence de chaque paramètre sur les limitations de l'observation expérimentale d'un interactome par rapport à l'interactome complet (à l'extrême gauche). Les liens et nœuds solides représentent des interactions physiques authentiques entre protéines dans l'interactome réel; en pointillé sont les interactions authentiques entre protéines qui ne sont pas détectées par l'expérience; en couleur, des interactions physiques artefactuelles.

Avec trois étudiants post-doctoraux, nous avons adapté ce cadre conceptuel à l'analyse de l'interactome du nématode *C. elegans.* Nos résultats ont été publiés dans le même numéro de *Nature Methods*, sous le titre « Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network » (Simonis, Rual et al. 2009) (**Document joint 5**). Nous avons construit expérimentalement une carte de 1816 interactions entre 1496 protéines. Afin de procurer une ressource plus complète à la communauté, nous avons appliqué des filtres de qualité à d'autres interactions Y2H déjà publiées par notre groupe de recherche et les avons intégrées à cette carte, le tout formant un réseau interactome de 3864 interactions entre 2528 proteines. Il s'agit d'un immense progrès, à comparer avec les ~450 interactions accumulées dans les bases de données recensant les interactions découvertes à petite échelle et publiées l'une après l'autre. De plus, grâce aux mesures de la proportion de faux négatifs et faux positifs que nous avons effectuées, nous avons pu estimer que le véritable réseau interactome du nématode comprend probablement autour de 116000 interactions. Leur identification exhaustive demeure donc un immense défi.

Perfectionnement de la cartographie et de l'analyse d'un réseau interactome experimental: application à Arabidopsis thaliana

Forte de l'expérience acquise lors de ces travaux sur le nématode, j'ai apporté des améliorations conséquentes à la méthodologie expérimentale et l'analyse des limitations des réseaux interactomes générés par Y2H. Ces améliorations ont été implémentées dans le cadre du premier réseau interactome jamais construit pour une plante (**Document Joint 6**). Centré sur l'organisme modèle pour les plantes, *Arabidopsis thaliana,* ce travail a été effectué avec Pascal Braun, chef de projet, et Matija Dreze, doctorant dirigeant une équipe d'expérimentateurs. Mes propositions bioinformatiques ont permis de :

- Limiter les erreurs de transfert de résultats entre l'expérimentateur et la base de données;
- Tracer systématiquement chaque étape du protocole afin de détecter les problèmes et les corriger informatiquement au besoin;
- Évaluer la part de variabilité expérimentale due à l'expérimentateur et due à la variabilité biologique;
- Optimiser la construction d'un ensemble de référence composé de 118 interactions pour augmenter sa fiabilité;
- Déterminer un paramètre unique de sensibilité expérimentale, plus précis que la multiplication de la sensibilité liée à la technique et la sensibilité liée à

l'échantillonnage;

- Confirmer par séquençage l'identité des protéines interagissant avant de valider les interactions.

De plus, avec l'aide du responsable du groupe bioinformatique du CCSB, Tong Hao, j'ai conçu une base de données retraçant toutes les étapes expérimentales et intégrant des données extérieures comme l'annotation du génome d'*Arabidopsis* ou des résultats publiés de génomique fonctionnelle. Nous avons pris soin d'optimiser cette base de données pour qu'elle soit facile à mettre à jour en cas de ré-annotation du génome ou de ré-actualisation des résultats expérimentaux. Grâce à ces modifications, la reproductibilité ainsi que la mesure des limitations de ce réseau interactome ont été améliorées par rapport aux réseaux interactomes publiés jusqu'à présent. Nous avons découvert plus de 6000 interactions, doublant ainsi le nombre d'interactions recensées à ce jour pour *Arabidopsis* (Aranda, Achuthan et al. ; Breitkreutz, Stark et al. 2008; Swarbreck, Wilks et al. 2008). Pourtant, nos mesures de la proportion de faux négatifs et de faux positifs pour ce réseau suggèrent que l'ensemble de toutes ces interactions couvre moins de 5% de la taille du véritable réseau interactome d'*Arabidopsis*. Comme pour le nématode, cartographier le réseau interactome entier d'une plante demeure donc un défi de taille pour les années à venir.

Évaluation informatique de la pertinence biologique des interactions biophysiques détectées par Y2H

Malgré leur haute qualité technique, les interactions Y2H représentent des événements biophysiques qui ne sont pas nécessairement pertinents *in vivo*. Par exemple, deux protéines ayant la capacité d'interagir peuvent ne jamais être exprimées dans la même cellule. Pour explorer cette question, nous avons comparé le niveau de co-expression de l'ARN de paires de gènes encodant des protéines en interaction dans un réseau Y2H à celui de paires de gènes aléatoirement choisies. Comme chez l'humain (Rual, Venkatesan et al. 2005) et la levure (Yu, Braun et al. 2008), les paires de gènes encodant des protéines en interaction chez le nématode et chez *Arabidopsis* sont significativement plus souvent co-exprimées que les paires aléatoirement choisies. Par ailleurs, nous avons combiné les interactions Y2H pour le nématode aux données d'expression au cours du développement de promoteurs d'environ 2000 gènes résultant des travaux dirigés par D. Dupuy (IECB, Bordeaux) (**Document Joint 7**). Cela nous a permis de predire les «territoires spatiotemporels» où 69 interactions pourraient avoir lieu *in vivo*.

Nous avons également mis en œuvre une autre approche pour évaluer la

pertinence biologique des interactions Y2H. Celle-ci consiste à calculer la proportion de paires de protéines ayant des caractéristiques biologiques communes en utilisant les annotations de la « Gene Ontology » (Ashburner, Ball et al. 2000). Pour le nématode comme pour *Arabidopsis*, nous avons observé un enrichissement significatif en paires de protéines ayant des annotations communes. Dans les deux cas, cet enrichissement est particulièrement élevé lorsque les annotations indiquent des fonctions moléculaires précises (**Figure 8A**, illustration pour *Arabidopsis*). Chez le nématode, nous avons de plus montré que les interactions physiques entre protéines informent sur des phénomènes biologiques plus spécifiques que d'autres types de données issues d'expériences à grande échelle (**Figure 8B**). Dans le cadre de l'étude du réseau Y2H d'*Arabidopsis*, nous avons montré que l'enrichissement en fonctions communes ne s'arrête pas à l'interaction individuelle : les protéines qui partagent des interacteurs, ou qui font partie du même module au sein du réseau, sont aussi enrichies en annotations partagées. Il apparaît donc clairement que les interactions reportées dans ces deux réseaux correspondent souvent à des événements biologiquement pertinents. Nous avons extrait quelques exemples d'interactions qui éclairent certains processus biologiques spécifiques au nématode et à *Arabidopsis*, et espérons que ces deux réseaux seront des ressources utiles à la communauté scientifique pour mieux comprendre la biologie moléculaire de ces deux organismes. Cela dit, nous ignorons la proportion des interactions Y2H qui correspondent à des interactions biophysiques qui n'ont jamais lieu *in vivo*. Différencier les interactions biologiquement pertinentes de ces « pseudo-interactions » représente un axe de recherches captivant pour le futur. En attendant, j'ai utilisé l'intégralité du réseau interactome d'*Arabidopsis* actuellement disponible pour répondre à des questions concernant les changements de son organisation topologique au cours des temps évolutifs. Ces études sont résumées dans le troisième et dernier chapitre de ce manuscrit.

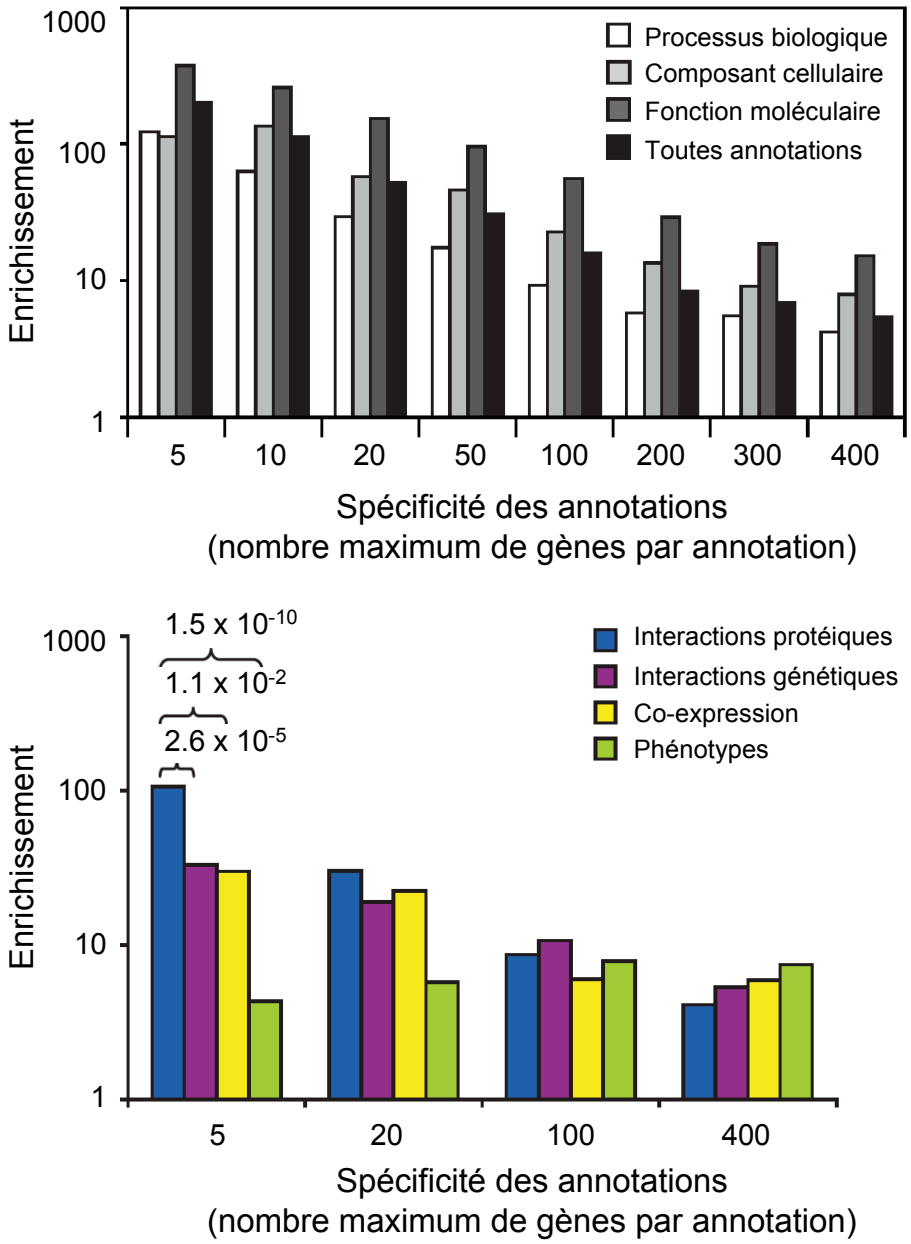


Figure 8 : Pertinence biologique des interactions physiques entre protéines détectées par double hybride en levure. Le graphe supérieur montre l'enrichissement en annotations GO partagées par les paires de protéines interagissant dans l'interactome expérimental d'*Arabidopsis* par rapport à des paires de protéines aléatoirement choisies. Le graphe inférieur montre l'enrichissement en annotations GO partagées par les paires de protéines interagissant physiquement, interagissant génétiquement, fortement co-exprimées, et similaires phénotypiquement chez le nématode. Dans les deux graphes, l'axe des abscisses représente la spécificité des annotations GO considérées (nombre maximum de gènes par annotation). Tous ces enrichissements sont significatifs ($p<0.05$) d'après le test de Fisher. Les valeurs de probabilités indiquées dans le graphe inférieur montrent que l'enrichissement en annotations GO partagées très précises (moins de 5 gènes par annotation) est significativement plus élevé pour les paires de protéines interagissant physiquement que pour les autres types de paires de protéines considérées. Voir les **documents joints 5 et 6** pour détails sur les données utilisées et leur traitement informatique.

DOCUMENT JOINT 4

Titre : Literature-Curated Interactions.

Auteurs : Cusick ME*, Yu H*, Smolyar A, Venkatesan K, Carvunis AR, Simonis N, Rual JF, Borick H, Braun P, Dreze M, Vandenhaute J, Galli M, Yazaki J, Hill DE, Ecker JR, Roth FP, Vidal M

Description : article de recherche original paru dans le magazine *Nature Methods* en 2009

Contribution : Ma contribution à cet article englobe exclusivement les analyses de données concernant *Arabidopsis thaliana.* J'ai aussi participé activement à la réalisation de l'*addendum*.

Literature-curated protein interaction datasets

Michael E Cusick[1,2,9], Haiyuan Yu[1,2,9], Alex Smolyar[1,2], Kavitha Venkatesan[1,2,8], Anne-Ruxandra Carvunis[1–3], Nicolas Simonis[1,2], Jean-François Rual[1,2,8], Heather Borick[1,2,8], Pascal Braun[1,2], Matija Dreze[1,2], Jean Vandenhaute[4], Mary Galli[5], Junshi Yazaki[5,6], David E Hill[1,2], Joseph R Ecker[5,6], Frederick P Roth[1,7] & Marc Vidal[1,2]

High-quality datasets are needed to understand how global and local properties of protein-protein interaction, or 'interactome', networks relate to biological mechanisms, and to guide research on individual proteins. In an evaluation of existing curation of protein interaction experiments reported in the literature, we found that curation can be error-prone and possibly of lower quality than commonly assumed.

An essential component of systems biology is discovery of the network of all possible physical protein-protein interactions (PPIs), the 'interactome' network[1–3]. There are two complementary ways to obtain comprehensive PPI information. One is to systematically test all pairwise combinations of proteins for physical interactions at proteome scale with a high-throughput assay[3]. The alternative is to curate all publications in the literature, each describing one (or a few) PPI(s) assayed at low throughput[4], and then make the curation accessible in interaction databases. As neither strategy can come close to allowing us to discover the full interactomes yet[5–7], the matter arises as to which strategy can best fill in the missing pieces.

High-throughput protein interaction assays

Two approaches are in frequent use for high-throughput mapping of protein interactions at proteome scale. Yeast two-hybrid assays attempt to identify binary interactions[8,9], whereas co-affinity purification followed by mass spectrometry identifies presence in a protein complex[10] but may not accurately determine the binary interactions between proteins within a complex[7]. Other technologies exist for mapping both binary interactions and presence in the same complex[11], but none can yet be routinely scaled up for high-throughput assays, although recently, a protein complementation assay allowed a large-scale mapping of the yeast interactome[12].

Curating protein interactions

Manual curation of protein interactions from literature began with pioneering curation for the yeast *Saccharomyces cerevisiae* by the Yeast Proteome Database (YPD)[13]. Those early efforts demonstrated that effective curation was possible and also broadly aimed to capture all types of functional and genomic information, not only PPIs. Genomic databases dedicated to a single model organism arose in parallel with genome sequencing projects, for example, the Saccharomyces Genome Database (SGD)[14] and The *Arabidopsis* Information Resource (TAIR)[15]. Although initially devoted to sequence information, many of these databases eventually added many types of literature-curated information, including PPI data. In time, the publications reporting PPIs exceeded the capacity of specialized genome databases and led to

[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA. [2]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [3]Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications de Grenoble (TIMC-IMAG), Unité Mixte de Recherche 5525 Centre National de la Recherche Scientifique (CNRS), Faculté de Médecine, Université Joseph Fourier, 38706 La Tronche Cedex, France. [4]Unité de Recherche en Biologie Moléculaire, Facultés Universitaires Notre-Dame de la Paix, 61 Rue de Bruxelles, 5000 Namur, Wallonia, Belgium. [5]Genomic Analysis Laboratory and [6]Plant Biology Laboratory, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA. [7]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue, Boston, Massachusetts 02115, USA. [8]Present addresses: Novartis Institutes for Biomedical Research, 250 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA (K.V.), Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA (J.-F.R.) and Department of Biological Sciences, 132 Long Hall, Clemson University, Clemson, South Carolina 29634, USA (H.B.). [9]These authors contributed equally to this work. Correspondence should be addressed to M.V. (marc_vidal@dfci.harvard.edu) or M.E.C. (michael_cusick@dfci.harvard.edu).

PUBLISHED ONLINE 30 DECEMBER 2008; DOI:10.1038/NMETH.1284

BOX 1 INTEROLOGS

Interologs are *in silico* predictions of protein interactions in one species between a pair of proteins whose orthologs are known to interact in another species[62–64]. The assumption that interologs are more likely true than not is widely held[65,66]. Recent evaluations have now revisited this assumption in several ways[24,67].

The most important question is where to draw the line for interspecies transfer. For instance, is mouse-human transfer close enough but more evolutionarily distant mammals not close enough? Actually, it is not the species relatedness but the sequence relatedness that really matters. Interolog transfers are only accurate for especially high sequence similarity[24,64]. Hence, interolog predictions with low sequence conservation should not be accepted, even between closely related species[24].

Investigations of intrinsic disorder in proteins have also unsettled the certainty that protein interactions are highly conserved. There are two types of interacting surfaces in proteins. Domain-domain interactions are more prevalent in stable protein complexes, whereas domain-disorder interactions are more transient[2,68,69]. Domain-disorder interactions evolve much faster than domain-domain interactions[70]. The proportion of protein interactions that are of the domain-disorder type versus the domain-domain type is not known, even approximately, for any species. Still, the likely considerable proportion of poorly conserved domain-disorder interactions means that the proportion of nonconserved interactions is substantial[24].

In the one experimental test of interologs so far, only one-third of the sample set of yeast interactions found by yeast two-hybrid were reproduced by yeast two-hybrid between the *C. elegans* orthologs[63]. Perhaps the large evolutionary distance between yeast and worm precluded a higher success rate, and mouse-human interologs might have a better success rate, but that supposition has not been experimentally tested.

In light of all these reappraisals, curation policies are changing. For instance, one interaction database has stopped transferring nonhuman interactions to human[19], a change from earlier practice[48]. Other interaction databases may follow suit. Alternatively, those interactions predicted by interolog extrapolation could be explicitly delineated in databases from those experimentally demonstrated, so the user could chose the appropriate data to examine. Either policy becomes complicated because species of the interactors are not often provided in publications[30,33]. Overall, it would seem best practice to only curate the species for which there is direct experimental evidence; in reality, doing so is difficult.

the creation of databases dedicated to PPIs, for example, the Munich Information Center for Protein Sequence (MIPS) protein interaction database[16], the Biomolecular Interaction Network Database (BIND)[17], the Database of Interacting Proteins (DIP)[18], the Molecular Interaction database (MINT)[19] and the protein Interaction database (IntAct)[20]. More recent PPI curation efforts, the Biological General Repository for Interaction Datasets (BioGRID)[21] and the Human Protein Reference Database (HPRD)[22], have attempted larger-scale curation of data from more manuscripts and more interactions.

High-throughput efforts versus literature curation

High-throughput approaches contrast in several attributes with literature-curation strategies (**Table 1**). Literature-curated collections represent the accumulation of thousands of small-scale, 'hypothesis-driven' investigations, whereas high-throughput experiments are 'discovery-based', designed to discover new biology without a priori expectations of what could be learned. Because literature-curated datasets are hypothesis-driven, biological functions of interacting proteins often, though not always, can be inferred from the actual study design. Discovery-based high-throughput datasets do not present this advantage, though function can sometimes be inferred through additional analyses[23]. Hypothesis-driven studies set up an inevitable study bias[7], in that what has been successfully investigated before tends to be investigated again, whereas high-throughput screens avoid study bias[24]. The completeness, or the portion of the proteome that has been tested for interactions[5], can be precisely estimated in a carefully designed high-throughput study[5,7,25], but this is not so even for the largest literature-curated datasets because negative results, the pairs tested but not found to interact, are almost never reported.

Estimating reliability—the portion of reported interactions that are valid (and hence reproducible)—is daunting. For high-throughput datasets, the introduction of an empirical framework for interactome mapping now allows experimental estimation of reliability parameters[5]. Previously, reliability of high-throughput datasets was routinely estimated by measuring the overlap with a reference set of gold-standard positives (GSP). Several caveats must be considered when constructing GSPs. The assays used to generate a GSP have to match as closely as possible the assays used to generate the experimental dataset, especially indirect co-complex versus binary representation[7]. A GSP should be as unbiased as possible, sampling all, or at least most, parts and processes of the cell[26], and a GSP must be of the highest reliability and reproducibility[27].

Literature-curated datasets are used for appraisal of the reliability of experimental PPI datasets, for predicting PPIs, for predicting other features such as protein function and for benchmarking data-mining methodologies[28–30]. In these efforts, the superior reliability of literature-curated PPI datasets, versus high-throughput datasets, is generally presumed. High-quality reference datasets of PPIs are integral for empirical estimation of the reliability and size of interactome maps[5,7,25,27]. Confidence in literature curation is accordingly a prerequisite for generating useful reference datasets. Whether literature-curated PPI datasets really have exceptional reliability has not been thoroughly investigated.

Table 1 | Comparison of strategies toward completing an interactome map

Attribute	High-throughput	Literature-curated
Investigation	Discovery-based	Hypothesis-driven
Functional inference	Determinable from network?	Determinable from study design?
Study bias	Unbiased	Biased
Completeness	Estimable	Inestimable
Reliability	Determinable	Indeterminable

Completeness and replication of literature-curated datasets

As PPIs supported by multiple publications should be more reliable than those supported by only a single publication, we assessed the proportion of multiply supported PPIs for yeast. We ranked the 11,858 literature-curated yeast PPIs in BioGRID[21] (LC-all downloaded in mid-2007). Only 25% of LC-all PPIs have been described in multiple publications (**Fig. 1**), with just 5% and 2% of these pairs described in ≥3 or ≥5 publications, respectively. More than 75% of LC-all PPIs were thus described in a single publication. Consistent with this low portion of multiply supported PPIs, experimental retests have demonstrated a significantly lower quality for singly supported versus multiply supported literature-curated PPIs for yeast[7].

Similar investigations for human and for *Arabidopsis thaliana* showed comparably low proportions of multiply supported PPIs. In the initial search space of ~7,000 × 7,000 genes for a first-draft human interactome mapping project, there are 4,067 binary literature-curated interactions[31]. Only 15% of these PPIs have been described in multiple publications (**Fig. 1**), with just 5% and 1% described in ≥3 or ≥5 publications, respectively. More than 85% of human PPIs in the literature-curated set are supported by a single publication, greater than the 75% for yeast. The set of *Arabidopsis* PPIs was collected from the only two protein-interaction databases that curate *Arabidopsis* protein interactions, TAIR[15] and IntAct[20]. The *Arabidopsis* PPI dataset has fewer interactions supported by data in multiple manuscripts than yeast or human (**Fig. 1**), with just 1% and 0.1% described in ≥3 or ≥5 publications, respectively, with 93% of available *Arabidopsis* literature-curated PPIs supported by data in only a single publication. All told, the number of PPIs supported by data in multiple publications is small.

Literature-curated datasets are reported to be composed primarily of small-scale experiments[21,32]. To assess the presumption that PPI databases are small-scale, we measured the proportion of total PPIs identified in high-throughput experiments. For yeast, we ranked the 8,933 interactions supported by data in a single publication by the number of distinct PPIs reported in each corresponding publication (**Fig. 2a**). More than 60% of protein pairs were curated from manuscripts that described more than 10 interactions, all extracted from 6% of all the manuscripts curated. One-third of the total interactions came from less than 1% of all manuscripts that each describe 100 or more interactions (**Fig. 2a**), which would reasonably be considered high-throughput. Thus, the yeast literature-curated dataset of PPIs supported by a single publication record is not composed solely of validated interactions from small-scale studies but has a marked portion of PPIs derived from high-throughput experiments. We similarly analyzed a dataset of human curated PPIs[31] and found that this human PPI dataset is predominantly low-throughput (**Fig. 2b**), possibly because at the time these PPIs had been downloaded from the databases few medium- to high-throughput experiments had been published. For *Arabidopsis,* the proportion of the total literature-curated interactions derived from medium to high-throughput manuscripts is about the same as for yeast (**Fig. 2c**). In sum, many available literature-curated PPI datasets are populated widely by PPIs from high-throughput experiments.

As an assessment of the completeness for literature-curated datasets is not possible (**Table 1**), we evaluated database overlaps as a surrogate for completeness, on the argument that different PPI databases should curate from the same set of PubMed reports. BioGRID reports the greatest completeness for yeast but is not yet a participating member of the International Molecular Exchange (IMEx) consortium[33,34], so we could not use this database for this analysis. The three IMEx members that do substantial curation of yeast PPIs (MINT, IntAct and DIP) had surprisingly low overlap of curated PPIs (**Fig. 3a**). That the overlap is so small after years of intense curation of protein interactions is reason for concern. The small overlap is not due to differential curation of high-throughput data, as removal of the six largest PPI reports[35–40] still left small overlaps, especially of IntAct with the other two databases (**Fig. 3a**).

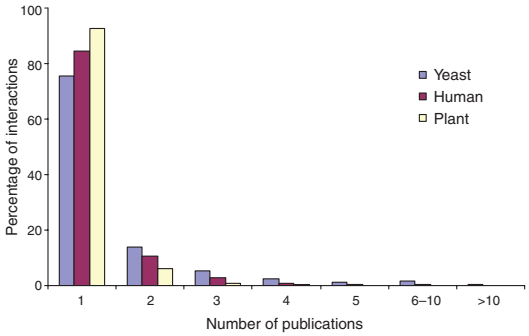


Figure 1 | Distribution of the number of published manuscripts supporting each interaction. Data are from the dataset of yeast protein interactions downloaded from the BioGRID[21] database, the literature-curated dataset of human protein interactions, and the literature-curated dataset of *Arabidopsis* protein interactions.

Are the small overlaps due to curating different manuscripts or to differential curation of data from overlapping sets of manuscripts? The answer seems to be that vastly different sets of manuscripts are curated because the curation of PubMed reports also shows small overlap (**Fig. 3b**). For multiply supported interactions (those reported in two or more published studies), the low overlap remains (**Fig. 3c**), though the number of interactions drops greatly. Hence even the most heavily investigated interactions, those most likely to be multiply curated, do not seem to be comprehensively covered. In sum, surrogate estimates of completeness of literature-curated datasets, at least for yeast, suggest that coverage of curated literature is far from comprehensive.

These investigations suggest, but in no way demonstrate, that literature-curated PPIs may not have the high reliability often attributed to them. There has not yet been any intensive investigation of the actual reliability of literature-curated PPI datasets. To do so, we recurated representative samples of existing literature-curated PPI datasets for three model organisms—yeast (*Saccharomyces cerevisiae*), human and plant (*Arabidopsis thaliana*)—and found that the literature curation of PPI publications can be less than impeccable.

Estimating curation reliability by recuration

For yeast, we recurated in detail 100 randomly selected pairs from the yeast dataset of singly supported interactions (**Fig. 1**). After evaluating several relevant criteria, we assigned each interaction a score of 0 (no confidence), 1 (low confidence or unsubstantiated) or 2 (substantiated or of high confidence) (see detailed protocol below).

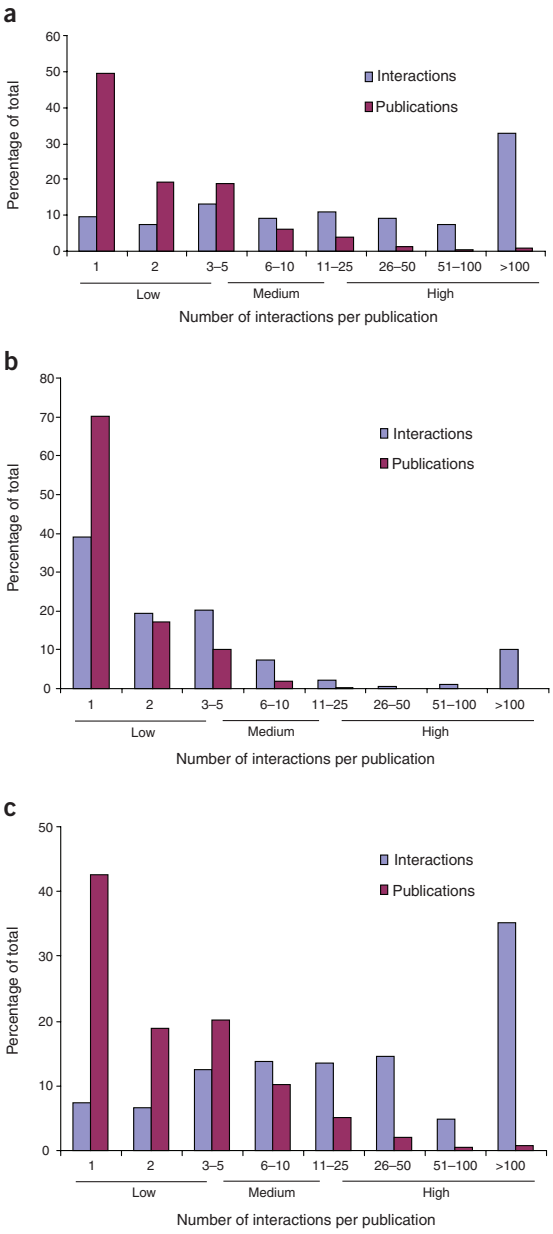


Figure 2 | Distribution of the publications in literature-curated datasets by the number of interactions reported in the publication. (**a**–**c**) Distribution in the yeast (**a**), human (**b**) and *Arabidopsis* (**c**) literature-curated PPI datasets supported by a single publication.

The results of this recuration (**Fig. 4a** and **Supplementary Table 1** online) showed that 25% of the sampled interactions could be substantiated whereas three-quarters were not. Of the interacting pairs in the sample, 35% were incorrectly curated. These observations explain the poor reliability, relative to high-throughput datasets, of the singly supported literature-curated dataset in both computational and experimental comparative analyses[7].

For human PPI recuration, we prepared two curation datasets. One was a presumed high-confidence literature-curated dataset of interactions (LC-multiple) within the initial search space of ~7,000 × 7,000 genes for a first-draft human interactome mapping project[31] corresponding to pairs reported two or more times (two different PubMed identifiers) and curated in two or more databases (the five databases used were HPRD[22], BIND[41], MINT[19], MIPS mammalian database[16] and DIP[18]). From within this small (275 multiply supported interactions) 'hypercore' set of protein interactions[31], 188 interactions were left for recuration, after excluding homodimers.

The other dataset was a lower-confidence literature-sampled dataset of 188 interactions, generated by randomly selecting interactions from the initial search space[5]. Most of these interactions have one publication linked to the interaction, but because sampling was random, several interactions had been reported in more than one publication.

In the LC-multiple recuration set, 38% of the initial curation unit values (defined in **Table 2**) were wrong (**Fig. 4b** and **Supplementary Table 2** online). The most common errors were wrong species (assignment to a species other than human (**Box 1**)) and absence of a binding experiment supporting the interactions. Although 40% of the human LC-multiple interactions were not supported by multiple publications after recuration, most of these interactions were supported in only one manuscript instead of two or more, perhaps constituting a 'secondary' dataset of reduced confidence (**Supplementary Table 2**).

For the presumably lower confidence literature-sampled dataset of 160 interacting pairs (after removing interactions that had more than one supporting publication), 45% of interactions were not validated (**Fig. 4c** and **Supplementary Table 3** online) and 55% were validated. Almost half of the randomly sampled interactions were not supported by recuration. The most common errors here were wrong species and wrong protein name (**Fig. 4c**).

Yeast and human have the largest amount of curated literature in interaction databases[21,42]. A model organism with fewer curated interactions might yield different results. We curated 100 higher-confidence protein interactions of *Arabidopsis* from the two interaction databases that curate *Arabidopsis*, TAIR[15] and IntAct[20]. The results were improved relative to the yeast or human results, as 6 interactions and 24 curation units were scored incorrect (**Supplementary Table 4** online and **Table 2**). We scored the 24 errors as follows: 9 as 'no binding experiment'; 6 as 'no binding partner'; 6 as 'indirect'; and 3 as 'wrong protein'. The improved results for *Arabidopsis* likely reflect a smaller research community whose members can maintain uniformity in gene and protein names[15].

Why is reliability of literature curation so low?

Our findings of large error rates in curated protein interaction databases, at least for yeast and human, are consistent with recent hints that the quality of literature-curated datasets may not be as high as widely perceived[23,29,43–45]. Perhaps occasionally curator error is responsible. However, we suggest that the errors are due not so much to curators but to the simple reality that extracting accurate information from a long free-text document can be extremely difficult. Gene name confusion is particularly thorny[30,46]. An example from our curated yeast sample illustrates the difficulties. A purification with

a tandem affinity purification tag with Vps71/Swc6 (slash separates synonymous approved names) as bait[47] pulls down a protein named Swc3, but double-checking this finds that the coresponding open reading frame is actually *SWC3* (locus name YAL011w), and not the *ALR1*/*SWC3* (locus name YOL130w) open reading frame curated in the database. A shared synonym thoroughly muddled the curation.

Common curation practice has been to score equally every interaction reported in a publication[21,48], even though common experimental practice consists of first screening for new interacting proteins, then focusing on and substantiating one or a few of the most interesting interactions while leaving the others aside. Perhaps more curator judgment is needed, applying higher ranking to verified interactions and lower ranking to unverified 'along for the ride' interactions. Users can then choose the confidence level suitable to their needs. Given the demands of systems biology, perhaps biological databases should no longer serve as mere repositories of data but should appraise data[49]. Recent small incremental steps at developing a confidence score for curated PPIs have been taken[50,51].

The difficulty of literature curation is often underappreciated[4,21,30]. The lack of formal representation of PPIs in published manuscripts makes it difficult, if not impossible, to extract the PPI data in usable form. Designation of the species of origin of the protein interactors, an absolutely critical piece of information, is often buried or lacking altogether; protein or gene name synonyms used in a particular manuscript are hard to trace back to the canonical protein or gene names, especially in older manuscripts; and standardized descriptions, sometimes all description, of the methods used are absent. Faced with these difficulties, the curator is forced either to omit the information altogether (curated false negative) or make an educated guess, even though guesses, albeit educated ones, are often erroneous (curated false positive). The small overlaps noted between curated yeast interactions in different databases (**Fig. 3**) might be due to differential treatments of potential curated false negatives.

Our observations that literature-curated datasets have inherent reliability difficulties should influence thinking about proper generation of positive reference sets[29]. Already the human positive reference sets generated in our sampled recuration efforts have proven useful in multiple investigations[5,27,52].

It is still rarely doubted that literature-curated interactions are better than datasets generated with any high-throughput technology[6,21,53,54]. Our findings lead us to argue otherwise. If rigorously carried out, high-throughput experimental PPIs can be of higher quality than literature-curated interactions[5,25,27].

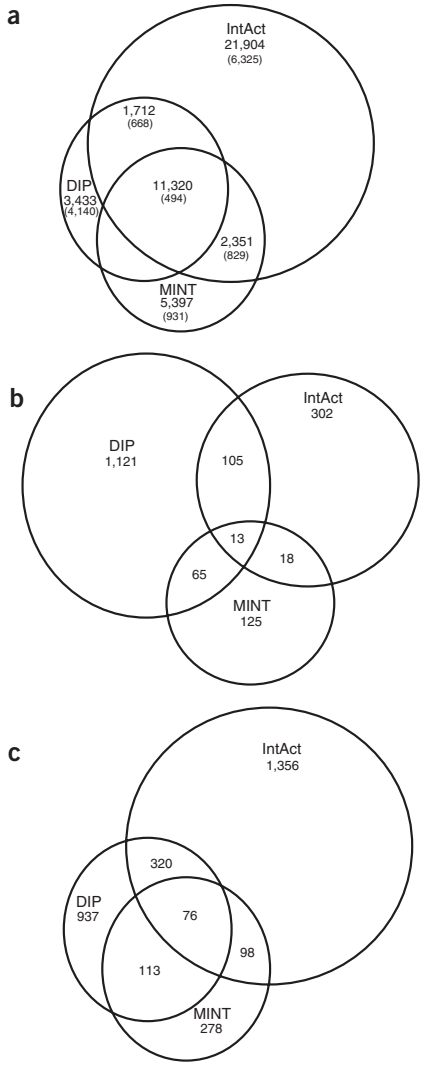


Figure 3 | Overlaps of reported curation for yeast PPIs. (**a**) Overlaps of the total number of reported binary PPIs or after removing the largest high-throughput yeast PPI reports (numbers in parentheses). (**b**) Overlaps of the PubMed reports curated. (**c**) Overlaps after removing multiply supported interactions.

Improving reliability of literature-curated PPI datasets

The difficulty of curation arises partly because PPI data are not submitted to databases in standardized format upon publication[55,56], unlike DNA-sequence or protein-structure data. The difficulty that curators have in extracting PPI information from manuscripts has led to the promulgation of the minimal information about a molecular interaction experiment (MIMIx) initiative[55]. MIMIx standardizes the presentation of PPI information in published manuscripts regarding species, protein names, methodological descriptions and protein identifiers, making it easier for curators to extract the pertinent information[33]. Once widely promulgated, which should come about sooner if the structured digital abstract[57,58] project gains traction, MIMIx will greatly improve curation such that the erroneous curation uncovered here will be lessened. Other minimal information initiatives for large-scale biology data are under development[59], and their development is wholeheartedly endorsed by the biocuration community so as to reduce curation error[30].

Our findings, although possibly critical of the quality of existing PPI curation, must not be used for quality evaluation of the underlying scientific literature. Actually, some PPI publications do warn of possible cross-contamination[60] or even occasionally provide heuristic confidence scores[61], warnings that should be taken into account in the curation.

Table 2 | Summary of curation results for human and *Arabidopsis*

Sampled dataset	Interaction units	Curation units[a]
Human LC-multiple	Correct: 172 (91.5%) Incorrect: 16 (8.5%)	Correct: 362 (62%) Incorrect: 223 (38%)
Human literature sampled	Correct: 88 (55%) Incorrect: 72 (45%)	Correct: 88 (55%) Incorrect: 72 (45%)
Arabidopsis	Correct: 94 (94%) Incorrect: 6 (6%)	Correct: 201 (89.3%) Incorrect: 24 (10.7%)

[a]For human a curation unit is an interaction reported in one publication regardless of the number of databases curating the interaction. An interaction reported in three distinct manuscripts and curated in two databases represents three curation units. For *Arabidopsis* a curation unit is an interaction reported in one publication or one database. An interaction reported in three distinct manuscripts and with all three curated in the two *Arabidopsis* PPI databases represents six curation units.

Curation protocols

Yeast PPI recuration. For each randomly selected protein pair, we read in detail the reporting manuscript (text, figures and supporting information), searching for all supporting information about the presumed interaction. We answered five questions for each protein pair. (i) Is there any information in the manuscript that supports the interaction? (ii) Has the experiment supporting the interaction been done at low throughput? As the perception persists that low-throughput experiments have greater reliability[21], knowing this is important. (iii) Are the interacting proteins mentioned together in the text? Lack of co-citation indicates that the authors did not actually focus on that particular interaction. (iv) Is the interaction supported by multiple methods? (v) Is the interaction likely direct? That is, did the method(s) used gauge binary interaction or membership in the same complex?Lastly, we assigned to each interacting pair an overall score of 0 (no confidence: no mention of the interacting pair, negative answer to the other four questions), 1 (low confidence: interacting pair is mentioned but the interaction is not substantiated by alternative methods) or 2 (high confidence: multiple validations by alternative methods).

Two different curators independently curated and scored all interactions. A third independent curator resolved the few scoring conflicts.

Human PPI recuration. We compiled the human PPI dataset as previously described[31]. First, we classified the method codes used by each database as binary (for example, two hybrid methods) or indirect (for example, co-affinity purification)[25]. Then, we selected only protein interactions with binary support for subsequent analysis[31]. The multiply supported literature-curated dataset comprised 585 curation units (**Table 2**) representing 188 PPIs, each reported in two or more publications and curated in two or more PPI databases. The dataset randomly selected from the full human literature-curated dataset[31] comprised 240 curation units representing 188 PPIs.

The types of information we collected during recuration were: the gene symbols and GeneID of each interactor; the associated PubMed identifier; the name and the identifier number of the interaction assay following the standard 'interaction detection method' vocabulary implemented in Proteomics Standards Initiative–Molecular Interactions[34]; the region of each protein used for the interaction assay (marked full-length if the entire protein sequence was used); the species for each interacting protein; and clarifying free-text comments used by the curator when needed. Interpretative fields included; an assessment of whether the interaction was bona fide, that is, not erroneous; an assessment of whether the interaction was indeed binary; and an error field, using a simple controlled vocabulary to classify erroneous curation units such as 'wrong protein', 'wrong species', 'no binding experiment', 'no binding partner' (interaction between the proteins is not shown), 'indirect' (no direct interaction is shown), 'redundant PubMed identifier' (some manuscripts (usually crystallographic structure determination manuscripts) have two distinct PubMed identifier

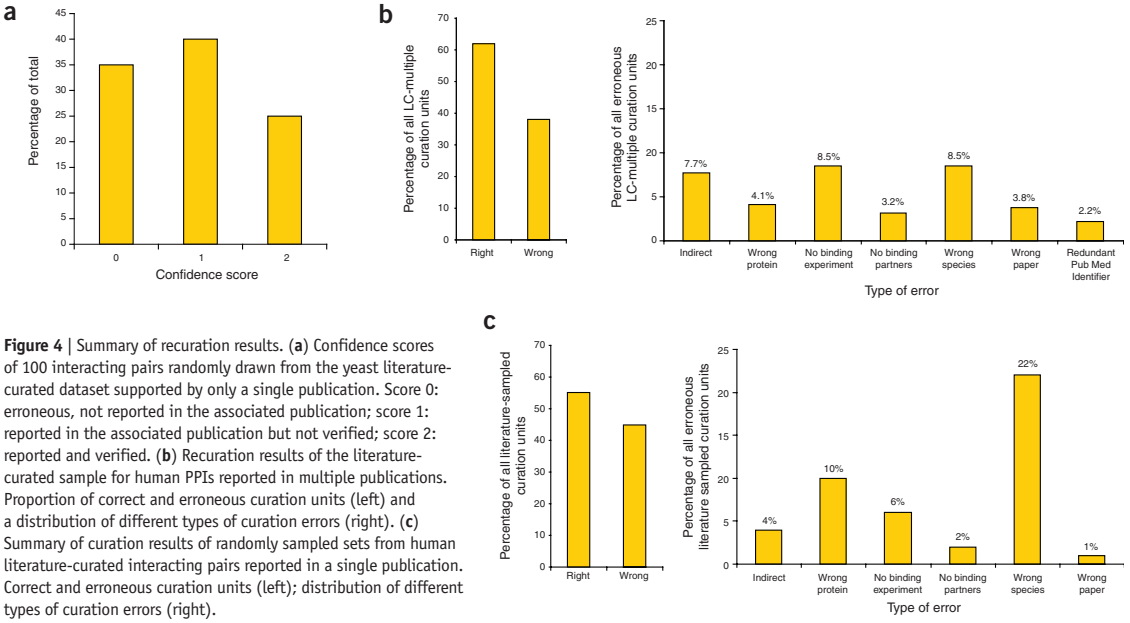


Figure 4 | Summary of recuration results. (**a**) Confidence scores of 100 interacting pairs randomly drawn from the yeast literature-curated dataset supported by only a single publication. Score 0: erroneous, not reported in the associated publication; score 1: reported in the associated publication but not verified; score 2: reported and verified. (**b**) Recuration results of the literature-curated sample for human PPIs reported in multiple publications. Proportion of correct and erroneous curation units (left) and a distribution of different types of curation errors (right). (**c**) Summary of curation results of randomly sampled sets from human literature-curated interacting pairs reported in a single publication. Correct and erroneous curation units (left); distribution of different types of curation errors (right).

numbers in PPI databases, and thus do not constitute two distinct manuscripts supporting an interaction).

If there was no information about the region of the protein responsible for the interaction, then the default we used was the full-length protein. If the species of the interacting proteins was not stated in a manuscript, a distressingly common occurrence, our default was to record the species as human. Thus, many interactions that did not involve human proteins might have been curated as human, so we may have underestimated the actual error rate. If the interaction was legitimate but one or the other protein partner was a species other than human, then we did not call this interaction bona fide. An interaction supported by multiple methods had to have just one bona fide and binary method to be recorded as legitimate; other methods apart from this one could be not binary or erroneous and not affect the final scoring.

Generally, we labeled yeast two-hybrid and other protein complementation assays as well as structural determinations as binary. We considered immunoprecipitation and co-affinity purification methodologies done *in vivo* that assess membership in the same complex not binary, whereas we considered those done *in vitro* with, for example, recombinant proteins as binary. If a tagged protein was heterologously expressed in a cell to pull down endogenous proteins, then we called such an interaction not binary. However, if both proteins of an interacting pair were heterologously expressed in a cell and shown to interact by, for example, pull down, then we considered such an interaction binary, as it is unlikely that an endogenous host protein mediates the interaction between the two heterologous proteins. If a co-immunoprecipitation done *in vivo* occurred in both orientations (protein A immunoprecipitation pulls down protein B and protein B immunoprecipitation pulls down protein A), then we judged this interaction as binary. As experimental procedures are often not described in sufficient detail to allow judgment of binary interaction, consistent policies in this regard were difficult to achieve.

Usually, we considered structural determinations to be binary, except for protein complexes of more than two proteins in which the interacting protein pairs did not actually contact each other in the solved structure. We scored solved structures that required a small third entity for crystal formation (for example, GTP, phosphatidylinositol) as binary, even though the interaction does not occur unless the small molecule is present.

A particular curation unit could have more than one error, though we counted only the most prominent error.

***Arabidopsis* PPI recuration.** For *Arabidopsis*, we defined high-confidence interactions as those supported by two manuscripts or by two databases. In the initial search space of an ongoing *Arabidopsis* interactome mapping project, we collected 100 such interactions. We chose the union (or) instead of the intersection (and) for *Arabidopsis*, in contrast to human, so that a sufficiently sized sample of interactions was available. Otherwise, curation policies were as for human, including the error codes, but adding the name and the identifier of the 'participant identification method' vocabulary implemented in PSI-MI[34].

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This work was supported by US National Human Genome Research Institute grants R01 HG001715 to M.V. and F.P.R., P50 HG004233 to M.V. and R01 HG003224 to F.P.R. by funds from the W.M. Keck Foundation to M.V. by an award (DBI-0703905) from the National Science Foundation to M.V., J.R.E. and D.E.H. and by Institute Sponsored Research funds from the Dana-Farber Cancer Institute Strategic Initiative to M.V. and CCSB. is a Chercheur Qualifié Honoraire from the Fonds de la Recherche Scientifique (FRS-FNRS, French Community of Belgium). We thank all members of CCSB for constructive discussions.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Cusick, M.E., Klitgord, N., Vidal, M. & Hill, D.E. Interactome: Gateway into systems biology. *Hum. Mol. Genet.* **14**, R171–R181 (2005).
2. Bader, S., Kuhner, S. & Gavin, A.C. Interaction networks for systems biology. *FEBS Lett.* **582**, 1220–1224 (2008).
3. Vidal, M. Interactome modeling. *FEBS Lett.* **579**, 1834–1838 (2005).
4. Roberts, P.M. Mining literature for systems biology. *Brief. Bioinform.* **7**, 399–406 (2006).
5. Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2008).
6. Stumpf, M.P. *et al.* Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. USA* **105**, 6959–6964 (2008).
7. Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
8. Parrish, J.R., Gulyas, K.D. & Finley, R.L. Jr. Yeast two-hybrid contributions to interactome mapping. *Curr. Opin. Biotechnol.* **17**, 387–393 (2006).
9. Ito, T. *et al.* Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol. Cell. Proteomics* **1**, 561–566 (2002).
10. Köcher, T. & Superti-Furga, G. Mass spectrometry-based functional proteomics: from molecular machines to protein networks. *Nat. Methods* **4**, 807–815 (2007).
11. Suter, B., Kittanakom, S. & Stagljar, I. Interactive proteomics: what lies ahead? *Biotechniques* **44**, 681–691 (2008).
12. Tarassov, K. *et al.* An in vivo map of the yeast protein interactome. *Science* **320**, 1465–1470 (2008).
13. Garrels, J.I. YPD–A database for the proteins of *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **24**, 46–49 (1996).
14. Hong, E.L. *et al.* Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.* **36**, D577–D581 (2008).
15. Swarbreck, D. *et al.* The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**, D1009–D1014 (2007).
16. Pagel, P. *et al.* The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**, 832–834 (2005).
17. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250 (2003).
18. Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
19. Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res.* **35**, D572–D574 (2007).
20. Kerrien, S. *et al.* IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.* **35**, D561–D565 (2007).
21. Reguly, T. *et al.* Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol.* **5**, 11 (2006).
22. Mishra, G.R. *et al.* Human protein reference database—2006 update. *Nucleic Acids Res.* **34**, D411–D414 (2006).
23. Myers, C.L., Barrett, D.R., Hibbs, M.A., Huttenhower, C. & Troyanskaya, O.G. Finding function: evaluation methods for functional genomic data. *BMC Genomics* **7**, 187 (2006).
24. Mika, S. & Rost, B. Protein-protein interactions more conserved within species than across species. *PLoS Comput. Biol.* **2**, e79 (2006).
25. Simonis, N. *et al.* Empirically-controlled mapping of the *Caenorhabditis elegans* protein-protein interaction network. *Nat. Methods* **6**, 47–54 (2008).
26. Jansen, R. & Gerstein, M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr. Opin. Microbiol.* **7**, 535–545 (2004).
27. Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* **6**, 91–97 (2008).
28. Bader, G.D. & Hogue, C.W. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat. Biotechnol.* **20**, 991–997 (2002).
29. Ramírez, F., Schlicker, A., Assenov, Y., Lengauer, T. & Albrecht, M. Computational analysis of human protein interaction networks. *Proteomics* **7**, 2541–2552 (2007).
30. Howe, D. *et al.* The future of biocuration. *Nature* **455**, 47–50 (2008).
31. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).

32. Peri, S. *et al.* Development of Human Protein Reference Database as an initial platform for approaching systems biology in humans. *Genome Res.* **13**, 2363–2371 (2003).
33. Orchard, S. *et al.* Submit your interaction data the IMEx way. A step by step guide to trouble-free deposition. *Proteomics* **7**, 28–34 (2007).
34. Kerrien, S. *et al.* Broadening the horizon - Level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.* **5**, 44 (2007).
35. Gavin, A.C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
36. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
37. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
38. Krogan, N.J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006).
39. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
40. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
41. Alfarano, C. *et al.* The Biomolecular Interaction Network Database (BIND) and related tools 2005 update. *Nucleic Acids Res.* **33**, D418–D424 (2005).
42. Mathivanan, S. *et al.* An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics* **7**, S19 (2006).
43. Gentleman, R. & Huber, W. Making the most of high-throughput protein-interaction data. *Genome Biol.* **8**, 112 (2007).
44. Mackay, J.P., Sunde, M., Lowry, J.A., Crossley, M. & Matthews, J.M. Protein interactions: is seeing believing? *Trends Biochem. Sci.* **32**, 530–531 (2007).
45. Mackay, J.P., Sunde, M., Lowry, J.A., Crossley, M. & Matthews, J.M. Response to Chatr-aryamontri *et al.*: Protein interactions: to believe or not to believe? *Trends Biochem. Sci.* **33**, 242–243 (2008).
46. Nelson, D.R. Gene nomenclature by default, or BLASTing to Babel. *Hum. Genomics* **2**, 196–201 (2005).
47. Krogan, N.J. *et al.* A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1. *Mol. Cell* **12**, 1565–1576 (2003).
48. Zanzoni, A. *et al.* MINT: a Molecular INTeraction database. *FEBS Lett.* **513**, 135–140 (2002).
49. Philippi, S. & Kohler, J. Addressing the problems with life-science databases for traditional uses and systems biology. *Nat. Rev. Genet.* **7**, 482–488 (2006).
50. Kiemer, L., Costa, S., Ueffing, M. & Cesareni, G. WI-PHI a weighted yeast interactome enriched for direct physical interactions. *Proteomics* **7**, 932–943 (2007).
51. Chatr-Aryamontri, A., Ceol, A., Licata, L. & Cesareni, G. Protein interactions: integration leads to belief. *Trends Biochem. Sci.* **33**, 241–242 (2008).
52. Boxem, M. *et al.* A protein domain-based interactome network for *C. elegans* early embryogenesis. *Cell* **134**, 534–545 (2008).
53. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
54. Batada, N.N., Hurst, L.D. & Tyers, M. Evolutionary and physiological importance of hub proteins. *PLoS Comput. Biol.* **2**, e88 (2006).
55. Orchard, S. *et al.* The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.* **25**, 894–898 (2007).
56. Hermjakob, H. *et al.* The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat. Biotechnol.* **22**, 177–183 (2004).
57. Ceol, A., Chatr-Aryamontri, A., Licata, L. & Cesareni, G. Linking entries in protein interaction database to structured text: the FEBS Letters experiment. *FEBS Lett.* **582**, 1171–1177 (2008).
58. Gerstein, M., Seringhaus, M. & Fields, S. Structured digital abstract makes text mining easy. *Nature* **447**, 142 (2007).
59. Taylor, C.F. *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* **26**, 889–896 (2008).
60. Stevens, S.W. *et al.* Composition and functional characterization of the yeast spliceosomal penta-snRNP. *Mol. Cell* **9**, 31–44 (2002).
61. Fromont-Racine, M., Rain, J.C. & Legrain, P. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.* **16**, 277–282 (1997).
62. Walhout, A.J. *et al.* Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
63. Matthews, L.R. *et al.* Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res.* **11**, 2120–2126 (2001).
64. Yu, H. *et al.* Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.* **14**, 1107–1118 (2004).
65. Ramani, A.K., Bunescu, R.C., Mooney, R.J. & Marcotte, E.M. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.* **6**, R40 (2005).
66. Sharan, R. *et al.* Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA* **102**, 1974–1979 (2005).
67. Levy, E.D. & Pereira-Leal, J.B. Evolution and dynamics of protein interactions and networks. *Curr. Opin. Struct. Biol.* **18**, 349–357 (2008).
68. Tompa, P. & Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* **33**, 2–8 (2008).
69. Fuxreiter, M., Tompa, P. & Simon, I. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* **23**, 950–956 (2007).
70. Beltrao, P. & Serrano, L. Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comput. Biol.* **3**, e25 (2007).

measured in PBS[4], which may result in an overestimation of photostability compared to commonly used live-cell imaging conditions. The use of media depleted of vitamins for fluorescence imaging of live cultured cells appears to be a simple and efficient way to improve the performance of some widely used fluorescent proteins in various ensemble and single-molecule applications[1,5,6].

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This work was supported by Russian Academy of Sciences (Molecular and Cell Biology program and Innovation and Development support 31-236), Howard Hughes Medical Institute (55005618) and Rosnauka grants 02.512.11.2216 and 02.512.12.2053. D.M.C. and K.A.L. are supported by President of the Russian Federation grants MK-6119.2008.4 and MD-2780.2009.4.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Alexey M Bogdanov[1], Ekaterina A Bogdanova[1], Dmitriy M Chudakov[1], Tatiana V Gorodnicheva[2], Sergey Lukyanov[1] & Konstantin A Lukyanov[1]

[1]Shemiakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia.
[2]Evrogen JSC, Moscow, Russia.
e-mail: kluk@ibch.ru

1. Shaner, N.C., Patterson, G.H. & Davidson, M.W. *J. Cell Sci.* **120**, 4247–4260 (2007).
2. Bogdanov, A.M. *et al. Nat. Chem. Biol.* **5**, 459–461 (2009).
3. Werner, R., Manthey, K.C., Griffin, J.B. & Zempleni, J. *J. Nutr. Biochem.* **16**, 617–624 (2005).
4. Shaner, N.C., Steinbach, P.A. & Tsien, R.Y. *Nat. Methods* **2**, 905–909 (2005).
5. Kohl, T. & Schwille, P. *Adv. Biochem. Eng. Biotechnol.* **95**, 107–142 (2005).
6. Fernandez-Suarez, M. & Ting, A.Y. *Nat. Rev. Mol. Cell. Biol.* **9**, 929–943 (2008).

Recurated protein interaction datasets

To the Editor: In their recent Perspective, Cusick *et al.*[1] state "…curation can be error-prone and possibly of lower quality than commonly assumed." Although we welcome rigorous scrutiny of curation efforts, Cusick *et al.*[1] had arrived at their conclusions by misunderstanding the difference between the reliability of experimental data supporting protein interactions and the correctness of the curation process itself.

The aim of the IntAct molecular interaction database (IntAct)[2], the Database of Interacting Proteins (DIP)[3], the Molecular Interaction database (MINT)[4], the *Arabidopsis* Information Resource (TAIR)[5] and the Biological General Repository for Interaction Datasets (BioGRID)[6] critiqued by Cusick *et al.*[1] is to collect and organize experiments supporting protein-protein interactions into a comprehensive set of accurately annotated experimental data. These databases allow the biological community facile and searchable access to a vast repository of biological interactions for many purposes ranging from individual hypothesis generation to functional annotation to biological network analysis. The transparent and full representation of interactions in the primary literature is an essential component of such a repository and is necessary to assess the reliability of published data. As databases support many different uses of their data, they aim to incorporate the complete data as presented in the source publications, rather than selecting evidence they consider more reliable or otherwise privileged. The use of detailed and well-defined controlled vocabularies for annotation allows the database users to efficiently select subsets of data according to criteria relevant for their particular use.

In contrast, Cusick *et al.*[1] define a set of criteria for a specific use restricted only to direct pairwise protein-protein interactions, which they refer to as 'binary' interactions. They evaluate literature-curated datasets against these criteria and then assert that failure to meet their criteria represents "incorrect curation." The criteria defined by Cusick *et al.*[1] vary slightly from species to species but aim to select only direct interactions with multiple independent supporting reports. While this is one valid use, other users might, for example, look for all observed interactions of a given protein, whether direct or indirect, to subsequently assess the supporting evidence by reading the supporting publications. Whereas protein-protein interaction databases may also use the term 'binary' when referring to pairs of interacting proteins, our usage of the term refers to any interaction pair and makes no judgment regarding whether the interaction is direct or indirect.

We strongly object to the notion that inclusion of an interaction with limited supporting evidence of a direct interaction represents a curation error. On the contrary, most interaction databases always fully curate a given publication and would consider it an egregious omission if only a subset of the protein interactions reported in a publication or its supplementary material would be contained in the database. When information—for example, species information—in a publication is ambiguous, database curators attempt to contact the authors and only leave out data if clarification cannot be obtained.

In response to the claims of Cusick *et al.*[1], we reanalyzed interactions presented in their paper to identify actual curation errors, defined as inconsistencies between the original published data and their representation in our databases. Details of our analysis are available in the **Supplementary Note**, and we reannotated versions of the original tables supplied by Cusick *et al.*[1] (**Supplementary Tables 1–3**). The actual curation error rate was, in fact, consistently under 10%.

For the yeast dataset, we confirmed 4 actual curation errors among the 100 sample interactions from BioGRID chosen by Cusick *et al.*[1]; the curation error rate of 4% is precisely the value originally reported for the dataset[7] and an order of magnitude lower than the claim by Cusick *et al.*[1]: "Of the interacting pairs in the sample, 35% were incorrectly curated." For comparison, we analyzed a subset of the BioGRID data that is also present in the DIP database and identified 1 actual curation error out of 29 shared records, that is, a similarly low error rate of 3%.

For the human dataset, of the 220 sampled interactions annotated in MINT, only 10 were curation errors, corresponding to a curation error rate of 4.5%. Similarly, only 4 out of 42 curation records reported in DIP contained errors, a 9% curation error rate, or one-fifth of the 45% curation error rate implied by Cusick *et al.*[1].

For the *Arabidopsis thaliana*, the IntAct dataset contained 3 actual curation errors in 183 curation records, resulting in an error rate of 2%, less than one-fifth of the 10.7% rate claimed by Cusick *et al.*[1] in their Table 2. For TAIR, the actual error rate was only 3%, or less than one-third of the rate claimed by Cusick *et al.*[1].

Accurate and detailed curation is an arduous process both in terms of individual curator expertise and curation time. To optimize the use of public funding, the member databases of the International Molecular Exchange Consortium (IMEx)[8] DIP, IntAct and MINT coordinate their curation efforts to avoid unnecessary redundancy,

as described on the consortium webpage (http://imex.sf.net/). The low overlap between IMEx interaction datasets noted by Cusick *et al.*[1] is not, as claimed, an indicator for undersampling of the interaction space, but rather demonstrates the success of the international collaboration within the IMEx consortium.

In summary, when appropriately considering only actual curation errors rather than subjective reliability criteria intended to identify only the subset of directly interacting protein pairs, our analysis demonstrated a surprisingly narrow spread of 2–9% curation errors across datasets from three different species curated by five different interaction databases. This analysis testified to the precision of interaction database curation and substantiated the case for coordinated international efforts to curate biological interactions.

Note: Supplementary information is available on the Nature Methods website.

Lukasz Salwinski[1,10], Luana Licata[2,10], Andrew Winter[3,10], David Thorneycroft[4], Jyoti Khadake[4], Arnaud Ceol[2], Andrew Chatr Aryamontri[2], Rose Oughtred[5], Michael Livstone[5], Lorrie Boucher[6], David Botstein[5], Kara Dolinski[5], Tanya Berardini[7], Eva Huala[7], Mike Tyers[3,6], David Eisenberg[1,8], Gianni Cesareni[2,9] & Henning Hermjakob[4]

[1]University of California, Los Angeles Department of Energy Institute for Genomics and Proteomics, Los Angeles, California, USA. [2]Instituto di Ricovero e Cura a Carattere Scientificio Fondazione Santa Lucia, Rome, Italy. [3]School of Biological Sciences, The University of Edinburgh, Edinburgh, UK. [4]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [5]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, USA. [6]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. [7]Department of Plant Biology, Carnegie Institution for Science, Stanford, California, USA. [8]Department of Chemistry and Biochemistry, Howard Hughes Medical Institute, University of California, Los Angeles, California, USA. [9]Department of Biology, University of Rome Tor Vergata, Rome, Italy. [10]These authors contributed equally to this work.

e-mail: hhe@ebi.ac.uk, cesareni@uniroma2.it, david@mbi.ucla.edu or m.tyers@ed.ac.uk

1. Cusick, M.E. *et al. Nat. Methods* **6**, 39–46 (2009).
2. Kerrien, S. *et al. Nucleic Acids Res.* **35**, D561–D565 (2007).
3. Salwinski, L. *et al. Nucleic Acids Res.* **32**, D449–D451 (2004).
4. Chatr-aryamontri, A. *et al. Nucleic Acids Res.* **35**, D572–D574 (2007).
5. Swarbreck, D. *et al. Nucleic Acids Res.* **36**, D1009–D1014 (2008).
6. Breitkreutz, B.J. *et al. Nucleic Acids Res.* **36**, D637–D640 (2008).
7. Reguly, T. *et al. J. Biol.* **5**, 11 (2006).
8. Orchard, S. *et al. Proteomics* **7** (Suppl. 1), 28–34 (2007).

*Editor's note: For the response by Cusick et al., please see the Addendum to their Perspective (Cusick, M.E. et al. Nat. Methods **6**, 934–935; 2009).*

Addendum: Literature-curated protein interaction datasets

Michael E Cusick, Haiyuan Yu, Alex Smolyar, Kavitha Venkatesan, Anne-Ruxandra Carvunis, Nicolas Simonis, Jean-François Rual, Heather Borick, Pascal Braun, Matija Dreze, Jean Vandenhaute, Mary Galli, Junshi Yazaki, David E Hill, Joseph R Ecker, Frederick P Roth & Marc Vidal

Nat. Methods **6, 39–46 (2009); published online 30 December 2008; addendum published after print 25 November 2009.**

We assessed literature-curated protein-protein interaction (PPI) datasets for the parameters of completeness, coverage and quality by several means, concluding that such datasets might be "possibly of lower quality than commonly assumed." A Correspondence[71] by members of the International Molecular Exchange Consortium (IMEx), while accepting many of our points, objected to our recuration exercise to assess quality, finding our criteria "subjective." We argue that the criteria were commonsensical and essentially capture how these databases are often described.

A wide swath of the scientific community, from computer scientists and engineers to physicists, systems biologists and molecular biologists, use literature-curated datasets as 'gold-standard' positive controls with the tacit understanding that this information is nearly perfect. Whether user impressions were formed from statements made by database authors[18–21] or not, belief that database entries accurately correspond to high-quality, direct physical interactions is widespread[6,72]. The standards we used to assess quality are generally accepted by the IMEx members, but one that remains problematic is the definition of binary interactions. A meaningful fraction of database users is under the impression that 'binary interaction' means direct pairwise PPIs, and that is the definition we tried to apply. The definition that the IMEx databases apply is that of 'binary representation', meaning any pairwise association between two entities, direct or indirect. Although technically correct from an informatics viewpoint, binary representation likely does not accurately reflect biophysical reality. To better match user expectations, one IMEx database has adjusted their website presentation to allow users to filter 'spoke expanded co-complexes' from binary interactions, although all reported interactions are initially classified as 'binary'.

Another widespread perception is that curated databases contain predominantly low-throughput interactions, whereas the reality is that curated databases have a substantial portion of interactions derived from high-throughput experiments (Fig. 2 in our Perspective). The point is not whether high-throughput interaction experiments are of worse or better quality than low-throughput experiments, but that greater transparency should be provided so that users can filter the data according to their needs.

As a result of applying the criteria that we did, based on the observations above, the error rates we reported reflected not only errors in curation but also how well the underlying data meet the standards set forth. The details for the yeast, human and plant recurations are available in the **Supplementary Note**.

Our efforts are aimed at alerting the scientific community that literature-curated interactions may need further scrutiny or classification to qualify as a 'gold standard' for users who are specifically interested in direct pairwise PPIs. Closer inspection will allow the community to be the ultimate judge of how useful these curation units turn out to be.

We updated our original Supplementary Table 2 on LC-multiple human recurated dataset to show the databases from which each interaction came (**Supplementary Table 1**). Almost 90% of interactions, and 95% of the problematic curation units, came from non-IMEX databases (HPRD[22] and BIND[17]). We had been requested to omit this information originally, but for IMEX databases there is minimal difference in error rates between our recuration and that of Salwinski *et al.*[71]. A download discrepancy, which IntAct has now mended so that it cannot recur, necessitated the recuration of the errors for the *Arabidopsis* curation (Supplementary Table 4 in our original Perspective). We now score the 24 curation errors as: 3 'no binding experiment' (formerly 9); 6 'no binding partner' (formerly 6); 11 'indirect' (formerly 6); 3 'wrong protein' (formerly 3); and 1 'wrong species' (formerly 0).

Unfortunately the download dates for the interaction data in our original Perspective were unclear or missing. The download date for the yeast interaction data was originally reported as mid-2007 but is actually early 2006. Human interaction data were downloaded from HPRD, BIND, MINT, MIPS and DIP in mid-2005, as described in ref. 31. *Arabidopsis* interaction data from IntAct and TAIR were first downloaded in February 2008. The second download, which we used in the analysis above, occurred in March 2009 when the download inconsistencies were pointed out to us.

Our contentions that literature-curated datasets are imperfect were corroborated by a paper published concurrently[73]. Especially telling was the observation in that paper that many "databases lack a substantial portion of PPIs, emphasizing the need to integrate multiple PPI databases"[73], a concern fully echoed by our original finding of low overlaps between curated PPI databases (Fig. 3 in our original Perspective). The problem of low overlaps should be mitigated once the IMEx exchange of curation between databases becomes implemented[33].

Other investigators have reported that literature-curated interaction datasets are less perfect than is widely presumed. In papers in *Trends in Biochemical Sciences*[44,45,51] the authors argued over a distressing lack of reproducibility of curated interactions and contended that "protein interactions reported in the literature and curated in interaction databases might not occur as presented." Other reports have questioned the presumed perfection of curated PPIs[23,29,43,74], even one report by several authors of Salwinski *et al.*[71]: "a comparison of publications curated by both MINT and IntAct between 2003 and 2005 revealed that the two databases annotated exactly the same interaction pairs in only 6 out of 52 publications"[75]. BioGRID now grants that provisions are not made for quality assessment in curation: "We make no judgement calls on the methods or even, within reason, the quality of the data themselves"[76]. Perhaps quality of the underlying data should in some way begin to be assessed, to match community expectations of curated data.

Curation to extract protein-protein interactions from the literature is absolutely critical to the advancement of systems biology and proteomics. Increased transparency and appropriate communication of what is currently available in curated datasets will ultimately help these efforts. Preliminary steps toward generating confidence scores have been reported for curated[50], predicted[77] and experimental[27] PPI datasets. These measures go in the right direction and their further development should be encouraged and appropriately funded.

Note: Supplementary information is available on the Nature Methods website.

71. Salwinski, L. *et al.* Recurated protein interaction datasets. *Nat. Methods* **6**, 860–861 (2009).
72. Lee, I., Li, Z. & Marcotte, E.M. An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*. *PLoS ONE* **2**, e988 (2007).
73. Wu, J. *et al.* Integrated network analysis platform for protein-protein interactions. *Nat. Methods* **6**, 75–77 (2009).
74. Hart, G.T., Ramani, A.K. & Marcotte, E.M. How complete are current yeast and human protein-interaction networks? *Genome Biol.* **7**, 120 (2006).
75. Chatr-aryamontri, A. *et al.* MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data. *Genome Biol.* **9**, S5 (2008).
76. Blow, N. Systems biology: untangling the protein web. *Nature* **460**, 415–418 (2009).
77. Geisler-Lee, J. *et al.* A predicted interactome for *Arabidopsis*. *Plant Physiol.* **145**, 317–329 (2007).

Corrigendum: Nanoscale live-cell imaging using hopping probe ion conductance microscopy

Pavel Novak, Chao Li, Andrew I Shevchuk, Ruben Stepanyan, Matthew Caldwell, Simon Hughes, Trevor G Smart, Julia Gorelik, Victor P Ostanin, Max J Lab, Guy W J Moss, Gregory I Frolenkov, David Klenerman & Yuri E Korchev
***Nat. Methods* 6, 279–281 (2009); published online 1 March, 2009; corrected after print 3 September 2009.**

In the version of this paper originally published, references to previous work on pulse mode SICM should have been included (Mann, S.A. *et al. J. Neurosci. Methods* **116**, 113–117, (2002) and Happel, P. *et al. J. Microsc.* **212**, 144–151 (2003)). These references were removed during shortening of the paper for publication and have been added back to the PDF and HTML versions of this article. The pulse mode technique reported in these previous papers has conceptual similarity to our hopping mode SICM, in that distance feedback control is not continuous; thus, it also solves the problem of probe-sample collision for large cellular structures. However, the pulse mode technique is considerably slower owing to a different feedback mechanism and does not perform at nanoscale resolution.

Erratum: 'Edgetic' perturbation of a *C. elegans* BCL2 ortholog

Matija Dreze, Benoit Charloteaux, Stuart Milstein, Pierre-Olivier Vidalain, Muhammed A Yildirim, Quan Zhong, Nenad Svrzikapa, Viviana Romero, Géraldine Laloux, Robert Brasseur, Jean Vandenhaute, Mike Boxem, Michael E Cusick, David E Hill & Marc Vidal
***Nat. Methods* 6, 843–849 (2009); published online 25 October, 2009; corrected after print 16 November 2009.**

In the version of this article initially published, the schematic in Figure 5a was misaligned. The error has been corrected in the HTML and PDF versions of the article.

Erratum: What's in a test?

Anonymous
***Nat. Methods* 6, 783 (2009); published online 29 October 2009; corrected after print 16 November 2009**

In the version of this article initially published, the name of Robert Cook-Deegan was misspelled. The error has been corrected in the HTML and PDF versions of the article.

DOCUMENT JOINT 5

Titre : Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network.

Auteurs : Simonis N*, Rual JF*, Carvunis AR*, Tasan M*, Lemmens I*, Hirozane-Kishikawa T, Hao T, Sahalie JM, Venkatesan K, Gebreab F, Cevik S, Klitgord N, Fan C, Braun P, Li N, Ayivi-Guedehoussou N, Dann E, Bertin N, Szeto D, Dricot A, Yildirim MA, Lin C, de Smet AS, Kao HL, Simon C, Smolyar A, Ahn JS, Tewari M, Boxem M, Milstein S, Yu H, Dreze M, Vandenhaute J, Gunsalus KC, Cusick ME, Hill DE, Tavernier J, Roth FP, Vidal M

Description : article de recherche original paru dans le magazine *Nature Methods* en 2009

Contribution : Co-premier auteur de cet article, j'ai participé à toutes les analyses bioinformatiques, pratiquement et/ou intellectuellement. JF Rual a dirigé toutes les expériences sauf le MAPPIT, réalisé par I Lemmens. N Simonis a dirigé toutes les analyses, et m'a encadrée. N Simonis, JF Rual et moi avons rédigé l'article, aidés surtout de ME Cusick, FP Roth, DE Hill et de M Vidal, qui a supervisé le projet. Les autres auteurs ont apporté diverses contributions intellectuelles et techniques.

Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network

Nicolas Simonis[1,2,11], Jean-François Rual[1,2,10,11], Anne-Ruxandra Carvunis[1–3,11], Murat Tasan[4,11], Irma Lemmens[5,11], Tomoko Hirozane-Kishikawa[1,2], Tong Hao[1,2], Julie M Sahalie[1,2], Kavitha Venkatesan[1,2,10], Fana Gebreab[1,2], Sebiha Cevik[1,2,10], Niels Klitgord[1,2,10], Changyu Fan[1,2], Pascal Braun[1,2], Ning Li[1,2,10], Nono Ayivi-Guedehoussou[1,2,10], Elizabeth Dann[1,2], Nicolas Bertin[1,2,10], David Szeto[1,2,10], Amélie Dricot[1,2], Muhammed A Yildirim[1,2,6], Chenwei Lin[1,2], Anne-Sophie de Smet[5], Huey-Ling Kao[7], Christophe Simon[1,2,10], Alex Smolyar[1,2], Jin Sook Ahn[1,2], Muneesh Tewari[1,2,10], Mike Boxem[1,2,8,10], Stuart Milstein[1,2,10], Haiyuan Yu[1,2], Matija Dreze[1,2,9], Jean Vandenhaute[9], Kristin C Gunsalus[7], Michael E Cusick[1,2], David E Hill[1,2], Jan Tavernier[5], Frederick P Roth[1,4] & Marc Vidal[1,2]

To provide accurate biological hypotheses and elucidate global properties of cellular networks, systematic identification of protein-protein interactions must meet high quality standards. We present an expanded *C. elegans* protein-protein interaction network, or 'interactome' map, derived from testing a matrix of ~10,000 × ~10,000 proteins using a highly specific, high-throughput yeast two-hybrid system. Through a new empirical quality control framework, we show that the resulting data set (Worm Interactome 2007, or WI-2007) was similar in quality to low-throughput data curated from the literature. We filtered previous interaction data sets and integrated them with WI-2007 to generate a high-confidence consolidated map (Worm Interactome version 8, or WI8). This work allowed us to estimate the size of the worm interactome at ~116,000 interactions. Comparison with other types of functional genomic data shows the complementarity of distinct experimental approaches in predicting different functional relationships between genes or proteins.

The interactome of an organism is the network formed by the complete set of binary physical interactions that can occur between all proteins. Low-throughput protein-protein interaction experiments are of considerable value in understanding cellular processes at the molecular level. However, the development of high-throughput approaches can substantially increase the pace and scale of discovery, while permitting the implementation of standardized and systematic quality control. Initial steps toward binary interactome mapping in metazoans have been undertaken[1–5], and the resulting partial interactome maps have (i) provided insights into the organization of biological networks, (ii) assisted in determining functions of many proteins and complexes and (iii) identified hundreds of connections to proteins associated with human diseases.

High-throughput interactome mapping is particularly needed for *C. elegans*, a widely used model organism for which the set of protein-protein interactions derived from small-scale experiments and accessible in public databases is limited to less than 500. The first proteome-scale version of the Worm Interactome (WI5)[3] combined several sources of protein-protein interaction data: literature-curated interactions, yeast two-hybrid (Y2H) 'module' maps each devoted to a specific biological process[1,6–11], 'interolog'

[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA. [2]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [3]Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications de Grenoble (TIMC-IMAG), Unité Mixte de Recherche 5525 Centre National de la Recherche Scientifique (CNRS), Faculté de Médecine, Université Joseph Fourier, 38706 La Tronche Cedex, France. [4]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue, Boston, Massachusetts 02115, USA. [5]Department of Medical Protein Research, Vlaams Instituut voor Biotechnologie, and Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, 3 Albert Baertsoenkaai, 9000 Ghent, Belgium. [6]Division of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, Massachusetts 02138, USA. [7]Center for Genomics and Systems Biology, Department of Biology, New York University, 100 Washington Square East, New York, New York 10003, USA. [8]Massachusetts General Hospital Center for Cancer Research, Building 149, 13th Street, Charlestown, Massachusetts 02129, USA. [9]Unité de Recherche en Biologie Moléculaire, Facultés Notre-Dame de la Paix, 61 Rue de Bruxelles, 5000 Namur, Belgium. [10]Present addresses: Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA (J.-F.R.), Novartis Institutes for Biomedical Research Inc., 250 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA (K.V.), School of Biomolecular and Biomedical Science, University College Dublin, Belfield, Dublin 4, Ireland (S.C.), Bioinformatics Program, Boston University, 705 Commonwealth Avenue, Boston, Massachusetts 02215, USA (N.K.), Wyeth Pharmaceuticals Inc., 35 Cambridgepark Drive, Cambridge, Massachusetts 02140, USA (N.L.), Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, USA (N.A.-G.), RIKEN Omics Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa 230-0045, Japan (N.B., C.S.), University of California San Francisco School of Medicine, 500 Parnassus Avenue, San Francisco, California 94143, USA (D.S.), Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA (M. Tewari), Utrecht University, Kruytgebouw N309, 8 Padualaan, 3584 CH Utrecht, The Netherlands (M.B.) and Alnylam Pharmaceutical, 300 Third Street, Cambridge, Massachusetts 02142, USA (S.M.). [11]These authors contributed equally to this work. Correspondence should be addressed to M.V. (marc_vidal@dfci.harvard.edu), F.P.R. (fritz_roth@hms.harvard.edu), J.T. (jan.tavernier@ugent.be) or D.E.H. (david_hill@dfci.harvard.edu).

RECEIVED 4 JUNE; ACCEPTED 29 OCTOBER; PUBLISHED ONLINE 14 DECEMBER 2008; DOI:10.1038/NMETH.1279

interactions—that is, predicted pairs of interactors whose respective orthologs interact in another organism[12]—and lastly, Y2H interactions derived from a high-throughput screen performed with ~2,000 metazoan proteins as baits[3] (WI-2004). WI5 represents a key resource for formulating biological hypotheses and investigating the properties of the *C. elegans* interaction network. However, WI5 includes nonbinary interactions derived from the literature, interologs not experimentally confirmed, and some lower-confidence Y2H interactions.

Our updated Worm Interactome map (WI8) implements several techniques and strategies that are critical for generating high-quality protein-protein interaction data on a proteomic scale. First, we expanded the worm interactome map by screening a matrix of ~10,000 × ~10,000 proteins. Second, we developed new standards to deliver a data set of very high quality. These standards involve a highly stringent, high-throughput yeast two-hybrid (HT-Y2H) assay, strict methods for filtering and updating existing data sets, independent measurement of technical quality, and evaluation of biological relevance. Because worm genome annotations are improved frequently, we updated previous protein-protein interaction data according to recent gene models. Finally, we empirically estimated the full size of the *C. elegans* interactome, through the implementation of a new interactome mapping framework based exclusively on protein-protein interaction data[13].

To extend the use of WI8 beyond protein-protein interaction analysis and to place WI8 into broader biological context, we integrated the resulting protein-protein interaction data with complementary data sets, such as physical and genetic interactions from curated literature, our interolog data set (**Supplementary Methods** online), phenotypic profiling data and a coexpression compendium. We also identified tissue localizations and developmental stages in which interacting pairs are most likely to be physiologically relevant whenever anatomical annotation[14] or spatiotemporal expression patterns[15] were available for both proteins.

Our new data set, WI-2007, consists of 1,816 high-confidence, binary, protein-protein interactions. We integrated previously published high-quality *C. elegans* binary protein-protein interactions with WI-2007 into the updated WI8 version of the worm interactome, providing 3,864 high-quality binary physical interactions between 2,528 proteins. WI8 was significantly enriched for functionally linked protein pairs, confirming its biological relevance and demonstrating the value of unbiased, large-scale Y2H screens in inferring protein function.

RESULTS

A new HT-Y2H data set

For this iteration of worm interactome mapping, we implemented a HT-Y2H strategy previously used for human interactome mapping[5]. We tested all open reading frames (ORFs) in the worm ORFeome version 1.1 (ref. 8) against one another (a ~10,000 × ~10,000 matrix), a search space corresponding to ~24% of the total search space for a comprehensive *C. elegans* interactome map, excluding variants due to polymorphism, alternative transcription or alternative splicing (**Fig. 1a**). We also ensured the quality of the new data set by using stringent conditions and controls described previously[5], including low expression of DNA-binding-domain and activation-domain fusion proteins (DB-X and AD-Y), multiple reporter genes to ensure high precision, removal of all a priori and *de novo* DB-X autoactivators, and individual retesting of each positive protein-protein interaction. The resulting set of 1,816 protein-protein interactions between 1,496 proteins (**Fig. 1b**) is called WI-2007.

Characterization of WI-2007

To assess the quality of our new data set and estimate the size of the complete worm interactome, we used a framework we recently developed[13], with a slightly different implementation relevant to the data available in *C. elegans*. This framework empirically measures several parameters to characterize a high-throughput binary protein-protein interaction data set: 'screening completeness', the fraction of the proteome-wide space tested in the experiment; 'precision', the proportion of interactions in the data set that are true biophysical interactions; 'sampling sensitivity', the fraction of all detectable interactions for a particular assay found under the sampling conditions, which corresponds here to the saturation of a single screen; and 'assay sensitivity', the proportion of all biophysical interactions that can be identified by an assay at saturation, as each assay can only detect a fraction of all true biophysical interactions.

To estimate these parameters we performed the following experiments. First, we used the mammalian protein-protein interaction trap technique (MAPPIT) to measure how a random sample of WI-2007 performed in an independent protein interaction detection assay compared to a positive reference set (cePRS-v1, manually recurated interactions from low-throughput studies) and a random reference set (ceRRS-v1, randomly chosen pairs in the search space of WI-2007). Second, we used the overlap between WI-2007 and our previous Y2H study in their common search space to quantify the saturation of our screen. Third, to evaluate the proportion of interactions that can be captured by our Y2H assay, we used the fraction of cePRS-v1 pairs recovered in a pairwise Y2H experiment and in WI-2007, as well as the proportion of widely conserved interologs found in WI-2007. Introducing these measurements into a Monte Carlo simulation (**Supplementary Methods**), we computed the four parameters in our framework, as well as the expected size of the worm interactome. According to this model, the screening completeness was 23.6%, the precision estimate 86% ± 16% (mean and s.d.), the sampling sensitivity 31% ± 8%, the assay sensitivity 16% ± 3% and the size of the worm interactome 115,600 ± 26,400 (**Fig. 1c**).

Given the potential bias in cePRS-v1 and in the set of ultra-conserved interologs toward interactions that are easy to detect, the associated assay sensitivity may be an overestimate. Thus, the predicted interactome size is likely to be a conservative estimate. The strength of this approach is that these calculations rely solely on protein-protein interactions, without depending on functional annotation or other types of genomic or proteomic data. Our estimate provides an endpoint for the worm interactome mapping project and can be used as a reference for evolutionary comparisons between interactome networks from different species.

A combined data set of high-quality binary interactions

To provide a set of integrated, high-quality, binary protein-protein interaction data for *C. elegans*, we employed higher stringency criteria and used updated WormBase (http://www.wormbase.org) gene models to reprocess the raw data from smaller scale Y2H screens encompassing proteins involved in vulval development[1],

protein degradation[6], DNA damage response[7], germline formation[9], TGF-β signaling pathway[11] and RNA interference[10], along with unpublished Y2H interactions (M. Tewari, N.A.-G. and J.S.A.; **Supplementary Methods**). This 'biological processes' subset (BPmaps) contains 554 protein-protein interactions.

WI8 is the union of WI-2004, WI-2007 and BPmaps. The consolidated WI8 network (**Fig. 2** and **Supplementary Table 1** online) contains 3,864 high-quality protein-protein interactions among 2,528 proteins. Approximately 40% of the interactions are newly identified, and the set excludes any lower-confidence interactions from previous studies[3]. The WI8 physical interaction network can be visualized on our website (http://interactome.dfci.harvard.edu/C_elegans/) using N-Browse[16] or VisANT[17].

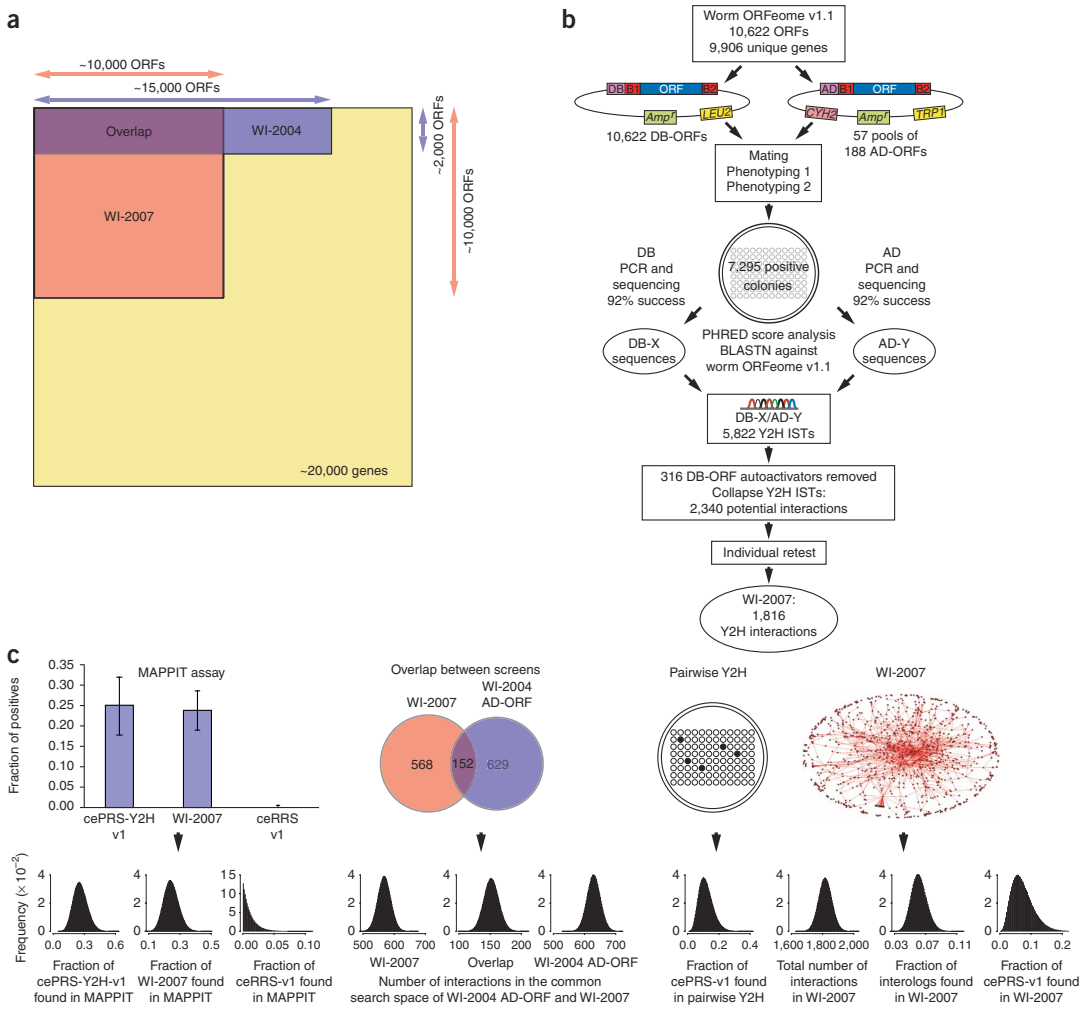


Figure 1 | Construction and characterization of WI-2007. (**a**) Search spaces of WI-2007 and WI-2004 relative to the whole proteome, three times larger for WI-2007 than WI-2004. (**b**) Pipeline used for WI-2007. ORFs from ORFeome v1.1 were transferred into DB and AD vectors by recombinational cloning, then transformed into yeast cells. Each bait was then mated with pools of 188 AD-ORFs. Two rounds of phenotyping were performed to isolate positive colonies, which were used to PCR-amplify DB-ORFs and AD-ORFs for sequencing, leading to the identification of 5,822 interaction sequence tags (ISTs). After excluding autoactivators and collapsing redundant ISTs corresponding to the same, nonoriented protein pair, each interaction was individually retested in an independent Y2H experiment to generate the final WI-2007 data set. (**c**) WI-2007 characterization. Ten measures are shown (left to right): proportions observed in MAPPIT of (i) cePRS-Y2H-v1, (ii) a random sample of WI-2007 and (iii) ceRRS-v1; number of interactions detected in the common search space of WI-2007 and WI-2004 (iv) in WI-2007, (v) in both screens and (vi) in WI-2004 AD-ORF; (vii) proportion of cePRS-v1 detected in an independent pairwise Y2H experiment; (viii) total number of interactions in WI-2007 and proportion recovered in WI-2007 of (ix) ultraconserved interologs and (x) cePRS-v1. The sampling errors on the ten measurements are modeled with beta distributions (bottom row). These distributions are then used in a Monte Carlo simulation to compute precision, sampling sensitivity, assay sensitivity and the total number of interactions in *C. elegans*, along with their associated error bars. Label on *y* axis (frequency) applies to all ten sampling distributions.

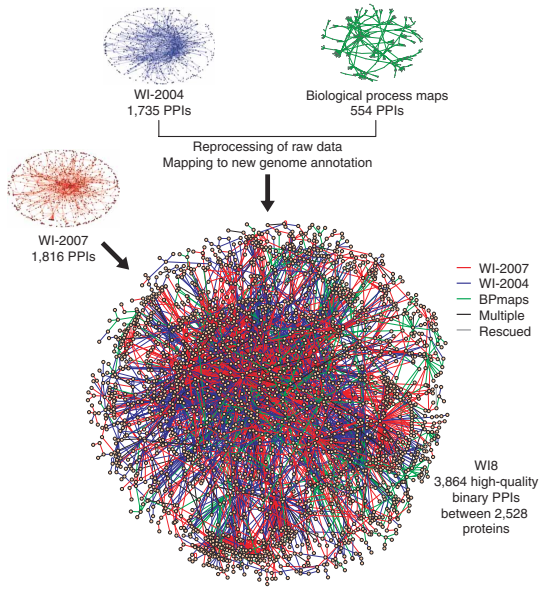


Figure 2 | WI8: an extended, high-quality, protein-protein interaction network. High-quality data on Y2H protein-protein interactions (PPIs) from WI-2007, WI-2004 and diverse medium-throughput biological processes based Y2H maps[1,6–11] were integrated into WI8. The color of the edge indicates the data set of origin: WI-2007, red; WI-2004, blue; biological process maps, green. Edges corresponding to more than one of these evidence types are shown in black, and edges corresponding to 'rescued' interactions—that is, supported by at least two lower-confidence pieces of evidence—in gray. Only the main giant component of the network (connected subgraph that contains the majority of the entire network's nodes) is shown.

Confirmed Y2H interactions may be 'biophysically true' interactions that do not actually occur *in vivo* if the involved proteins are not present at the same time and place within a multicellular organism, or are not present with the proper post-translational modifications. We evaluated the overall biological relevance of WI8 by assessing the degree to which interacting pairs share Gene Ontology annotation terms—that is, can be considered as functionally linked. A Gene Ontology term may be specific or broad, depending on the number of genes to which it is assigned. We therefore defined four different thresholds of functional specificity: less than or equal to 5, 20, 100 and 400 annotated genes per Gene Ontology term. For all three component subsets of WI8, we compared the degree of functional linkage with that of binary interactions derived from the literature (LCI binary; **Supplementary Methods** and **Supplementary Table 2** online), normalizing for protein composition bias of each of these subsets. All data subsets showed a high enrichment for both broad and specific functional linkage (**Fig. 3a**), suggesting high biological relevance. The degree of functional linkage among WI-2007 was similar to or exceeded the literature enrichment at each functional specificity limit tested.

Various interactions in WI8 provide new biological information. For example, EBP-1, a microtubule-binding protein whose homologs are involved in a variety of microtubule-mediated processes[18], interacts with several proteins involved in microtubule dynamics, including UNC-14, a protein required for axon growth and sex myoblast migration[19], VAB-8, a kinesin-like protein required for axon outgrowth and cell migration[20], and RSA-2, a protein specifically required for microtubule outgrowth from centrosomes and for spindle assembly[21].

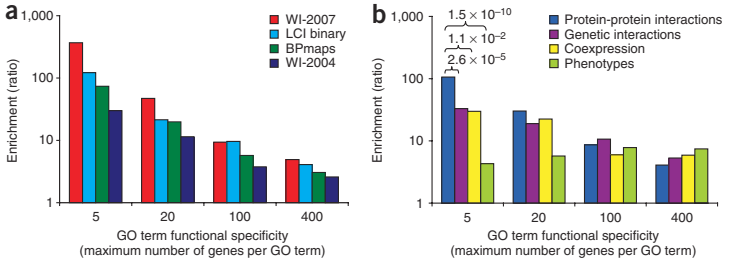


Figure 3 | Biological relevance. Enrichment represents the frequency of functional linkage of protein or gene pairs expressed as a multiple of the value for random pairs and was plotted against functional specificity groupings. The maximum number of genes associated with a particular Gene Ontology (GO) term was used as an estimate of the functional specificity (5, 20, 100 or 400 genes). (**a**) Enrichment for functional relationships in different components of the WI8 data set and in the LCI binary data set. (**b**) Functional relationship enrichments for distinct types of experimental evidence. *P*-values assessing the difference between protein-protein interactions and other types of evidence are shown for very specific Gene Ontology terms (terms with a maximum of five genes).

Integrated functional network

Integration of diverse large-scale data sets was previously used to demonstrate the coordination of interconnected yet distinct molecular machines involved in worm early embryogenesis[22]. Another recent publication[23] describes Bayesian integration of functional linkages into a single network, weighting each type of evidence according to a reference data set (benchmark). Such an approach can be a valuable resource leading to interesting hypotheses, but is highly dependent on the benchmark, which can strongly bias the predictions (**Supplementary Discussion** online). In contrast, we chose to provide an unweighted data set that (i) does not artificially bias the network toward highly-studied proteins, (ii) allows the user to select their own threshold for some types of linkages (for example, correlation coefficient with expression data), (iii) separates each type of experimental evidence and (iv) does not rely on an inevitably biased benchmark.

We integrated WI8 with five different sources of evidence for functional relationships: (i) mRNA coexpression data available in WormBase (**Supplementary Table 3** online); (ii) RNAi phenotypes from RNAiDB[24] (**Supplementary Table 4** online); (iii) genetic interactions curated in WormBase; (iv) interolog interactions and (v) all binary and nonbinary protein-protein interactions from our literature-curated data set (LCI; **Supplementary Methods**). This integrated network involves 178,151 links between 6,176 genes and can be visualized online using N-Browse (http://interactome.dfci.harvard.edu/C_elegans/).

Table 1 | Overlap between data sets from the integrated functional network

	WI8		LCI		Interologs		Genetic interactions		Phenotypes	
	E	*P*	*E*	*P*	*E*	*P*	*E*	*P*	*E*	*P*
LCI	182.3	1.01×10^{-37}								
Interologs	91.4	1.13×10^{-212}	145.6	5.89×10^{-75}						
Genetic interactions	23.9	1.59×10^{-14}	66.9	1.17×10^{-72}	24.1	6.58×10^{-58}				
Phenotypes	3.0	5.33×10^{-3}	4.6	1.02×10^{-3}	3.0	1.27×10^{-16}	3.3	3.83×10^{-6}		
Coexpression	2.5	1.20×10^{-8}	2.6	3.20×10^{-3}	3.2	5.01×10^{-103}	1.6	1.61×10^{-1}	1.6	1.09×10^{-21}

Enrichment (*E*, expressed as a multiple) and significance (*P*-values) of the overlaps between distinct functional data sets. The enrichment is defined as the number of pairs shared between two data sets divided by the expected random number of shared pairs, and the significance is assessed by Fisher's exact test.

We compared the biological relevance of each type of data from the integrated network by calculating the enrichment for functional linkage, as described before for protein-protein interaction data sets (**Fig. 3b**). All the analyzed data sets showed highly significant enrichment for functional linkage ($P < 2.5 \times 10^{-3}$). Notably, among the analyzed data sets, physical interactions seemed to be the best predictors of highly specific shared Gene Ontology terms, whereas pairs sharing phenotypes showed the highest enrichment for less specific functional linkages. The phenotypic profiles used in this study were gross phenotypes, and more precise phenotypic observations would probably be better predictors for more precise functions but worse predictors for more global functions. Similarly, linkages from expression data were derived from a wide range of experimental conditions; such data could be a better predictor of more specific linkages, if a set of experimental conditions targeting a particular process had been used. This observation reflects how these different data sets address biological questions at different levels, in the same way that sequence and structure similarity are better predictors of whether proteins exert the same enzymatic activity than of whether they belong to the same pathways[25].

Next we examined the overlap between component networks of each type. We observed significant overlap for almost all combinations of component networks (**Table 1**). WI8, LCI, interologs and genetic interactions showed more overlap with one another than coexpression or phenotypic correlation with any other data set. The strong association between the two physical interaction data sets and interologs (LCI and WI8 confirmed 56 and 194 predicted interologs, including 49 and 147 heterodimers, respectively) was expected, and it confirmed that many interactions are conserved during evolution. LCI shared higher overlaps with phenotypically correlated pairs, genetic interactions and interologs than WI8, most likely because lower-throughput assays often test physical interactions that are enriched a priori for a common phenotype or are known to have interacting orthologs. Still, WI8 substantiated 57 pairs of genes with high coexpression among a wide range of experimental conditions, 9 pairs of genes with similar RNAi phenotypic profiles and 14 pairs of genetically interacting genes ("shared edges" section at http://interactome.dfci.harvard.edu/C_elegans/).

Although significant and informative, these overlaps remain relatively low (**Supplementary Table 5** online). This can be explained by lack of 'screening completeness' of most data sets; that is, most of these data sets are not genome or proteome wide. Indeed, more than 60% of genes/proteins in the network (the term 'genes/proteins' is used to reflect the mixed nature of the network, built from links between both genes and proteins) are present in one data set only, whereas less than 5% are present in four data sets or more. Furthermore, most of the screens that have led to the generation of these data sets (including our Y2H screens) are far from saturation and are probably limited by low sampling sensitivity in addition to inherent limitations of each assay; that is, precision and assay sensitivity. Finally, a perfect overlap is not expected because of intrinsic differences in the nature of the biological attributes measured in these data sets.

Module-scale biological networks

Module-scale biological subnetworks can be extracted from the integrated network by selecting 'seed' genes/proteins known to be associated with a specific process and then expanding by selecting neighboring genes/proteins. For example, using as seeds genes/proteins implicated in RNA-binding processes (**Fig. 4a**), nearly all genes/proteins in the expanded set are linked to several RNA-binding genes/proteins and are connected by at least two types of relationships. Most of these linked genes/proteins were thus predicted as functionally related to RNA binding, and several (for example sup-12) were already annotated or predicted by sequence similarity to be associated with RNA binding within WormBase. Other genes/proteins have annotations consistent with RNA binding. For example, T26A8.4 encodes a protein predicted to be part of the CPSF subcomplex of the Polyadenylation Factor I complex through clusters of eukaryotic orthologous groups (KOGs)[26] and is orthologous to yeast Caf120, which is part of the conserved Ccr4-Not transcriptional regulatory complex involved in mRNA initiation, elongation and degradation.

When expanding from a seed set of genes/proteins involved in cell fusion (**Fig. 4b**), almost all added genes/proteins are linked to more than one in the seed set, with many links supported by more than one evidence source. For example, unc-62 had phenotypic correlation with seed genes/proteins nhr-25, lin-29 and ceh-20; physical, genetic and interolog links with ceh-20; and interolog links with mab-5. In contrast to the RNA-binding subnetwork, where most links were physical interactions with few pairs being supported by more than one evidence type, in this example most links were either phenotypic or genetic interactions, and many physical interactions were supported by other evidence.

Notably, WI-2007 contains physical interactions between proteins not previously linked to one another, but at a network distance of two in the integrated network (**Supplementary Fig. 1** online). In the RNA-binding network, for example, star-2 and mec-8, which are known to be indirectly linked through sup-12, were found to directly interact. We found 1,157 new 'triangle closures' of

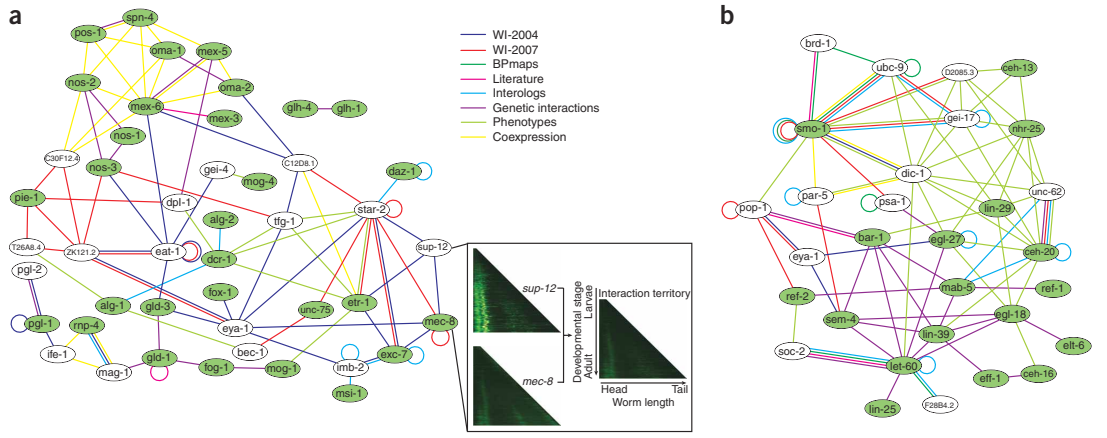


Figure 4 | Examples of multiple-evidence subnetworks. The networks represent relationships among genes/proteins from several evidence sources, color-coded as indicated. Genes and their products are labeled using an unitalicized lower-case version of the standard *C. elegans* three-letter system to reflect the inclusion of links between both proteins and genes. (**a**) Genes/proteins related to RNA binding. Green ellipses are genes/proteins annotated as 'RNA-binding' in WormBook[31]; white ellipses are genes/proteins linked to RNA-binding genes/proteins by at least one protein-protein interaction from WI8 and one other piece of evidence. The inset shows the chronograms of *sup-12* and *mec-8* (left) and their predicted spatiotemporal pattern of interaction (right). The chronograms represent the absolute GFP intensity measured (increasing values coded black-green-yellow-white) using reporter constructs with the indicated promoter, along the worm length (*x* axis) and as a function of development stage (*y* axis)[15]. (**b**) Genes related to cell fusion. Green ellipses are genes/proteins annotated as 'cell fusion' in WormBook[32]; white ellipses are genes/proteins linked to cell fusion genes/proteins by a protein-protein interaction from WI8 and one other type of evidence.

this kind (viewable within the "intersections" and "display" sections of http://interactome.dfci.harvard.edu/C_elegans/).

From 'static' map to spatiotemporal interactome

Spatiotemporal expression patterns for ~2,000 worm genes have recently become available through large-scale studies of worms carrying endogenous promoters driving expression of GFP[14,15]. Examination of the resulting GFP intensity patterns informs the question of where (tissue) and when (developmental stage) promoters are activated. The GFP profiles can be sorted according to developmental stage by worm length and aligned, forming a 'chronogram' representation[15].

We performed computational 'chronogram intersection' of the spatiotemporal expression patterns corresponding to two interacting proteins and used these to infer a potential 'interaction territory' (**Supplementary Figs. 2** and **3** online). We also inferred interaction territories on the basis of explicit anatomical annotations[14] for interacting proteins. We identified 111 common anatomical annotations and generated 69 chronogram intersections for protein-protein interactions from WI8 (viewable within the "localization" section of http://interactome.dfci.harvard.edu/C_elegans/). Examples from the RNA-binding subnetwork (**Fig. 4a**) included common interaction territories for SUP-12 and MEC-8 (**Fig. 4a**, inset), MEC-8 and EXC-7, and MEP-1 and MOG-4 through chronogram intersections, and for 21 more interactions through anatomical annotations. Although this GFP-based technique has limitations related to resolution and coverage, these examples provide a glimpse of how integrating spatiotemporal expression information could eventually allow extraction of tissue-specific subnetworks corresponding to pathways, functional modules or protein complexes, once the technology improves and more data become available.

DISCUSSION

We describe the implementation of an integrated strategy for generating high-confidence networks based on a highly stringent HT-Y2H assay combined with a quality control framework[13], thus achieving a step along the path to completion of the *C. elegans* interactome. Our estimated size of the complete *C. elegans* biophysical interactome is approximately 116,000 interactions, considering only a single protein isoform per gene. Although WI8 provides 3,864 interactions, 96%–97% of the interactome remains untouched because of lack of screening completeness as well as incomplete sampling and assay sensitivity. From the overlap of two independent HT-Y2H screens, we estimate that a single high-throughput screen can capture ~30% of the detectable interactions and thus would need to be repeated several times to reach saturation. Even at saturation, some interactions may not be detectable by Y2H because of intrinsic limitations of the assay—for example, proteins may not be imported into the nucleus, proper folding may not occur because of the fusion with the DNA-binding or activation domains, or interactions may require post-translational modifications or cofactors not present within *S. cerevisiae*. We estimate the proportion of interactions detectable with our HT-Y2H system (assay sensitivity) at approximately 16%.

Several approaches under development, involving optimization of the experimental setup[27] or systematic ORF fragmentation[28], should improve the assay sensitivity in future interactome mapping projects. However, achieving comprehensive mapping of the interactome will require use of various assays with complementary assay sensitivities. For example, experiments conducted in mammalian cells may uncover some interactions missed by Y2H, but fail to find others because some interactions do not occur under the conditions tested[27]. In addition to improving sensitivity, further cloning efforts will have to be undertaken to increase the screening

completeness of future interactome mapping projects. WI8 represents an early milestone toward uncovering the complete interactome network, yet it is to our knowledge the most comprehensive and reliable protein-protein interactions data set available today for *C. elegans*.

METHODS

Y2H screening. We mated 94 individual *MATα* MaV203 DB-ORF yeast strains, in a 96-well format, with the same *MATa* MaV103 AD-188ORFs mini-library on solid medium containing yeast extract, peptone and dextrose (YPD). Each DB-ORF 96-well plate was individually mated to all AD-ORFs compiled into 57 AD-188ORFs pools. After overnight growth at 30 °C, we transferred the colonies to plates containing synthetic complete (SC) yeast medium lacking leucine, tryptophan and histidine and containing 20 mM 3-aminotriazole (3AT) to select for diploids that showed elevated expression of the *GAL1::HIS3* Y2H marker. The same cells were transferred in parallel onto SC medium lacking leucine (SC–L) and containing 3AT and cycloheximide (SC–L+3AT+CYH). The pAD-dest-CYH vector contains the *CYH2* negative-selection marker, which allows plasmid shuffling on cycloheximide-containing media. This step is crucial to eliminate autoactivators that can arise during Y2H selection. Autoactivators show a $3AT^+$/3AT-CYH^+ phenotype, whereas genuine positives show a $3AT^+$/3AT-CYH^- phenotype in this assay. We picked approximately 180,000 positive colonies from $3AT^+$/3AT-CYH^- spots into a second-generation set of 96-well plates for further phenotypic screening.

Scoring Y2H assays. Consolidated and regrown $3AT^+$/3AT-CYH^- colonies were transferred to both Sc–L+3AT and Sc–L+3AT+CYH plates to confirm *GAL1::HIS3* transcriptional activity, and to YPD to determine *GAL1::lacZ* transcriptional activity using a β-galactosidase filter assay. We selected colonies that retested $3AT^+$/3AT-CYH^- and tested positive at levels equal or higher to that of the control DB-RB/AD-E2F interaction pair in our Y2H control set. Of the original ∼180,000 $3AT^+$/3AT-CYH^- colonies, 7,295 passed this double phenotypic test and represented Y2H positives. We also systematically tested all DB-ORFs for autoactivation by growth on solid SC-L+3AT medium, identifying all strong autoactivators and removing them from further consideration as baits in Y2H.

Yeast PCR and IST sequencing. We performed PCR amplifications on all Y2H-positive colonies to individually amplify DB-ORFs and AD-ORFs. The products from the PCR were purified and used as templates in a cycle-sequencing reaction to obtain two interaction sequence tags (ISTs) per Y2H positive.

WI-2007 IST analysis. The quality of the ISTs obtained by sequencing was measured by moving a sliding window of 10 base pairs to define the portion of the IST that had an average PHRED (http://www.phrap.com/phred/) score of 10 or higher over at least 10% of their length. We aligned all sequences against the worm ORFeome v1.1 database (http://worfdb.dfci.harvard.edu/) and remapped them to WormBase version WS150. We retained only those 5,822 showing a BLASTN *E*-value $E \leq 10^{-20}$. We collapsed all IST pairs corresponding to the same unordered gene locus pair.

Pairwise Y2H verification. We verified all Y2H interactions by mating fresh individual *MATα* MaV203 DB-ORF yeast cells with their corresponding individual *MATa* MaV103 AD-ORF yeast cells. For genes with more than one clone in the worm ORFeome v1.1, we used the clone with the highest similarity to the IST sequenced in the high-throughput screen for the retest. We tested the resulting diploids for their ability to activate two out of the three Y2H reporter genes. Of the 2,340 potential interactions, 78% (1,816) successfully passed this Y2H retest.

Reference literature data sets: PRS and RRS. To evaluate WI-2007 interactions, we assembled a positive reference set (PRS) and a random reference set (RRS) of binary interactions. We manually recurated physical interactions derived from low-throughput studies in the curated literature, both to ensure high quality and to verify evidence that the interactions were direct and binary, producing the *C. elegans* positive reference set version 1 (cePRS-v1), including 53 worm binary protein-protein interactions. Another 94 pairs selected randomly from the set of ∼50,000,000 pairs among proteins represented as clones in the worm ORFeome v1.1 constituted the *C. elegans* random reference set version 1 (ceRRS-v1). To overcome potential biases of MAPPIT compared to Y2H interactions (the two assays may not be completely independent), we selected only the 47 cePRS-v1 pairs that have been detected by Y2H (cePRS-Y2H-v1) to compute the precision.

MAPPIT assay. In this system, the bait is fused to a STAT recruitment-deficient, homodimeric cytokine receptor and the prey protein is fused to functional STAT recruitment sites (gp130). An interaction between bait and prey allows the activation of a ligand-dependent signal transduction pathway, which controls the activation of a luciferase marker. MAPPIT was performed as described[29] with minor changes. We transfected plasmids into human 293T cells in 96-well plates using a calcium phosphate protocol[30]. Transfected cells were cultured for 24 h in Dulbecco's Modified Eagle's Medium supplemented with 10% fetal bovine serum and then stimulated with erythropoietin (R&D Systems) or left untreated for another 24 h, followed by measurement of luciferase activity in triplicate. For details of the use of MAPPIT to evaluate the Y2H data set, see **Supplementary Methods**.

Functional linkage estimation. The enrichment of a particular data set is expressed as an odds ratio—the number of distinct pairs (excluding homomeric interactions) sharing at least one Gene Ontology term (at a given functional specificity threshold) divided by the number of pairs expected at random. Significance of enrichment was calculated using a one-sided Fisher's exact test. We estimated the space of possible gene pairs as all unordered pairs between the genes in the input data set to account for specific biases of each data set, and then restricted this space to pairs in which both genes have one or more annotations at the considered functional specificity level. The number of genes associated with a particular Gene Ontology term was used as an estimate of the functional specificity, and we calculated the enrichments for several functional specificity levels (5, 20, 100 and 400). Differences between enrichments were assessed using an independent, two-sample *t*-test. **Supplementary Figures 4** and **5** online detail the separate branches of the Gene Ontology.

Additional methods. Detailed descriptions of the cloning and transformation steps, MAPPIT scoring, WI-2007 characterization through Monte Carlo simulation, reprocessing of BPmaps and WI-2004 data, overlap between component networks, module-scale subnetwork extraction, and chronogram intersections, as well as LCI, interologs, genetic interactions, coexpression, phenotypic similarity and anatomical annotation data sets, are available in **Supplementary Methods**. WI8 is provided with MIMIX specifications as **Supplementary Data 1** online. The integrated functional network is available as **Supplementary Data 2** online.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank F. Piano and members of the Cancer Center for System Biology and the Vidal laboratory for discussions, A. Petcherski from WormBase for assistance with worm genetic interactions, and Z. Hu for VisANT assistance. The worm interactome project was supported by grants from the US National Institutes of Health—R01 HG001715 (M.V. and F.P.R.), R01 HG003224 (F.P.R.), F32 HG004098 (M. Tasan), T32 CA09361 (K.V.)—a University of Ghent grant GOA12051401 (J.T.), and the Fonds Wetenschappelijk Onderzoek – Vlaanderen (FWO-V) G.0031.06 (J.T.). I.L. was supported by a postdoctoral fellowship from the FWO-V. K.C.G. and H.-L.K. were supported by US Department of the Army Award W81XWH-04-1-0307 and the State of New York's Science and Tech Resources contract C040066. M.V. is a Chercheur Qualifié Honoraire from the Fonds de la Recherche Scientifique (FRS-FNRS, French Community of Belgium).

AUTHOR CONTRIBUTIONS

J.-F.R., N.S. and A.-R.C. coordinated experiments and data analysis. J.-F.R., T.H.-K., J.M.S., F.G., S.C., P.B., N.L., N.A.-G., E.D., D.S., A.D., C.S., M.V., H.Y., M.B., S.M., M.D., M. Tewari and J.S.A. performed the high-throughput ORF cloning and Y2H screens. I.L., A.-S.d.S., P.B. and J.T. conducted the MAPPIT experiments. N.S., A.-R.C., M. Tasan, T.H., N.K., K.V., C.F., N.B., M.A.Y., C.L., A.S., H.-L.K. and K.C.G. performed the computational analyses. M. Tasan, N.S., C.F., A.-R.C., H.-L.K. and K.C.G. adapted or built the website and visualization tools. N.S., A.-R.C., J.-F.R., M.E.C., J.V., F.P.R. and M.V. wrote the manuscript. M.V. conceived the project. D.E.H., J.T., F.P.R. and M.V. co-directed the project.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Walhout, A.J. *et al.* Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
2. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
3. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004).
4. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
5. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
6. Davy, A. *et al.* A protein-protein interaction map of the *Caenorhabditis elegans* 26S proteasome. *EMBO Rep.* **2**, 821–828 (2001).
7. Boulton, S.J. *et al.* Combined functional genomic maps of the *C. elegans* DNA damage response. *Science* **295**, 127–131 (2002).
8. Reboul, J. *et al. C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* **34**, 35–41 (2003).
9. Walhout, A.J. *et al.* Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Curr. Biol.* **12**, 1952–1958 (2002).
10. Kim, J.K. *et al.* Functional genomic analysis of RNA interference in *C. elegans*. *Science* **308**, 1164–1167 (2005).
11. Tewari, M. *et al.* Systematic interactome mapping and genetic perturbation analysis of a *C. elegans* TGF-β signaling network. *Mol. Cell* **13**, 469–482 (2004).
12. Matthews, L.R. *et al.* Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs." *Genome Res.* **11**, 2120–2126 (2001).
13. Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat. Methods* advance online publication, doi:10.1038/nmeth.1280 (7 December 2008).
14. Hunt-Newbury, R. *et al.* High-throughput in vivo analysis of gene expression in *Caenorhabditis elegans*. *PLoS Biol.* **5**, e237 (2007).
15. Dupuy, D. *et al.* Genome-scale analysis of *in vivo* spatiotemporal promoter activity in *Caenorhabditis elegans*. *Nat. Biotechnol.* **25**, 663–668 (2007).
16. Kao, H.L. & Gunsalus, K.C. Browsing multidimensional molecular networks with the generic network browser (N-Browse). *Curr. Protoc. Bioinformatics* Ch. 9, Unit 9 11 (2008).
17. Hu, Z., Mellor, J., Wu, J. & DeLisi, C. VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics* **5**, 17 (2004).
18. Motegi, F., Velarde, N.V., Piano, F. & Sugimoto, A. Two phases of astral microtubule activity during cytokinesis in *C. elegans* embryos. *Dev. Cell* **10**, 509–520 (2006).
19. Branda, C.S. & Stern, M.J. Mechanisms controlling sex myoblast migration in *Caenorhabditis elegans* hermaphrodites. *Dev. Biol.* **226**, 137–151 (2000).
20. Wolf, F.W., Hung, M.S., Wightman, B., Way, J. & Garriga, G. vab-8 is a key regulator of posteriorly directed migrations in *C. elegans* and encodes a novel protein with kinesin motor similarity. *Neuron* **20**, 655–666 (1998).
21. Schlaitz, A.L. *et al.* The *C. elegans* RSA complex localizes protein phosphatase 2A to centrosomes and regulates mitotic spindle assembly. *Cell* **128**, 115–127 (2007).
22. Gunsalus, K.C. *et al.* Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* **436**, 861–865 (2005).
23. Lee, I. *et al.* A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nat. Genet.* **40**, 181–188 (2008).
24. Gunsalus, K.C., Yueh, W.C., MacMenamin, P. & Piano, F. RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects. *Nucleic Acids Res.* **32**, D406–D410 (2004).
25. Wilson, C.A., Kreychman, J. & Gerstein, M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**, 233–249 (2000).
26. Tatusov, R.L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
27. Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* advance online publication, doi:10.1038/nmeth.1281 (7 December 2008).
28. Boxem, M. *et al.* A protein domain-based interactome network for *C. elegans* early embryogenesis. *Cell* **134**, 534–545 (2008).
29. Eyckerman, S. *et al.* Design and application of a cytokine-receptor-based interaction trap. *Nat. Cell Biol.* **3**, 1114–1119 (2001).
30. Lemmens, I., Lievens, S., Eyckerman, S. & Tavernier, J. Reverse MAPPIT detects disruptors of protein-protein interactions in human cells. *Nat. Protoc.* **1**, 92–97 (2006).
31. Lee, M.-H. & Schedl, T. RNA-binding proteins. in *WormBook* (ed. Blumenthal, T.) doi:10.1895/wormbook.1.7.1 (2006).
32. Podbilewicz, B. Cell fusion. in *WormBook* (eds. Kramer, J.M. & Moerman, D.G.) doi:10.1895/wormbook.1.7.1 (2006).

DOCUMENT JOINT 6

Titre : Signatures of Darwinian Network Evolution in a Plant Interactome Map

Auteurs : The Arabidopsis interactome mapping consortium; Pascal Braun, Anne-Ruxandra Carvunis, Benoit Charloteaux, Matija Dreze, Joseph R. Ecker, David E. Hill, Frederick P. Roth, Marc Vidal.

Description : Article décrivant la construction et l'analyse d'une carte d'interactions physiques entre protéines de la plante *Arabidopsis thaliana*, actuellement en révision chez le magazine *Science.* Me contacter pour le matériel supplémentaire, volontairement omis ici car contenant ~80 pages.

Contribution : J'ai dirigé toutes les analyses bioinformatiques de ce projet depuis le début (2008). Matija Dreze a dirigé toutes les expériences sauf le wNAPPA, réalisé par Mary Galli. Matija Dreze et moi avons développé de nombreuses innovations expérimentales ensemble. Benoit Charloteaux a travaillé avec moi intensément sur la partie évolution. Tong Hao et moi avons apporté un support bioinformatique constant à la réalisation des expériences. Marc Vidal et Joseph R. Ecker sont les directeurs des principaux laboratoires impliqués. Pascal Braun a supervisé le projet dans son ensemble. Les autres membres du consortium, dont la liste et l'ordre restent à déterminer, ont apporté diverses contributions intellectuelles et techniques.

Signatures of Darwinian Network Evolution in a Plant Interactome Map

Arabidopsis Interactome Mapping Consortium*

*Lists of participants and affiliations appear at the end of the main text.

ABSTRACT

Plants evolved unique features pertaining to their central role in ecosystems. How complex cellular networks underlie plant-specific molecular functions remains vastly under-explored. Here we describe the first proteome-wide binary protein-protein interaction map for the interactome network of a plant. Our dataset contains ~6,200 highly reliable interactions between ~2,700 *Arabidopsis thaliana* proteins. This represents a ten-fold increase over what is currently available from equally reliable small-scale literature-curated information for that organism. A global organization of plant biological processes emerges from community analyses of the resulting network, and large numbers of novel hypothetical functional links between proteins and pathways appear. Dynamic interaction rewiring following gene duplication events supports a Darwinian model for interactome networks evolution. This and future plant interactome maps should facilitate systems approaches to understand plant biology and improve crops.

Plants are central components of most ecosystems because of their ability to convert solar energy into biological energy. Plants provide food, fuel, and fiber for human existence. A better understanding of the mechanisms controlling plant processes, such as growth, development, and responses to biotic and abiotic stresses, is of vital importance in addressing current and future agronomical and environmental challenges (*1*). Classical genetic and molecular studies have established mechanisms underlying many genotype-to-phenotype relationships for a variety of plant systems. Yet, more than ten years after the publication of the first plant genome sequence for the reference plant *Arabidopsis thaliana* (*2*) (hereafter Arabidopsis), at least 60% of its protein-coding genes remain functionally uncharacterized.

Genotype-to-phenotype relationships are mediated by macromolecules physically interacting in highly complex and dynamic "interactome" networks. Knowledge of these networks is thus indispensable for a global understanding of genotype-to-phenotype relationships (*3*). Despite a wealth of molecular and genetic data, there is a noticeable lack of experimentally determined protein-protein interactions for Arabidopsis, and more generally of systematically generated datasets for any plant system (Fig. 1A; fig. S1; tables S1 and S2). This is in stark contrast to well-studied non-plant species, where proteome-scale interactome maps have provided a rich resource with which to understand both systems level biological organization and detailed molecular mechanisms. Here we describe a systematic, proteome-scale protein-protein interactome map for Arabidopsis and demonstrate its value to uncover novel aspects of plant biology, systems organization, and evolution.

An experimental high-quality systematic binary interactome map for a plant

We first assembled a Gateway compatible collection (*4*) of ~8,500 open reading frames (ORFs) for Arabidopsis, representing ~32% of all predicted protein-coding genes (Fig. 1B; table S3; Supporting Online Material (SOM) I)(*5*). All pair-wise combinations of proteins encoded by these ORFs ("space 1") were tested using a high-throughput binary interactome mapping pipeline based on the yeast two-hybrid (Y2H) system (SOM II)(*6*). To further increase data quality, we improved the methodology used so far for generating interactome datasets for other organisms by implementing three modifications (Fig. 1C)(*7-10*): (i) the entire Y2H selection step was completed twice to increase sampling depth, (ii) the phenotypes of each candidate Y2H pair were verified four times independently by different experimenters; each was required to score positive at least three times to ensure reproducibility (fig. S2), and (iii) the identity of each verified Y2H pair was confirmed by DNA sequencing. Confirmed Y2H protein pairs were assembled into a dataset named "Arabidopsis Interactome version 1, main screen" ("AI-1_{MAIN}"), containing 5,664 binary interactions between 2,661 proteins (table S4).

We evaluated the quality and overall coverage of AI-1_{MAIN} with a recently developed empirical interactome mapping framework (*8*). The proportion of true biophysical interactions (*precision*) was determined by testing a set of 249 randomly chosen interactions from AI-1_{MAIN} (4.5% of all detected Y2H interactions) in a benchmarked 96-well microtiter plate based validation assay called "well Nucleic Acid Programmable Protein Array" (wNAPPA; SOM III)(*11, 12*). For benchmarking, we assembled a positive reference set of well-documented Arabidopsis protein-protein interactions from the literature (AtPRS-v1)(*13*) and a random reference set of protein pairs (AtRRS-v1) (fig. S3; table S5; SOM IV). These constitute a public resource for standardizing Arabidopsis binary protein-protein interaction assays (http://interactome.dfci.harvard.edu/A_thaliana/index.php?page=2010anm_download, http://signal.salk.edu/cdna/gateway/PRS-RRS_Salk.txt). Under optimal experimental conditions, we found that the validation rate of AI-1_{MAIN} subset pairs was statistically

indistinguishable from those of AtPRS-v1 (Fig. 1D), with values comparable to the benchmarks of other validation assays using reference sets for worm, human and yeast (*7-9*). AI-1_{MAIN} is therefore of similar quality to highly reliable small-scale interactions from the literature (Fig. 1D; fig. S4; table S5; SOM III and IV).

To estimate the proportion of the complete Arabidopsis protein-protein interactome covered by AI-1_{MAIN}, we experimentally measured three other framework parameters (*8*): (i) *assay sensitivity*, the proportion of all binary interactions that are detectable with our implementation of Y2H; (ii) *sampling sensitivity*, the fraction of all detectable interactions observed in our mapping experiment, *i.e.* its degree of saturation; and (iii) *screening completeness,* the fraction of all possible pair-wise protein combinations that were tested in our experiment.

To determine our Y2H assay sensitivity, we applied the pipeline verification step protocol (Fig. 1C) to each AtPRS-v1 pair. Combined with the results of the interactome mapping pipeline (Fig. 1C), this experiment led to the detection of 43 of the 118 AtPRS-v1 pairs, corresponding to an assay sensitivity of 36.4% ± 4.4% (mean ± standard deviation; Fig. 1D; table S5; SOM IV). To measure the sampling sensitivity, we repeated six iterations of the pipeline selection step (Fig. 1C) on a defined subspace within space 1 (Fig. 1B; table S6). The resulting dataset ("AI-1_{REPEAT}") contained 1,066 interactions between 673 proteins (tables S4 and S7). The total number of observed interactions increased with each iteration, but did not reach saturation (Fig. 1E). The corresponding saturation curve resembled the one observed for the 31 AtPRS-v1 pairs embedded in the subspace (Fig. 1E; SOM IV), indicating that both known and newly identified interactions are similarly affected by sampling in our experiment. Taking the AtPRS-v1 pairs and all AI-1_{REPEAT} pairs independently, we modeled the sampling sensitivity of AI-1_{MAIN}. Both yielded an estimate of ~37% saturation after two screens (fig. S5; SOM IV). The product of assay and sampling sensitivity, reflecting the overall sensitivity of our experiment, was in close agreement with the proportion of AtPRS-v1 pairs detected after two screens (13.4% ± 2.2% and 15.7% ± 3.8%, respectively, mean ± standard deviation; SOM IV). Altogether, we therefore estimate that the complete Arabidopsis protein-protein interactome contains 300,000 ± 80,000 binary interactions (mean ± standard deviation; SOM IV). While this represents a larger interactome size than estimated for human, worm or yeast with the same framework, the interaction density in all four species appears similar (5-10 interactions per 10,000 possible pairs)(*7-9*).

AI-1 increases ten times the number of highly reliable interactions for Arabidopsis

The integration of both binary interactions and protein complex associations reported in public databases for Arabidopsis yielded 5,270 protein pairs (SOM V). We assembled these into a literature-curated interaction dataset (LCI), from which we extracted a subset of ~4,250 binary interactions (LCI_{BINARY}) and a subset of ~600 "high-quality" binary interactions (LCI_{CORE}) consisting of binary interactions described in ≥2 papers and/or by ≥2 methods (fig. S1; tables S1 and S4)(*13)*. Although AI-1_{MAIN} and LCI_{BINARY} are small-world networks containing similar numbers of proteins (nodes) and interactions (edges), their topology is significantly different with shorter distances between proteins and a lower clustering for AI-1_{MAIN} (fig. S6). This is likely due to inspection biases inherent to LCI datasets (*9*). The overlap of LCI_{BINARY} with AI-1_{MAIN} is significantly larger than with randomized networks, but slightly smaller than expected given the completeness and sensitivity of AI-1_{MAIN} (Fig. 1, F and G; SOM V). In contrast, the overlap between AI-1_{MAIN} and the higher quality LCI_{CORE} falls exactly within the range predicted by the framework parameters indicating that AI-1_{MAIN} is of the same quality as well documented interactions reported in the literature, as shown in other respect in Fig. 1D. The smaller overlap with LCI_{BINARY} therefore suggests a lower quality for less documented LCI interactions (Fig. 1G),

as previously described (*13, 14*).

We combined AI-1_{MAIN} and AI-1_{REPEAT} into a single dataset named AI-1, containing 6,205 interactions between 2,774 proteins. The union of LCI_{BINARY} and AI-1 contains 10,361 interactions between 4,439 proteins, covering ~3.5% of the projected complete Arabidopsis interactome (table S4). Altogether, AI-1 doubles the number of experimentally detected Arabidopsis protein-protein interactions and increases it ten times compared to the ~600 highly reliable LCI_{CORE} interactions (*13, 14*).

Overlap of AI-1 with other biological relationships

To evaluate the biological relevance of AI-1 interactions, we first tested the extent to which the level of mRNA transcripts encoding interacting protein pairs correlates across a compendium of expression arrays (SOM V). As for interactome maps of other organisms (*7, 9*), the observed correlation is significantly higher for AI-1_{MAIN} interacting protein pairs than for control pairs of proteins present in AI-1_{MAIN} but not found to interact (hereafter "non-interacting pairs"; Fig. 2A; SOM V). We also assessed to what extent interacting proteins share Gene Ontology (GO) annotations. Compared to non-interacting pairs of proteins, interacting proteins are more often involved in the same biological processes, localize to the same sub-cellular compartments, and most notably share molecular functions, with GO annotations assigned to only a few proteins (precise GO annotations, Fig. 2B; SOM V). In addition, interacting proteins share a precise GO mutant phenotype four times more often than non-interacting pairs (Fig. 2C). Lastly, proteins that share interactors (but do not themselves directly interact) are also enriched in common precise GO annotations, especially when sharing more than half of their interaction partners (Fig. 2D; SOM V). Together these trends support the overall biological relevance of interactions in AI-1.

Similar to the whole Arabidopsis proteome, but in obvious contrast to proteins in LCI (fig. S7), two thirds of proteins in AI-1 lack GO annotations altogether or lack precise GO annotations (Fig. 2E). Therefore, AI-1 will be a powerful resource to better characterize these proteins (Fig. 2E), as exemplified in Fig. 3.

New hypotheses pertaining to plant signaling networks

We explored the extent to which our new interactome map can aid the development of hypotheses and can reveal novel features of plant signaling networks. Stratification of interactions with orthogonal functional data can expand current knowledge and uncover unexpected biological relationships at the scale of individual proteins, pathways, and whole networks (*15, 16*).

An important goal of systems analysis is the understanding of direction and rate of information flow through networks. Protein-protein interactions are not directed unless functional relationships such as "A modifies B" can be inferred based on orthogonal information. Kinase and phosphatase interactions with their respective substrates can be used as an example to illustrate this approach. In AI-1, 220 proteins interact with a protein kinase or phosphatase and are thus potential substrates. Among the interactions between these proteins, the 38 involving a protein for which phosphorylation has been experimentally demonstrated suggest particularly appealing candidates. The recovery of the known MKK4-MPK6 interaction indicates that this approach can indeed identify genuine kinase-substrate pairs (*17*). Within space 1 we double the number of putative kinase/phosphatase-substrate interactions and provide compelling starting points for a deeper functional characterization of the involved proteins (Fig. 3A; SOM V).

Another example of directed functional relationships involves ubiquitinating enzymes and their substrates. The ubiquitin system is greatly expanded in plants, with ~1400 potential E3 ubiquitin ligases (E3s) in Arabidopsis (*18*), and likely plays a critical

role in all aspects of plant physiology. Nonetheless, the specific targets of most E3s, and consequently a systems level understanding of ubiquitin signaling complexity, remain elusive. AI-1 quadruples the number of binary interactions in the ubiquitination cascade for Arabidopsis and increases it 7-fold within space 1 (Fig. 3B). Like for kinase-substrate pairs, directional information flow can be inferred within components of the ubiquitin-ligase system, *e.g.* from E2s to E3s, and between E3s and potential substrates. The 32 interactions between E3s and 20 proteins experimentally shown to be ubiquitinated appear to form an intricate network, where some putative targets interact with several E3s and many E3s interact with the same putative target suggesting a combinatorial control system (Fig. 3B; SOM V). Thus, our data point towards a high regulatory complexity within the ubiquitin system and provide starting points for exploration of this critical signaling system.

Plant hormones are key regulators of development and mediate responses to a wide variety of environmental stimuli. For example, jasmonic acid (JA) serves as a primary signal in the regulation of plant reproductive development and of defense against necrotrophic pathogens (*19*). Current understanding of the transcriptional response to JA from the perspective of protein-protein interactions is limited to interactions between transcriptional repressors containing a JA ZIM-domain (JAZ) and the transcription factor MYC2 (Fig. 3C). Given the involvement of JA in a wide range of biological processes, it has been hypothesized that other transcription factors are regulated through their interaction with JAZ proteins (*20*). In agreement with this hypothesis, we found seven other transcription factors binding to JAZ proteins in AI-1 (Fig. 3C). Moreover, we identified three JAZ-related proteins (*21*) that also interact with transcription factors, two of which also bind a JAZ protein, suggesting a complex transcriptional regulation module.

Auxin is another plant hormone critical for nearly all developmental and defense processes in plants (*19*). In the auxin signaling pathway, AUX/IAA proteins mediate the transcriptional repression of response genes via physical interactions with the TUP/Groucho like-repressor TOPLESS (TPL), through an ethylene-response-factor-associated amphiphilic repression (EAR) motif (*22*). Twelve interactions in AI-1, including seven not described in the literature, connect either TPL or TPL-related 3 (TPR3) to ten AUX/IAA proteins (Fig. 3D), each containing an EAR motif. While only two non-AUX/IAA interactors of TPL have been reported (*23, 24*), we found 19 such interactors in AI-1 (all novel), of which 15 contain a predicted EAR motif supporting the validity of these interactions ($P < 10^{-24}$, hypergeometric test)(*25*). This indicates that TPL is involved in a much larger signaling network than currently appreciated.

In addition to its role in auxin response, TPL is implicated in JA signaling by indirect association with JAZ proteins via the adaptor protein NINJA (*24*). Expanding on this finding in AI-1, JAZ5 and JAZ8 directly interact with TPL, likely via their EAR motifs (Fig. 3D). Because of TPL involvement in both auxin and JA signaling, and because of the large number of proteins containing a predicted EAR motif in Arabidopsis (*25*), it has been suggested that TPL plays a general role as a transcriptional co-repressor in other hormone-signaling pathways (*24*). In agreement with this hypothesis, in AI-1 TPL interacts with a transcriptional regulator of ethylene response (ERF9) and known regulators of salicylic acid signaling (NIMIN-2, NIMIN-3; Fig. 3D). Furthermore, we find that the JA responsive co-repressor JAZ9 interacts with components of the gibberellin response pathways, the DELLA proteins RGA1 and RGA2, a connection which likely has a similar function as the recently described competitive inhibition of JAZ1 by DELLA proteins (Fig. 3E)(*26*). Together, these observations suggest that various co-repressors, adaptors and transcription factors may assemble in a modular way to integrate simultaneous inputs from several hormones.

Overall, we demonstrate the power of combining a systematic high-quality interactome map with orthogonal biological information to develop novel hypotheses about

individual proteins and to uncover unexpected network complexity even in well-characterized signaling systems. To facilitate the use of AI-1 for future hypothesis generation, we provide several resources, including a table with 26 types of protein annotations from published large-scale studies or databases, 23 subsets of interactions including predicted interactions overlapping with AI-1, and an online search tool (Table 1; tables S8 and S9; SOM V; http://interactome/A_thaliana/index.php?page=display).

Communities in the AI-1_{MAIN} network

Systematically derived proteome-wide datasets such as AI-1_{MAIN} provide an opportunity to investigate the global structural organization of biological systems. In many networks, communities of tightly interconnected components that function together can be identified (*27*). We used a recently developed edge clustering approach (*28*) to identify communities in AI-1_{MAIN} and investigated their biological relevance. Edge clustering approaches, in contrast to more commonly used node clustering methods, use edges, *i.e.* protein-protein interactions, as elements to identify communities and can therefore naturally assign proteins to more than one community. This is necessary and conceptually appealing as many proteins participate in different biological functions (*29*).

We identified 26 communities containing more than five proteins in AI-1_{MAIN} (Fig. 4A; fig. S8; SOM VI) and investigated their biological relevance. We found that ~90% of these communities were enriched in at least one GO annotation (all $P < 0.05$; Fig. 4A; fig. S8; table S10; SOM VI), whereas the vast majority of negative control networks randomized by degree-preserving edge shuffling harbored no such meaningful communities ($P = 0.01$; Fig. 4A ; fig S9).

The biological validity of AI-1_{MAIN} communities was further supported by a detailed inspection at the level of individual proteins (figs. S10-35). For example, the "brassinosteroid signaling/phosphoprotein-binding" community contains several 14-3-3 proteins known to regulate brassinosteroid signaling (fig. S10). Consistent with the tendency of 14-3-3 proteins to interact with phosphorylated partners (*30*), this community is enriched in phosphorylated proteins ($P = 0.005$, Fisher's exact test). Furthermore, the interactions between the 14-3-3 proteins and the ABA-induced AREB3 transcription factor are corroborated by previous findings in barley (*31*), and suggest that plant 14-3-3 proteins mediate multiple hormone signaling pathways.

Several communities, such as "transcription" and "nucleosome assembly", share proteins indicating linked biological processes (fig. S36). Particularly striking is the large "transmembrane transport" community sharing thirteen and six proteins with the "vesicle mediated transport" and "water transport" communities, respectively (Fig. 4B). These shared proteins are bridged via four well-connected proteins within the "transmembrane transport" community, including two membrane-tethered NAC-type transcription factors, ANAC089 and NTL9. Transcription factors in this family are thought to be activated by release from the cellular membrane by endopeptidase- or ubiquitin-mediated cleavage (*32*). Interactions corresponding to both mechanisms are found in the "transmembrane transport" community (fig. S37). While NTL9 has previously been implicated in osmotic stress signaling (*33*), ANAC089 is uncharacterized. The similar position bridging three network communities and the high fraction of common interactors (42%) of ANAC089 and NTL9 suggest a related function for these proteins, which is supported by upregulation of ANAC089 mRNA levels in guard cells and other tissues in response to osmotic insults or abscisic acid (figs. S38 and S39)(*34-36*).

Together, the significant enrichment of shared GO annotations within communities, literature-based inspection of intra-community relationships, and examination of community boundaries, support the relevance of the communities identified in AI-1_{MAIN}. We anticipate that more comprehensive interactome maps will reveal extensive structural

organization. Though the view remains incomplete, a picture of the systems organization of plant interactome networks begins to emerge from network analyses of AI-1_{MAIN}.

Signatures of Darwinian interactome network evolution

Genotype-to-phenotype relationships are, at least in part, mediated through physical and functional interactions between genes and gene products. Yet the interplay between natural selection and interactome networks remains vastly under-explored. We addressed this issue in the context of gene duplication, a major driving force of evolutionary novelty (*37*). Following duplication of individual genes, chromosomal segments or whole genomes, most of the resulting gene copies are lost relatively rapidly (*38*). The few paralogous pairs that persist are thought to confer selective advantages usually attributed to a range of phenomena including sub- and neo-functionalization (*39*). This idea is consistent with the non-constant rates of divergence observed in predicted protein sequences. Initially high following gene duplication, and indication of a transient relaxation of selective pressure, these rates subsequently decline as selective pressure tightens on retained paralogs (*38, 40, 41*).

In striking contrast with such "rapid-then-slow" protein sequence divergence, the rate of protein-protein interaction rewiring in interactome network graphs is usually modeled as a constant (Fig. 5A)(*42-44*). This implicitly assumes that protein-protein interactions evolve randomly following duplication (SOM VII). However, the "duplication-divergence" model suggests that sub- and neo-functionalization occur through interactome network rewiring, *i.e.* interaction losses and/or acquisitions (*45, 46*).

Previous efforts to quantify the rate of interaction rewiring following duplication relied on datasets that contained few paralogous pairs. Because the genome of Arabidopsis contains a high fraction of duplicated genes compared to well-studied non-plant species (73% and ~40%, respectively) and because AI-1_{MAIN} is larger than datasets of similar quality generated for human (*10*) or yeast (*9*), it contains ~10 times more pairs of paralogous proteins (fig. S40). The 1,882 paralogous pairs in AI-1_{MAIN} span a wide range of apparent interaction rewiring as measured for each pair by the fraction of shared interactors (Fig. 5B).

Because the ultimate downstream effect of paralog evolution is their functional divergence, we investigated network rewiring of proteins encoded by paralogous gene pairs classified in a recent publication as having "no", "some", or "high" functional divergence on the basis of morphological consequences of knocking out either one or both members of these pairs (*47*). For the 17 pairs in AI-1_{MAIN} for which comparative phenotypic data was available, the fraction of shared interactors was predictive of this functional divergence classification (Fig. 5B). This supports the validity of our approach and the relevance of our data to study paralog evolution dynamic.

To study the rate of interaction rewiring following duplication, we divided AI-1_{MAIN} paralogous pairs into four age groups corresponding to the time elapsed since the duplication event based on comparative genomics ("time-since-duplication"; SOM VII). In addition, since dataset incompleteness can often induce an illusion of divergence (fig. S41)(*43*), we corrected for this artifact by applying concepts of our empirically controlled quantitative framework (*8*) and estimated an expected upper bound to normalize the fraction of observed common interactors between paralogous proteins (fig. S41; SOM VII). We then measured and corrected the average fraction of shared interactors of paralogous pairs in the four age groups (Fig. 5C). This fraction decreases over evolutionary time, showing evidence of substantial divergence yet also exhibiting retention of ancestral history in the interactome, as the oldest paralogous pairs share on average a higher fraction of interactors than non-paralogous proteins pairs ($P < 2.2 \times 10^{-16}$, Mann-Whitney U-test). Despite variations in the number of paralogous pairs per protein family in AI-1_{MAIN},

the observed trend is largely independent of family size (fig. S42). Paralogous pairs originating from recent whole genome duplications share slightly but significantly more interactors than other paralogous pairs (Fig. 5D; fig. S43; SOM VII). The small size of the effect is surprising given that paralogs retained following whole-genome duplications are thought to be maintained essentially to preserve gene dosage balance and would therefore be expected to share a greater fraction of their ancestral interactions (*48*). Overall, interaction divergence following duplication exhibits a robust trend of decline with increasing age of paralogous pairs (SOM VII).

We then investigated the dynamic of this decline. Rather than following an exponential decay, expected theoretically for a constant rate of interaction loss (SOM VII), the observed dynamic follows a trend akin to a power-law decay, mirroring the sequence divergence of these paralogous pairs: rapid at first, followed by a slower divergence (Fig. 5C). As rapid-then-slow protein sequence divergence reflects evolutionary selective pressure and paralog retention dynamic (*38, 41*), a rapid-then-slow interaction divergence suggests that interactome network rewiring occurs in a Darwinian fashion and directly contributes to paralog retention. Since sequence changes can simultaneously affect both protein-protein interactions and other aspects of protein function, *e.g.* enzymatic activity, we cannot formally exclude that to some extent evolutionary selected phenotypes would be independent of the protein-protein interaction changes seen here but it is unlikely that the overall trends are caused solely by such indirect effects.

Asymmetric sub-functionalization has been proposed to confer selective advantage (*49*). Thus paralog retention could also depend on asymmetric interactome rewiring, *i.e.* uneven distribution of specific interactors between paralogous proteins. Asymmetric divergence of paralogs has been reported based on sequence comparison (*40*), gene expression (*50, 51*), genetic interactions (*52*) and protein-protein interactions (*49*), although the last suffered from limited sample sizes and technical caveats. We investigated the extent and dynamic of asymmetric protein-protein interaction divergence between Arabidopsis paralogs (SOM VII). We accounted for experimental limitations described in our empirical framework, and for each paralogous pair we measured asymmetry as the deviation from an equal number of specific interactors (perfect symmetry, Fig. 5E; SOM VII). Approximately 13% of paralogous pairs in AI-1_{MAIN} were found to be significantly asymmetric as compared to a binomial randomization control (empirical $P < 0.05$; Fig. 5, E and F; figs. S44-S45; SOM VII). This fraction, as well as the level of asymmetry, was independent of protein family size and did not show any obvious time dependence (Fig. 5F; figs. S46-S47). Thus, the protein-protein interaction asymmetry of paralogs likely appeared before our most recent time-since-duplication point and is maintained throughout evolutionary time as corroborated by the more rapid interaction divergence of asymmetric pairs compared to all paralogous pairs (fig. S48). Similarly, a rapid-then-stable establishment of asymmetry has been independently proposed for a human gene co-expression network (*51*).

The high interaction divergence of the most recent paralogous pairs is surprising, considering their relatively low sequence divergence. While they share less than half of their interactors (41%), and ~13% of them are significantly asymmetric, their sequences are >60% identical on average (Fig. 5C; figs. S49 and S50). This contrast is consistent with the common understanding that protein-protein interactions are only one of many constraints limiting sequence changes during evolution. One extreme example of interaction rewiring under strong sequence constraints is observed in the actin family, represented by six proteins in AI-1_{MAIN} (15 pairs). Each pair shares >90% sequence identity, yet the extent to which each pair shares interactors differs substantially. Collectively the actin family exhibits time-dependent interaction rewiring as more globally observed in AI-1_{MAIN} (Fig. 5G).

Together with observations of fast rewiring of bacterial transcriptional networks (*53*), our data invite speculation that edge rewiring could be faster than node evolution in interactome networks. The non-constant rate of interaction rewiring provides insight into the Darwinian evolution of interactome networks and their topology (*54*). Whether this rewiring is merely a consequence of non-constant paralog sequence divergence or is a primary driver remains an open question.

Conclusion

We described the first empirically determined high-quality protein-protein interaction map for a plant interactome network. Our dataset should not only hasten the functional characterization of unknown gene products, including those with potential biotechnological utility, but should also enable systems level investigations of genotype-to-phenotype relationships in the plant kingdom through the lens of network science. Just as network approaches have facilitated the understanding of several human pathologies (*16, 55-58*), we expect that AI-1 may help illuminate mechanisms and strategies by which plants cope with environmental and pathogenic challenges (Mukhtar *et al.*, co-submitted).

The paradigms established here are compatible with models in which the interactome network constrains and shapes sequence evolution. Studying sequence variation (conservation, mutation, evolution rate) has shed light on how natural selection drives evolution. Explorations of interaction variation will similarly broaden the understanding of interactome network evolution whether in the context of duplication or in trans-kingdom comparative interactomics.

FIGURE LEGENDS

Fig. 1. AI-1 doubles the experimental coverage of the *Arabidopsis thaliana* interactome. (**A**) Fraction of the Arabidopsis, human (*Homo sapiens*) and yeast (*Saccharomyces cerevisiae*) proteome with binary protein-protein interactions reported in IntAct (*59*) (table S1). (**B**) Space 1 relative to all possible Arabidopsis ORF pairs. Gray inset: subspace screened in the repeat experiments (Fig. 1E). (**C**) Experimental pipeline used for large-scale interactome mapping. To obtain AI-1_{MAIN} and AI-1_{REPEAT}, the Y2H selection step was completed twice on space 1 and six times on the subspace. Right: representative examples and schematizations. (**D**) Validation by wNAPPA. Top: fraction of pairs of AtPRS-v1, AtRRS-v1 or a random subset from AI-1_{MAIN} that scored positive in wNAPPA and associated precision across a range of scoring thresholds (SOM III). Gray area: optimal score range. Dashed line: z-score of 1.5. Bottom: fraction of pairs of AtPRS-v1, AtRRS-v1 or a random subset from AI-1_{MAIN} that scored positive in Y2H or in wNAPPA at a z-score of 1.5 (SOM III). Error bars: standard error of the proportion. *P*-values: one-sided two-sample *t*-tests. AtPRS-v1 pairs are significantly more often detected than AtRRS-v1 pairs in wNAPPA ($P = 2 \times 10^{-8}$, one-sided two-sample *t*-test) and in Y2H ($P < 2.2 \times 10^{-16}$, one-sided Fisher's exact test). (**E**) Sensitivity of the interactome mapping strategy measured by screening six times the subspace (Fig. 1B). Average number of Y2H interactions detected for AtPRS-v1 pairs and for all pairs as a function of the number of screening iterations (SOM III). Error bars: standard deviation. (**F**) Distribution of the number of common interactions between LCI_{BINARY} and sets of 5,664 protein pairs randomly extracted from space 1 (black), or randomized networks with structure topology and protein composition identical to AI-1_{MAIN} (red). Empirical *p*-value (SOM V). (**G**) Influence of framework parameters on AI-1_{MAIN} overlap with literature. Bars with full lines, darker colors: observed counts. Dashed bars, light colors: expected counts given framework parameters. AtPRS-v1 pairs were removed from LCI_{CORE} and LCI_{BINARY} for these analyses. Error bars: two standard deviations from the expected counts. Observed

counts of interactions from the literature detected in AI-1_{MAIN} are not largely different from expected, $P = 0.047$ (LCI_{BINARY}), $P = 0.48$ (LCI_{CORE}), z-test. Network cartoons at the bottom represent the impact of the framework parameters on the number of interactions from the literature (full network, left) that we expect to detect experimentally (incomplete network, right).

Fig. 2. AI-1 is enriched in biologically relevant protein-protein interactions. (**A**) Distribution of mRNA expression Pearson correlation coefficients (PCC) over 1,436 arrays (SOM V) for AI-1_{MAIN} interacting and non-interacting pairs. Inset: percentage of AI-1_{MAIN} interacting and non-interacting pairs with PCC > 0.75, *P*-value of one-sided Fisher's exact test; error bars: standard error of the proportion. P_{MW}: *p*-value of Mann-Whitney U-test. (**B**) Enrichments in shared Gene Ontology (GO) annotations for AI-1_{MAIN} interacting pairs (schematized on top) versus non-interacting pairs, in the three branches of the GO vocabulary and as a function of GO annotation breadth (SOM V). All enrichments are statistically significant by Fisher's exact test ($P < 0.05$). (**C**) Fraction of AI-1_{MAIN} interacting and non-interacting pairs sharing a precise GO mutant phenotype (schematized on the left), relative to pairs where both members have a GO mutant phenotype in the biological process branch with a breadth ≤50 ($n = 308$). Error bars: standard error of the proportion. (**D**) Enrichment in shared precise GO annotations (of breadth ≤50) in the indicated GO categories for protein pairs sharing interactors in AI-1_{MAIN} (schematized on top). All directly interacting pairs were excluded. All enrichments are statistically significant by Fisher's exact test ($P < 0.05$). (**E**) Proportion of proteins in AI-1 with and without GO annotation of breadth ≤50. Proteins without such GO annotation are classified as either directly interacting or sharing more than half of their interactors with a protein with such a GO annotation, or both, or neither.

Fig. 3. AI-1 expands concepts of plant signaling pathways. (**A**) Putative phosphorylation signaling subnetwork extracted from $LCI_{BINARY/SPACE1}$ and AI-1. Bar plot: number of protein-protein interactions between kinases/phosphatases and phosphorylated proteins in LCI_{BINARY} and AI-1 (outside and within space 1). (**B**) Putative ubiquitination subnetwork extracted from LCI_{BINARY} and AI-1. Bar plot: number of protein-protein interactions between proteins in the ubiquitination cascade in LCI_{BINARY} and AI-1 (outside and within space 1). (**C**) Aspects of jasmonic acid-mediated transcriptional regulation suggested by protein-protein interactions from LCI_{BINARY} and AI-1. Literature interactions from LCI_{BINARY} and (*60*). (**D**) Novel aspects of transcriptional co-repression mediated by TOPLESS (TPL) and TPL-related 3 (TPR3) through physical interactions with EAR-motif containing proteins in LCI_{BINARY} and AI-1. Literature interactions from LCI_{BINARY} and (*22, 24, 60*). (**E**) Protein-protein interactions in AI-1 support the proposed role of TPL as a global transcriptional co-repressor in multiple hormone-mediated transcriptional responses.

Fig. 4. Systems level organization in AI-1_{MAIN}. (**A**) Link communities in a degree-preserving edge shuffling randomized network (top) and in AI-1_{MAIN} (bottom). Colored regions indicate communities enriched in the GO annotations summarized by the indicated terms (table S10). The randomized network is one example from fig. S9. Only the largest component of each network is shown. GA: gibberellic acid, JA: jasmonic acid, TCA: tricarboxylic acid. (**B**) Example of communities interfaces.

Fig. 5. Signatures of interactome network evolution. (**A**) Schema of rewiring of paralogous protein interactions over time according to the duplication-divergence model. (**B**) Top: distribution of the fraction of shared interactors between paralogous proteins in AI-1_{MAIN}. Bottom: average fraction of interactors shared between pairs of paralogous proteins with no, some, and high functional divergence. Error bars: standard error of the mean (SOM

VII). *P*-value: one-sided Kendall correlation test. (**C**) Corrected average fraction of shared interactors and average protein sequence identity between pairs of paralogous proteins as a function of the estimated Δ time since duplication. Inset: data plotted in log-log scale. Error bars: standard error of the mean (SOM VII). Dashed black line: corrected average fraction of shared interactors of non-paralogous pairs. Full lines: power-law fits; dotted red line: exponential fit to the corrected fraction of shared interactors of paralogous pairs. myrs: million years. (**D**) Corrected average fraction of interactors shared between pairs of paralogous proteins originating from a recent polyploidy events as compared to other paralogous protein pairs of the same age. Error bars: standard error of the mean (SOM VII). (**E**) Thirteen percent of paralogous pairs show significant asymmetric divergence. Top: schema presenting principles of calculations. Bottom: maximum versus minimum number of specific interactors of paralogous pairs in AI-1_{MAIN} and of sample randomization pairs. Dashed diagonal: perfect symmetry. Asymmetry significance assessed by binomial randomizations (empirical P <0.05; SOM VII). (**F**) Fraction of significantly asymmetric pairs (empirical P < 0.05; SOM VII) as a function of the estimated Δ time since duplication. Error bars: standard error of the proportion. myrs: million years. (**G**) Evolution of protein-protein interactions within the actin family. Left: interactions of six actin proteins in AI-1_{MAIN}. Right: sequence identity and fraction of shared interactors for each of the 15 pairs of actin proteins as a function of the estimated Δ time since duplication. Colored arrows: groups of proteins from the same time-since-duplication event. myrs: million years.

Table 1. Examples of interaction types found in AI-1. First column: protein category 1. Second column: protein category 2. Third column: number of interactions in AI-1 between proteins of categories 1 and 2. Four categories of AI-1 interactions, from top to bottom: (i) interactions involving a protein containing one of the 10 most represented domains in AI-1 proteins; (ii) interactions involving proteins evolutionary conserved within the plant kingdom; (iii) interactions between biochemically connected categories of proteins; (iv) interactions computationally predicted and experimentally supported by AI-1. SOM V describes the methods for extracting interactions in all four categories.

ARABIDOPSIS INTERACTOME MAPPING CONSORTIUM.

Authorship of this paper should be cited as "Arabidopsis Interactome Mapping Consortium".
Participants are arranged by group then listed in alphabetical order (AO), except for Chairs, co-Chairs and Project Leaders when indicated.
Correspondence and request for materials should be addressed, to M.V. (marc_vidal@dfci.harvard.edu); J.R.E. (ecker@salk.edu); P.B. (pascal_braun@dfci.harvard.edu).
Matija Dreze, Anne-Ruxandra Carvunis, Benoit Charloteaux, Mary Galli, Samuel J. Pevzner and Murat Tasan contributed equally to this work.

Steering group (AO): Pascal Braun[1,2] (Chair), Anne-Ruxandra Carvunis,[1,2,4] Benoit Charloteaux,[1,2†] Matija Dreze,[1,2,3] Joseph R. Ecker,[5,9] David E. Hill,[1,2] Frederick P. Roth,[1,8¤] Marc Vidal[1,2].

ORFeome group: Mary Galli[5] (Project Leader), Padmavathi Balumuri,[10] Vanessa Bautista,[5] Jonathan D. Chesnut,[10] Chris de los Reyes,[5] Patrick Gilles,[10] Christopher J. Kim,[5] Rosa C. Kim,[10] Uday Matrubutham,[10] Jyothika Mirchandani,[10] Eric Olivares,[10] Suswapna Patnaik,[10] Rosa Quan,[5] Gopalakrishna Ramaswamy,[10] Paul Shinn,[5] Geetha M. Swamilingiah,[10] Stacy Wu,[5] Joseph R. Ecker[5,9] (Chair).

Interactome data acquisition group: Matija Dreze[1,2,3] (Project Leader), Danielle Byrdsong,[1,2] Amélie Dricot,[1,2] Melissa Duarte,[1,2] Fana Gebreab,[1,2] Bryan J. Gutierrez,[1,2] Andrew MacWilliams,[1,2] Dario Monachello,[11] M. Shahid Mukhtar,[12‡] Matthew M. Poulin,[1,2] Patrick Reichert,[1,2] Viviana Romero,[1,2] Stanley Tam,[1,2] Selma Waaijers,[1,2§] Evan M. Weiner,[1,2] Marc Vidal[1,2] (co-Chair), David E. Hill[1,2] (co-Chair), Pascal Braun[1,2] (Chair).

NAPPA interactome validation group: Mary Galli[5] (Project Leader), Anne-Ruxandra Carvunis,[1,2,4] Michael E. Cusick,[1,2] Matija Dreze,[1,2,3] Viviana Romero,[1,2] Frederick P. Roth,[1,8¤] Murat Tasan,[8] Junshi Yazaki,[9] Pascal Braun[1,2] (co-Chair), Joseph R. Ecker[5,9] (Chair).

Bioinformatics and analysis group: Anne-Ruxandra Carvunis[1,2,4] (Project Leader), Yong-Yeol Ahn,[1,13] Albert-László Barabási,[1,13] Benoit Charloteaux,[1,2†] Huaming Chen,[5] Michael E. Cusick,[1,2] Jeffery L. Dangl,[12] Matija Dreze,[1,2,3] Joseph R. Ecker,[5,9] Changyu Fan,[1,2] Lantian Gai,[5] Mary Galli,[5] Gourab Ghoshal,[1,13] Tong Hao,[1,2] David E. Hill,[1,2] Claire Lurin,[11] Tijana Milenkovic,[14] M. Shahid Mukhtar,[12‡] Samuel J. Pevzner,[1,2,6,7] Natasa Przulj,[15] Sabrina Rabello,[1,13] Edward A. Rietman,[1,2] Frederick P. Roth,[1,8¤] Balaji Santhanam,[1,2] Robert J. Schmitz,[9] William Spooner,[16] Joshua Stein,[16] Murat Tasan,[8] Jean Vandenhaute,[3] Doreen Ware,[16,17] Pascal Braun[1,2] (co-Chair), Marc Vidal[1,2] (Chair).

Writing group (AO): Pascal Braun[1,2] (Chair), Anne-Ruxandra Carvunis,[1,2,4] Benoit Charloteaux,[1,2†] Matija Dreze,[1,2,3] Mary Galli,[5] Marc Vidal[1,2] (co-Chair).

[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, MA 02215, USA.
[2]Department of Genetics, Harvard Medical School, Boston, MA 02115 USA.
[3]Unité de Recherche en Biologie Moléculaire, Facultés Universitaires Notre-Dame de la Paix, 5000 Namur, Wallonia, Belgium.
[4]Computational and Mathematical Biology Group, TIMC-IMAG, CNRS UMR5525 and Université de Grenoble, Faculté de Médecine, 38706 La Tronche cedex, France.
[5]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA.
[6]Biomedical Engineering Department, Boston University, Boston, MA 02215, USA.
[7]Boston University School of Medicine, Boston, MA 02118, USA.
[8]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA.
[9]Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA.
[10]Life Technologies, Carlsbad, CA 92008, USA.
[11]Unité de Recherche en Génomique Végétale (URGV), UMR INRA/UEVE - ERL CNRS 91057 Evry Cedex, France.
[12]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.
[13]Center for Complex Network Research (CCNR), Department of Physics, Northeastern University, Boston, MA 02115, USA.
[14]Department of Computer Science and Engineering, University of Notre Dame, IN 46556, USA.
[15]Department of Computing, Imperial College London SW7 2AZ, UK.
[16]Cold Spring Harbor Laboratory (CSHL), Cold Spring Harbor, NY 11724, USA.
[17]United States Department of Agriculture, Agricultural Research Service (USDA ARS), Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, NY 14853,

USA.

[†]Present address: Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liège, 4000 Liège, Wallonia, Belgium.
[¤]Present address: Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S3E1, Canada and Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, Ontario M5G1X5, Canada
[‡]Present address: Department of Biology, University of Alabama at Birmingham, Birmingham, AL 35294, USA.
[§]Present address: University of Utrecht, 3508 TC Utrecht, The Netherlands.

ACKNOWLEDGEMENTS:

The authors wish to acknowledge Drs. Philip Benfey, Haiyuan Yu and Magnus Nordborg as well as members of the Center for Cancer Systems Biology (CCSB) for helpful discussions. This work was supported by grant NSF0703905 to M.V., J.R.E., D.E.H.; by grants NSF0520253, NSF0313578 to J.R.E.; by the Institute Sponsored Research funds from the Dana-Farber Cancer Institute Strategic Initiative to M.V. and CCSB; by grant R01 HG001715 to M.V., D.E.H. and F.P.R.; Canada Excellence Research Chairs (CERC) Program and Canadian Institute for Advanced Research Fellowship to F.P.R.; James S. McDonnell Foundation (JSMF) grant JSMF 220020084 – 21st Century Initiative in Studying Complex Systems to A.-L.B.; M.T. is supported by NIH NRSA Fellowship HG004098; D.M. and C.L. are supported by grant AGRONOMICS LSHG-CT-2006-037704 from the 6th Framework Program of the European Commission to C.L.; M.S.M. and J.L.D. were supported by NIH GM066025 to J.L.D.; R.J.S. is supported by NIH NRSA Fellowship F32-HG004830, J.S. and W.S. are supported by NSF grant 0703908; D.W. is supported by USDA 1907-21000-030; M.V. is a ''Chercheur Qualifié Honoraire'' from the Fonds de la Recherche Scientifique (FRS-FNRS, French Community of Belgium).

REFERENCES

1. National Research Council of the National Academies, *A New Biology for the 21st Century*. (The National Academies Press, Washington, D.C., 2009).
2. The Arabidopsis Genome Initiative, *Nature* **408**, 796 (2000).
3. M. Vidal, *FEBS Lett.* **579**, 1834 (2005).
4. A. J. Walhout *et al.*, *Methods Enzymol.* **328**, 575 (2000).
5. K. Yamada *et al.*, *Science* **302**, 842 (2003).
6. M. Dreze *et al.*, *Methods Enzymol.* **470**, 281 (2010).
7. N. Simonis *et al.*, *Nat. Methods* **6**, 47 (2009).
8. K. Venkatesan *et al.*, *Nat. Methods* **6**, 83 (2009).
9. H. Yu *et al.*, *Science* **322**, 104 (2008).
10. J. F. Rual *et al.*, *Nature* **437**, 1173 (2005).
11. P. Braun *et al.*, *Nat. Methods* **6**, 91 (2009).
12. N. Ramachandran *et al.*, *Nat. Methods* **5**, 535 (2008).
13. M. E. Cusick *et al.*, *Nat. Methods* **6**, 39 (2009).
14. A. L. Turinsky, S. Razick, B. Turner, I. M. Donaldson, S. J. Wodak, *Database (Oxford)* **2010**, baq026 (2010).
15. M. Vidal, *Cell* **104**, 333 (2001).
16. M. A. Pujana *et al.*, *Nat. Genet.* **39**, 1338 (2007).
17. H. Wang, N. Ngwenyama, Y. Liu, J. C. Walker, S. Zhang, *Plant Cell* **19**, 63 (2007).
18. R. D. Vierstra, *Nat. Rev. Mol. Cell Biol.* **10**, 385 (2009).

19. A. Santner, M. Estelle, *Nature* **459**, 1071 (2009).
20. L. Katsir, H. S. Chung, A. J. Koo, G. A. Howe, *Curr. Opin. Plant Biol.* **11**, 428 (2008).
21. B. Vanholme, W. Grunewald, A. Bateman, T. Kohchi, G. Gheysen, *Trends Plant Sci.* **12**, 239 (2007).
22. H. Szemenyei, M. Hannon, J. A. Long, *Science* **319**, 1384 (2008).
23. M. Kieffer *et al.*, *Plant Cell* **18**, 560 (2006).
24. L. Pauwels *et al.*, *Nature* **464**, 788 (2010).
25. S. Kagale, M. G. Links, K. Rozwadowski, *Plant Physiol.* **152**, 1109 (2010).
26. X. Hou, L. Y. Lee, K. Xia, Y. Yan, H. Yu, *Dev. Cell* **19**, 884 (2010).
27. S. Fortunato, *Phys. Rep.* **486**, 75 (2010).
28. Y. Y. Ahn, J. P. Bagrow, S. Lehmann, *Nature* **466**, 761 (2010).
29. C. J. Jeffery, *Mol. Biosyst.* **5**, 345 (2009).
30. D. Bridges, G. B. Moorhead, *Sci. STKE* **2005**, re10 (2005).
31. P. J. Schoonheim *et al.*, *Plant J.* **49**, 289 (2007).
32. P. J. Seo, S. G. Kim, C. M. Park, *Trends Plant Sci.* **13**, 550 (2008).
33. H. K. Yoon, S. G. Kim, S. Y. Kim, C. M. Park, *Mol. Cells* **25**, 438 (2008).
34. J. Kilian *et al.*, *Plant J.* **50**, 347 (2007).
35. D. Winter *et al.*, *PLoS One* **2**, e718 (2007).
36. Y. Yang, A. Costa, N. Leonhardt, R. Siegel, J. Schroeder, *Plant Methods* **4**, 6 (2008).
37. G. C. Conant, K. H. Wolfe, *Nat. Rev. Genet.* **9**, 938 (2008).
38. M. Lynch, J. S. Conery, *Science* **290**, 1151 (2000).
39. H. Innan, F. Kondrashov, *Nat. Rev. Genet.* **11**, 97 (2010).
40. D. R. Scannell, K. H. Wolfe, *Genome Res.* **18**, 137 (2008).
41. F. A. Kondrashov, I. B. Rogozin, Y. I. Wolf, E. V. Koonin, *Genome Biol.* **3**, (2002).
42. E. D. Levy, J. B. Pereira-Leal, *Curr. Opin. Struct. Biol.* **18**, 349 (2008).
43. S. Maslov, K. Sneppen, K. A. Eriksen, K. K. Yan, *BMC Evol. Biol.* **4**, 9 (2004).
44. A. Presser, M. B. Elowitz, M. Kellis, R. Kishony, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 950 (2008).
45. R. Pastor-Satorras, E. Smith, R. V. Sole, *J. Theor. Biol.* **222**, 199 (2003).
46. A. Vazquez, A. Flammini, A. Maritan, A. Vespignani, *ComPlexUs* **1**, 38 (2003).
47. K. Hanada, T. Kuromori, F. Myouga, T. Toyoda, K. Shinozaki, *PLoS Genet.* **5**, e1000781 (2009).
48. M. Freeling, *Annu. Rev. Plant Biol.* **60**, 433 (2009).
49. A. Wagner, *Mol. Biol. Evol.* **19**, 1760 (2002).
50. T. Casneuf, S. De Bodt, J. Raes, S. Maere, Y. Van de Peer, *Genome Biol.* **7**, R13 (2006).
51. W. Y. Chung, R. Albert, I. Albert, A. Nekrutenko, K. D. Makova, *BMC Bioinformatics* **7**, 46 (2006).
52. B. Vandersluis *et al.*, *Mol. Syst. Biol.* **6**, 429 (2010).
53. A. E. Mayo, Y. Setty, S. Shavit, A. Zaslaver, U. Alon, *PLoS Biol.* **4**, e45 (2006).
54. A. L. Barabasi, Z. N. Oltvai, *Nat. Rev. Genet.* **5**, 101 (2004).
55. Q. Zhong *et al.*, *Mol. Syst. Biol.* **5**, 321 (2009).
56. I. W. Taylor *et al.*, *Nat. Biotechnol.* **27**, 199 (2009).
57. E. J. Rossin *et al.*, *PLoS Genet.* **7**, e1001273 (2011).
58. T. Milenkovic, V. Memisevic, A. K. Ganesan, N. Przulj, *J. R. Soc. Interface* **7**, 423 (2010).
59. S. Kerrien *et al.*, *Nucleic Acids Res.* **35**, D561 (2007).
60. H. S. Chung, G. A. Howe, *Plant Cell* **21**, 131 (2009).

Figure 1A

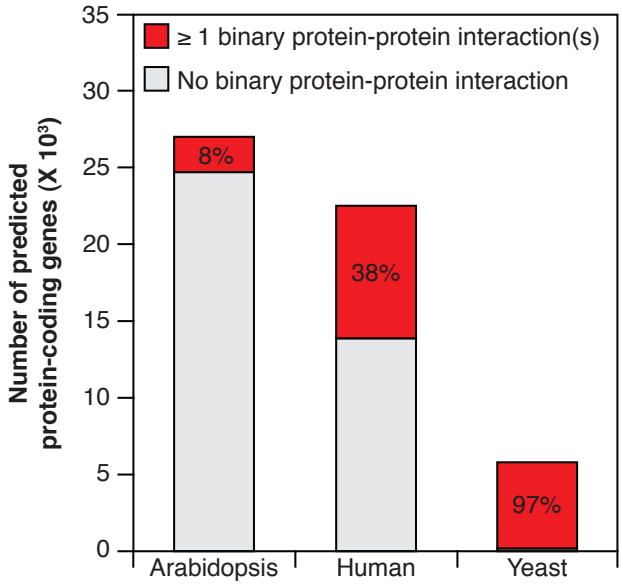


Figure 1B

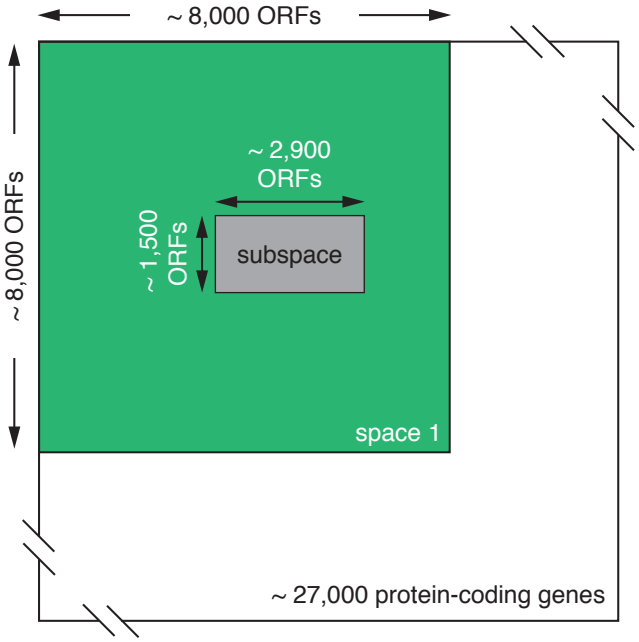


Figure 1C

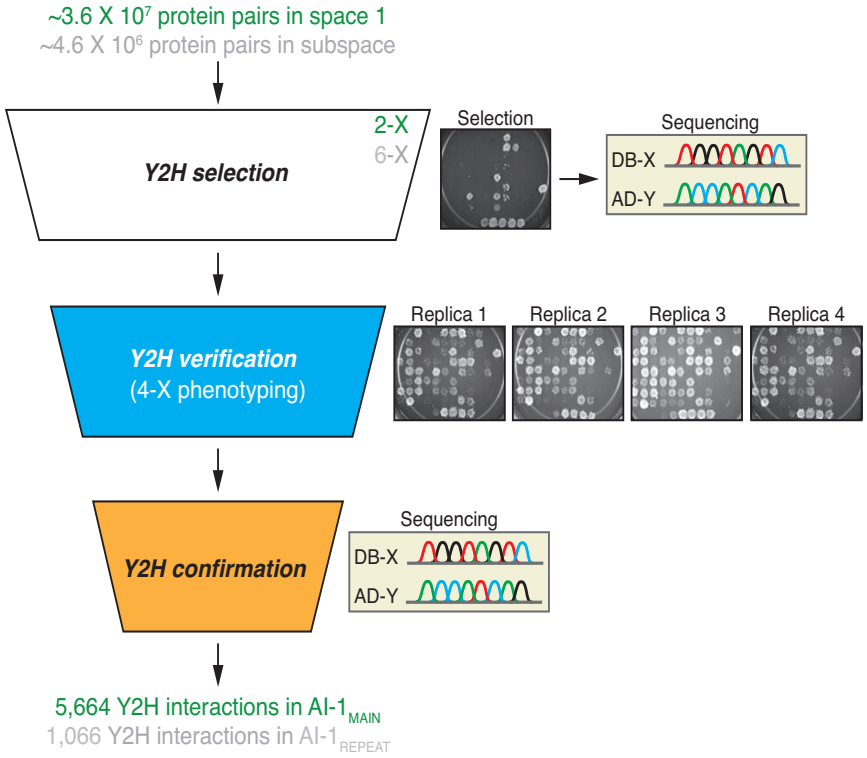
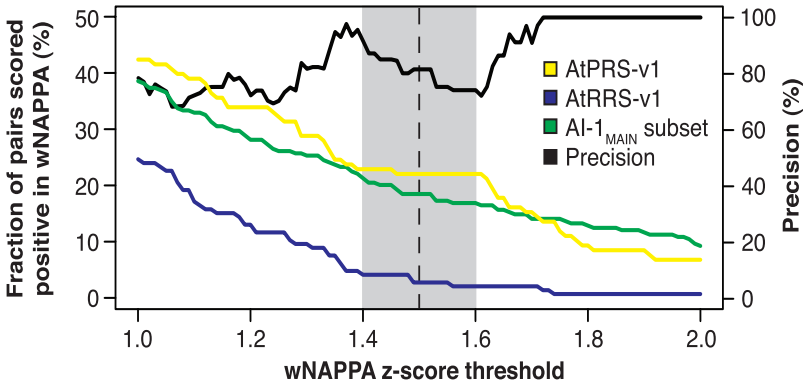

~3.6 X 10^7 protein pairs in space 1
~4.6 X 10^6 protein pairs in subspace
Y2H selection
2-X
6-X
Selection
Sequencing
DB-X
AD-Y
Y2H verification
(4-X phenotyping)
Replica 1
Replica 2
Replica 3
Replica 4
Y2H confirmation
Sequencing
DB-X
AD-Y
5,664 Y2H interactions in AI-1 MAIN
1,066 Y2H interactions in AI-1 REPEAT

Figure 1D



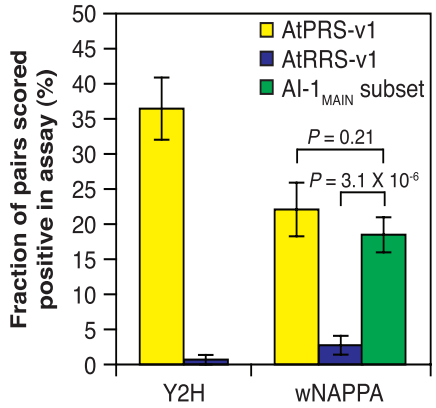


Figure 1E

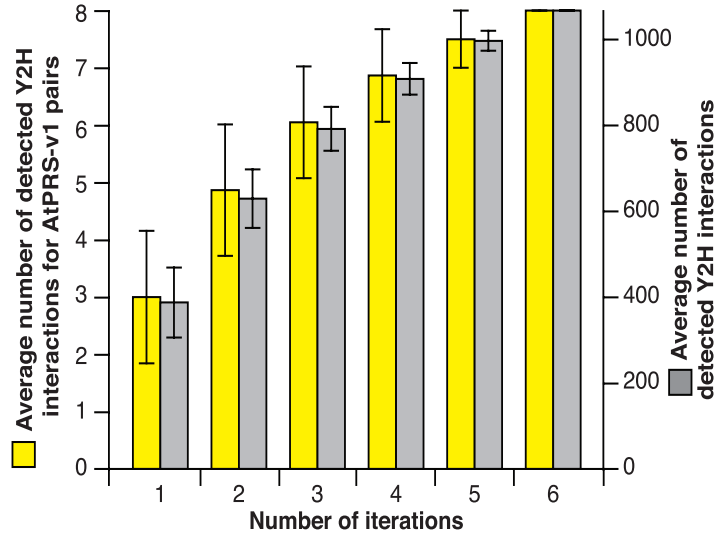


Figure 1F

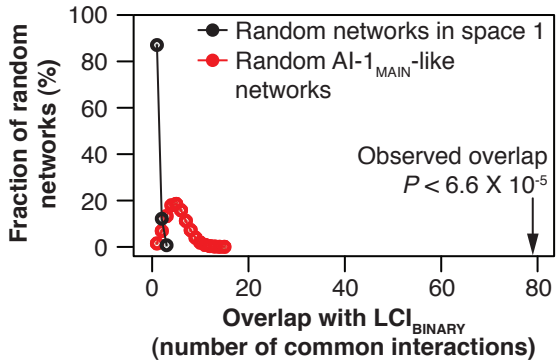


Figure 1G

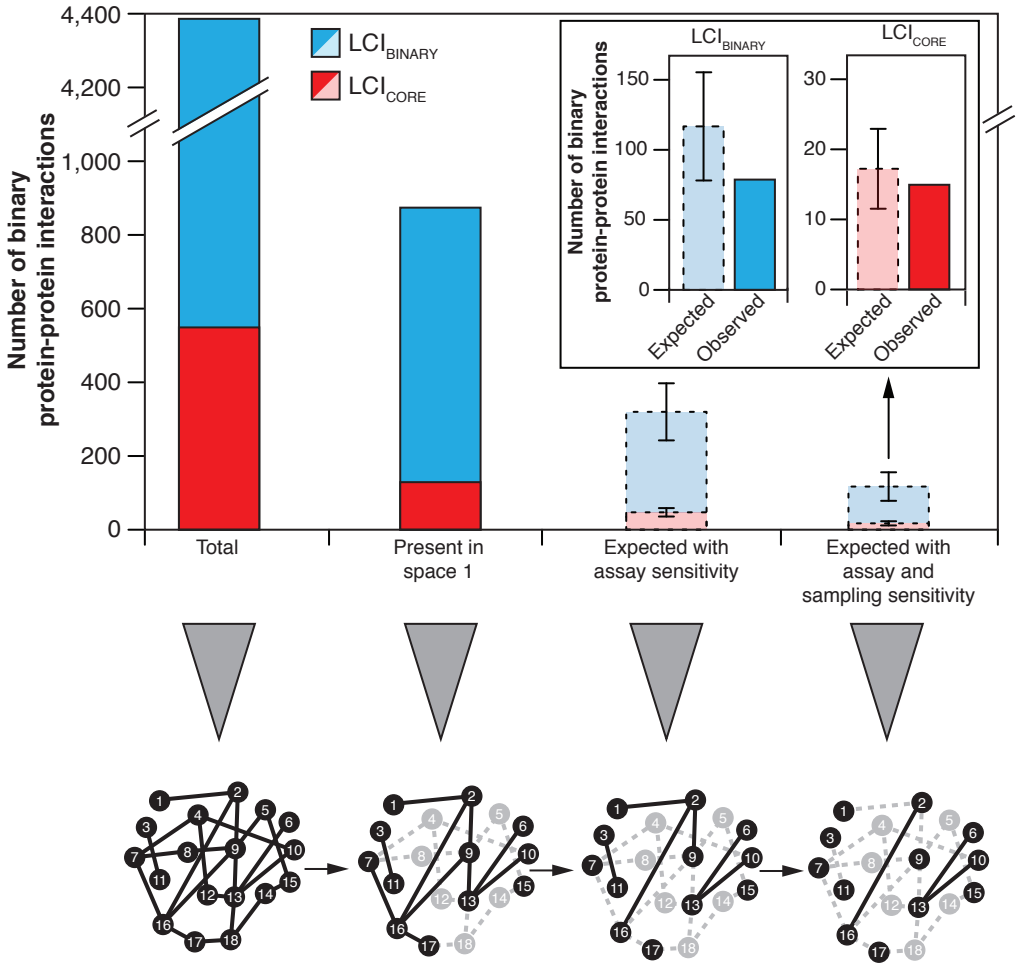


Figure 2A

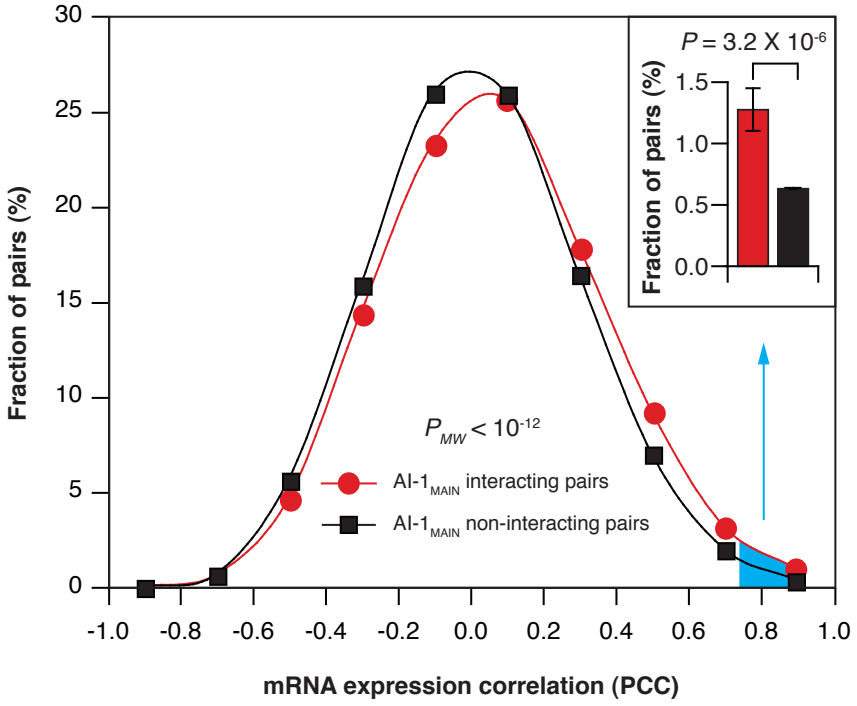


Figure 2B

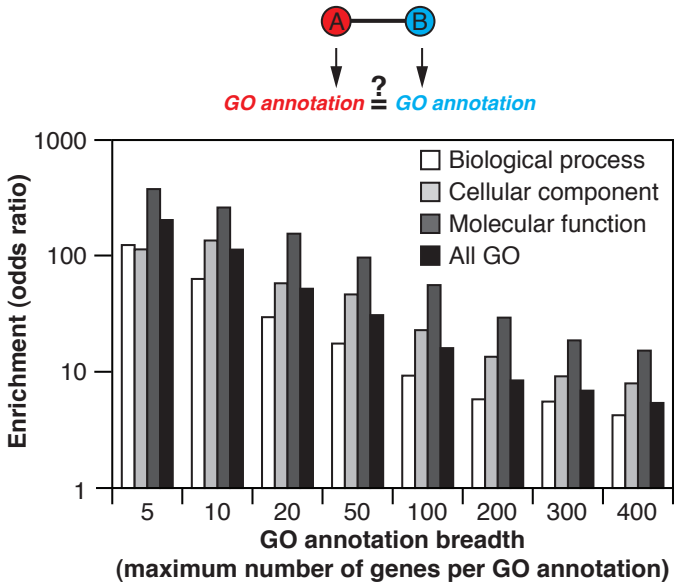


Figure 2C

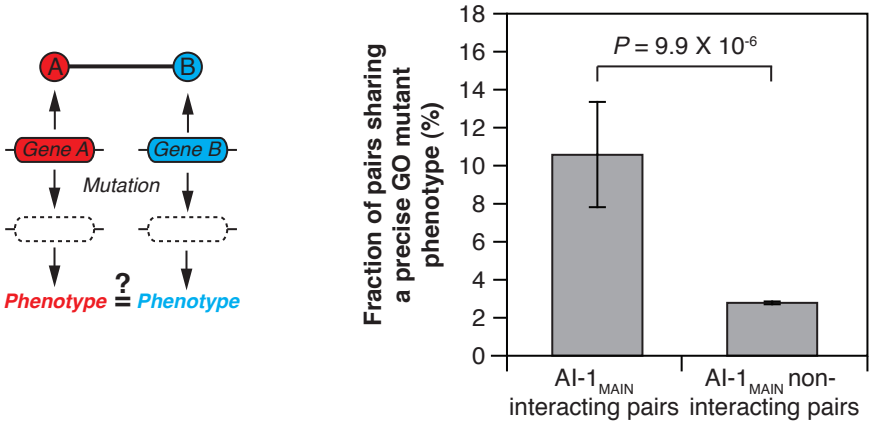


Figure 2D

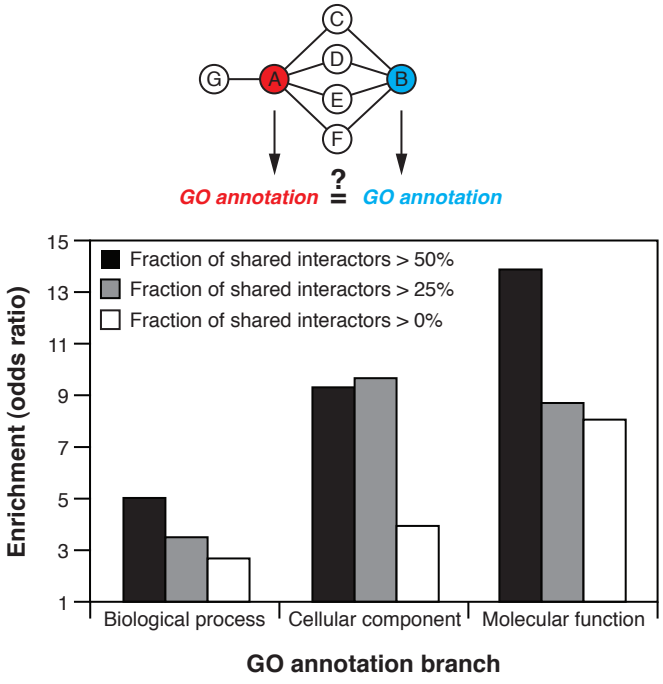
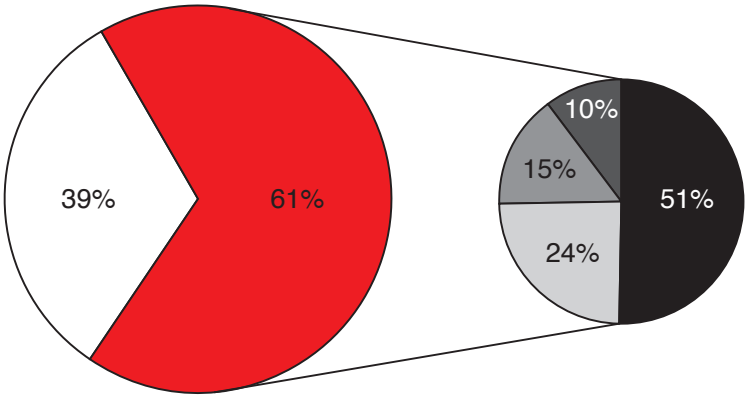


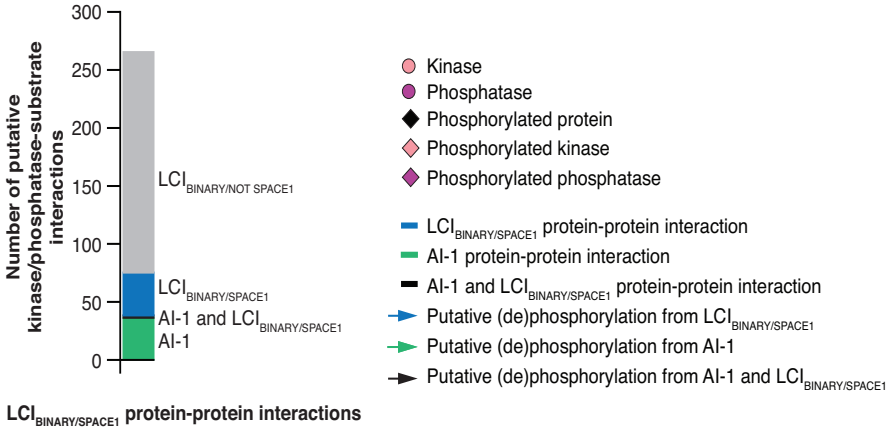
Figure 2E



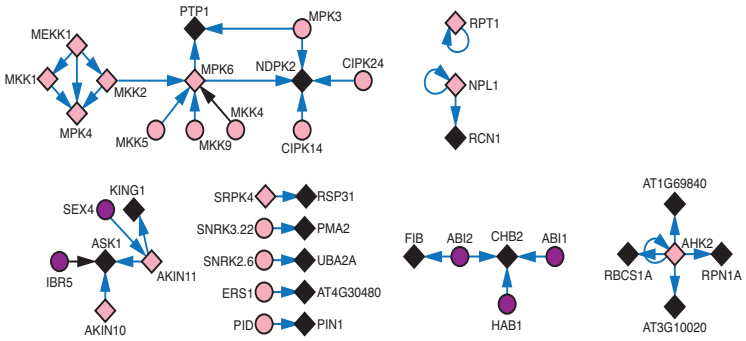
Protein with precise GO annotation (describing ≤ 50 proteins)
Protein with broad (describing > 50 proteins) or no GO annotation

Interacts with protein with precise GO annotation (Fig. 2B, 2C)
Shares ≥ 50% of interactors with protein with precise GO annotation (Fig. 2D)
Both
Neither

Figure 3A



LCI$_{BINARY/SPACE1}$ protein-protein interactions



AI-1 or LCI$_{BINARY/SPACE1}$ protein-protein interactions

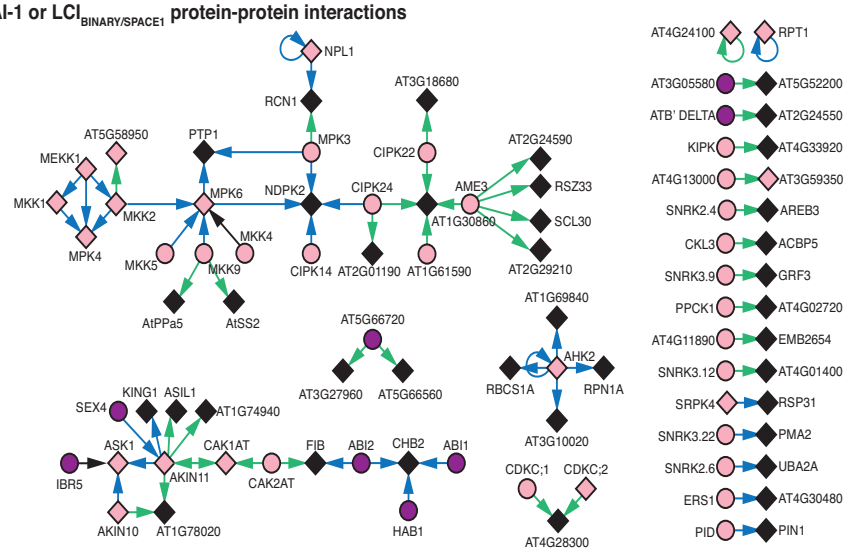
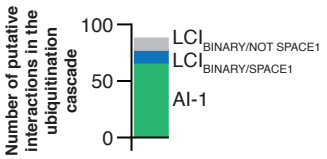
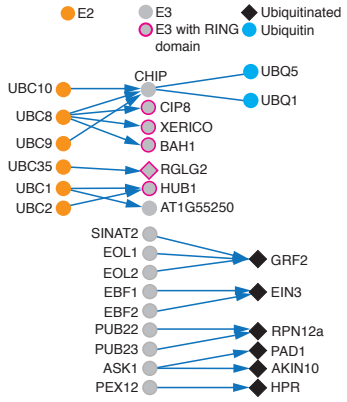


Figure 3B



- LCI_{BINARY} protein-protein interaction
- AI-1 protein-protein interaction
- LCI_{BINARY} interaction directed in the context of the ubiquitination cascade
- AI-1 interaction directed in the context of the ubiquitination cascade

LCI_{BINARY} protein-protein interactions



AI-1 or LCI_{BINARY} protein-protein interactions

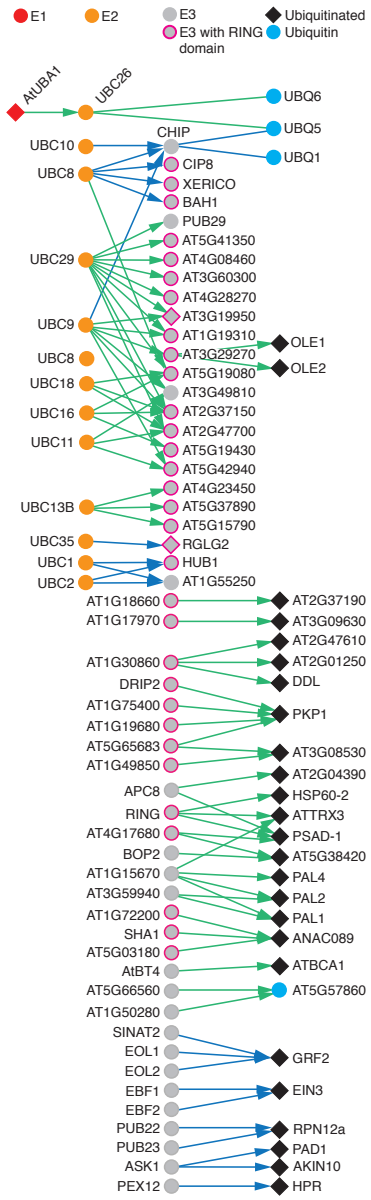


Figure 3C

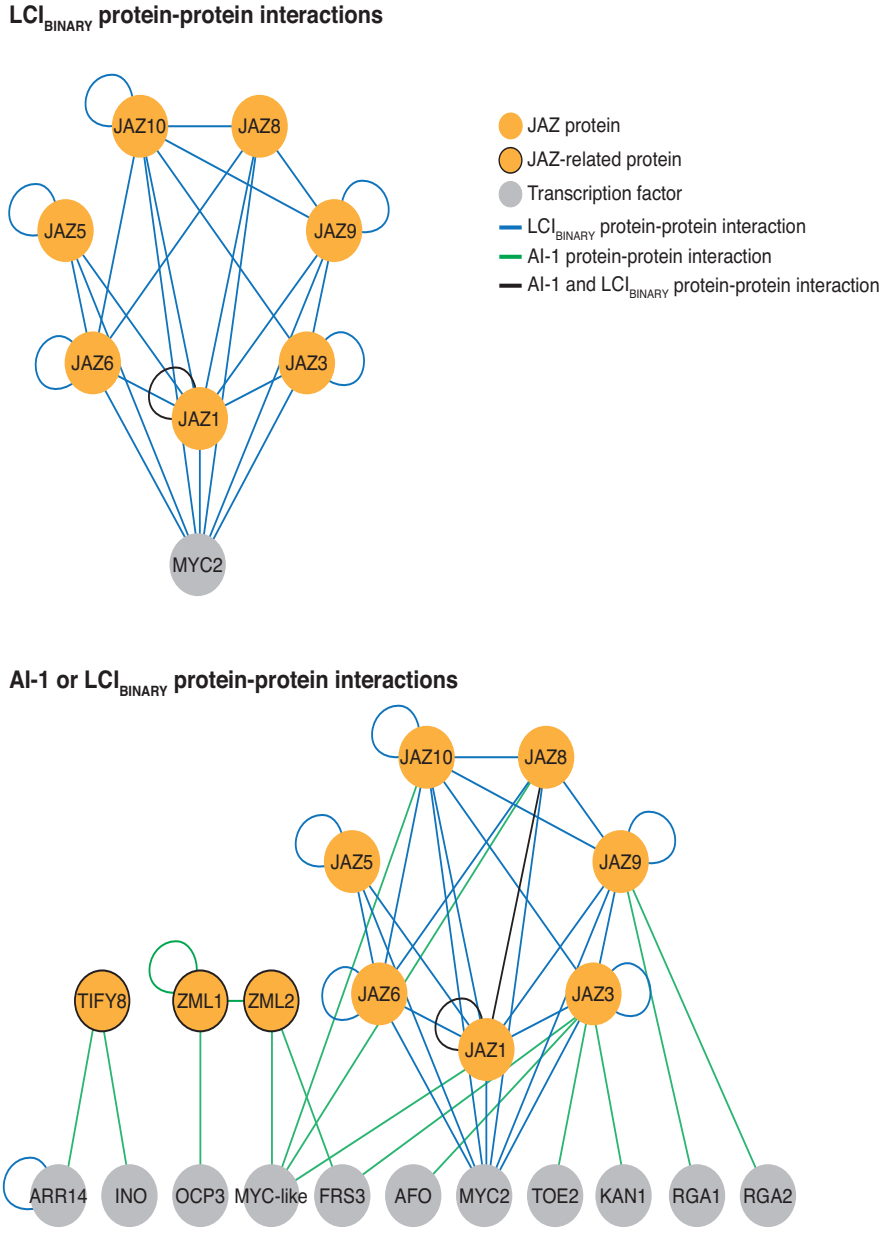

LCI$_{BINARY}$ protein-protein interactions
JAZ10
JAZ8
JAZ5
JAZ9
JAZ6
JAZ3
JAZ1
MYC2
JAZ protein
JAZ-related protein
Transcription factor
LCI$_{BINARY}$ protein-protein interaction
AI-1 protein-protein interaction
AI-1 and LCI$_{BINARY}$ protein-protein interaction
AI-1 or LCI$_{BINARY}$ protein-protein interactions
JAZ10
JAZ8
JAZ5
JAZ9
TIFY8
ZML1
ZML2
JAZ6
JAZ3
JAZ1
ARR14
INO
OCP3
MYC-like
FRS3
AFO
MYC2
TOE2
KAN1
RGA1
RGA2

Figure 3D

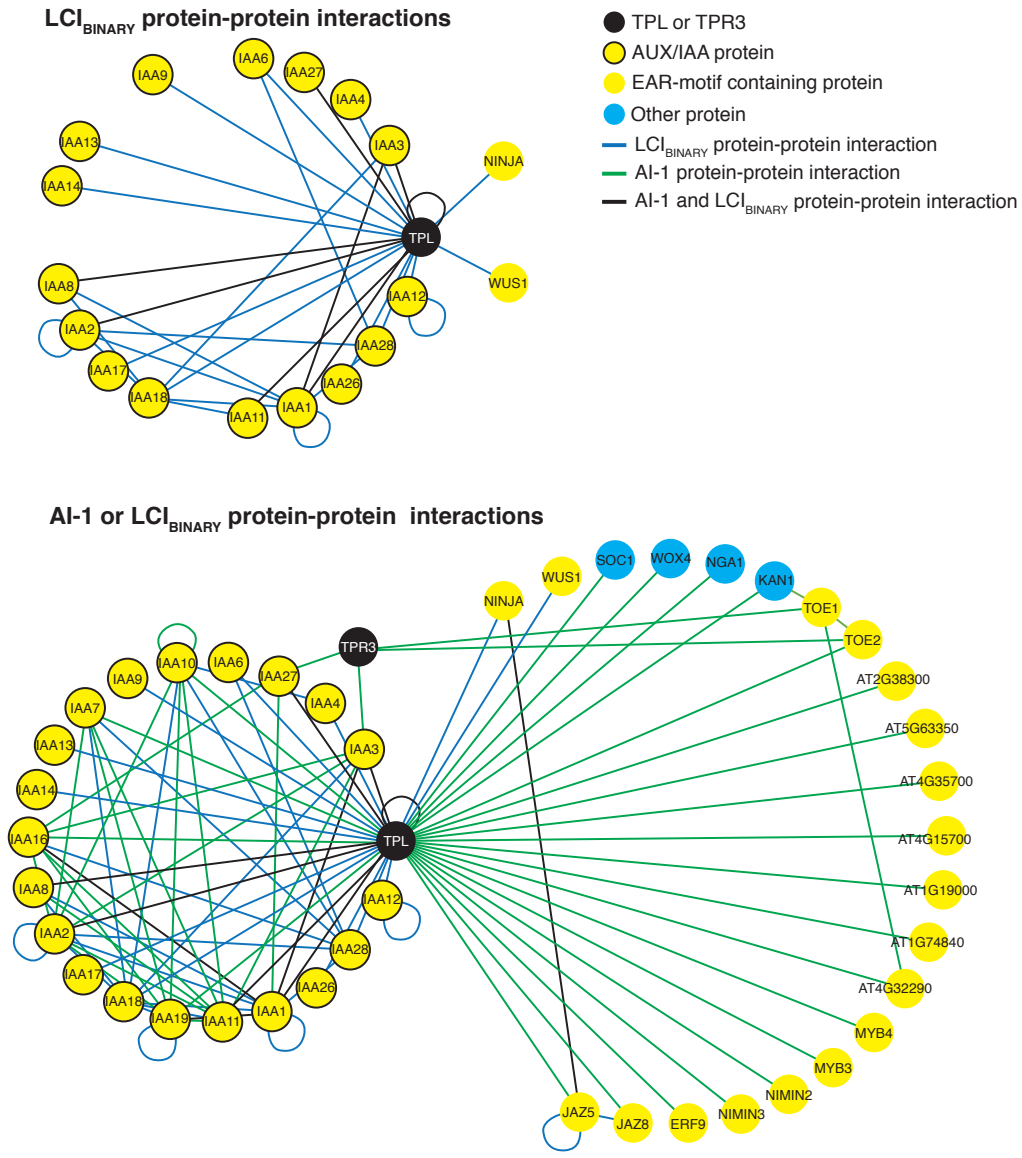

LCI BINARY protein-protein interactions
TPL or TPR3
AUX/IAA protein
EAR-motif containing protein
Other protein
LCI BINARY protein-protein interaction
AI-1 protein-protein interaction
AI-1 and LCI BINARY protein-protein interaction
IAA9
IAA6
IAA27
IAA4
IAA13
IAA3
NINJA
IAA14
TPL
IAA8
WUS1
IAA12
IAA2
IAA28
IAA17
IAA26
IAA18
IAA1
IAA11
AI-1 or LCI BINARY protein-protein interactions
WUS1
SOC1
WOX4
NGA1
KAN1
NINJA
TOE1
TOE2
TPR3
IAA10
IAA6
IAA9
IAA27
AT2G38300
IAA7
IAA4
AT5G63350
IAA13
IAA3
AT4G35700
IAA14
AT4G15700
IAA16
TPL
AT1G19000
IAA8
IAA12
AT1G74840
IAA2
IAA28
AT4G32290
IAA17
IAA26
MYB4
IAA18
IAA19
IAA1
IAA11
MYB3
NIMIN2
JAZ5
JAZ8
ERF9
NIMIN3

Figure 3E

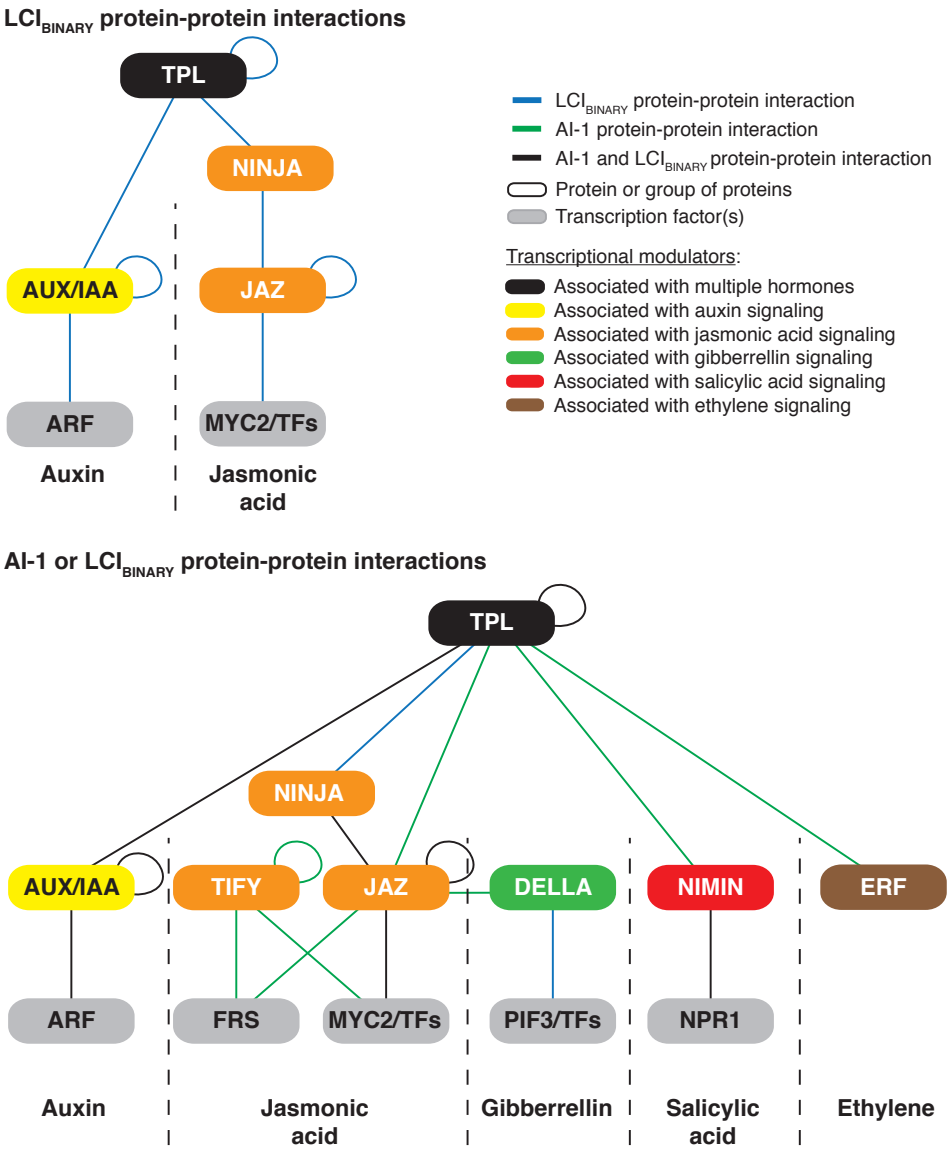

LCI$_{BINARY}$ protein-protein interactions
TPL
NINJA
AUX/IAA
JAZ
ARF
MYC2/TFs
Auxin
Jasmonic acid
LCI$_{BINARY}$ protein-protein interaction
AI-1 protein-protein interaction
AI-1 and LCI$_{BINARY}$ protein-protein interaction
Protein or group of proteins
Transcription factor(s)
Transcriptional modulators:
Associated with multiple hormones
Associated with auxin signaling
Associated with jasmonic acid signaling
Associated with gibberellin signaling
Associated with salicylic acid signaling
Associated with ethylene signaling
AI-1 or LCI$_{BINARY}$ protein-protein interactions
TPL
NINJA
AUX/IAA
TIFY
JAZ
DELLA
NIMIN
ERF
ARF
FRS
MYC2/TFs
PIF3/TFs
NPR1
Auxin
Jasmonic acid
Gibberellin
Salicylic acid
Ethylene

Table 1

Protein 1	Protein 2	No. of interactions
Arm (Multiple functions)	Any protein	145
Mov34 (Multiple functions)	Any protein	167
Myb (Transcription)	Any protein	171
TPR (Multiple function)	Any protein	175
zf-C3HC4 (Signaling and protein stability)	Any protein	193
Pkinase (Signaling)	Any protein	198
TCP (Transcription)	Any protein	257
MIP (Transport)	Any protein	248
RRM (RNA processing)	Any protein	258
NAM (Transcription)	Any protein	407
DUF-containing	Any protein	727
Plant-specific	Any protein	2,212
Ortholog in rice	Ortholog in rice	1,959
Ortholog in maize	Ortholog in maize	1,439
Ortholog in tomato	Ortholog in tomato	2,158
Ortholog in cyanobacteria	Ortholog in cyanobacteria	717
Co-evolving with protein 2	Co-evolving with protein 1	231
Kinases	Phosphorylated protein	33
Phosphatases	Phosphorylated protein	5
E2	E3	29
E3	Ubiquitinated protein	32
Transcription factor	Transcription factor	175
Predicted interactor 1	Predicted interactor 2	408

Figure 4A

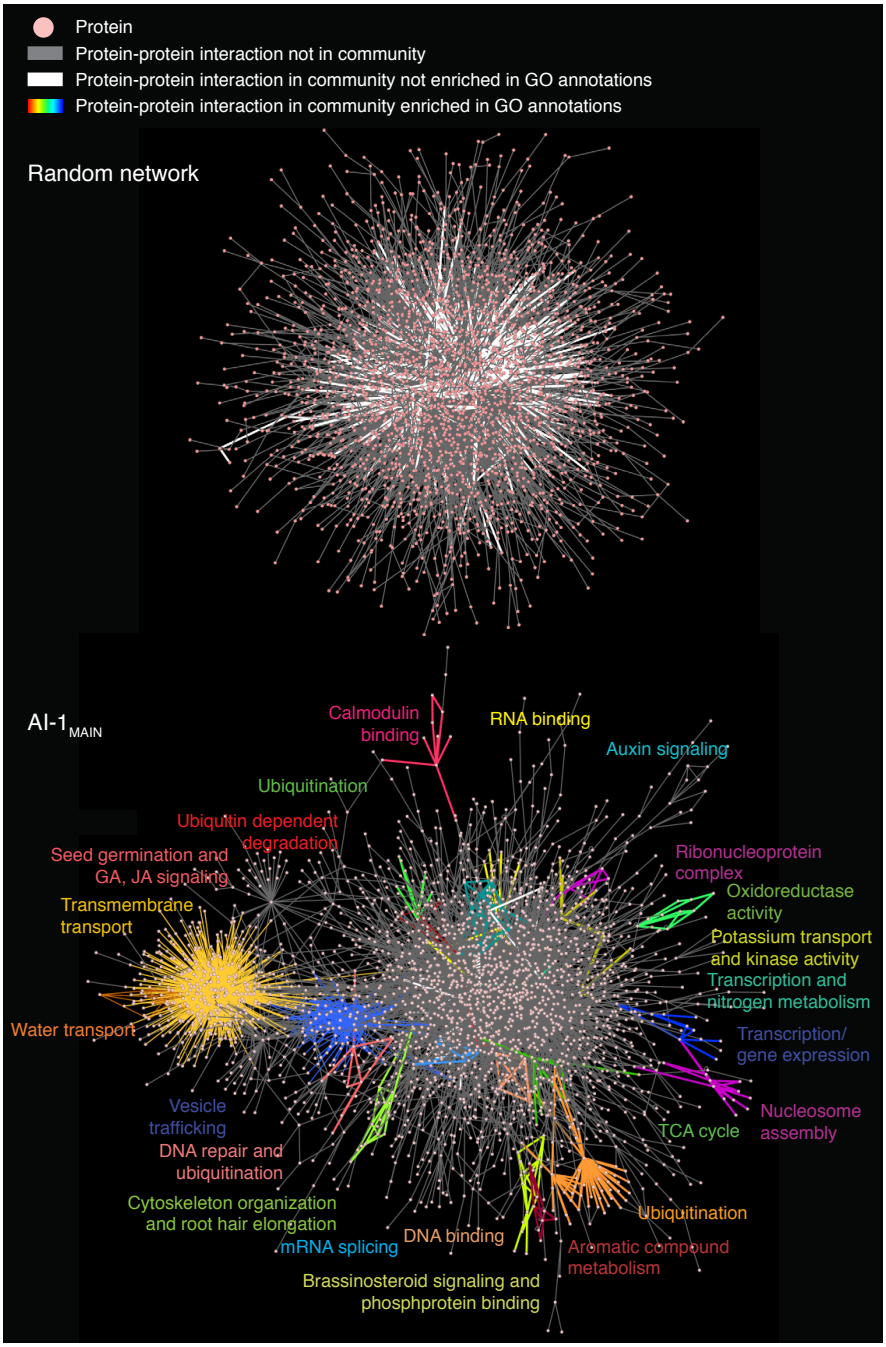

Protein
Protein-protein interaction not in community
Protein-protein interaction in community not enriched in GO annotations
Protein-protein interaction in community enriched in GO annotations
Random network
AI-1MAIN
Calmodulin binding
RNA binding
Auxin signaling
Ubiquitination
Ubiquitin dependent degradation
Seed germination and GA, JA signaling
Ribonucleoprotein complex
Oxidoreductase activity
Transmembrane transport
Potassium transport and kinase activity
Transcription and nitrogen metabolism
Water transport
Transcription/ gene expression
Vesicle trafficking
Nucleosome assembly
TCA cycle
DNA repair and ubiquitination
Cytoskeleton organization and root hair elongation
Ubiquitination
DNA binding
mRNA splicing
Aromatic compound metabolism
Brassinosteroid signaling and phosphprotein binding

Figure 4B

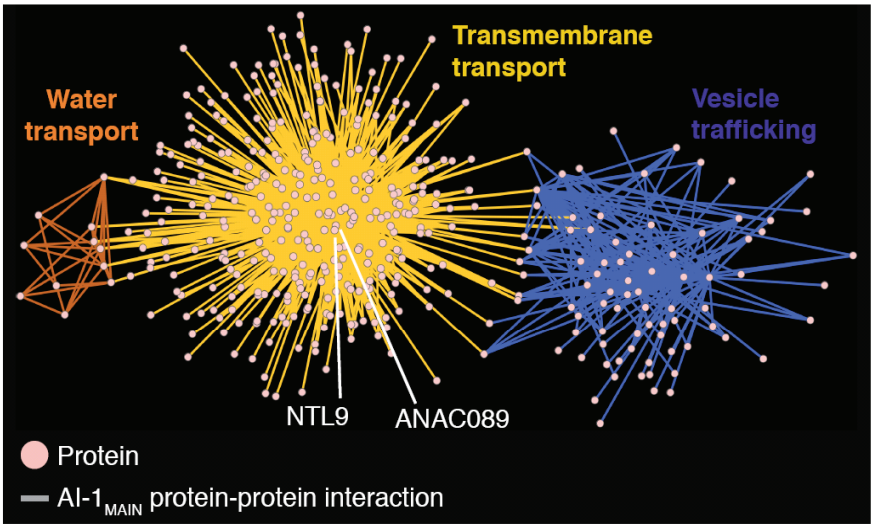


Figure 5A

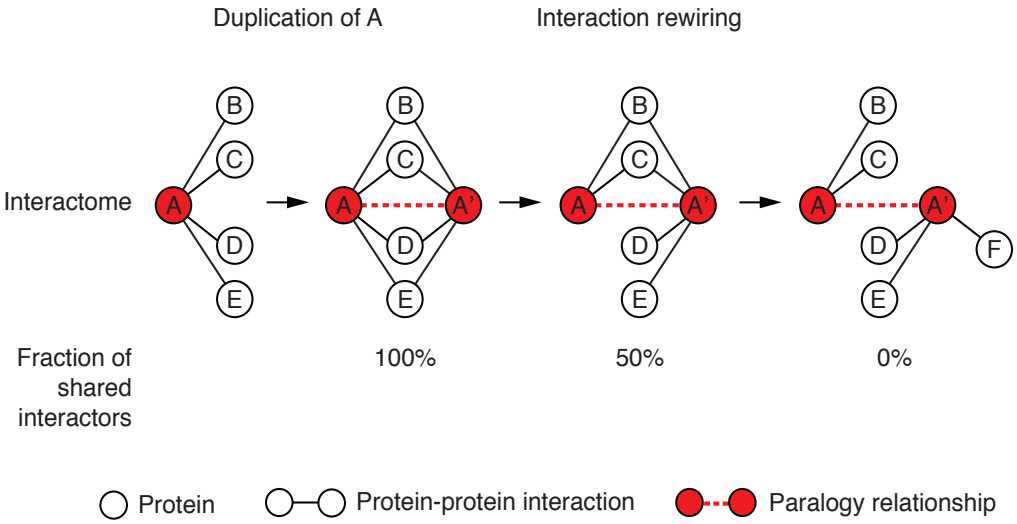
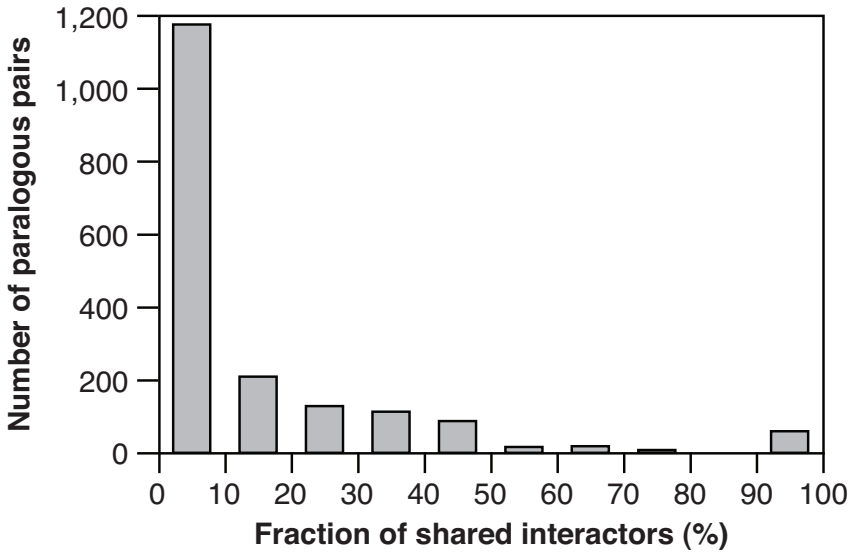


Figure 5B


Number of paralogous pairs
1,200
1,000
800
600
400
200
0
0
10
20
30
40
50
60
70
80
90
100
Fraction of shared interactors (%)

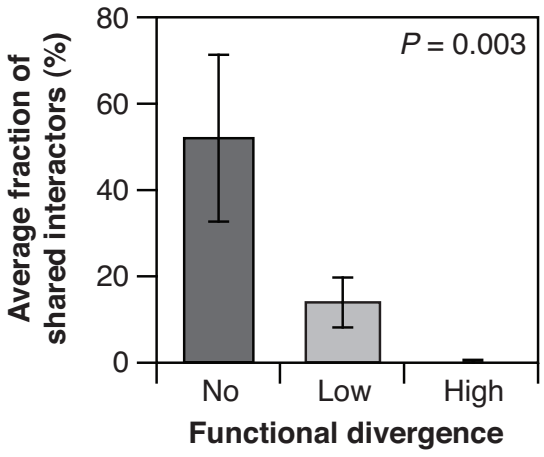

Average fraction of shared interactors (%)
80
60
40
20
0
P = 0.003
No
Low
High
Functional divergence

Figure 5C

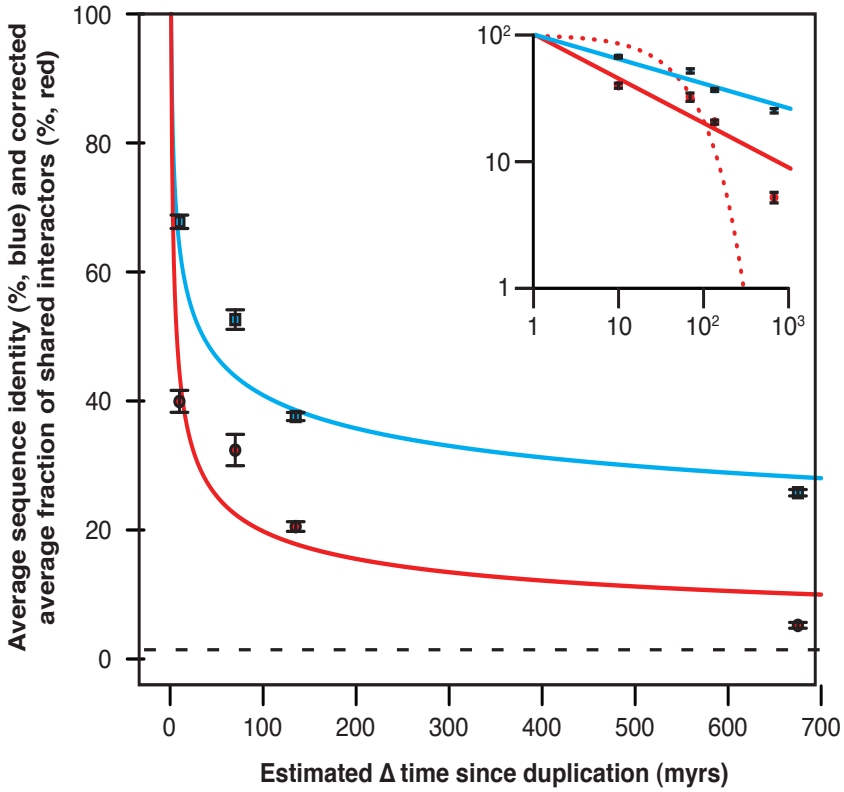


Figure 5D

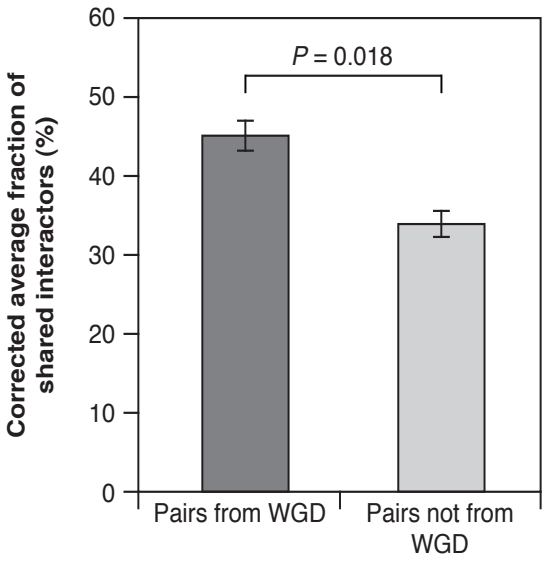


Figure 5E

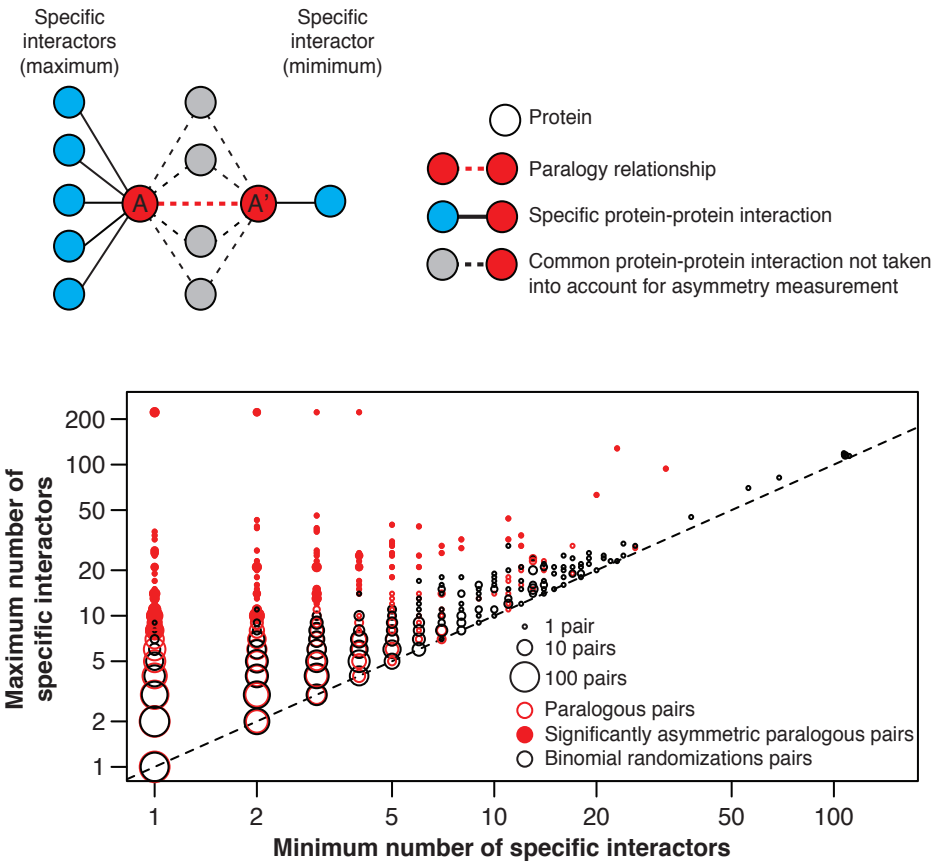


Figure 5F

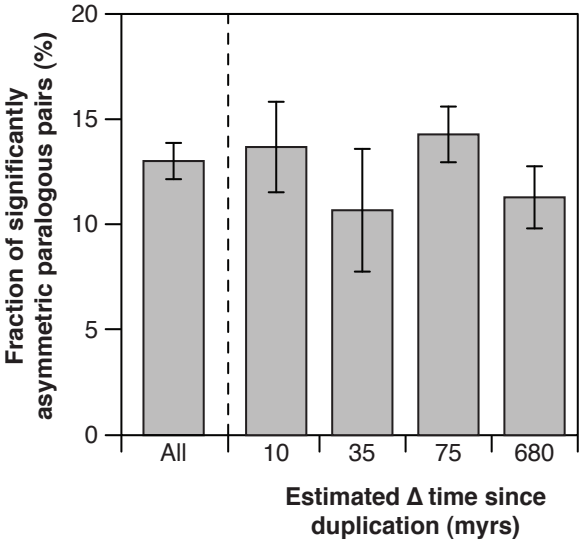
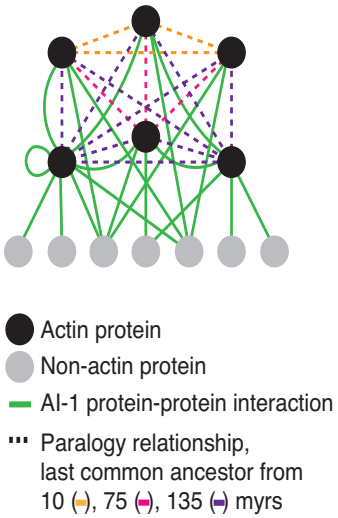
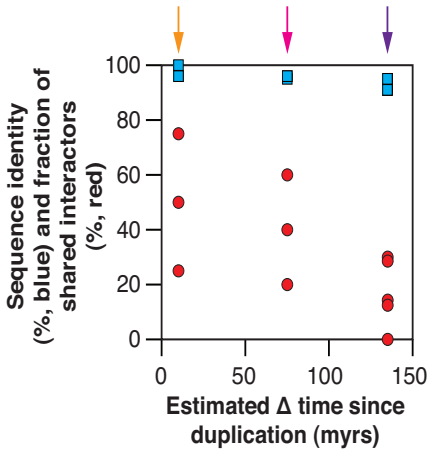


Figure 5G


Actin protein
Non-actin protein
AI-1 protein-protein interaction
Paralogy relationship, last common ancestor from 10 (), 75 (), 135 () myrs


Sequence identity (%, blue) and fraction of shared interactors (%, red)
100
80
60
40
20
0
0
50
100
150
Estimated Δ time since duplication (myrs)

DOCUMENT JOINT 7

Titre : Genome-scale analysis of *in vivo* spatiotemporal promoter activity in *Caenorhabditis elegans*.

Auteurs : Dupuy D*, Bertin N*, Hidalgo CA*, Venkatesan K, Tu D, Lee D, Rosenberg J, Svrzikapa N, Blanc A, Carnec A, Carvunis AR, Pulak R, Shingles J, Reece-Hoyes J, Hunt-Newbury R, Viveiros R, Mohler WA, Tasan M, Roth FP, Le Peuch C, Hope IA, Johnsen R, Moerman DG, Barabási AL, Baillie D, Vidal M.

Description : article de recherche original paru dans le magazine *Nature Biotechnology* en 2007

Contribution : D Dupuy, N Bertin et CA Hidalgo ont dirigé ce projet. J'ai apporté un soutien bioinformatique à l'analyse fonctionnelle de groupes de gènes par D Dupuy, sous la supervision de K Venkatesan. M Vidal a supervisé le projet.

nature biotechnology

Genome-scale analysis of *in vivo* spatiotemporal promoter activity in *Caenorhabditis elegans*

Denis Dupuy[1,10], Nicolas Bertin[1,2,10], César A Hidalgo[1,3,10], Kavitha Venkatesan[1], Domena Tu[4], David Lee[4], Jennifer Rosenberg[1], Nenad Svrzikapa[1], Aurélie Blanc[1], Alain Carnec[1], Anne-Ruxandra Carvunis[1], Rock Pulak[5], Jane Shingles[6], John Reece-Hoyes[6], Rebecca Hunt-Newbury[7], Ryan Viveiros[7], William A Mohler[8], Murat Tasan[9], Frederick P Roth[9], Christian Le Peuch[2], Ian A Hope[6], Robert Johnsen[4], Donald G Moerman[7], Albert-László Barabási[1,3], David Baillie[4] & Marc Vidal[1]

Differential regulation of gene expression is essential for cell fate specification in metazoans. Characterizing the transcriptional activity of gene promoters, in time and in space, is therefore a critical step toward understanding complex biological systems. Here we present an *in vivo* spatiotemporal analysis for ~900 predicted *C. elegans* promoters (~5% of the predicted protein-coding genes), each driving the expression of green fluorescent protein (GFP). Using a flow-cytometer adapted for nematode profiling, we generated 'chronograms', two-dimensional representations of fluorescence intensity along the body axis and throughout development from early larvae to adults. Automated comparison and clustering of the obtained *in vivo* expression patterns show that genes coexpressed in space and time tend to belong to common functional categories. Moreover, integration of this data set with *C. elegans* protein-protein interactome data sets enables prediction of anatomical and temporal interaction territories between protein partners.

During development, cell type determination depends on the activation and/or repression of specific subsets of genes[1,2]. One of the requirements to fully understand the molecular networks driving cell differentiation in metazoans is to characterize the *in vivo* expression state of the genome at each differentiation step, that is, to determine what specific set of genes is activated in which specific cell type at what stage of development[3].

C. elegans is a unique multicellular model for such *in vivo* global gene expression studies. Its transparent body and nearly invariant cell lineage[4–6] enable precise, cell-by-cell analysis of promoter activity throughout development in transgenic animals carrying *promoter::GFP* reporter constructs[6–8]. With the completion of the *C. elegans* genome sequence[9], it is conceivable to develop genome-wide analysis of *in vivo* promoter activity for the complete set of promoters. The collection of all obtained patterns, referred to here as a 'localizome' data set, should provide important insight into the functional organization of gene regulation at the genome scale (refs. 10–13, and http://elegans.bcgsc.ca/perl/eprofile/index).

High-magnification fluorescence microscopy has been used to characterize expression at the single-cell level in *C. elegans*. However, this kind of analysis is extremely time consuming and labor intensive[13]. Moreover, because rapid microscopic examination can be performed on only a limited number of animals, the obtained results can be strongly affected by stochastic variations among individuals.

RESULTS

Reconstitution of time-lapse expression patterns

To study *promoter::GFP* expression at the level of a large population of animals with a quantitative read-out, we used a 'complex object parametric analysis and sorter' (COPAS) instrument equipped with a profiler system that analyzes up to ~100 animals/s. This system generates fluorescent emission profiles along the antero-posterior axis of the *C. elegans* body. By analyzing large numbers of animals of all sizes and ages at high throughput, we generated a digitized overview of the promoter activity throughout post-embryonic development.

For each transgenic line analyzed, fluorescence profiles were acquired for thousands of nematodes from a mixed-stage culture. For each worm in the population we then converted the corresponding profile into a color-coded representation of the fluorescence

[1]Center for Cancer Systems Biology (CCSB), and Department of Cancer Biology, Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, 44 Binney Street, Boston, Massachusetts 02115, USA. [2]Centre de Recherche en Biochimie Macromoléculaire, Centre National de la Recherche Scientifique FRE 2593, 1919 Route de Mende, 34293 Montpellier Cedex 5, France. [3]Center for Complex Network Research, Department of Physics, University of Notre Dame, 225 Nieuwland Science Hall, Notre Dame, Indiana 46556, USA. [4]Department of Molecular Biology and Biochemistry, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia V5A 1S6, Canada. [5]Union Biometrica, 84 October Hill Road, Holliston, Massachusetts 01746, USA. [6]Institute of Integrative and Comparative Biology, University of Leeds, Clarendon Way, Leeds LS2 9JT, West Yorkshire, UK. [7]Department of Zoology, The University of British Columbia, 6270 University Boulevard, Vancouver, British Columbia V6T 1Z4, Canada. [8]Department of Genetics and Developmental Biology and Center for Cell Analysis and Modeling, University of Connecticut Health Center, 263 Farmington Avenue, Farmington, Connecticut 06030, USA. [9]Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, and Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to M.V. (marc_vidal@dfci.harvard.edu), D.B. (baillie@sfu.ca) or A.-L.B. (alb@nd.edu).

Received 14 December 2006; accepted 13 April 2007; published online 7 May 2007; doi:10.1038/nbt1305

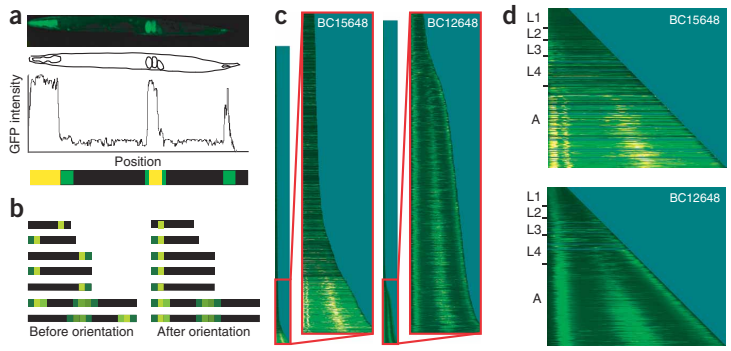


Figure 1 Generating post-embryonic developmental chronograms. (**a**) For each worm, the longitudinal GFP intensity profile is converted into a bar where the length corresponds to the animal's size and the color codes for the fluorescence intensity. (**b**) Profiles are sorted by size and oriented to match their neighbors. (**c**) The complete collection of oriented expression profiles reflects the size distribution of a given population. Represented here are the profiles gathered for the strains BC15648 and BC12648, expressing GFP under the control of *F25B5.1* and *ptl-1* promoters, respectively. The expansions on the right display the profiles of larger animals. (**d**) Averaging profiles of identical size produce chronograms with a standardized shape. Approximate larval stages and adult transitions are indicated on the *y*-axis. The color code represents the absolute GFP intensity measured (increasing values as black-green-yellow-white).

intensity (**Fig. 1a**). After orienting these profiles (**Fig. 1b**), we assembled them so that the short rows at the top represent L1 larvae, whereas the bottom rows correspond to fully-grown adults (**Fig. 1c**). As the distribution of worm sizes varies dramatically between analyzed populations (**Fig. 1c**), we generated images in which each row represents the average of all worms of a given size (**Fig. 1d**). If no animal of a given size was found, the corresponding row was skipped. As these normalized images provide an overview of GFP expression in time throughout post-embryonic development (the length of the worm being a proxy for its age), we refer to them as 'chronograms'.

Several examples of chronograms corresponding to well-characterized GFP-expressing strains are shown in **Figure 2**. Tissue-specific signatures corresponding to major organs are clearly identifiable on the chronograms, even when expression is restricted to a small number of cells, such as olfactory neurons (**Fig. 2b**). An advantage of the chronogram is that it provides information on the temporal expression of the reporter. For example, the onset of *egl-15* expression during the L4 stage is clearly visible in **Figure 2f**.

We acquired chronograms for 1,992 GFP strains which, after accounting for redundancy in transgene content, report the activity of the proximal promoter of 1,610 unique predicted gene loci (**Supplementary Fig. 1** online)[13,14]. We further analyzed 876 chronograms for which the average signal was above background (see "Chronogram signal level classification" in **Supplementary Methods** online). Several factors explain the absence of significant signal in the other 734 chronograms. Most of the strains (79.2%) analyzed carry extrachromosomal arrays of the *promoter::GFP* construct, which may have a transmission rate too low to enable population-scale analysis of expression[15]. Moreover, the presence of extrachromosomal arrays may sometimes cause a deleterious phenotype because, for example, of transcription factor titration effects[16] and may thus be counter-selected. Alternatively, some of the promoters may drive expression at levels below the detection threshold of the instrument under the conditions used here.

Extraction of tissue-specific signatures

We developed an averaging method to extract the characteristic features of sets of chronograms, corresponding either to strains that share individual anatomic annotations defined by microscopic observation[13], or to genes that belong to expression clusters derived from microarray experiments[17]. To prevent chronograms with strong expression from dominating weaker ones, we divided each image by its mean intensity. The average chronogram of all 315 strains annotated as expressed in the pharynx (**Fig. 3a**) displayed a distinctive anterior expression pattern, with a clearly visible separation between the two pharyngeal bulbs, a feature already seen in the *myo-2::GFP* chronogram (**Fig. 2e**), representing the expression of a promoter specific to pharyngeal muscle cells. Furthermore, a late-onset mid-body signal can be seen, corresponding to GFP expression in embryos *in utero*, which is consistent with the timing of pharynx development. The average image corresponding to the 150 strains that scored positive for expression in the nerve ring (**Fig. 3b**) shows a pattern clearly distinct from the pharyngeal one (**Fig. 3a**), with a narrower signal starting at the level of the isthmus of the pharynx. This average nerve ring pattern replicates the signature observed in the chronograms of *unc-119::GFP* (**Fig. 2a**) and *lim-4::GFP* (**Fig. 2b**), both representing expression in the nerve ring.

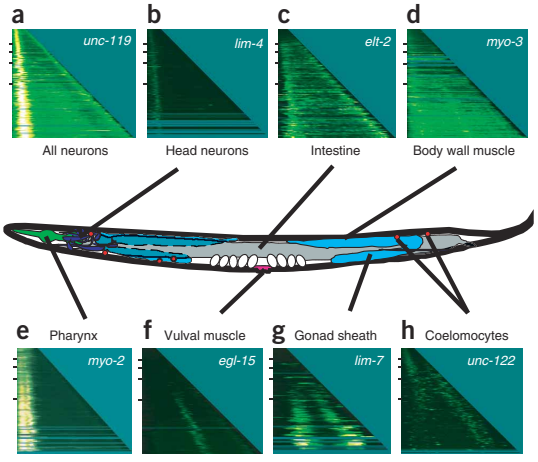


Figure 2 Visualization of tissue-specific expression. (**a**–**h**) Strains known to express GFP in all neurons (*unc-119*, strain IM324 (ref. 36)) (**a**), olfactory neurons (*lim-4*, strain OH96 (ref. 37)) (**b**), intestine (*elt-2*, strain MR142 (ref. 38)) (**c**), body wall and vulval muscles (*myo-3*, strain PD4251 (ref. 39)) (**d**), pharynx (*myo-2*, strain PD4790) (**e**), vulval muscles (*egl-15*, strain NH2447 (ref. 40)) (**f**), gonad sheath cells (*lim-7*, strain OH172 (ref. 41)) (**g**) and coelomocytes (*unc-122*, strain OH910 (ref. 42)) (**h**). For each strain a chronogram is shown in relation to a diagram of the anatomy of a worm.

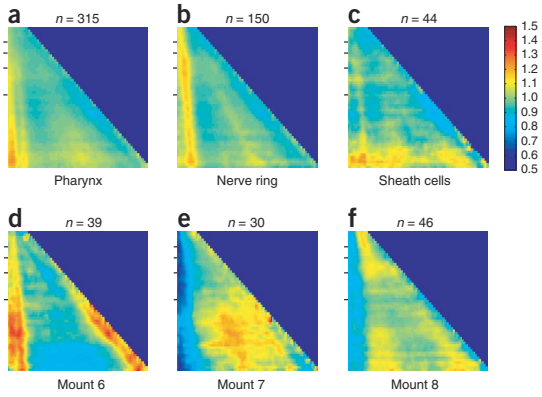


Figure 3 Feature extraction from multiple images. (**a**–**c**) Average chronograms of strains with reporter expression sharing a common anatomic annotation. (**d**–**f**) Strains with reporter fusions for genes corresponding to transcripts clustered in the same topomap mountain[17]. Mounts 6, 7 and 8 are enriched for genes expressed in neurons, germline and intestine respectively. *n* indicates the number of individual chronograms used to generate the average image. The color scale indicates the level of expression of a given position on the chronogram of the group considered relative to the average of all chronograms.

A caveat of this averaging method is that obtaining a robust tissue-specific signature depends on the availability of large numbers of images, because expression is more frequently observed in multiple tissues than in a single one. For example, only 44 strains were annotated with expression in the gonad sheath cells. Because most of them also expressed GFP in a variety of other tissues, the extracted signal (**Fig. 3c**) was not strictly restricted to the tissue-specific signature observed in the *lim-7::GFP* chronogram (**Fig. 2g**).

We also used this averaging approach to extract common features of chronograms corresponding to genes belonging to expression clusters generated based on a compendium of over 500 microarray experiments[17]. The average images associated with certain clusters ('topomap mountains') showed expression patterns that fit with their associated annotations, such as for neurons (mount 6, **Fig. 3d**), germ line (mount 7, **Fig. 3e**) and intestine (mount 8, **Fig. 3f**). However, most topomap mountains, even when sufficiently represented in our data set, did not display any characteristic pattern, suggesting that microarray clusters may not generally constitute a

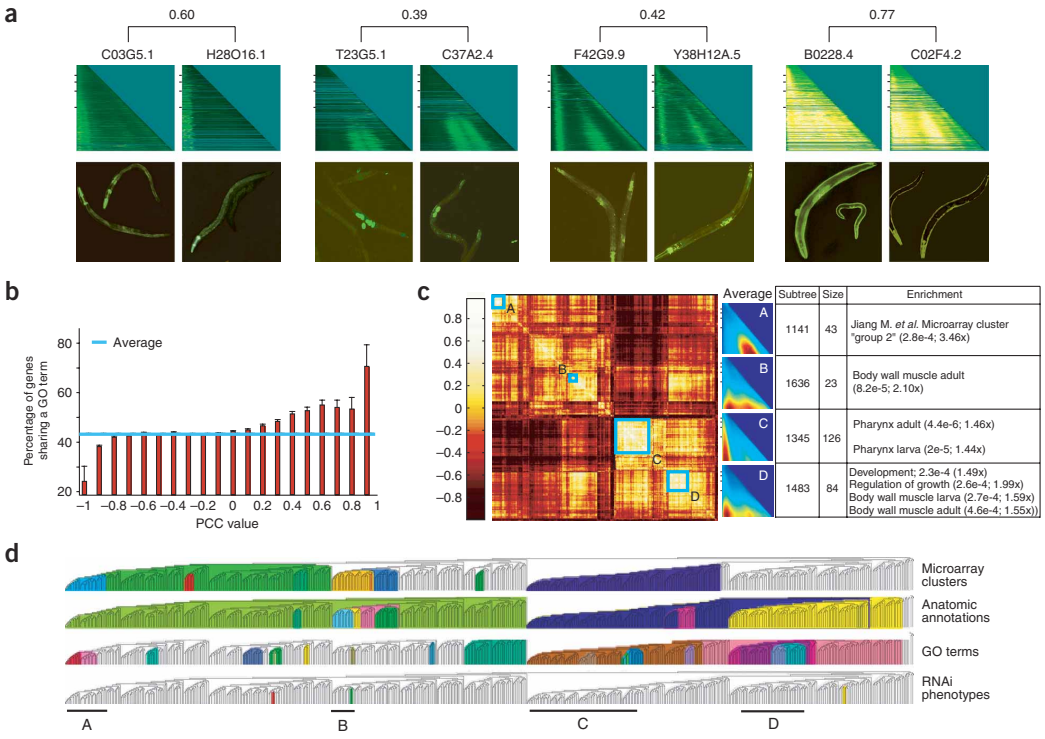


Figure 4 *In vivo* expression pattern clustering. (**a**) Pairs of highly correlated chronograms were selected from the PCC matrix. For each strain, the chronogram and a whole animal fluorescence micrograph are presented. (**b**) Correlation between chronograms PCC and GO terms. (**c**) Clustered matrix of 876 chronograms based on spatial and temporal correlation of expression. Average images of chronograms grouped in some of the neighbor joining tree branches (A–D) are presented as well as their enrichment in specific GO terms, microarray expression cluster[17,23] and/or anatomical annotation. (**d**) Overview of subtrees enrichments in four functional annotation types. Distinct colors indicate distinct categories of enriched terms in each class (**Supplementary Fig. 3** presents an enlarged version of this figure and the list of enriched subtrees is available in **Supplementary Table 1**).

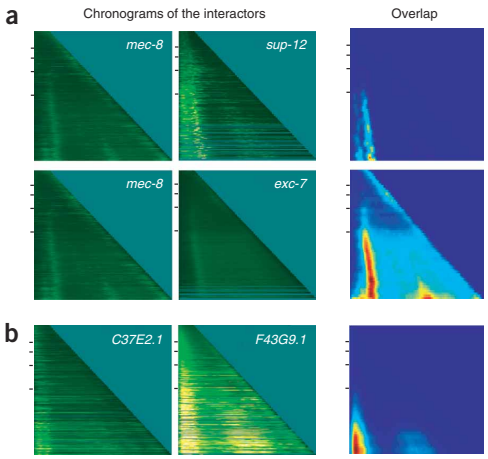


Figure 5 Localization of protein-protein interactions. (**a**) SUP-12 and EXC-7 each physically interact with MEC-8 (ref. 18) in high-throughput yeast two-hybrid. Although all three proteins are implicated in mRNA splicing, their expression patterns are distinct. Overlapping the chronograms of genes encoding interacting proteins can provide insights into where and when the proteins may interact. (**b**) C37E2.1 and F43G9.1 encoding the gamma and alpha subunits of the isocitrate dehydrogenase, respectively, are direct interactors in the WI5 protein-protein interaction map[18] and their chronograms show a high-level of correlation (PCC = 0.53).

good predictor for coexpression in the same tissue. For example, genes grouped in categories such as 'development', 'heat shock', 'aging' or 'Dauer' could accomplish their function in these processes in a wide range of distinct tissues and therefore not be colocalized despite their apparent coexpression.

Chronogram clustering

The digital nature of the chronograms allows quantification of the similarity between them. We calculated Pearson's correlation coefficient (PCC) between all pairs among the 876 chronograms with detectable signal. For a randomly sampled set of pairs within the top 95th percentile of the PCC distribution, we verified that a high PCC value between the chronograms correlated with a strong visual resemblance between the fluorescent micrographs of the corresponding animals (**Fig. 4a** and **Supplementary Fig. 2** online). We also examined whether chronogram similarity correlates with common Gene Ontology (GO) annotations, as has previously been observed between pairs of physically interacting proteins[18], coregulated transcripts[17] (from microarray expression profiling) and genes that share RNAi phenotypes[18–21]. Indeed, two genes whose corresponding chronograms were highly correlated tended to share GO annotations more often than expected by chance ($P = 2e^{-44}$, Fisher exact test; **Fig. 4b**), further demonstrating the power of chronogram comparisons.

We used average-linkage hierarchical clustering[22] to group genes based on the spatiotemporal activity of their promoter throughout post-embryonic development (**Fig. 4c**). As examples, we show four clusters, each corresponding to different branches of the hierarchical tree. For each cluster, we show an average chronogram image and list biological attributes enriched among genes within the cluster. Cluster A displayed significant ($P = 2.8e^{-4}$) enrichment in genes overexpressed in hermaphrodites relative to males and in adults relative to the other stages of development[23]. The average expression pattern observed for cluster A was consistent with these enrichments, with a signal located in the mature uterus and/or developing embryos. Noticeably, cluster A also contained two subtrees that were enriched for the GO terms 'DNA metabolism' and 'mitotic cell cycle' (**Fig. 4d** and **Supplementary Fig. 3** online). This is consistent with expression during early embryogenesis, a developmental stage during which most cell divisions occur. Cluster B was significantly ($P = 8.2e^{-5}$) enriched for genes expressed in adult body wall muscle, consistent with the average image obtained for this cluster, which showed expression appearing in the bigger adult animals. Cluster C enrichment in genes expressed in the pharynx of both larvae and adult animals was reflected by the corresponding average image, which was highly similar to the average chronogram of all strains annotated as expressed in the pharynx (**Fig. 3a**). Finally, cluster D was enriched in genes associated with the regulation of growth rate. It was also enriched in genes expressed in the body wall muscle, like cluster B, which explain the resemblance of their average image. Interestingly, features shared by groups of genes that are distant in the neighbor-joining tree were visible as bright, off-diagonal regions in the matrix (**Fig. 4c**).

Noticeably, the number of subtrees presenting significant ($P < 0.001$) enrichment for a given RNAi phenotype was much lower than for the other functional categories explored (**Fig. 4d**, **Supplementary Table 1** online). This is probably because only 10–12% of *C. elegans*–predicted genes are associated with any RNAi phenotype[18–21,24–26], whereas most are associated with GO terms and/or microarray expression data.

Localizome and interactome

High-throughput interaction mapping techniques, such as yeast two-hybrid, identify physical protein-protein interactions that may occur *in vivo*, but can not determine in what cells or tissues the two proteins in question are coexpressed for the interaction to happen. Comparison of microarray profiles can indicate coregulation between genes but, as most microarray data are derived from whole-animal RNA extraction, it remains possible that their expression occurs in distinct tissues. In contrast, chronograms can be used to define putative common expression territories of interacting proteins by observing the overlap between their expression patterns. MEC-8 was shown to be able to interact with SUP-12 and EXC-7 by high-throughput yeast two-hybrid[18], and all three proteins share the GO annotation 'RNA-binding protein', suggesting that they could function together within a single macromolecular complex. However, the three reporter strains for these genes displayed dissimilar expression patterns with chronograms that overlapped only partially (**Fig. 5a**), suggesting that these proteins function independently of each other *in vivo*, or might interact to function together in only a few cells. This interpretation is consistent with previous studies that indicate independent functions for these proteins. All three proteins have been shown to be involved in regulating tissue-specific alternate splicing, but act in different cells on distinct targets: MEC-8 in the maturation of *unc-52* mRNA in the hypodermis and unknown additional transcripts in neurons[27,28]; SUP-12 on *unc-60* mRNA in muscle cells; and EXC-7 on *sma-1* mRNA in the excretory cells and neurons, respectively[29–31]. On the other hand, strong spatiotemporal expression correlation associated with a protein-protein interaction may indicate a strong functional correlation. For 48 protein pairs from the worm interactome WI5 (ref. 18), the corresponding promoter pairs were assayed in this localizome data set. One of these chronogram pairs (C37E2.1-F43G9.1) scored in the top 95th percentile of the PCC distribution, indicating highly overlapping expression patterns (**Fig. 5b**). Based on the data from these two

unbiased data sets, one could hypothesize a functional association of the two corresponding proteins. Indeed, C37E2.1 and F43G9.1 appear to encode the gamma and alpha subunits of the isocitrate dehydrogenase, respectively.

As the coverage of both worm localizome and interactome maps improve, integration of the two maps will provide more insight into the specific organization of tissue-specific macromolecular networks. For example, we could distinguish two classes of interaction relationships: 'committed' (where the two partners form an obligatory heterodimer and both interactors are always expressed together) and 'uncommitted' (where the interaction is not necessary for the individual proteins to function and both partners may have different expression patterns), a distinction analogous to that proposed between party and date hubs in the yeast interactome[32].

DISCUSSION

The *C. elegans* adult is composed of 959 somatic cells, each identifiable by high-resolution microscopy. However, analyzing individual gene expression patterns at single-cell resolution is currently too cumbersome to be accomplished for all 19,000 genes in *C. elegans*. The high-throughput classification of expression patterns we devised can be viewed as a first step toward more refined annotations and as a complement to traditional microscopic analyses. For example, several clear patterns visible in the chronograms led us to reexamine the corresponding strains and to annotate tissue-specific expression that had been missed in the initial assessment (**Supplementary Fig. 4** online). Chronograms capture the time-lapse expression pattern for post-embryonic development and are quantitative digital data that can be analyzed using mathematical tools similar to those used for microarray data analysis. This approach combines the throughput necessary for genome-wide promoter activity analysis in transgenic worm strains, with a data format suitable for large-scale analyses and for comparisons with other genome-wide data sets.

Chronograms simultaneously provide coarse-grained spatial resolution along the anterior-posterior axis of the animal (length) and high temporal resolution throughout post-embryonic development (time). A limitation of our approach is the lack of the other two spatial dimensions (width and height), which precludes distinguishing between tissues located in the same cross-section (for example, vulval muscles and vulval neurons). We anticipate, however, that future optical developments of the COPAS profiler, or other systems, may provide enhancements in this regard.

Chronograms can be used to identify potential spatiotemporal territories where functional interactions between gene products are most likely to happen. Generating expression annotations for all protein interactions will add spatiotemporal information to the worm interactome network[18]. Localizome mapping will thus help characterize the dynamic aspects of the functional interactions between gene products in *C. elegans*.

METHODS

Worm profiling. For each analyzed strain, 20–30 transgenic animals were placed on 10-cm NGM agar plates seeded with *Escherichia coli* strain OP50 and left to proliferate at 20 °C. Upon exhaustion of the bacterial lawn, the mixed-stage population was washed out and analyzed using a COPAS profiler (Union Biometrica) Individual profiles were acquired until all animal sizes were represented.

Chronogram generation. As the animals pass through the profiler either tail or head first, we oriented all profiles relative to one another by using an automated method based on the best fit with their neighbors as determined by Pearson's correlation coefficient (PCC).

Clustering. Before calculating the correlation between chronograms we reduced their complexity by applying a linear filter to smooth out the signal and eliminate high-frequency noise. Image compression and PCC calculation were performed in Matlab. Images displaying no detectable variation from background were excluded from the clustering analysis (details in **Supplementary Methods** section "Hierarchical clustering of the chronograms"). The neighbor-joining tree was calculated using average-linkage hierarchical clustering[22]. Tree illustrations were performed with the tree editor TreeDyn (http://www.treedyn.org/)[33].

Cluster enrichments. The significance of enrichment for gene attributes within a list of genes induced by each given subtree was calculated using the cumulative hypergeometric distribution[34]. False discovery rates were associated with each observed nominal *P*-value according to an empirical null distribution of nominal *P*-values calculated similarly from 100 random permutations of all genes within the same tree structure (full details in **Supplementary Methods** in the section "Extraction of functionally enriched subtrees").

Strains origin. Most transgenic lines analyzed here were generated either by a modification of the microinjection method described by Mello *et al.* (1991) where 5′ regulatory DNA::GFP constructs and *dpy-5*(+) plasmid (pCeh-361) and selection for rescue of the Dpy-5 mutant phenotype[35], or by microprojectile bombardment[14]. A small fraction was provided by the Caenorhabditis Genetics Center. Individual strain origins are available for each expression patterns on the localizome webpage (see below) and at Wormbase (http://www.wormbase.org/).

Data availability. Data collected in the course of this project are available on Wormbase for batch download of both raw data and processed chronograms. We also created a searchable Localizome database that is freely accessible through the Internet and that provides the user with the anatomic annotation defined by microscopic observation and the associated chronogram. For any query gene, the web interface also displays the best chronogram matches, providing a convenient way to identify genes with similar expression patterns (http://vidal.dfci.harvard.edu/localizome/).

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was funded by the National Cancer Institute (NCI 4 R33 CA097516-02)(M.V.), Genome British Columbia, the Canadian Institute of Health Research and Genome Canada (D.B. and D.G.M.), the NetWork Bench National Science Foundation (IIS-0513650), the Muscular Dystrophy Association, the National Institutes of Health (NIH) (1 P20 CA11300-01, A.L.B. and HD43156, W.A.M.) and the Helen Kellogg Institute for International Studies (C.A.H.R.). F.P.R. was supported in part by NIH grant HG003224. M.T. was supported by NIH NRSA Fellowship HG004098. Some nematode strains used in this work were provided by the *Caenorhabditis Genetics* Center, which is funded by the NIH National Center for Research Resources. Thanks to Tracey Clingingsmith and Abigail Bird for their indispensable assistance and to Michael Cusick and Mike Boxem for careful proofreading of the manuscript.

AUTHOR CONTRIBUTIONS

Transgenic animals were generated by ballistic transformation by J.S. and J.R.-H. under the supervision of I.A.H., and by microinjection by D.T. and D.L. under the supervision of R.J. and D.B., anatomic annotations of the strains were performed by R.H.-N. and R.V. under the supervision of D.G.M. The chronogram concept was conceived by W.A.M. and implemented by D.D., N.B. and C.A.H. D.D. generated the Gateway *promoter::GFP* constructs with the help of J.R., and performed the profiling experiments with the technical support of R.P. Computational analyses were performed by N.B., C.A.H., K.V., A.-R.C., A.C. and M.T. under the supervision of D.D., F.P.R., C.L., A.-L.B. and M.V. Lab support was provided by J.R., N.S. and A.B. The manuscript was written by D.D., N.B., C.A.H., A.-L.B. and M.V. The project was conceived and codirected by D.B. and M.V.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Davidson, E.H. *et al.* A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).
2. Inoue, T., Wang, M., Ririe, T.O., Fernandes, J.S. & Sternberg, P.W. Transcriptional network underlying *Caenorhabditis elegans* vulval development. *Proc. Natl. Acad. Sci. USA* **102**, 4972–4977 (2005).
3. Cooper, S.J., Trinklein, N.D., Anton, E.D., Nguyen, L. & Myers, R.M. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.* **16**, 1–10 (2006).
4. Sulston, J.E., Schierenberg, E., White, J.G. & Thomson, J.N. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100**, 64–119 (1983).
5. Sulston, J.E. & Horvitz, H.R. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev. Biol.* **56**, 110–156 (1977).
6. Kimble, J. & Hirsh, D. The postembryonic cell lineages of the hermaphrodite and male gonads in *Caenorhabditis elegans*. *Dev. Biol.* **70**, 396–417 (1979).
7. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W. & Prasher, D.C. Green fluorescent protein as a marker for gene expression. *Science* **263**, 802–805 (1994).
8. Bao, Z. *et al.* Automated cell lineage tracing in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **103**, 2707–2712 (2006).
9. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
10. Vidal, M. A biological atlas of functional maps. *Cell* **104**, 333–339 (2001).
11. Dupuy, D. *et al.* A first version of the *Caenorhabditis elegans* Promoterome. *Genome Res.* **14**, 2169–2175 (2004).
12. Hobert, O. PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic *C. elegans*. *Biotechniques* **32**, 728–730 (2002).
13. McKay, S.J. *et al.* Gene expression profiling of cells, tissues, and developmental stages of the nematode *C. elegans*. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 159–169 (2003).
14. Reece-Hoyes, J.S. *et al.* Insight into transcription factor gene duplication from *Caenorhabditis elegans* Promoterome-driven expression patterns. *BMC Genomics* **8**, 27 (2007).
15. Mello, C.C., Kramer, J.M., Stinchcomb, D. & Ambros, V. Efficient gene transfer in *C. elegans*: extrachromosomal maintenance and integration of transforming sequences. *EMBO J.* **10**, 3959–3970 (1991).
16. Toms, N., Cooper, J., Patchen, B. & Aamodt, E. High copy arrays containing a sequence upstream of *mec-3* alter cell migration and axonal morphology in *C. elegans*. *BMC Dev. Biol.* **1**, 2 (2001).
17. Kim, S.K. *et al.* A gene expression map for *Caenorhabditis elegans*. *Science* **293**, 2087–2092 (2001).
18. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004).
19. Gunsalus, K.C. *et al.* Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* **436**, 861–865 (2005).
20. Ge, H., Liu, Z., Church, G.M. & Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* **29**, 482–486 (2001).
21. Rual, J.F. *et al.* Toward improving *Caenorhabditis elegans* phenome mapping with an ORFeome-based RNAi library. *Genome Res.* **14**, 2162–2168 (2004).
22. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
23. Jiang, M. *et al.* Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **98**, 218–223 (2001).
24. Kamath, R.S. *et al.* Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237 (2003).
25. Simmer, F. *et al.* Genome-wide RNAi of *C. elegans* using the hypersensitive *rrf-3* strain reveals novel gene functions. *PLoS Biol.* **1**, E12 (2003).
26. Fernandez, A.G. *et al.* New genes with roles in the *C. elegans* embryo revealed using RNAi of ovary-enriched ORFeome clones. *Genome Res.* **15**, 250–259 (2005).
27. Lundquist, E.A. *et al.* The *mec-8* gene of *C. elegans* encodes a protein with two RNA recognition motifs and regulates alternative splicing of *unc-52* transcripts. *Development* **122**, 1601–1610 (1996).
28. Spike, C.A., Davies, A.G., Shaw, J.E. & Herman, R.K. MEC-8 regulates alternative splicing of *unc-52* transcripts in *C. elegans* hypodermal cells. *Development* **129**, 4999–5008 (2002).
29. Anyanful, A. *et al.* The RNA-binding protein SUP-12 controls muscle-specific splicing of the ADF/cofilin pre-mRNA in *C. elegans*. *J. Cell Biol.* **167**, 639–647 (2004).
30. Loria, P.M., Duke, A., Rand, J.B. & Hobert, O. Two neuronal, nuclear-localized RNA binding proteins involved in synaptic transmission. *Curr. Biol.* **13**, 1317–1323 (2003).
31. Fujita, M. *et al.* The role of the ELAV homologue EXC-7 in the development of the *Caenorhabditis elegans* excretory canals. *Dev. Biol.* **256**, 290–301 (2003).
32. Han, J.D. *et al.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93 (2004).
33. Chevenet, F., Brun, C., Banuls, A.L., Jacq, B. & Christen, R. TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* **7**, 439 (2006).
34. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281–285 (1999).
35. Thacker, C., Sheps, J.A. & Rose, A.M. *Caenorhabditis elegans dpy-5* is a cuticle procollagen processed by a proprotein convertase. *Cell. Mol. Life Sci.* **63**, 1193–1204 (2006).
36. Kim, S. & Wadsworth, W.G. Positioning of longitudinal nerves in *C. elegans* by nidogen. *Science* **288**, 150–154 (2000).
37. Sagasti, A., Hobert, O., Troemel, E.R., Ruvkun, G. & Bargmann, C.I. Alternative olfactory neuron fates are specified by the LIM homeobox gene *lim-4*. *Genes Dev.* **13**, 1794–1806 (1999).
38. Kostic, I. & Roy, R. Organ-specific cell division abnormalities caused by mutation in a general cell cycle regulator in *C. elegans*. *Development* **129**, 2155–2165 (2002).
39. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
40. Harfe, B.D. *et al.* Analysis of a *Caenorhabditis elegans* Twist homolog identifies conserved and divergent aspects of mesodermal patterning. *Genes Dev.* **12**, 2623–2635 (1998).
41. Hall, D.H. *et al.* Ultrastructural features of the adult hermaphrodite gonad of *Caenorhabditis elegans*: relations between the germ line and soma. *Dev. Biol.* **212**, 101–123 (1999).
42. Bulow, H.E. & Hobert, O. Differential sulfations and epimerization define heparan sulfate specificity in nervous system development. *Neuron* **41**, 723–736 (2004).

CHAPITRE 3 : A LA RECHERCHE DES *PRINCIPES EVOLUTIFS* DE L'INTERACTOME

Dans le chapitre précédent, j'ai montré que le réseau interactome que nous avons construit pour *Arabidopsis* (« *Arabidopsis* Interactome 1 », ou AI-1) représente un ensemble d'interactions biophysiques de très bonne qualité technique et enrichi en interactions pertinentes biologiquement. Cependant, AI-1 contient une quantité encore inconnue de pseudo-interactions et ne représente qu'un échantillon limité du véritable réseau interactome d'*Arabidopsis*. En étant bien consciente des atouts comme des limitations de ce réseau, je l'ai analysé à la lumière de la théorie de l'évolution pour vérifier une hypothèse centrale de la biologie systémique.

Une hypothèse centrale de la biologie systémique propose que les relations entre génotype et phénotype soient sous-tendues par un ensemble de réseaux moléculaires dynamiques au sein de la cellule. L'ensemble des interactions physiques entre protéines constitue l'un de ces réseaux, mais pas le seul : les protéines interagissent aussi avec des acides nucléiques, des métabolites, des lipides... Ces différentes facettes des fonctions moléculaires des protéines s'additionnent comme autant de contraintes sur les séquences qui les encodent. D'après la théorie de l'évolution, les phénotypes résultants exercent une rétroaction sur les génotypes via l'action de la sélection naturelle. Intuitivement, l'hypothèse centrale de la biologie systémique vue sous un angle évolutif impliquerait donc que les pressions de sélection qui façonnent les génotypes le fassent à travers le remaniement dynamique de réseaux cellulaires (**Figure 9A**). Les travaux exposés dans ce chapitre valident en partie cette hypothèse en mettant en évidence qu'une évolution de l'interactome semble avoir accompagné l'évolution d'*Arabidopsis*.

Rôle de la sélection naturelle dans l'évolution des interactions physiques entre protéines : le cas des duplications de gènes

Les travaux présentés dans ce sous-chapitre et la partie correspondante du **Document Joint 6** ont été effectués en collaboration avec Benoit Charloteaux (alors au laboratoire de Marc Vidal), Murat Tasan (au laboratoire de Fritz Roth, Harvard Medical School), Sabrina Rabello et Gourab Ghoshal (au laboratoire de Laszlo Barabasi, Northeastern University).

Pour déterminer si le remaniement des interactions physiques entre protéines joue un rôle évolutif, nous nous sommes placés dans le contexte très étudié des duplications de gènes, phénomène fréquent et universel associé à l'innovation moléculaire et à la spéciation depuis plus de 40 ans (Ohno 1970). De nombreuses

études ont montré que les séquences initialement identiques de gènes issus du même ancêtre par duplication (paralogues) divergent rapidement, puis continuent à diverger mais plus lentement (Innan and Kondrashov ; Lynch and Conery 2000; Kondrashov, Rogozin et al. 2002; Scannell and Wolfe 2008). Ce changement de taux d'évolution, « rapide-puis-lent », est une signature de l'action de la sélection naturelle, relâchée après duplication, puis resserrée sur les quelques paralogues maintenus dans le génome (la plupart disparaissent). Au niveau de l'interactome, le modèle le plus simple propose que les protéines paralogues partagent initialement tous leurs interacteurs, puis que la fraction d'interacteurs partagés décroît suite à des pertes et gains d'interactions (**Figure 9B**). Nombre de modèles théoriques ont estimé, et des travaux empiriques ont tenté de mesurer, la valeur du taux de re-câblage des profils d'interactions (Pastor-Satorras, Smith et al. 2003; Vazquez 2003; Ispolatov, Krapivsky et al. 2005; Beltrao and Serrano 2007; Evlampiev and Isambert 2008; Presser, Elowitz et al. 2008), supposée constante au cours du temps. En l'absence de réseau interactome adapté à une telle étude, leurs conclusions demeuraient limitées (Wagner 2001; Wagner 2003; Maslov, Sneppen et al. 2004). Or, le postulat que le taux de re-câblage des interactions serait constant tandis que les séquences paralogues divergent de manière rapide-puis-lente suppose implicitement que ce re-câblage est aléatoire et indépendant de l'évolution fonctionnelle des protéines. Autrement dit, déterminer si la divergence des profils d'interactions de protéines paralogues a lieu à taux constant ou non permettrait de savoir si le re-câblage de l'interactome joue un rôle évolutif ou non.

La construction d'AI-1 et la mesure de ses limitations expérimentales (**chapitre 2**) nous ont permis d'observer empiriquement le taux de divergence des profils d'interactions de protéines paralogues (**Document Joint 6**). D'une part, la grande taille d'AI-1 couplée à la haute fréquence de gènes dupliqués dans le génome d'*Arabidopsis*, aboutit à la présence de près de 2000 paires de protéines paralogues dans AI-1. Ce nombre est environ dix fois plus élevé que celui qu'on observe dans les réseaux interactomes de qualité comparable disponibles pour l'humain (Rual, Venkatesan et al. 2005) ou la levure (Yu, Braun et al. 2008). D'autre part, puisque nous avons mesuré la sensibilité expérimentale d'AI-1 (chapitre 2), nous avons pu normaliser la fraction d'interacteurs partagés par deux protéines par rapport à une limite supérieure reflétant le manque de saturation de notre expérience. Pour vérifier la validité de nos mesures, nous les avons comparées avec la divergence fonctionnelle de paires de paralogues, résultat de leur évolution. En utilisant des mesures phénotypiques comparatives recensées par Hanada et ses collègues (Hanada, Kuromori et al. 2009), nous avons observé comme prédit que les protéines paralogues dont les mutants sont phénotypiquement redondants partagent la

plupart de leurs interacteurs, ce qui n'est pas le cas de protéines paralogues dont les mutants ont des phénotypes distincts. AI-1 contient donc de l'information pertinente sur l'évolution des paralogues.

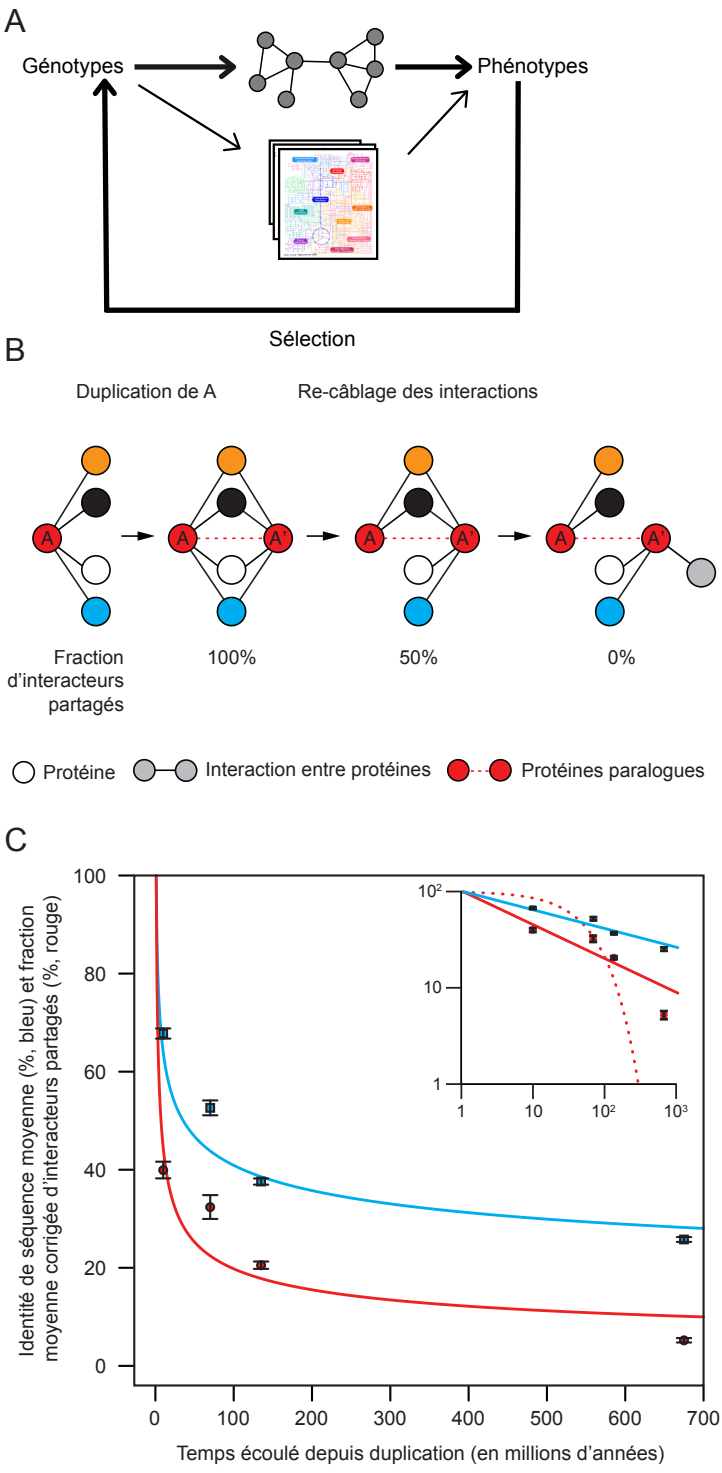


Figure 9 : Evolution de l'interactome d'*Arabidopsis thaliana*

A. Une hypothèse centrale de la biologie systémique vue sous l'angle évolutif. Les interactions physiques entre protéines (schéma de réseau) représentent l'un des nombreux intermédiaires (voies métabolique en illustration) entre génotypes et phénotypes. Le chapitre 3 étudie si le re-câblage des interactions physiques entre protéines est soumis aux lois de la sélection naturelle, dans le cas des duplications de gènes (premier sous-chapitre) et de l'évolution de la résistance aux pathogènes (second sous-chapitre).

B. Modèle de re-câblage des interactions entre protéines paralogues au cours du temps.

C. Fraction moyenne corrigée d'interacteurs partagés (cercles rouges) et identité de séquences protéiques moyenne (carres bleus) entre paires de protéines paralogues en fonction du temps écoulé depuis la duplication. Encart : représentation à l'échelle logarithmique. Lignes pleines : loi de puissance ajustée aux données ; ligne rouge en pointillés : loi exponentielle ajustée au re-câblage des interactions entre protéines paralogues au cours du temps. La divergence des profils d'interactions de protéines paralogues peut être approximée par une loi de puissance (somme des carrés résiduels = 85) bien mieux que par une loi exponentielle (somme des carrés résiduels = 144).

En utilisant des résultats récents de génomique comparative datant phylogénétiquement les duplications de gènes d'*Arabidopsis* (Hedges, Dudley et al. 2006; Vilella, Severin et al. 2009)(gramene.org), nous avons pu mesurer la proportion moyenne d'interacteurs partagés par des protéines paralogues en fonction du temps écoulé depuis qu'une duplication leur a donné naissance. Nos résultats (**Figure 9C**) montrent que la fraction d'interacteurs partagés par les protéines paralogues diminue à un rythme qui s'apparente plus à une loi de puissance qu'à l'exponentielle attendue si le re-câblage avait lieu à taux constant. Cette observation ne semble pas être influencée par la magnitude des duplications (grandes familles de gènes, duplication du génome entier). L'évolution des interactions physiques entre protéines, rapide-puis-lente, suit donc un profil similaire (même loi mais exposant différent) à celle des séquences de gènes dupliqués, et n'a pas lieu à un taux constant. Cette corrélation nous permet de rejeter l'hypothèse selon laquelle le re-câblage de l'interactome est indépendant des pressions de sélection qui façonnent la divergence fonctionnelle des paralogues. Il est donc fort probable que, au moins dans le contexte de l'évolution par duplication de gènes chez *Arabidopsis*, la sélection naturelle agisse, entre autres, directement ou indirectement, à travers le re-câblage des interactions physiques entre protéines.

À l'avenir, afin d'analyser plus finement l'évolution de l'organisation de l'interactome, je propose de comparer les vitesses de divergence des profils d'interactions des protéines paralogues en fonction des types de pression de sélection s'exerçant sur leurs séquences, ainsi qu'en fonction de leurs catégories fonctionnelles. Par ailleurs, lorsque des réseaux interactomes contenant suffisamment de paires de protéines paralogues seront disponibles pour d'autres organismes, il sera intéressant d'explorer si nos observations sont valides en dehors des plantes. *Arabidopsis* et les plantes à fleurs en général possèdent bien plus de paralogues que les autres phylums, pour des raisons probablement liées à leur mode de reproduction. Cette capacité à tolérer les duplications est supposément à l'origine de l'explosion du nombre d'espèces angiospermes. Mieux comprendre et pouvoir prédire la réponse des systèmes cellulaires de plantes aux duplications serait extrêmement utile à l'industrie agronomique (Leitch and Leitch 2008).

Les résultats résumés dans ce sous-chapitre et exposés dans le **Document Joint 6** suggèrent la possibilité que, comme prédit par l'hypothèse centrale de la biologie systémique, le réseau interactome « évolue » au sens darwinien du terme. Le sous-chapitre suivant et le **Document Joint 8** abordent la même question fondamentale mais de façon complémentaire. En collaboration avec Shahid Mukhtar (alors au laboratoire de Jeff Dangl, Université de Caroline du Nord), j'ai montré que des

millions d'années de co-évolution entre le système immunitaire d'*Arabidopsis* et divers phyto-pathogènes se reflètent sur l'organisation de l'interactome.

Attaques ciblées et défenses gardées: la course aux armes entre phyto-pathogènes et le système immunitaire d'une plante

Comprendre les mécanismes d'attaque et de défense entre les plantes et leurs pathogènes représente un enjeu crucial pour l'agriculture durable et l'équilibre de l'environnement. Vingt ans de recherches fondamentales utilisant *Arabidopsis* comme organisme modèle ont abouti à un concept surnommé « zigzag», qui décrit les relations dynamiques entre plantes et phyto-pathogènes (Nishimura and Dangl). La détection de molécules du « non-soi » par des récepteurs membranaires déclenche une première ligne de défense immunitaire générique, le « zig ». De nombreux pathogènes sont capables d'injecter à l'intérieur des cellules végétales des « effecteurs » moléculaires qui, affaiblissant l'effet du zig, augmentent la susceptibilité à l'infection : le « zag ». En réponse, des récepteurs intracellulaires semblent pouvoir reconnaître le « soi modifié » et provoquer une seconde vague de défense immunitaire, un second zig. Le degré de résistance de la plante au pathogène dépend de l'amplitude de ces deux lignes de défense, ainsi que de l'amplitude de l'attaque.

Indépendamment de ces mécanismes physiologiques complexes, la science des réseaux propose une élégante prédiction. Des travaux théoriques datant de dix ans ont montré que, lorsque la distribution des degrés (nombre de connections par composant) d'un réseau suit une loi de puissance (*i.e*. un petit nombre de composants, les *hubs*, établissent de nombreuses connections tandis que la plupart des composants font peu de connections), la structure de ce réseau est robuste aux « erreurs » aléatoires en général, mais en revanche très sensible aux « attaques » dirigées contre les *hubs* (Albert, Jeong et al. 2000). La distribution des degrés des réseaux interactomes semble suivre une loi de puissance (Jeong, Mason et al. 2001), même si on ne peut pas en être certain au vu de leur couverture limitée ((Han, Dupuy et al. 2005) et chapitre 2). En conséquence, les pathogènes maximiseraient leur virulence en ciblant les *hubs*.

La science des réseaux et la physiologie moléculaire se rencontrent donc dans la question fascinante des interactions hôtes-pathogènes. Comment concilier le modèle zigzag et la prédiction de la science des réseaux ? En accord avec cette prédiction, plusieurs études ont montré que les protéines de pathogènes, viraux ou bactériens, interagissent préférentiellement avec les *hubs* du réseau interactome (Dyer, Neff et al. ; Calderwood, Venkatesan et al. 2007; de Chassey, Navratil et al.

2008). Ceci étant, comment déterminer s'il ne s'agit pas d'un artefact technique ? Si les *hubs* étaient des protéines « collantes » interagissant de manière non spécifique avec de nombreux partenaires, comme il a été suggéré (Hart, Ramani et al. 2006), leurs interactions avec des protéines de pathogènes n'auraient rien de surprenant, ni rien d'intéressant. Par ailleurs, le modèle zigzag repose sur des hypothèses qui pour la plupart n'ont pas été démontrées systématiquement. Pour permettre l'étude des interactions plantes-pathogènes à la lumière de la science des réseaux, il était donc nécessaire de cartographier le premier interactome du système immunitaire végétal.

Nous avons couplé ce projet avec la construction d'AI-1, en collaboration avec les groupes de Jeff Dangl et Jim Beynon (Université de Warwick). Nos collaborateurs ont cloné ~60 effecteurs injectés dans la plante par la bactérie *P. syringae* (*Psy*), ~100 par l'oomycète *Hyaloperonospora arabidopsidis* (*Hpa*), ainsi que 3 types de protéines ayant un rôle dans le système immunitaire connu ou prédit : ~180 domaines cytoplasmiques de récepteurs membranaires (RLKs), ~140 domaines N-terminaux de récepteurs intracellulaires (NB-LRRs), et ~80 protéines de fonction variées impliquées dans la défense (Defense). Tous ces clones (les « appâts ») ont été systématiquement testés en Y2H entre eux, et contre les 8430 protéines utilisées pour AI-1 (« Space I », les « proies »), en même temps que les 8430 protéines ont été testées contre elles-mêmes. J'ai trié informatiquement les résultats de cette expérience pour en extraire deux réseaux : AI-1, et un réseau contenant toutes les interactions impliquant au moins un appât. Ce second réseau comprend 1358 interactions entre 926 protéines, dont 83 effecteurs pathogéniques, 170 protéines ayant un lien connu ou prédit avec l'immunité, et 673 proies (**Figure 10A**). J'ai proposé d'augmenter la connectivité de ce réseau en intégrant toutes les interactions entre ces 926 protéines décrites dans la littérature ou provenant d'AI-1. Cela aboutit à un nouveau réseau, comprenant 3148 interactions physiques, constituant la première carte du système immunitaire d'une plante en interaction avec des pathogènes (en anglais « first plant pathogen immune network », ou PPIN-1) (**Figure 10B**). Grâce à cette carte, nous avons pu tester plusieurs hypothèses clés proposées par la science des réseaux comme par la physiologie moléculaire.

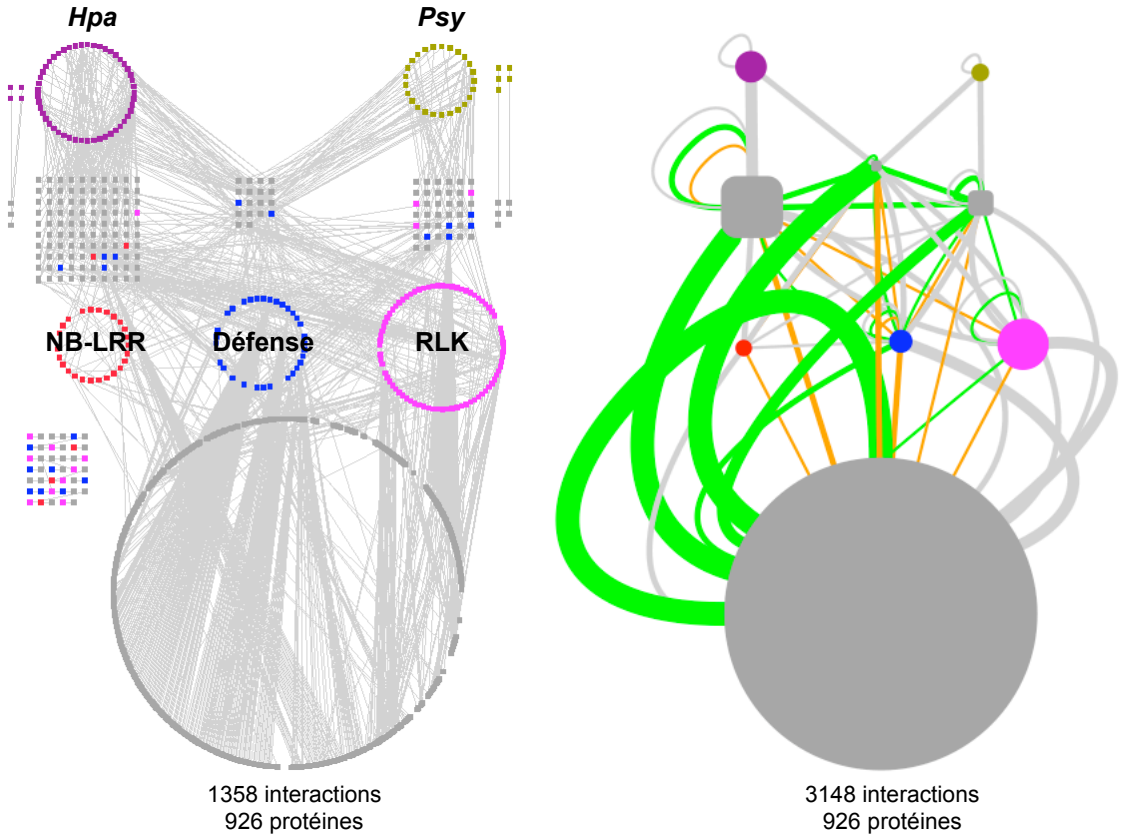


Figure 10: Réseau interactome plante-pathogènes (PPIN-1).
A. Présentation en étages du résultat d'un crible double hybride; les appâts sont représentés par des nœuds en couleurs, les proies en gris. Chaque interaction est représentée par un lien gris.
B. La connectivite du réseau (A) est augmentée par l'addition d'interactions d'AI-1 et de la littérature. Dans cette schématisation, la taille des nœuds est proportionnelle au nombre de protéines dans chaque catégorie; l'épaisseur des liens représente le nombre d'interactions entre chaque groupe de protéines, avec 1-10 interactions = 1 (unité arbitraire), 11-100 interactions = 2, 101-250 interactions = 4 et >250 interactions = 8. Les liens gris résument les interactions de (A), les verts ceux d'AI-1 et de la littérature, et les oranges ceux communs à (A) et à AI-1 ou la littérature.
Dans (A) et (B), PPIN-1 est organisé en quatre niveaux. En haut, les effecteurs des deux pathogènes sont regroupés en deux cercles, un pour la bactérie (*Psy*) et l'autre pour l'oomycète (*Hpa*). Au deuxième niveau en partant du haut, les protéines d'*Arabidopsis* interagissant directement avec au moins un de ces effecteurs sont assemblées en trois rectangles selon si elles interagissent avec des effecteurs de l'un, de l'autre ou des deux pathogènes. Aux troisième niveau se trouvent les trois types de protéines d'*Arabidopsis* utilisées comme appâts. Pour simplifier, les rectangles du deuxième niveau en partant du haut sont entièrement gris dans (B). Le niveau inférieur représente les protéines proies qui n'interagissent avec aucun effecteur.

Attaques ciblées. L'une des métaphores militaires décrivant les relations hôtes-pathogènes propose que les effecteurs se livrent à des attaques ciblées contre les protéines-clés de l'hôte afin de prendre contrôle de la cellule, ainsi que contre les défenses de l'hôte pour les affaiblir. Si les théories évolutives et les connaissances actuelles sont en faveur de ce modèle, il n'a jamais été réellement démontré. Nous avons comparé le nombre de protéines d'*Arabidopsis* avec lesquelles les effecteurs interagissent dans PPIN-1 avec le nombre auquel on s'attendrait si les effecteurs établissaient le même nombre de connections mais avec des protéines aléatoirement choisies parmi toutes celles de PPIN-1 et AI-1. Nos simulations prédisent que, si l'hypothèse des attaques ciblées était fausse, les effecteurs interagiraient avec 320 protéines d'*Arabidopsis* en moyenne, dont ~1% partagées par *Hpa* et *Psy*. En fait, dans PPIN-1 les effecteurs ont ciblé de manière répétitive seulement 165 protéines, dont 10% partagées par *Hpa* et *Psy*. Un tel degré de convergence de la part de deux espèces de pathogènes éloignées d'environ un milliard d'années d'évolution apporte un argument fort en faveur de l'hypothèse des attaques ciblées.

Défenses gardées. Le modèle zigzag repose sur une autre hypothèse aux consonances militaires, selon laquelle certaines protéines, appelées « R » pour résistance, sont capables de détecter le soi modifié et de déclencher une réponse immunitaire de forte amplitude, comme les « gardes » d'une citadelle (Dangl and Jones 2001). Cette hypothèse s'oppose au modèle « gène pour gène », où la plante aurait développé des récepteurs pour chaque effecteur pathogénique possible. Selon ce second modèle, les protéines R devraient interagir directement avec les effecteurs, alors que selon l'hypothèse des défenses gardées, ces interactions devraient être indirectes. Dans PPIN-1, les interactions entre effecteurs et protéines R (ici fragments de NB-LRRs) se sont avérées majoritairement indirectes, donc en accord avec l'hypothèse des défenses gardées.

Course aux armes. Le système immunitaire est probablement en co-évolution permanente avec les mécanismes de virulence des pathogènes qui tentent de déjouer les défenses de l'hôte. Chez les plantes, cette hypothèse est soutenue à l'échelle moléculaire par le fait que les gènes R (NB-LRRs) évoluent à un taux plus rapide (Caldwell and Michelmore 2009) que le reste des gènes d'*Arabidopsis*. En collaboration avec Jonathan Moore (Université de Warwick), j'ai montré qu'il en va de même pour les 673 proies que nous avons découvertes avec PPIN-1, ce qui appuie l'hypothèse de la course aux armes.

Ciblage des *hubs*. Comme prédit par la science des réseaux, les cibles des effecteurs de phyto-pathogènes dans PPIN-1 sont en général des *hubs* dans AI-1. Pour vérifier qu'il ne s'agit pas d'un artefact du Y2H, mes collaborateurs ont testé expérimentalement la résistance immunitaire de plantes mutées pour chacune

des 18 protéines ciblées par *Hpa* et *Psy*, dont les 16 présentes dans AI-1 sont des *hubs* ayant plus de 10 interacteurs. Neuf de ces mutants sont significativement plus susceptibles à l'infection que la plante sauvage, et sept sont plus résistants. Sept mutants choisis au hasard ont aussi été testés, mais n'ont presenté aucun phénotype immunitaire. Le ciblage des *hubs* par les pathogènes n'est donc pas un artefact du Y2H.

Comme je viens de le montrer, la construction et l'analyse de PPIN-1 nous ont permis de tester rigoureusement plusieurs hypothèses clés liées à l'organisation du système immunitaire végétal. Deux observations supplémentaires, sans relation avec les hypothèses mentionnées ci-dessus, ouvrent peut-être la voie vers de nouveaux axes de recherche dans ce domaine. Premièrement, les protéines de PPIN-1 qui ne sont pas les cibles des effecteurs de pathogènes sont elles aussi fortement connectées dans notre réseau systématique AI-1. Cette observation, qui ne prend pas en compte les interactions biaisées de PPIN-1, peut suggérer que les cibles font partie intégrante d'une machinerie de défense qui serait elle-même fortement connectée, ou bien que la machinerie de défense est entremêlée avec d'autres processus cellulaires centraux. Deuxièmement, l'expression des protéines qui interagissent avec des récepteurs est étonnamment stable dans des contextes de défense, tandis que les récepteurs eux-mêmes sont fortement régulés. Ceci pourrait signifier que l'amplitude de la réponse immunitaire dépend principalement de l'abondance des récepteurs, ou que les récepteurs peuvent s'associer au reste du réseau cellulaire, indépendamment du système immunitaire.

En conclusion, nos travaux démontrent que des effecteurs de pathogènes séparés par un milliard d'années d'évolution convergent sur et manipulent des machines intracellulaires extrêmement connectées entre elles et au reste de l'interactome. Réciproquement, le systeme immunitaire d'*Arabidopsis* a donc probablement évolué de manière à utiliser un nombre limité de protéines pour défendre l'organisme contre des pathogènes différents. Ces résultats, avec ceux présentés dans la première partie du chapitre 3, démontrent clairement que l'histoire évolutive d'*Arabidopsis* et l'organisation topologique de son interactome sont fortement interdépendants.

DOCUMENT JOINT 8

Titre : Independently Evolved Virulence Effectors Converge onto Cellular Hubs in a Plant Immune System Network.

Auteurs : M. Shahid Mukhtar*, Anne-Ruxandra Carvunis*, Matija Dreze*, Petra Epple*, Jens Steinbrenner, Jonathan Moore, Murat Tasan, Mary Galli, Tong Hao, Marc T. Nishimura, Samuel J. Pevzner, Susan E. Donovan, Lila Ghamsari, Santhanam Balaji, Viviana Romero, Matthew M. Poulin, Fana Gebreab, Bryan J. Gutierrez, Stanley Tam, Christopher J. Harbort, Nathan McDonald, Lantian Gai, Huaming Chen, EU Effectoromics Consortium, Frederick P. Roth, David E. Hill, Joseph R. Ecker, Marc Vidal, Jim Beynon, Pascal Braun, Jeffrey L. Dangl.

Description : Article décrivant la construction et l'analyse d'une carte d'interactions physiques entre protéines de la plante *Arabidopsis thaliana* et protéines de deux pathogènes, actuellement en révision chez le magazine *Science*.

Contribution : J'ai dirigé toutes les analyses bioinformatiques de ce projet depuis le début (2009) et en ai implémenté la grande majorité. SM Mukhtar et J Steinbrenner ont généré les clones des appâts. M Dreze a dirigé les expériences de double hybride en levure. P Epple a testé les phénotypes mutants presentés dans la dernière figure. SM Mukhtar et moi avons rédigé le manuscrit, avec l'aide principalement de J Dangl qui a supervisé le projet. J Dangl et P Braun ont dirigé le projet en général. Les autres auteurs, dont la liste et l'ordre restent à déterminer, ont apporté diverses contributions intellectuelles et techniques.

Independently Evolved Virulence Effectors Converge onto Cellular Hubs in a Plant Immune System Network

M. Shahid Mukhtar,[1*†] Anne-Ruxandra Carvunis,[2,3,4*] Matija Dreze,[2,3,5*] Petra Epple,[1*] Jens Steinbrenner,[6] Jonathan Moore,[7] Murat Tasan,[8] Mary Galli,[9] Tong Hao,[2,3] Marc T. Nishimura, [1] Samuel J. Pevzner,[2,3,10,11] Susan E. Donovan,[6‡] Lila Ghamsari,[2,3] Santhanam Balaji,[2,3] Viviana Romero,[2,3] Matthew M. Poulin,[2,3] Fana Gebreab,[2,3] Bryan J. Gutierrez,[2,3] Stanley Tam,[2,3] Christopher J. Harbort,[1§] Nathan McDonald,[1] Lantian Gai,[9] Huaming Chen,[9] EU Effectoromics Consortium, Frederick P. Roth,[2,12] David E. Hill,[2,3] Joseph R. Ecker,[9,13] Marc Vidal,[2,3] Jim Beynon,[6,7¶] Pascal Braun,[2,3¶] Jeffrey L. Dangl[1,14,15,16¶]

[1]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. [2]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [3]Department of Genetics, Harvard Medical School, Boston, MA 02115 USA. [4]Computational and Mathematical Biology Group, TIMC-IMAG, CNRS UMR5525 and Université de Grenoble, Faculté de Médecine, 38706 La Tronche cedex, France. [5]Unité de Recherche en Biologie Moléculaire, Facultés Universitaires Notre-Dame de la Paix, 5000 Namur, Wallonia, Belgium. [6]School of Life Sciences, Warwick University, Wellesbourne, Warwick, CV35 9EF, UK. [7]Warwick Systems Biology Centre, Coventry House, University of Warwick, Coventry, CV4 7AL, UK. [8]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA. [9]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA. [10]Biomedical Engineering Department, Boston University, Boston, MA 02215, USA. [11]Boston University School of Medicine, Boston, MA 02118, USA. [12]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA. [13]Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92027, USA. [14]Curriculum in Genetics and Molecular Biology, [15]Carolina Center for Genome Science, [16]Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

* These authors contributed equally to this project.

† present address: Department of Biology, CH106, University of Alabama at Birmingham, 1300 University Blvd., Birmingham, AL 35294, USA.

‡ present address: ADAS Boxworth Research Centre, Boxworth, Cambridgeshire CB23 4NN, United Kingdom.

§ present address: Max Planck Institute for Infection Biology, Chariteplatz 1, 10117 Berlin, Germany

present address: Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S3E1, Canada and Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, Ontario M5G1X5, Canada

¶ To whom correspondence should be addressed: Emails: dangl@email.unc.edu, pascal_braun@dfci.harvard.edu, Jim.beynon@warwick.ac.uk

Author contributions:

MSM: initiation of project, lead for experimental design and data analyses
A-RC: lead on all bioinformatics analyses, database design
MD: co-lead on all Y2H analyses, experimental quality control, co-data analysis
PE: validation of 18 common pathogen effector targets, data for Fig. 4A, B and table S9
JS: network data analysis
JM: evolutionary analysis of pathogen targets and mRNA expression analyses
MT: statistical analysis
MG: Q/C of Y2H data by wNAPPA
TH: database design and analysis of all IST sequencing traces
MTN: provided experimentally validated type III effector clones
SJP: statistical analysis of connectivity shown in Fig. 2B
SED: participation in Y2H screening of *H. arabidopsidis* ORFs and making groups of *H. arabidopsidis* ORFs
LG: Y2H screening of *H. arabidopsidis* ORFs and verification of interactions
SB: orthologs of effector targets
VR, MMP, FG, BJG, ST: Y2H pipeline
CJH: validation of Y2H *P. syringae* data and assistance with Fig. 4
NM: genotyped all validation mutants for Fig. 4
LG, HC: web-based visualization of PPIN-1
DEH: oversight of CCSB operations
FPR: oversight of statistical analysis of wNAPPA data
JRE: oversight Q/C of Y2H data by wNAPPA, provision of *A. thaliana* ORFs
MV: discussion and oversight of analyses
JB: planning of project, organization of *Hpa* candidate effector clones
PB: conception and planning of project, oversight of Y2H and Q/C experiments, and statistical analyses
JLD: conception and planning of project, oversight of biological validation experiments, and analyses

One sentence summary: A plant immune system interactome identifies a subnetwork of plant proteins that are repeatedly targeted by independently evolved virulence effectors from unrelated pathogens.

Abstract

Plants generate effective responses to microbial infection by recognizing both conserved and variable pathogen-encoded molecules. Pathogens manipulate plant defenses by deploying virulence effector proteins into host cells, where they interact physically with host proteins and alter their functions. We investigated how effector-mediated subversion of host defense, and the plant immune system response to it, are organized on a systems level. We generated a plant-pathogen immune system protein interaction network using virulence effectors from two pathogens that span the eukaryote-eubacteria divergence, three classes of plant immune system proteins and over 8,000 full-length Arabidopsis proteins. We noted striking convergence of pathogen effectors onto highly interconnected host proteins, and indirect connections between pathogen effectors and plant immune receptors, confirming predictions from network science and plant immunology. We validated our findings by reverse genetics and demonstrated plant immune system functions for 15 of 17 tested host proteins that interact with effectors from both pathogens. Thus, pathogens from different kingdoms deploy independently evolved virulence proteins that target a limited set of well connected protein hubs to facilitate their diverse life strategies.

Introduction

Interactions between microbes and their hosts are complex and dynamic. Their outcomes depend upon the ability of microbes to cause disease and the ability of the host to mount effective defense responses. Microbial virulence proteins (effectors), host cell surface receptors, intracellular signaling molecules and transcriptional regulators form the molecular battleground for confrontations between pathogens and hosts. Plants recognize pathogens through two major classes of receptors. Initially, plants sense microbes via perception of conserved Microbial-Associated Molecular Patterns (MAMPs) by pattern-recognition receptors (PRRs) located on the cell surface. This first level of recognition results in MAMP-triggered immunity (MTI), which is sufficient to fend off most microbes (*1*). To counter PRR-based detection and MTI, evolutionarily unrelated groups of plant pathogens have independently evolved mechanisms to secrete and deliver effector proteins into host cells (*2, 3*). These effectors interact with cellular host targets, and modulate MTI or host metabolism in a manner conducive to pathogen proliferation and dispersal (*2, 4, 5*). Plants deploy a second set of polymorphic intracellular immune receptors to recognize specific effectors. Nearly all are members of the nucleotide binding site-leucine rich repeat (NB-LRR) protein family, analogous to animal innate immune NLR proteins (*6, 7*). NB-LRR proteins can be activated upon direct recognition of an effector, or indirectly by the action of a specific effector protein on a host target (*2, 4, 5*). NB-LRR activation causes Effector-Triggered Immunity, or ETI, essentially a high amplitude MTI response that results in robust disease resistance responses that often include localized host cell death and systemic signaling (*2, 5*).

While increasingly sophisticated mechanistic details of pathogen infection and plant immune response are emerging, current data do not enable a systems level synthesis. We therefore systematically mapped protein-protein interactions between proteins from the reference plant *Arabidopsis thaliana* (hereafter Arabidopsis) and effector proteins from two pathogens: the Gram-negative bacterium *Pseudomonas syringae* (*Psy*) and the obligate biotrophic oomycete *Hyaloperonospora arabidopsidis* (*Hpa)*. These two pathogens are separated by at least 1 billion years of evolution, and have vastly different strategies to colonize plants. Our work was guided by the hypothesis that, despite evolution of

independent virulence mechanisms, these two pathogens would deploy effectors to manipulate a largely overlapping set of core cellular MTI machinery to successfully colonize the plant (*4, 5*).

Construction of a high quality plant-pathogen protein interaction network.

To build a plant-pathogen network, we started from 552 immune-related plant proteins and pathogen effector proteins including: experimentally validated *Psy* effector proteins (*8*) and candidate effectors from *Hpa* (*9*); putative signaling domains from two major classes of Arabidopsis immune receptors: i) N-termini of NB-LRR disease resistance proteins; ii) cytoplasmic domains of leucine-rich repeat (LRR)-containing receptor like kinases (RLKs), a subclass of PRRs; iii) known signaling components or targets of pathogen effectors (defense proteins) (**SOM**). For simplicity, we refer to the plant proteins from these sub-classes as 'immune proteins' (**Fig. 1A**, **table S1** and **SOM**). We mapped the network of binary (direct) interactions between these 552 immune-related proteins as well as between these and 8,430 full-length Arabidopsis proteins (Space 1) used to construct the Arabidopsis Interactome 1 (AI-1) using the same high-throughput yeast two-hybrid pipeline (*10, 11*); **SOM**).These experiments resulted in a dataset of 1,358 interactions among 926 proteins, including 83 pathogen effectors, 170 immune proteins, and 673 Arabidopsis Space 1 proteins (hereafter 'immune interactors') (**Fig. 1B, table S2** and **SOM**). Because this network was acquired using the identical experimental pipeline used to build AI-1, we estimate its *precision* (proportion of true biophysical interactions) to approximate that of AI-1, ~80% (*10*) (**SOM**). To obtain a more comprehensive view of the network, we integrated this new dataset with experimentally defined interactions from AI-1 and literature curated interactions (LCI; *10*) resulting in a network of 3,148 interactions among the same 926 proteins (**Fig. 1C**, **fig. S1**, **table S2**). We refer to this derived network as the 'plant-pathogen immune network' (PPIN-1).

PPIN-1 is easily viewed as four layers (**Fig. 1B**: the experimental network; **Fig. 1C** and **fig. S1**: the derived PPIN-1). The top layer contains effector proteins from both pathogens; the second layer consists of unique and shared interactors of those effectors ('effector targets'); the third layer depicts the three previously defined classes of Arabidopsis immune proteins: NB-LRR, defense and RLK proteins; and the fourth layer consists of the remaining Space 1 immune interactors.

Of the 673 immune interactors, only 66 were among the 975 Space 1 proteins with a Gene Ontology (GO) annotation related to immunity ('GO-immune proteins'; **table S3**) ($P > 0.05$, **table S4**). This lack of enrichment has several possible explanations, including technical limitations of both large- and small-scale experiments (*11-14*), and a lack of knowledge about the plant immune system. While 239 of the 673 proteins interacted with a GO-immune protein in the systematically mapped subset of AI-1 termed AI-1_{MAIN} (*10*), 368 were neither GO-immune annotated nor previously known to interact with a GO-immune annotated protein (**fig. S2**). These represent attractive starting points for future research.

We identified 165 interactors of effector proteins in PPIN-1, compared to ~20 described previously (*15*). While the function of most of these putative targets (hereafter, targets) is unknown, they are enriched in GO annotations for transcriptional regulation, nuclear localization and regulation of metabolism (**SOM** and **table S5**). Directed hypothesis testing revealed enrichments in GO-immune annotations, and plant hormone-related annotations (**SOM** and **table S4**) (*16*). Angiosperm-specific proteins are over-represented among the effector interactors that are present in Space 1, in comparison to all of Space 1

(hypergeometric P = 0.0007; **table S4**). We divided the network into 10 non-overlapping groups for comparisons (**table S6**). Genes encoding effector targets were enriched for differential expression in defense and immune-related contexts, as were those encoding immune receptors, defense proteins and interactors of defense proteins (**Fig. 1D, fig. S3**, **table S7** and **SOM**). By contrast, the expression of genes encoding interactors specific to immune receptors tended to be stable under these conditions (**Fig. 1D, fig. S3**, and **table S7**), even though a significant fraction of receptor genes are differentially regulated (**Fig. 1D**). This could suggest that pathogen detection sensitivity is specifically modulated via transcriptional regulation of receptor genes (*17, 18*). In addition, receptor genes might also be co-regulated with, and their products associate with, proteins unrelated to the defense machinery (see below).

PPIN-1 proteins evolve faster than those of AI-1.

The LRR domains of both plant immune receptor classes exhibit footprints of positive selection (*2, 19*). Since it is often supposed that host-pathogen 'arms races' drive adaptive evolution of immune system genes, we evaluated the evolution rate of the 673 genes encoding immune interactors (d_N/d_S ratio, **Fig. 1B, 1C; fig. S1**). These 673 non-receptor proteins are evolving very slowly overall, suggesting functional constraint and consequent purifying selection. They nevertheless collectively exhibit a significantly higher evolution rate than proteins of the AI-1_{MAIN} network (P < 0.01; **Fig. 1E**). In contrast, this was not the case for control groups of genes encoding hormone-related proteins (*16*) (**fig. S4A**) or metabolic enzymes (*20, 21*) (**fig. S4B**). Hence, even the non-receptor proteins from PPIN-1 evolve faster than other functionally related protein groups, and proteins in AI-1_{MAIN} in general.

Pathogen effectors converge onto highly connected proteins in the plant interactome.

In order to test the hypothesis that effectors from evolutionarily diverse pathogens converge onto a limited set of defense related host targets and molecular machines (*4, 5*), we compared the number of effector targets identified by PPIN-1 to the number of targets that would have been expected if effector proteins interacted with Arabidopsis proteins at random ('random targets'). PPIN-1 defined 165 direct effector targets; 18 of these were targeted by effectors from both pathogens (**Fig. 1B, C**) **(Fig. 2A**, left panel; **SOM**). In contrast, simulations using randomly picked targets identified an average of 320 effector targets, of which less than 1% would be targeted by effectors from both pathogens (P < 0.001, empirical p-value, **fig. S5**; **SOM**; an example is shown in **Fig. 2A** right panel). Further, we investigated the connectivity between the 137 observed effector targets also present in AI-1_{MAIN}. These are connected by 139 interactions in AI-1_{MAIN} (P < 6.7 x 10^{-5}, empirical p-value; **Fig. 2B**, left panel). In contrast, these proteins would form an average of only 22 (maximum 59) connections in a randomly rewired network with the same structure (**fig. S6**; example shown in **Fig. 2B**, right panel). Collectively, these data support our hypothesis that diverse pathogens deploy virulence effectors that converge onto a limited set of host cellular machines.

These observations raise the question of what proteins constitute attractive targets for pathogen effectors. Scale-free networks are resilient to random perturbations, but sensitive and easily destabilized by targeted attack on their hubs (the most highly connected protein nodes in the network) (*22*). While AI-1_{MAIN}, like many other biological networks, is not perfectly scale-free, simulations demonstrate that it shares the property of being resilient to

random perturbations but sensitive to targeted removal of its hubs (**fig. S7**). Consistent with this, we found that the number of interaction partners (degree) of the effector targets present in AI-1_{MAIN} was indeed significantly higher than proteins in AI-1_{MAIN} that are not in PPIN-1 (**Fig. 2C**). Remarkably, 7 of the 15 hubs of degree greater than 50 ($hubs_{50}$) in AI-1_{MAIN} were targeted by effectors from both pathogens ($P = 3.9 \times 10^{-13}$, **table S4**), and 14 of the 15 $hubs_{50}$ were targeted by effectors from at least one pathogen ($P = 6.9 \times 10^{-18}$, **table S4**). Simulations demonstrate that an attack on the experimentally identified effector targets is much more damaging to the network structure than an attack on the same number of randomly selected proteins (**fig. S7**).

We investigated the network connectivity of PPIN-1 protein groups (**table S6**). We found that categories other than effector targets also displayed a higher connectivity than proteins in AI-1_{MAIN} that are not in PPIN-1 (**Fig. 2C**), and that PPIN-1 proteins as a whole are more connected than other proteins in AI-1_{MAIN} (**Fig. 2D**). As a consequence, immune interactors form a highly connected cluster in the plant interactome (**fig. S8**; **fig. S9A**), which is not the case for hormone-related proteins or metabolic enzymes (**fig. S9B, C**) (*16, 20, 21*). Thus, PPIN-1 proteins as a whole, and effector targets in particular, are highly connected nodes within the overall plant network. Because deletion of yeast genes encoding hub proteins tends to cause multiple phenotypes (*12*), we consider it unlikely that most PPIN-1 proteins are exclusively devoted to immune function. Our data are compatible with a model in which pathogen effectors target highly connected cellular proteins, and the immune system is consequently driven via evolutionary forces to integrate with and respond to perturbations of the cellular network at large.

The plant response: guarding high value targets.

We investigated how the plant detects attacks on hubs by examining the position of predicted immune receptors with respect to other proteins in the network. We found that only 2 out of 30 NB-LRR immune receptor fragments present in PPIN-1 directly interacted with a pathogen effector ($P = 0.04$, **table S4**). Conversely, 24 of the 52 Space 1 interactors of NB-LRRs, including 7 of the 15 $hubs_{50}$ proteins, were targets of pathogen effectors ($P = 4.6 \times 10^{5}$ and $P = 8 \times 10^{-12}$, respectively; **table S4**). This finding suggests that these NB-LRR protein fragments interact preferentially with effector targets, rather than with effectors. Our admittedly incomplete dataset provides experimental and statistical support for the Guard Hypothesis, which proposes that NB-LRR proteins monitor the integrity of key cellular proteins and are activated when pathogen effectors act to generate 'modified self' molecules (*4, 5*), at least for the 30 NB-LRR proteins present in PPIN-1. Furthermore, in PPIN-1 only 4 of 90 putative RLK receptors interacted directly with a pathogen effector ($P = 10^{-5}$, **table S4**), but 46 of 162 Space 1 proteins interacting with RLKs were also effector targets ($P = 0.02$, **table S4**). This systems level observation presents a contrast to the direct perturbation of PRR-RLK kinase function by at least two well studied *Psy* type III effectors (*23*). In sum, our observations are consistent with the view that pathogen effectors are mostly *indirectly* connected to at least the host immune receptors represented in PPIN-1.

Effector targets and immune receptors participate in diverse potential protein machines.

A central tenet of the Guard Hypothesis is that effectors from evolutionarily diverse pathogens will converge onto shared host targets (*4, 5*). We determined that many effector targets are hubs, and thus likely to be part of various molecular machines. To illustrate

this, we modeled 'hypothetical complexes' of two, three, or four experimentally connected PPIN-1 proteins (**Fig. 3**, **table S8**). The high interconnectivity of effector targets results in their membership in many hypothetical complexes. Among the 105 interconnected effector targets, we found: i) that the 18 proteins targeted by effectors from both pathogens were involved in 304 hypothetical complexes of *Psy* effector – Arabidopsis protein – *Hpa* effector (**Fig. 3A, B**). Similarly, we noted the following hypothetical complex classes: ii) 226 consisting of *Psy* effector-*Psy* target-*Psy* target-*Psy* effector; iii) 674 consisting of *Hpa* effector-*Hpa* target-*Hpa* target-*Hpa* effector; and iv) 613 consisting of *Psy* effector-*Psy* target-*Hpa* target-*Hpa* effector (**Fig. 3A, B, table S8**). Ninety-one of the 105 effector targets involved in these hypothetical complexes are present in AI-1_{MAIN} and they have an average degree of 29 (compared to an average degree of 4.8 and 2.6 for PPIN-1 and non-PPIN-1 proteins, respectively, in AI-1_{MAIN}). We also found 19 and 41 proteins interacting specifically with *Psy* and *Hpa* effectors, respectively (**Fig. 3A B**), although this apparent pathogen specificity may reflect the limited sensitivity of our experimental pipeline (*10*). In addition, we assembled a striking number of hypothetical complexes where pathogen effectors indirectly interacted with either an RLK (856) or NB-LRR (274) receptor domain via an Arabidopsis protein (**Fig. 3C**). Further, single Arabidopsis proteins mediated hypothetical complexes between a cytoplasmic RLK domain and an NB-LRR N-terminus in 154 cases (**Fig. 3C**). These hypothetical complexes form a rich dataset for future functional testing.

Experimental validation of host proteins targeted by multiple pathogen effectors.

We focused our initial functional validation efforts on the 18 proteins targeted by effectors from both pathogens (**Fig. 1B, Fig. 3A**). This subset includes seven of the 15 $hubs_{50}$ proteins from AI-1_{MAIN}. We assayed whether these effector targets function to positively regulate host defense (mutation leads to enhanced host susceptibility), negatively regulate host defense (mutation leads to enhanced host resistance) or function to facilitate infection (mutation also leads to enhanced host resistance). We discovered enhanced disease susceptibility to two different *Hpa* isolates, Emwa1 and Emoy2, for nine of 17 loci for which insertion mutants were available (*24, 25*) (**Fig. 4A**, **table S9, top; fig. S10A**). Mutants in the eight remaining loci did not exhibit enhanced disease susceptibility. However, at least six of these eight did exhibit enhanced disease resistance to the virulent *Hpa* isolate Noco2 (**Fig. 4B**), based on three experiments (**table S9**). Hence, 15 of 17 proteins targeted by effectors from both pathogens, including all seven of the 15 $hubs_{50}$ proteins in this group, have mutant phenotypes consistent with immune system functions. As a control, seven unrelated mutant lines inoculated with 30,000 spores/ml of *Hpa* isolate Emwa1 did not exhibit altered disease resistance (means: Col-0=1.3 +/- 0.2; seven mutant lines = 0.8–1.8 +/- 0.3; *rpp4*= 16.1 +/- 0.7).

We were surprised that these 17 proteins did not generally express pleiotropic morphological phenotypes, which can confound pathogen tests. One did: CSN5a (At1g22920) is a subunit of the COP9 signalosome and is targeted by multiple effectors from both pathogens. It also interacted with N-termini of NB-LRR proteins and cytoplasmic domains of RLKs (**table S8**). The morphological consequences of *csn5a* pleiotropy can be suppressed by reducing the expression of either of the two Arabidopsis CUL3 subunits (*26, 27*). We found that *csn5a-2 cul3a* seedlings displayed enhanced disease resistance compared to controls following infection with virulent *Hpa* (**Fig. 4C**). These results correlated with infection-triggered over-accumulation of PR-1 protein, a common marker for MTI, in *Hpa* (**Fig. 4D**) or *Psy* infected (**fig. S10B**) *csn5a-2 cul3a* plants, compared to Col-0 (note the absence of ectopic PR1 expression before infection in both cases). Hence

our observed enhanced disease resistance phenotype of *csn5a* is not due to its pleiotropic morphological phenotypes.

We also validated prefoldin 6 (PFD6; At1g29990; **Fig. 4E, F**), because it interacted with the known defense regulator EDS1 (Enhanced Disease Susceptibility 1), and two bacterial effectors (**table S8** and **SOM**). We tested whether *pfd6-1* expressed altered MTI by assaying for alterations in flagellin- (flg22 peptide)-induced disease resistance. Bacterial growth in flg22 pre-treated leaves of Col-0 plants was 10-20 times less than that in mock pre-treated leaves, reflecting successful MTI. Importantly, flg22-induced MTI was compromised in *pfd6-1* plants (**Fig. 4E**). We tested the induction of several well defined molecular MTI markers by real time RT-PCR. The transcript levels of all these genes were highly induced within 45 minutes of treatment with flg22 in Col-0 plants. Transcriptional induction of these markers was abolished in *fls2* and largely impaired in *pfd6-1* (**Fig. 4F**). These results unveil the importance of PFD6 in MTI, suggest a link between its known functions and FLS2 PRR receptor function (*28*) (**SOM**), and provide an example of alternative functional assays to further query PPIN-1. Collectively, results in **Fig. 4** validate PPIN-1, and confirm that pathogen effectors target host proteins required for effective defense or pathogen fitness.

Conclusions

We established a plant immune interactome network containing 673 Arabidopsis proteins; most of these have not been implicated previously in immune system function. Our analyses reveal that oomycete and bacterial effectors separated by ~1 billion years of evolution can target an overlapping subset of plant proteins that are extremely well connected hubs in the cellular network. Our functional validation supports the concept that effectors from evolutionarily diverse pathogens converge onto and manipulate interconnected host machinery to suppress effective host defense and facilitate pathogen fitness. We predict that many of these hubs will also be targets of additional independently evolved effectors from other plant pathogens. Our data are also consistent with indirect connections of pathogen effectors to at least the NB-LRR immune receptors present in PPIN-1, as proposed in the Guard Hypothesis. The high degree of the effector targets, and their demonstrable immune functions, argue against a 'decoy' role for these proteins. While the concept of cellular decoys specifically evolved to intercept pathogen effectors is attractive, and likely true in at least one case in the plant immune system (*29*), these are expected to have few, if any, additional cellular functions and as such would likely exhibit a low degree. Targeting of network hubs by pathogen proteins has been reported in other organisms, most prominently the targeting of human hubs by pathogenic viruses and bacteria (*30-34*), but convergence onto shared targets was not addressed in these cases.

Our results bridge previously independent concepts from plant immunology (effectors should target common proteins) and network science (hubs should be targets of disruption) in a fascinating manner. This confluence raises an interesting question: Are hubs targets because of their inherent protein properties, such as a biochemical propensity to engage in diverse physical interactions, or because of their likely roles as multi-functional units in different cellular processes? In network science, attacks on networks are modeled by deletion of nodes, whereas in cellular systems like infected cells, proteins dynamically interact and functionally modify each other in complex and subtle ways. Future mechanistic and theoretical studies based on PPIN-1 will focus on the relationship between hub targeting and response to infection. Derivation of general rules regarding the organization and function of host cellular machinery required for effective defense against

microbial infection, and detailed mechanistic understanding of how pathogen effectors manipulate these machines to increase their fitness, will facilitate improvement of immune systems.

References cited

1. C. Zipfel, *Curr. Opin. Plant Biol.* **12**, 414 (2009).
2. P. N. Dodds, J. P. Rathjen, *Nat. Rev. Genet.* **11**, 539 (2010).
3. T. Boller, S. Y. He, *Science* **324**, 742 (2009).
4. J. L. Dangl, J. D. Jones, *Nature* **411**, 826 (2001).
5. J. D. Jones, J. L. Dangl, *Nature* **444**, 323 (2006).
6. E. Lukasik, F. L. Takken, *Curr. Opin. Plant Biol.* **12**, 427 (2009).
7. G. van Ooijen *et al.*, *J. Exp. Bot.* **59**, 1383 (2008).
8. D. Baltrus *et al.*, *Plos Pathog.*, (*in revision*).
9. L. Baxter *et al.*, *Science*, (*in press*).
10. M. Dreze *et al.*, (*co-submitted*).
11. M. Dreze *et al.*, *Methods Enzymol.* **470**, 281 (2010).
12. H. Yu *et al.*, *Science* **322**, 104 (2008).
13. M. E. Cusick *et al.*, *Nat. Methods* **6**, 39 (2009).
14. P. Braun *et al.*, *Nat. Methods* **6**, 91 (2009).
15. J. D. Lewis, D. S. Guttman, D. Desveaux, *Semin. Cell Dev. Biol.* **20**, 1055 (2009).
16. Z. Y. Peng *et al.*, *Nucleic Acids Res.* **37**, D975 (2009).
17. C. Zipfel *et al.*, *Nature* **428**, 764 (2004).
18. X. Tan *et al.*, *BMC Plant Biol.* **7**, 56 (2007).
19. T. B. Sackton *et al.*, *Nat. Genet.* **39**, 1461 (2007).
20. P. Zhang *et al.*, *Plant Physiol.* **138**, 27 (2005).
21. Y. Jaillais, J. Chory, *Nat. Struct. Mol. Biol.* **17**, 642 (2010).
22. R. Albert, H. Jeong, A. L. Barabasi, *Nature* **406**, 378 (2000).
23. T. Boller, *Cell Host Microbe* **4**, 5 (2008).
24. J. M. Alonso *et al.*, *Science* **301**, 653 (2003).
25. A. Sessions *et al.*, *Plant Cell* **14**, 2985 (2002).
26. G. Gusmaroli, S. Feng, X. W. Deng, *Plant Cell* **16**, 2984 (2004).
27. G. Gusmaroli, P. Figueroa, G. Serino, X. W. Deng, *Plant Cell* **19**, 564 (2007).
28. S. Robatzek, D. Chinchilla, T. Boller, *Genes Dev.* **20**, 537 (2006).
29. R. A. van der Hoorn, S. Kamoun, *Plant Cell* **20**, 2009 (2008).
30. M. A. Calderwood *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **104**, 7606 (2007).
31. B. de Chassey *et al.*, *Mol. Syst. Biol.* **4**, 230 (2008).
32. P. Uetz *et al.*, *Science* **311**, 239 (2006).
33. M. D. Dyer *et al.*, *PLoS One* **5**, e12089 (2010).
34. M. D. Dyer, T. M. Murali, B. W. Sobral, *PLoS Pathog.* **4**, e32 (2008).
35. P. Shannon, *et al.,* Genome Res. (2003) **13,** 2498.
36. K. Tsuda, M. Sato, J. Glazebrook, J. D. Cohen, F. Katagiri, *Plant Journal* **55**, 1061 (2008).

37. This work was funded by NIH grant GM-066025 and DOE grant FG02-95ER20187 to J.L.D.; BBSRC grants E024815 and G015066 to J.B.; NSF grant 0703905 to M.V., J.R.E., D.E.H.; NIH grant P50-HG004233 to M.V.; and NSF grants 0520253, 0313578 and 0726408 to J.R.E. We gratefully acknowledge the NSF funded ABRC and SIGNAL projects for seeds and clones, respectively. We thank Dr. Laura Baxter (Warwick Systems Biology, United Kingdom) for Arabidopsis/Papaya ortholog identification, Dr. Benoit Charloteaux (CCSB, Boston, USA) for assisting in some bioinformatics analyses, Dr. Birgit Kemmerling (University of Tuebingen,

Germany) for several RLK clones not contained in SIGNAL, and Dr. Chris Somerville and Dr. Ying Gu (UC Berkeley, USA), Tesfaye Mengiste (Purdue University, USA) and Dr. Xing-Wang Deng (Yale University, USA) for mutant seeds. The EU Effectoromics Consortium was funded by EU ERAPG and includes: Adriana Cabral and Guido van den Ackerveken (Utrecht University, The Netherlands); Jaqueline Bator, Ruslan Yatusevich, Shinpei Katou and Jane Parker (Max Planck Institute for Plant Breeding Research, Cologne, Germany); Georgina Fabro and Jonathan Jones (The Sainsbury Laboratory, Norwich, United Kingdom); Mary Coates and Tina Payne (University of Warwick, Warwick, United Kingdom). M.V. is a ‘‘Chercheur Qualifié Honoraire’’ from the Fonds de la Recherche Scientifique (FRS-FNRS, French Community of Belgium).

Figure legends

Fig. 1: Construction of a high quality plant-pathogen immune network (PPIN-1).

(**A**) Pathogen effectors and immune proteins used as baits to query Arabidopsis Space 1 proteins (grey). Effectors were from the bacterial pathogen *P. syringae* (*Psy*; gold) and the oomycete pathogen *H. arabidopsidis* (*Hpa*; purple). Plant immune proteins included N-terminal domains of NB-LRR immune receptors (red), cytoplasmic domains of leucine-rich repeat (LRR)-containing receptors like kinase (RLK), a subclass of Pattern recognition receptors (pink) and literature curated defense proteins (blue). The number of proteins corresponding to each sub-class is listed next to each category. The resulting Y2H dataset and protein compositions are listed on the right.

(**B**) Layered representation of the experimental plant immune network. Nodes (proteins) are colored as in (A). Edges (grey) represent protein-protein interactions. Interactions that are not connected to the network involving *Hpa* or *Psy* effectors are indicated next to their relevant protein categories in the first and second layers. Grid at left denotes individual interactions involving proteins other than pathogen effectors.

(**C**) Connectedness among the 926 experimental immune network proteins is increased by integration with interactions from AI-1 and literature curation (LCI) to generate PPIN-1 (**fig. S1**). Grey edges: interactions from (A); Green edges: added from AI-1 plus LCI; Orange edges: interactions common to (A) and AI-1 plus LCI. Color-coded nodes represent the sub-classes of proteins listed in (B). Node size is proportional to the number of proteins in each category. Edge thickness in (C) is an arbitrary unit where 1-10 interactions = 5 (an arbitrary number for edge thickness in Cytoscape (*35*), 11-100 interactions = 10, 101-250 interactions = 20 and >250 interactions = 40.

(**D**) PPIN-1 contains groups enriched and depleted in proteins encoded by genes differentially expressed in defense context. Nodes represent subsets of proteins in PPIN-1. Node size is proportional to the number of proteins in the group. Nodes are connected by grey edges (numbers next to each edge) of width proportional to the number of Y2H interactions from **Fig. 1B** that connect them. The grey octagonal node represents pathogen effectors, the round nodes Arabidopsis protein groups. Red nodes are enriched, the green node depleted, and the grey nodes unchanged for proteins encoded by genes differentially expressed (DE) in defense contexts. A blue node border signifies that proteins in this group have a significantly higher degree distribution than AI-1_{MAIN} proteins that are not in PPIN-1 ($P < 0.05$ according to a Mann-Whitney test; see **SOM** and **fig. S9**), while

non-colored node borders indicate few or no proteins from this group are in AI-1_{MAIN}. The list of proteins corresponding to each of the 10 non-overlapping groups is provided in **table S6**.

(**E**) Immune interactors display signatures of increased rates of evolution compared to the AI-1_{MAIN} proteome. Proportions of the Arabidopsis 673 immune interactors (**Fig. 1A, B**), and all proteins present in the AI-1_{MAIN} network, with the noted d_N/d_S ratios. A Kolmogorov-Smirnov test shows that distributions of d_N/d_S differ significantly between AI-1_{MAIN} proteins and PPIN-1 immune interactors ($P < 0.01$). d_N/d_S values are computed between Arabidopsis proteins and their Papaya orthologs. Black and pink bars represent AI-1_{MAIN} and PPIN-1 immune interactors, respectively.

Fig. 2: Interconnected effector targets converge onto cellular hubs.

(**A**) Measurement of convergence of pathogen effectors onto cellular targets. Left: observed connectivity between 248 experimentally determined nodes (effectors plus their targets) defines 341 edges. Right: An example of random connectivity (out of 1000 simulations; **fig. S5A**) between the same number of hypothetical effectors with the same degree as the real effectors, and their random targets chosen among proteins present in AI-1 and PPIN-1 requires 402 nodes to accommodate 341 edges. Nodes represent proteins and are colored as in (A). Grey edges: protein-protein interactions from (A) (left), or their simulated equivalent (right). Node size is proportional to the number of connections made by the node. The smallest node has one interaction =10 (an arbitrary number for size in cytoscape (*33*) and the biggest node has 29 interactions =290.

(**B**) Pathogen effector targets are highly interconnected. Left: the observed connectivity in AI-1_{MAIN} between effectors plus their targets present in AI-1 is 220 nodes defining 448 edges. Right: an example of random connectivity (out of 15,000; **fig. S5B**) in AI-1_{MAIN} between targets of pathogen effectors generates only 326 edges with the same 220 nodes. Nodes represent proteins and are colored as in (A). Grey edges: protein-protein interactions from (A). Green edges: protein-protein interactions from AI-1_{MAIN} (left), or their simulated equivalent (right).

(**C**) PPIN-1 proteins are cellular hubs. Average degree (number of interactors) in AI-1_{MAIN} of the protein groups noted **(Fig. 1A**) and proteins in AI-1_{MAIN} that are not in PPIN-1. All groups of proteins from PPIN-1 have a significantly higher degree than non-PPIN-1 proteins in AI-1_{MAIN} (** $P < 0.0001$ Mann-Whitney tests). Vertical error bars represent standard error of the mean.

(**D**) All network hubs from AI-1_{MAIN} are PPIN-1 proteins. Relative frequency of degree in AI-1_{MAIN} of i) the 632 PPIN-1 proteins present in AI-1_{MAIN} (pink); and ii) the remaining 2,029 proteins in AI-1_{MAIN} (black). Group (i) shows a significantly higher degree distribution than group (ii); Mann-Whitney test ($P = 1.9 \times 10^{-103}$). The vertical line corresponds to degree of 50.

Fig. 3: Hypothetical complexes in PPIN-1

(**A**) The PPIN-1 sub-network of pathogen effector proteins and their Arabidopsis targets. Color coded nodes (proteins) are protein sub-classes from **Fig. 1A**. Grey edges: experimental interactions from **Fig. 1B**. Green edges: added interactions from AI-1 and

LCI (**fig. S1**). From the total of 165 effector targets, 105 interact with at least one other target, while 41 and 19 interact only with *Hpa* or *Psy* effectors, respectively.

(B) Schematic representation of hypothetical complexes involving effectors and effector targets in PPIN-1 (data extracted from A). Number of proteins (top) and number of interacting pairs (bottom) are indicated for each category of hypothetical complexes.

(C) Schematic representation of novel hypothetical complexes involving immune receptors. The numbers for each category are listed on top.

Fig. 4: Functional validation of host proteins targeted by effectors from both pathogens.

(**A**) Nine host proteins targeted by effectors from both pathogens are required for complete immune function. 12 day old seedlings were inoculated with the avirulent *Hpa* isolates Emwa1 (E1) or Emoy2 (E2). The number of asexual sporangiophores per cotyledon was determined at 5 days post-inoculation (dpi); black, mean number of Emwa1 sporangiophores per cotyledon; 30,000 spores/ml inoculum; red, mean number of Emoy2 sporangiophores per cotyledon; 40,000 spores/ml inoculum). Col-0 and *rpp4* are resistant and susceptible controls for both *Hpa* isolates. The *eds16* mutant is a control for compromised MTI (*36*). For means +/- two times standard error, sample size, additional alleles and independent repetitions see table S9.

(**B**) At least six host proteins targeted by effectors from both pathogens are required for maximal pathogen colonization. 12 day old seedlings were inoculated with 30,000 spores/ml of the virulent *Hpa* isolate Noco2. The number of sporangiophores per cotyledon was determined at 4 dpi. Ws and Col-0 represent the resistant and susceptible controls. For means +/- two times standard error, sample size, additional alleles and independent repetitions see table S9.

(**C**) The *csn5a-2 cul3a* double mutant exhibits enhanced resistance to *Hpa* isolate Emco5. Number of asexual sporangiophores were counted at 5 dpi on at least 50 cotyledons for each of the indicated genotypes. Col-0 and Ler were susceptible and resistant controls, respectively.

(**D**) Western blots with anti-PR1 and anti-CSN5 on crude leaf extract of the indicated genotypes from untreated and infected with *Hpa* Emco5 for 2 days. Ponceau S stain verifies equal loading.

(**E**) Bacterial growth (colony forming unit - CFU, expressed on a log scale) following flg22 (right) or mock treatment (water, left) of leaves of the indicated genotypes followed 24 hours later by infection with *Pto* DC3000. Bacterial growth was assessed at 3 dpi. Error bars represent the mean ± two times standard error of four replicates.

(**F**) Expression of MTI-responsive genes in *pfd6-1*. Relative transcript levels of MTI-responsive genes were determined by quantitative RT-PCR using cDNA generated from leaves treated with flg22 for 45min. The expression values were normalized using the expression level of the UBQ5 as an internal standard. Normalized values of MTI-responsive genes in Col-0 are arbitrarily adjusted to 1. Error bars represent average ± standard deviation of at least two replicates.

Fig. 1

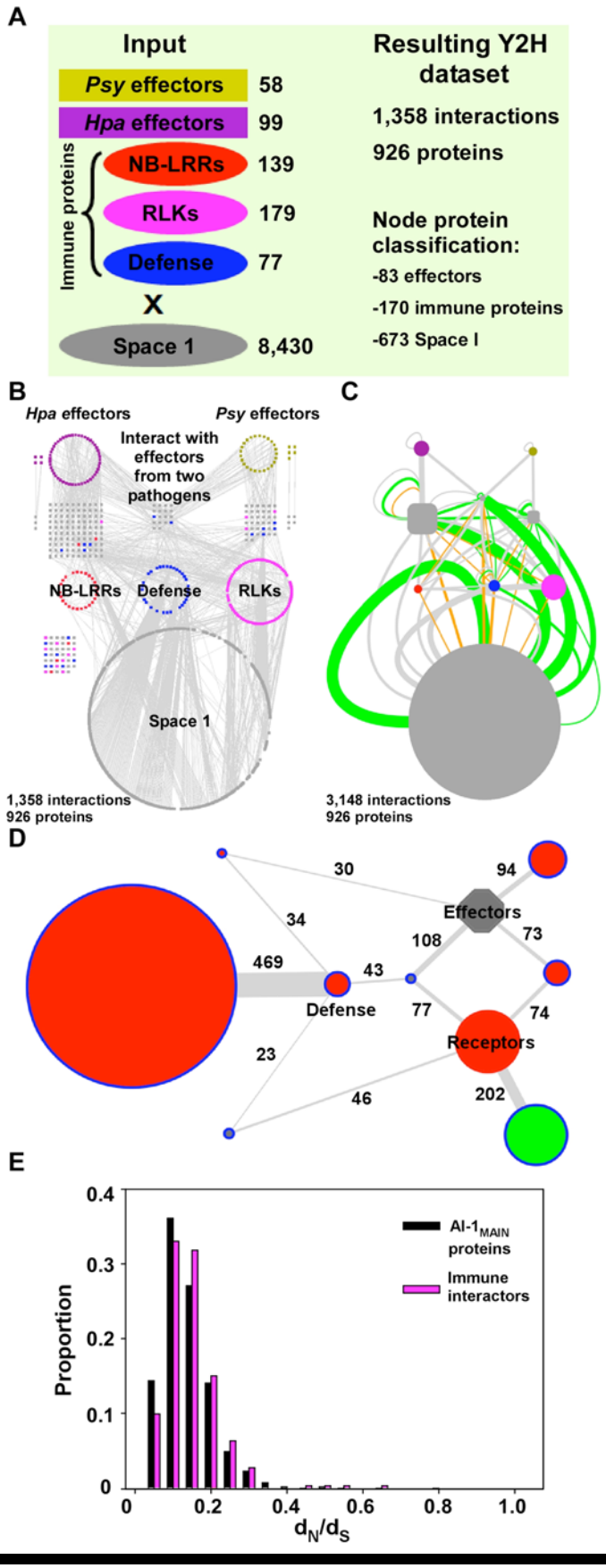

A
Input
Psy effectors 58
Hpa effectors 99
Immune proteins
NB-LRRs 139
RLKs 179
Defense 77
X
Space 1 8,430
Resulting Y2H dataset
1,358 interactions
926 proteins
Node protein classification:
-83 effectors
-170 immune proteins
-673 Space I
B
Hpa effectors
Psy effectors
Interact with effectors from two pathogens
NB-LRRs
Defense
RLKs
Space 1
1,358 interactions
926 proteins
C
3,148 interactions
926 proteins
D
30
94
Effectors
34
108
73
469
43
Defense
77
74
Receptors
23
46
202
E
Proportion
0.4
0.3
0.2
0.1
0
AI-1MAIN proteins
Immune interactors
0
0.2
0.4
0.6
0.8
1.0
dN/dS

Fig. 2

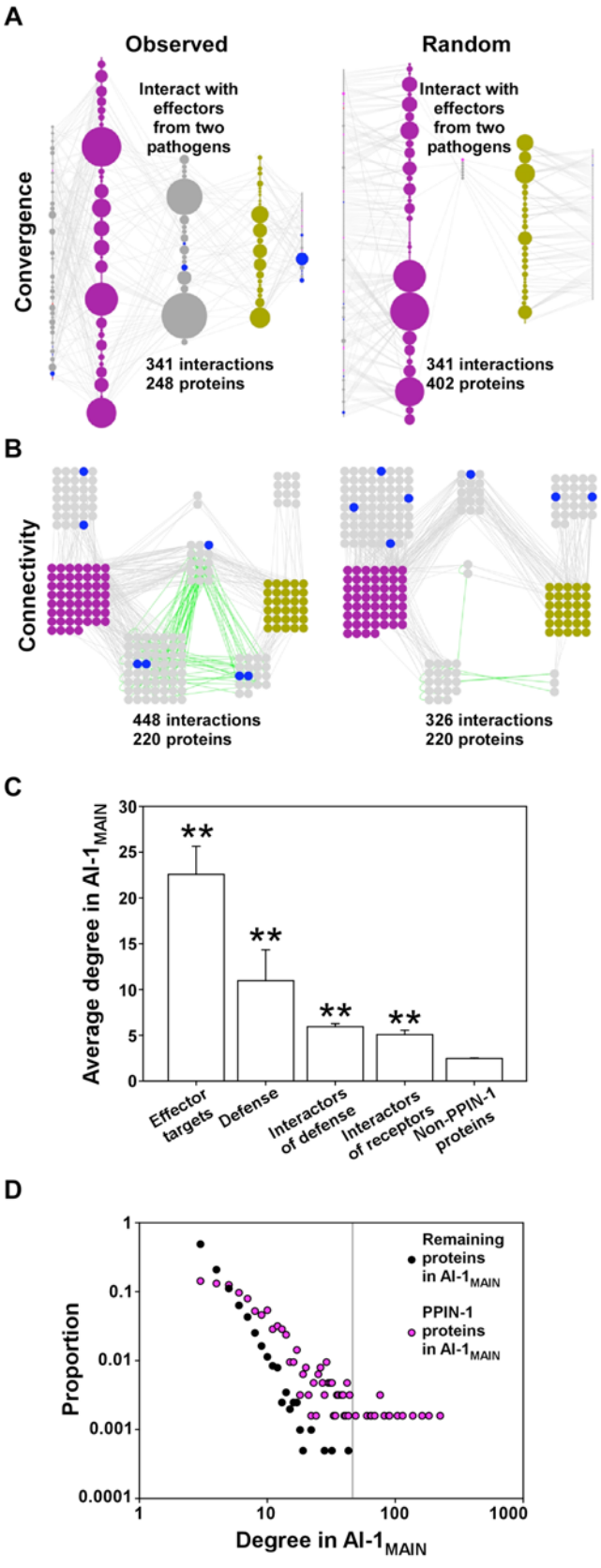

A
Observed
Random
Interact with effectors from two pathogens
Interact with effectors from two pathogens
Convergence
341 interactions
248 proteins
341 interactions
402 proteins
B
Connectivity
448 interactions
220 proteins
326 interactions
220 proteins
C
Average degree in AI-1MAIN
30
25
20
15
10
5
0
**
**
**
**
Effector targets
Defense
Interactors of defense
Interactors of receptors
Non-PPIN-1 proteins
D
Proportion
1
0.1
0.01
0.001
0.0001
1
10
100
1000
Degree in AI-1MAIN
Remaining proteins in AI-1MAIN
PPIN-1 proteins in AI-1MAIN

Fig. 3

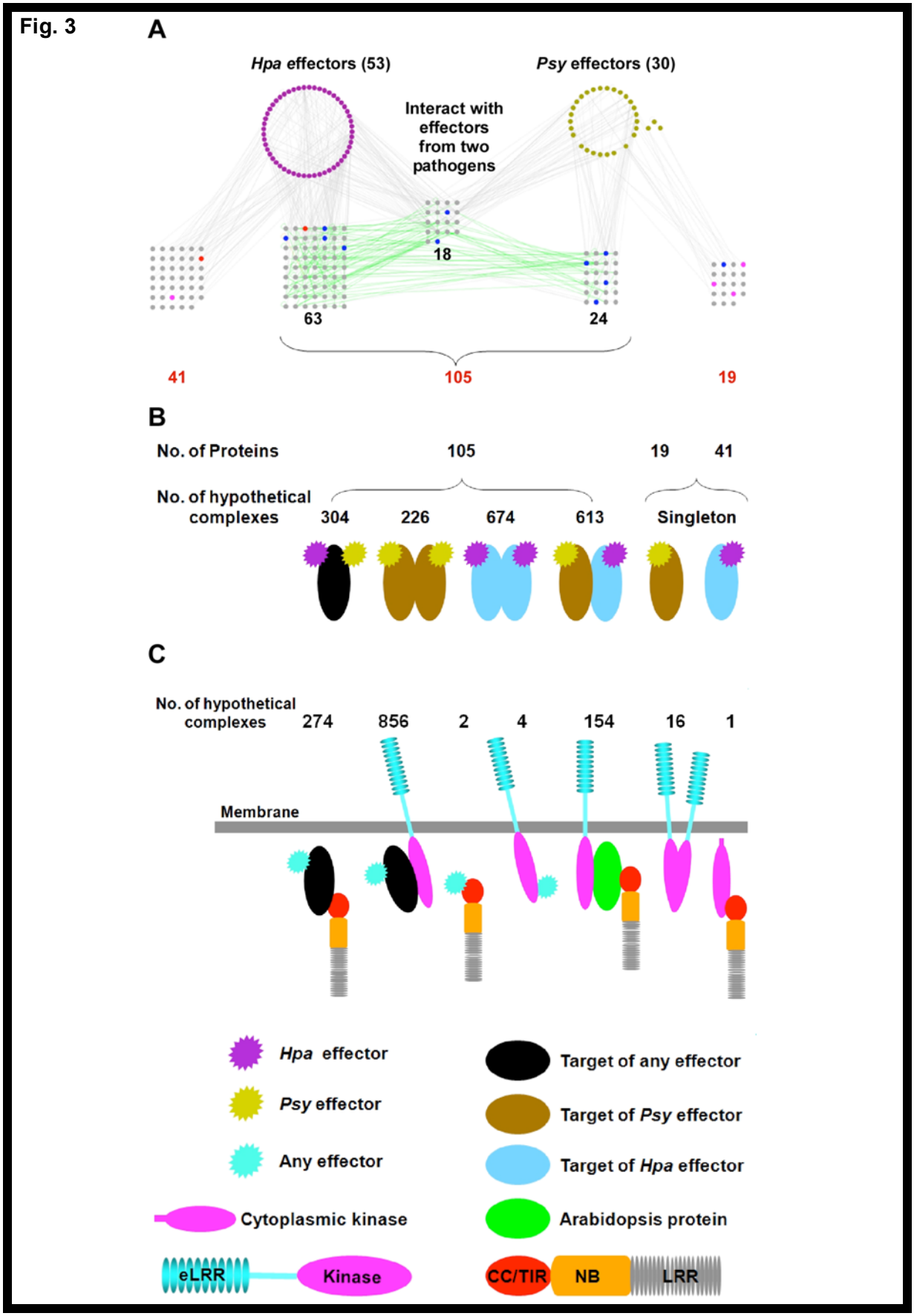
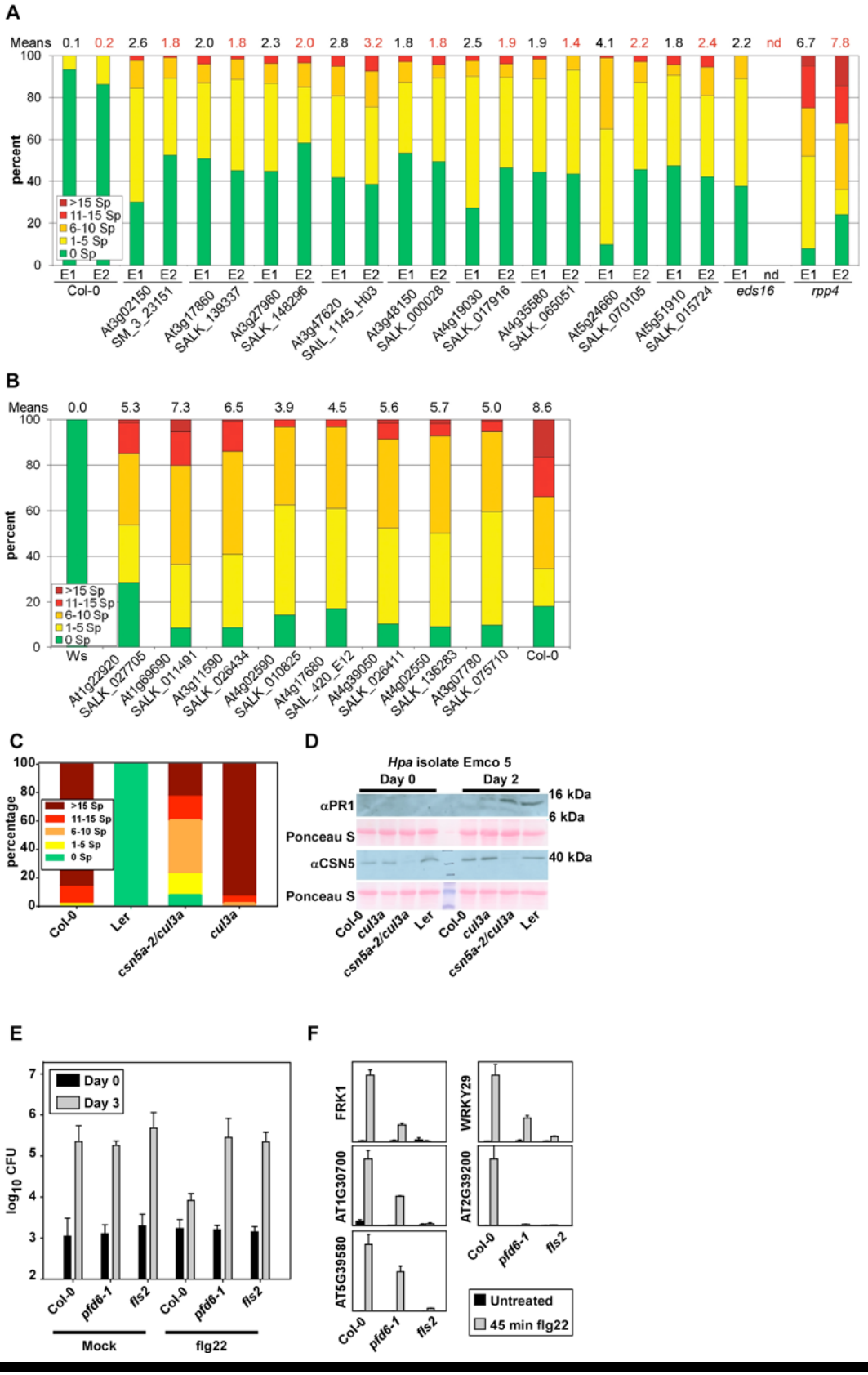

A
Hpa effectors (53)
Psy effectors (30)
Interact with effectors from two pathogens
18
63
24
41
105
19
B
No. of Proteins
105
19
41
No. of hypothetical complexes
304
226
674
613
Singleton
C
No. of hypothetical complexes
274
856
2
4
154
16
1
Membrane
Hpa effector
Psy effector
Any effector
Cytoplasmic kinase
eLRR
Kinase
Target of any effector
Target of Psy effector
Target of Hpa effector
Arabidopsis protein
CC/TIR
NB
LRR

Fig. 4


A
Means 0.1 0.2 2.6 1.8 2.0 1.8 2.3 2.0 2.8 3.2 1.8 1.8 2.5 1.9 1.9 1.4 4.1 2.2 1.8 2.4 2.2 nd 6.7 7.8
percent
>15 Sp
11-15 Sp
6-10 Sp
1-5 Sp
0 Sp
E1 E2
Col-0
At3g02150 SM_3_23151
At3g17860 SALK_139337
At3g27960 SALK_148296
At3g47620 SAIL_1145_H03
At3g48150 SALK_000028
At4g19030 SALK_017916
At4g35580 SALK_065051
At5g24860 SALK_070105
At5g51910 SALK_015724
eds16
rpp4
B
Means 0.0 5.3 7.3 6.5 3.9 4.5 5.6 5.7 5.0 8.6
Ws
At1g22920 SALK_027705
At1g69690 SALK_011491
At3g11590 SALK_026434
At4g02590 SALK_010825
At4g17680 SAIL_420_E12
At4g39050 SALK_026411
At4g02550 SALK_136283
At3g07780 SALK_075710
Col-0
C
percentage
Col-0
Ler
csn5a-2/cul3a
cul3a
D
Hpa isolate Emco 5
Day 0
Day 2
αPR1
16 kDa
6 kDa
Ponceau S
αCSN5
40 kDa
Ponceau S
Col-0 cul3a csn5a-2/cul3a Ler Col-0 cul3a csn5a-2/cul3a Ler
E
log10 CFU
Day 0
Day 3
Col-0 pfd6-1 fls2 Col-0 pfd6-1 fls2
Mock
flg22
F
FRK1
AT1G30700
AT5G39580
WRKY29
AT2G39200
Col-0 pfd6-1 fls2
Untreated
45 min flg22

Materials and Methods:

Selection and cloning of genes encoding Arabidopsis immune proteins.

1- Cytoplasmic domains of leucine-rich repeat (LRR)-containing receptor like kinases (RLKs), a subclass of pattern-recognition receptors (PRRs). The Arabidopsis genome encodes 610 PRRs including 216 RLKs. They consist of an extracellular LRR domain, a transmembrane domain and a cytoplasmic serine/threonine kinase domain (*1*). Several LRR domains have been shown to directly bind ligand, while the kinase domain is vital for downstream signaling. Only a limited number of RLKs have been ascribed any function, and only a few of these are reported to act as immune receptors (*2-5*). Transcriptome data suggests that the overwhelming majority of LRR-Kinase genes are down-regulated by pathogen effector delivery, compared to expression following PRR stimulation (*3, 6, 7*). We included almost a family-wide collection (179) of cytoplasmic domains of leucine-rich repeat (LRR)-containing RLKs (**Fig. 1A, table S1**).

2- N-terminal domains of NB-LRR proteins. Nucleotide binding site-leucine rich repeats (NB-LRR) proteins are closely related to animal NOD-like receptor or CATERPILLAR proteins and form the major R protein class in Arabidopsis with ~150 members (*8-10*). The NB-LRR family of proteins is further subdivided based on the presence of an N-terminal Toll/Interleukin-1 Receptor (TIR) or coiled-coil (CC) motif. NB-LRR proteins likely exist in resting state intra- and intermolecular folded conformers. N-terminal domains of NB-LRRs are thought to be negatively regulated by the LRR domain. Effector and/or effector-target binding is thought to relieve this intra-molecular repression and allow nucleotide binding and signal competence. Moreover, N-terminal domains of NB-LRRs can be involved in association with either cellular targets of effector action or with recruitment of downstream signaling components (*8-10*). Thus, we cloned sequences encoding N-terminal domains of 139 NB-LRRs (see below). In some cases, we also included full-length or other than N-terminal domains of an NB-LRR. In total, we included 147 clones corresponding to 139 loci (**Fig. 1A, table S1**).

3- Effector molecules from the bacterial pathogen *Pseudomonas syringae* (*Psy*) and the oomycete pathogen *Hyaloperonospora arabidopsidis* (*Hpa*). These effector proteins are critical virulence determinants that target host proteins. While Gram-negative bacteria use the type III secretion system to deliver type-III effectors, little is known about the mechanism(s) of delivery of oomycete effectors into host cells (*11*). Oomycete cytoplasmic effectors are modular proteins that carry N-terminal signal peptides followed by conserved motifs, notably the RXLR and LXLFLAK motifs (*12*). We cloned 101 coding sequences for translocation confirmed *Psy* type III effectors from 16 different bacterial strains (*13*). Moreover, we also cloned coding sequences for 130 *Hpa* RXLR/LXLFLAK candidate effector proteins from 17 different isolates (Ahco2, Aswa1, Bico1, Bico5, Cala2, Cand5, Emco5, Emoy2, Emwa1, Hiks1, Hind2, Hind4, Maks9, Noks1, Waco5, Waco9, Wela3). The *Hpa* candidate effectors were cloned from spore cDNA from the predicted signal peptide cleavage site (or otherwise stated in table S1) to the predicted stop codon (*14*). For network analyses, we collapsed alleles and domain subclones corresponding to the same effector protein and thus generated 58 and 99 effector groups for *Psy* and *Hpa*, respectively (**Fig. 1A, table S1**).

4- Defense proteins. We included 92 clones corresponding to 77 known signaling components and previously described host targets of pathogen effectors. Collectively, we refer to this sub-class as “defense proteins” (**Fig. 1A, table S1**).

5- Cloning of genes encoding immune proteins. Total RNA was isolated from Arabidopsis (ecotype Columbia-0, unless noted) leaves using Trizol reagent (Invitrogen). First-strand cDNA was synthesized using RETROscript reverse transcriptase (Ambion).The cDNA products were amplified using AccuPrime. *Pfx* DNA Polymerase (Invitrogen). For cloning of *Psy* type III effector and *Hpa* RxLR effector encoding genes, DNA was isolated from the appropriate strain/isolate and used for PCR. The PCR primers contained the *att*B1 and *att*B2 or CACC sequences for cloning PCR products into pDONR series of vectors or pENTR-D-TOPO series of vectors, respectively by Gateway BP recombinational cloning (Invitrogen). For the RLKs constructs, Gateway entry clones were obtained from ABRC unless mentioned otherwise, consisting of the cytoplasmic kinase domain (juxtamembrane region, catalytic kinase domain and carboxy terminal region) of an LRR RLK in the vector pDONR/zeo from Invitrogen. Stock numbers of the RLKs clones that obtained from ABRC are given in table S1. Domains of NB-LRR proteins were predicted by TAIR. DNA sequences upstream of NB-ARC encoding region were considered N-terminal region (for both CC and TIR) and cloned. pENTR clones were transferred into pDEST-DB (bait; DB-ORFs) and pDEST-AD (prey; AD-ORFs) vectors by Gateway LR recombinational cloning (Invitrogen) according to the manufacturer's instructions.

Yeast Two-Hybrid (Y2H) Screening.

The detailed procedure for Y2H screening is described in Dreze *et al.* (*15*). This strategy was applied to identify interactions in a systematic manner between pathogen effectors, NB-LRRs, RLKs, defense proteins and proteins in Space 1. Briefly, the yeast strains Y8930 (*MAT*α) and Y8800 (*MAT***a**) were transformed with individual DB-ORFs and AD-ORFs, respectively resulting in DB-X and AD-Y yeast strains. Prior to Y2H selections, each DB-X yeast strain was examined for auto-activation of the *GAL1-HIS3* reporter gene in the absence of any AD-Y. All yeast strains showing elevated expression of the *GAL1-HIS3* reporter gene were removed from the collection while non auto-activating DB-X strains were used for Y2H selections. In a first step, each DB-X yeast strain was tested for possible interaction against mini-libraries of 192 AD-Y yeast strains. The phenotype of the resulting primary positives was then tested again in a second step and only those whose phenotype could be confirmed (secondary positives) were retained for identification of DB-X and AD-Y pairs by end-read sequencing of PCR products amplified directly from yeast cells. In a third and final step, the phenotype of these candidate Y2H interactions was then verified in a pairwise manner (1 DB-X vs 1 AD-Y). The phenotype of each candidate Y2H interaction was tested four times, by four different experimenters, to increase reproducibility and reliability of interactions. In addition, at each of the three processing steps, we tested the phenotype of DB-X yeast strains for possible spontaneous DB-X auto-activation events and removed them when detected. We also tested the phenotype of each AD-Y yeast strains for infrequent yet possible AD-Y auto-activation and removed them when observed. In sum, only pairs whose phenotype could be verified at least three times out of four and that did not show any auto-activation were considered Y2H interactors. Since the identification of candidate Y2H interactions is done based on end-read sequences, it is difficult to differentiate nearly identical sequences, i.e. alleles, hence to assign interactors to the correct allele. To circumvent this problem, in the third step of

the Y2H strategy we systematically verified, within group of alleles, each interactor against each allele.

The extremely low number of available literature-curated effector-Arabidopsis interactions (*16*) precluded the construction of a reliable positive reference set for use in estimating the pathogen-host-specific interaction quality. Previous interactome screens with ORF clones from various organisms (*17-21*) were estimated to have similar precision, supporting the notion that the reliability of this screen technology is species-independent. For full details on the precision computation, please refer to the accompanying paper.

Statistical analyses.

Probabilities for the following overlaps between pairs of datasets were estimated using a hypergeometric test:

- Immune interactors and GO-immune annotated proteins from Space 1
- GO-immune proteins and effector targets
- Hormone-related proteins and effector targets
- $Hubs_{50}$ and effector targets
- Indirect connections between effectors and receptors (NB-LRRs and RLKs)
- Effector targets with angiosperm only orthologs and more broadly conserved orthologs.
- Proteins encoded by differentially expressed defense (DE) genes and proteins in PPIN-1 (all and subgroups)

Contingency tables for these tests are presented in **tables S4** and **S7.**

GO-term (*22*) enrichments for the effector target proteins were estimated using the FuncAssociate R library (*23*), with a false discovery rate cutoff of 20%, using proteins in $AI\text{-}1_{MAIN}$ as a reference set to control for ORF collection and Y2H potential biases. Only targets that were not in the original immune protein sets were considered for this analysis. The results describing more than 10% of these effector targets and representing an enrichment of more than 1.25 are presented in **table S5**.

Identification of ortholog clusters

We identified ortholog clusters between proteins in Arabidopsis and other species pairwise using the InParanoid resource (*24*). We chose the following species to provide a broad taxonomic sample between Arabidopsis and more distant taxa: *Populus trichocarpa, Oryza sativa, Sorghum_bicolor, Physcomitrella patens, Cyanidioschyzon merolae, Ostreococcus tauri, Chlamydomonas reinhardtii, Thalassiosira pseudonana, Neurospora_crassa, Saccharomyces_cerevisiae, Caenorhabditis_elegans, Drosophila_melanogaster, Homo_sapiens, Eschericia coli* K12. For each Arabidopsis gene we identified its most distant ortholog(s), and its 'phylogenetic footprint' according to the above taxonomic sampling scheme. To control for homoplasy, we identified instances of phylogenetic footprints being sparsely populated but over-represented in our data, in other words, cases where an Arabidopsis gene apparently had a very distant ortholog but no or very few orthologs in other intermediate species, and where such footprints occurred at least 1.5 times more frequently than expected given the overall proportions of orthologs we found in each species. In these cases, we removed the most distant ortholog from the footprint. The main effect of this filtering was that a number of genes with apparent orthologs in animal taxa, but none in other more closely related taxa, were reassigned to being Arabidopsis-specific genes.

Having identified a definitive list of 'most distant ortholog(s)' for each Arabidopsis gene, where we also had evidence of orthologs being present in more closely related taxa, we classified each Arabidoipsis gene as being 'angiosperm-specific' (where no ortholog was found, or where the most distant ortholog was found in *P. trichocarpa, O. sativa* or *S.bicolor*), or 'more broadly conserved' (where orthologs were found in angiosperms and also in more distant taxa). Angiosperm-specific proteins are over-represented among the effector targets that are present in Space 1, in comparison with all of Space 1, (hypergeometric $P = 0.0007$; **table S4**).

List of immune-related GO-terms (table S3).
These GO-terms were chosen by the authors in order to best describe the plant immune system and in particular the immune proteins used in the experiment.

Network analyses.
All calculations and simulations regarding network properties were performed using the R implementation of igraph (*26*). All network representations were drawn using Cytoscape (*27*).

d_N/d_S measurement.
Gene models of orthologs between Arabidopsis and Papaya genes were identified by aligning coding sequences using BLAST (*28*) and selecting reciprocal best hits. Only one-to-one orthologs were processed further. Coding sequences of orthologous gene pairs were aligned in protein space using a custom workflow employing Clustal (*29*). The ratio of non-synonymous to synonymous changes in coding sequences (d_N/d_S) was estimated for each aligned gene pair using the maximum likelihood codon substitution model method of (*30*) in the PAML package codeml program (*31*), allowing d_N/d_S to vary between branches, and estimating kappa and omega parameters from data. Values of d_S were found saturated in a small fraction of orthologs, so subsequent analyses were carried out both with and without the 15% of ortholog pairs for which $d_S>5$. Removal of these d_S-saturated orthologs did not qualitatively affect our results, but increased the observed significance levels. Significant differences between distribution of d_N/d_S within each subgroup of genes in the pathogen network, and distribution of d_N/d_S among all genes in AI-1_{MAIN}, were identified using a Kolmogorov-Smirnov test of significance. Poplar orthologs were also tried, as a basis for d_N/d_S calculations, and results were found positively correlated with those calculated from Papaya orthologs ($r=0.84$ for immune network genes), but were found to be d_S–saturated in many more cases. Papaya was ultimately used as its genome is fully sequenced, it shares membership of the Brassicales with Arabidopsis so d_N/d_S represents more recent evolutionary events, and the two are sufficiently diverged for d_N/d_S to be estimated, but not so distant that d_S is saturated, in most cases.

Differentially Expressed (DE) genes.
Results were mined from nine previously published studies of transcriptional responses of Arabidopsis to pathogen or other immune system related perturbations (**table S7** (*3, 32-39*). Priority was given to well-referenced studies, employing the Affymetrix ATH1 GeneChip array, encompassing overall a broad range of different perturbations. Lists of probes showing significant up and down regulation in each experimental condition were compiled, using criteria for significance employed in the respective original study. Probe lists were mapped to the TAIR7 genome, and filtered to only include genes corresponding to unique proteins in the AI-1_{MAIN} network with a TAIR7 gene model. Subgroups of proteins in PPIN-1 were tested for enrichment or depletion of differentially regulated genes in individual experiments, and for the number of times genes were differentially regulated

across all experiments, in comparison to proteins in AI-1_{MAIN}, using hypergeometric test (**table S7**).

Plant materials and growth conditions.

We used *Arabidopsis thaliana* Columbia (Col-0) unless mentioned otherwise. Insertion mutants are listed in **table S9.** Three additional T-DNA knock-out lines for At3g47620 (TCP14), *tcp14-4* (GK-861-G08), *tcp14*-5 (GK-611-C04) and *tcp14-6* (SAIL_1145_H03)) were obtained from the European Arabidopsis Stock Center (NASC;(*40, 41*). *tcp14-1*, *tcp14-2* and *tcp14*-3 are previously described (*42*). *cul3a* (SALK_050756) and *csn5a-2* (SALK_027705)/*cul3a* (SALK_050756) double mutant were provided by Xing-Wang Deng (*43, 44*) and the *pfd6.1* (CS16396) mutant was a gift from Chris Somerville's lab (*45*). Plants were grown under short day conditions (9 hrs light, 21°C; 15 hrs dark, 18°C except for *tcp14* mutants, which were grown under 10 hrs light 14 hrs dark cycle).

***Hyaloperonospora arabidopsidis* (*Hpa*) isolates, infections, and growth assays.**
Hyaloperonospora arabidopsidis (*Hpa*) isolates Emwa1, Emoy2, Emco5, Noco 2, or Noks1 were propagated on the susceptible Arabidopsis ecotypes Ws-2, Oy-1 and Col-0, respectively (*46,47*) 12 day old seedlings were infected with conidiospores suspended in water at the appropriate concentration (30,000 spores/ml, *Hpa* Emwa1; 40,000 spores/ml, *Hpa* Emoy2; 30,000 spores/ml *Hpa* Noco2; 50,000 spores/ml *Hpa* Emco5) and three week old adult plants were infected with 10,000 spores/ml, *Hpa* Noks1 (*48*). Plants were kept covered with a lid to increase humidity and grown at 20°C with a 9 hr light period.

Sporangiophores were counted on cotyledons at 4 or 5 dpi as described (*49*). The number of sporangiophores per cotyledon was determined and percentages were calculated as described (*49*). For Noks1, the number of spores per fresh weight of 10 plants was calculated at 6 dpi (6 replications per experiment).

Bacterial infection experiments.
The flg22-dependent MTI experiment was performed as reported in (*6*). Briefly, four-week old plants were first injected with 1uM flg22 for 24h (-1 day). flg22-injected leaves were infiltrated with a concentration of 10^5 cfu/ml (OD: 0.0002) of *Psy via* a needle-less syringe. Plants were covered for ~24h post-inoculation with a lid. Leaf discs were cored from the infiltrated area at the day of infiltration (0 dpi) and 3 dpi, ground in 10 mM $MgCl_2$, and serially diluted to measure bacterial numbers. For each sample, four leaf discs were pooled six times per data point (24 leaf discs total).

Detailed rationale for validation of Prefoldin 6 (PFD6; At1g29990).
PFD is a heterohexameric molecular chaperone family of proteins found in the cytosol of archaea and eukaryotes but absent from bacteria and required in protein folding complexes (*50*)**.** Prefoldin works in combination with other molecular chaperonins to correctly fold nascent proteins. Prefoldin 6 interacts with EDS1, an essential component of both MTI and some ETI pathways that also plays a key role in salicylic acid dependent plant defense signaling pathways (*51*). Both HopAO1 and AvrPto, which also interact with Prefoldin 6, can suppress MTI (*2, 4, 5, 16*). Prefoldin 6 is required for normal microtubule dynamics and organization in *Arabidopsis* (*45*). The *pfd6-1* mutant exhibits a range of microtubule defects, including hypersensitivity to oryzalin, defects in cell division, cortical array organization, and microtubule dynamicity. Oryzalin is a dinitroaniline herbicide that sequesters tubulin dimers (*45*) Moreover, flg22-dependent endocytosis of the FLS2 LRR-K PRR was inhibited by oryzalin (*52*).

Real time RT-PCR and Western blotting

Total RNA was isolated using Trizol reagent (Invitrogen) from 100 mg fresh tissue that was treated with 1uM flg22 for 45 minutes via syringe-infiltration. DNase treatment was performed using the DNA-*free* reagent (Ambion) for 20 min at 37ºC, according to the manufacturer's instructions. 3 mg RNA was used as starting template material for first strand cDNA synthesis using RETROscript reverse transcriptase (Ambion). Real time PCR was performed using the primers corresponding to MTI-responsive markers as indicated in Fig. 4C. For detection of PR-1 accumulation, total protein extracts were prepared from leaf tissue infected with *Hpa* isolates Emco5 for 2 days as described above or injected with virulent bacterial strain *Pto* DC3000 at a concentration of 10^5 cfu/ml via a needless syringe for 24h. Western blots were performed by using standard methods (*53*). Anti-PR1 serum (gift of Dr. Robert A. Dietrich, Syngenta, Research Triangle Park, NC) was used at a dilution of 1:10,000. Anti-CSN5 (Z04911; BML-PW8365-0100) detects both CSN5a and CSN5b subunits, and was used according to the manufacturer' instructions (BIOMOL).

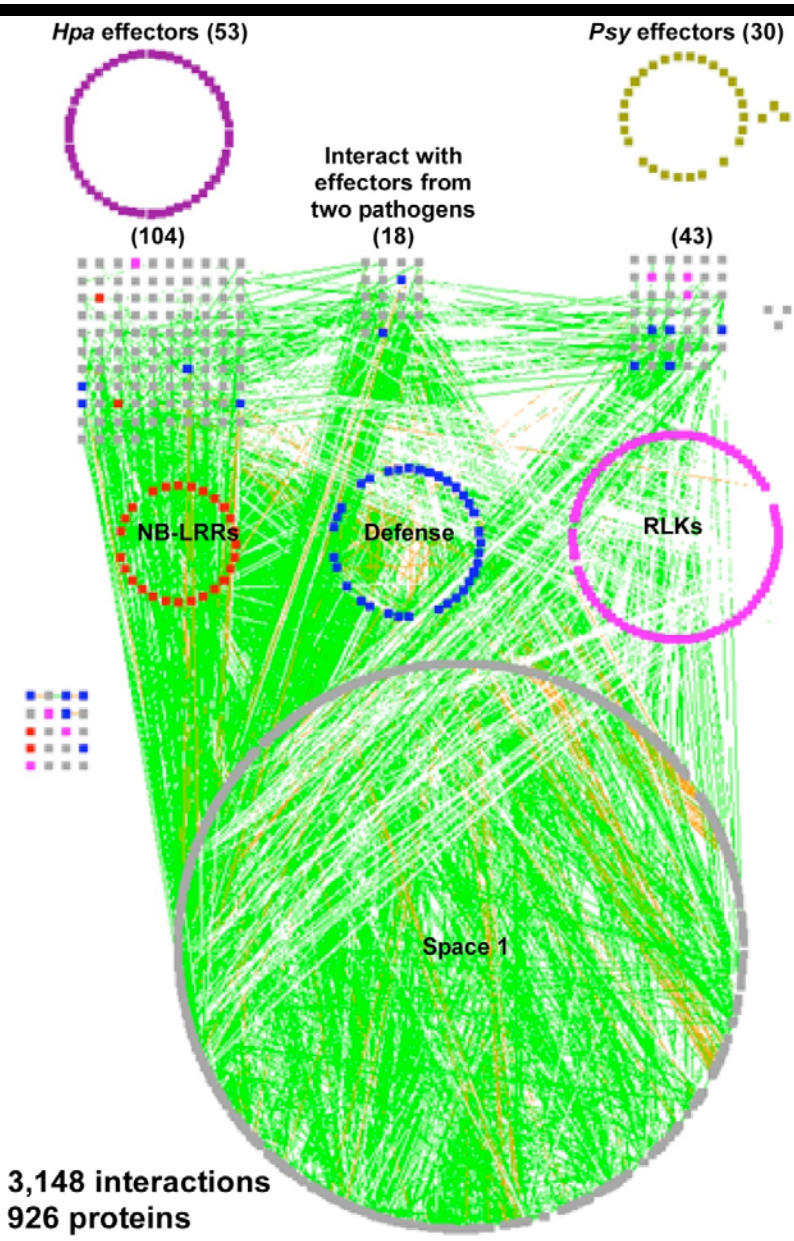


Fig. S1 PPIN-1 is densely connected. The network shown in **Fig. 1B** was integrated with the HT-Y2H network AI-1 (*15*) and with a compendium of protein interactions assembled from TAIR (*54*), IntAct (*55*) and BIOGRID (*56*) named the literature-curated interactions (LCI) network (see (*15*) for assembly methods). Nodes (representing proteins) are colored according to protein sub-classes in **Fig. 1A**. Number of nodes corresponding to each protein sub-class is indicated. Edges represent protein-protein interactions. Plant immune interactions (grey edges) from **Fig. 1A** are omitted in **fig. S1** to emphasize the connectivity acquired from AI-1. Green edges: added interactions from AI-1 and LCI. Orange edge: immune, AI-1 and LCI. Individual interactions that are not connected in the layered network involving *Psy* effectors are indicated next to their relevant protein categories in first and second layers. Grid at left denotes individual interactions involving proteins other than pathogen effectors. Note that the number of individual interactions in the grid is decreased in **fig. S1** compared to **Fig. 1B** due to increased connectivity by integrating HT-Y2H network AI-1 data.

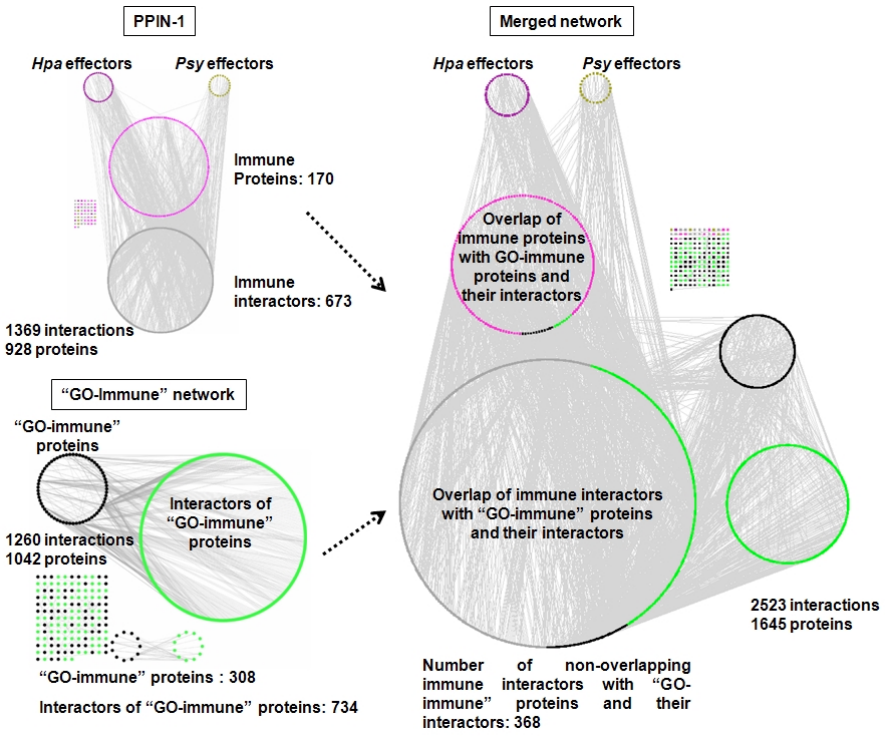


fig. S2 The overlap between plant immune network and "GO-immune" network suggests that at least 368 novel proteins play a role in plant immunity.
PPIN-1 (top left panel) is represented here in three layers: effectors of both pathogens (top), plant immune proteins (pink; middle) and the 673 immune interactors from Space 1 (grey; bottom). "GO-immune" network (bottom left panel) is derived from 308 AI-1_{MAIN} proteins annotated as "GO-immune proteins" (black) and their 734 'interactors of GO-immune proteins' (green). Merging of PPIN-1 and the "GO-immune" network (right panel) maps the non-overlapping proteins between the two networks (368). Individual interactions that are not connected in the layered networks are present in the grid of their respective network.

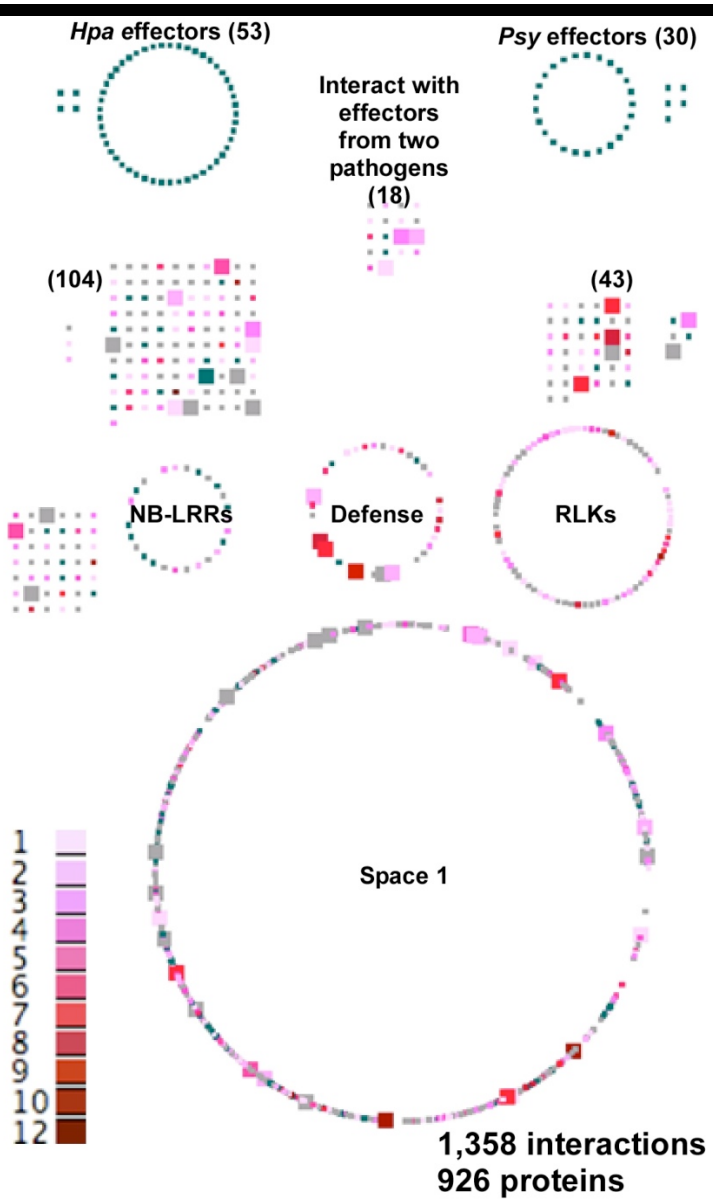


fig. S3 PPIN-1 is enriched in differentially expressed (DE) genes from defense-related experiments.
PPIN-1 proteins in layered layout (as in **Fig.1B**) but re-colored at a scale from pale pink (1) to dark red (12) according to the number of occurrences of proteins in differentially expressed (DE) genes lists from defense-related transcriptome experiments. Grey: proteins not in DE genes lists. Green: no information present for that particular locus/pathogen effector in defense-related transcriptome. Number of nodes related to several protein sub-class is listed. Larger node squares represent proteins GO-annotated as 'hormone response'. Plant immune interactions (grey edges) from **Fig. 1B** are omitted in **fig. S2** to emphasize the node colors and sizes.

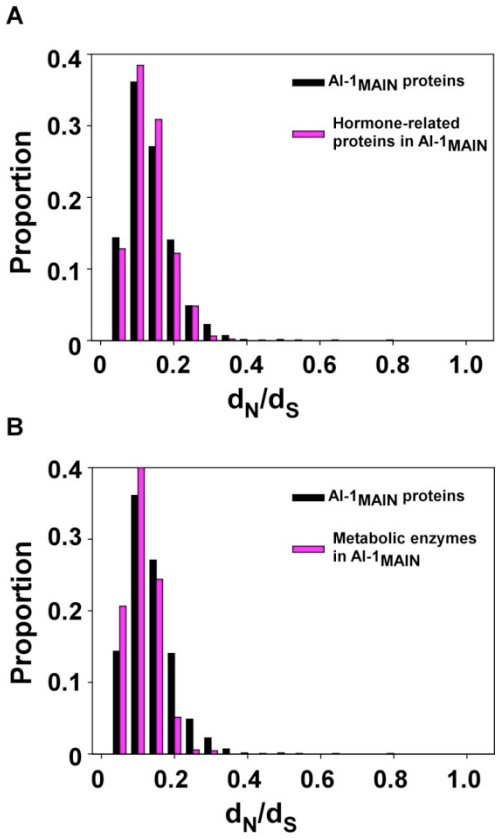


fig. S4 Controls for the evolution rate of PPIN-1 proteins.
In order to verify the specificity of our observation that immune interactors in PPIN-1 evolve faster than other proteins in AI-1_{MAIN}, we performed the same analysis on hormone-related proteins (*25*) and metabolic enzymes (*57*).

(**A**) Relative frequency of d_N/d_S between hormone-related proteins (*25*) in AI-1_{MAIN} and all proteins present in the AI-1_{MAIN} network. A Kolmogov-Smirnov test shows that these distributions are not statistically different. d_N/d_S values are computed between Arabidopsis proteins and their Papaya orthologs. Black and pink bars represent AI-1_{MAIN} and hormone-related proteins, respectively.

(**B**) Relative frequency of d_N/d_S between the metabolic enzymes present in AI-1_{MAIN} and all proteins present in the AI-1_{MAIN} network. A Kolmogov-Smirnov test shows that metabolic enzymes evolve slower than all proteins in AI-1_{MAIN} ($p<1e-22$) d_N/d_S values are computed between Arabidopsis proteins and their Papaya orthologs. Black and pink bars represent AI-1_{MAIN} and metabolic enzymes, respectively.

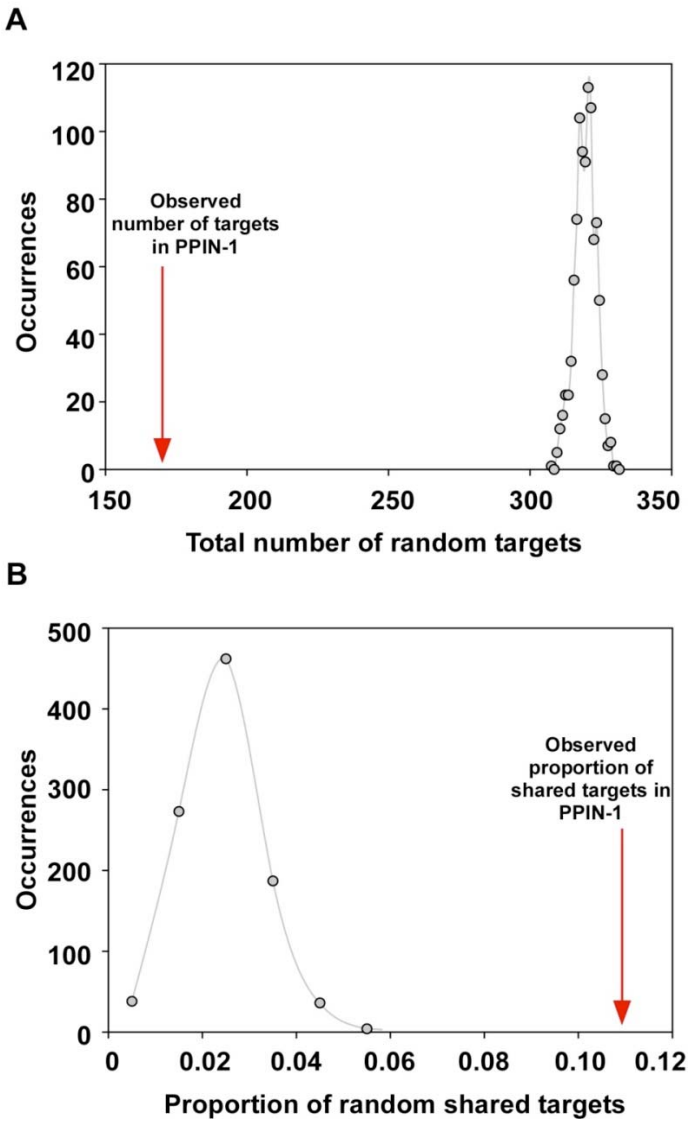


fig. S5 Computational simulations of random targeting of Arabidopsis proteins by effector targets: convergence.
In order to estimate how many Arabidopsis targets of pathogen effectors would be expected at random, we performed 1000 computer simulations as follows. We considered the union of all PPIN-1 and AI-1 Arabidopsis proteins as "Y2H-amenable". We counted the number of Arabidopsis proteins each effector connected to in PPIN-1 and we then selected the same number of proteins randomly from the set of Y2H-amendable proteins. After having repeated this process for all effector proteins in PPIN-1, we counted (1) the total number of Arabidopsis proteins selected by this process (total number of random targets, **fig. S5A**) and (2) the proportion of these random targets found connected by effectors from the two pathogen species by chance (proportion of random shared targets, **fig. S5B**).

(**A**) Distribution of the total number of random effector targets in 1000 simulations. The red arrow at 165 represents the observed number of effector targets in PPIN-1 (**Fig. 1B, 3A**). The observed number of targets is significantly smaller than random expectation ($p < 0.001$), consistent with convergence of effectors onto a specific set of proteins.

(**B**) Distribution of the proportion of random shared targets in 1000 simulations. The red arrow at 0.11 (18 / 165) represents the observed proportion of shared targets in PPIN-1 (**Fig. 1B, 3A**). The observed proportion of shared targets is significantly larger than random expectation ($p < 0.001$), also supporting convergence of effectors onto a specific set of proteins.

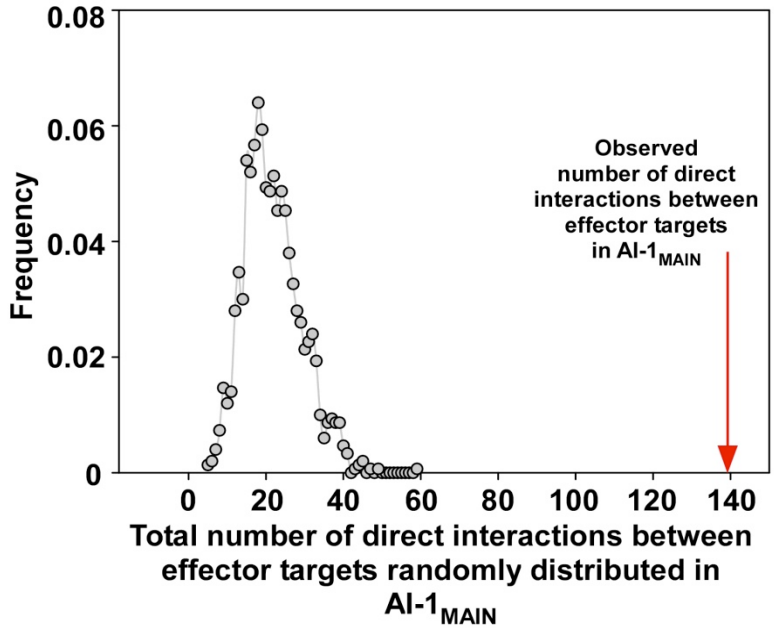


fig. S6 Computational simulations of random targeting of Arabidopsis proteins by effector targets: connectivity.

In order to estimate how many direct interactions between effector targets would be expected at random, we considered that the number of effector targets remains constant (as opposed to **fig. S5**), but that they are randomly distributed in AI-1_{MAIN}. We performed 15000 simulations where we shuffled the names of the proteins in AI-1_{MAIN} while keeping the network structure intact, and counted the number of direct interactions between targets in the shuffled network.

fig. S6 shows the distribution of number of direct connections between effector targets in the shuffled networks ranging from 5 to 59. The red arrow at 139 represents the observed number of direct interactions between effector targets present in AI-1_{MAIN}. Observed number of direct interactions between effector targets is significantly larger than random expectation ($p<0.00015$), suggesting high connectivity between effector targets and thus the existence of defense molecular machines.

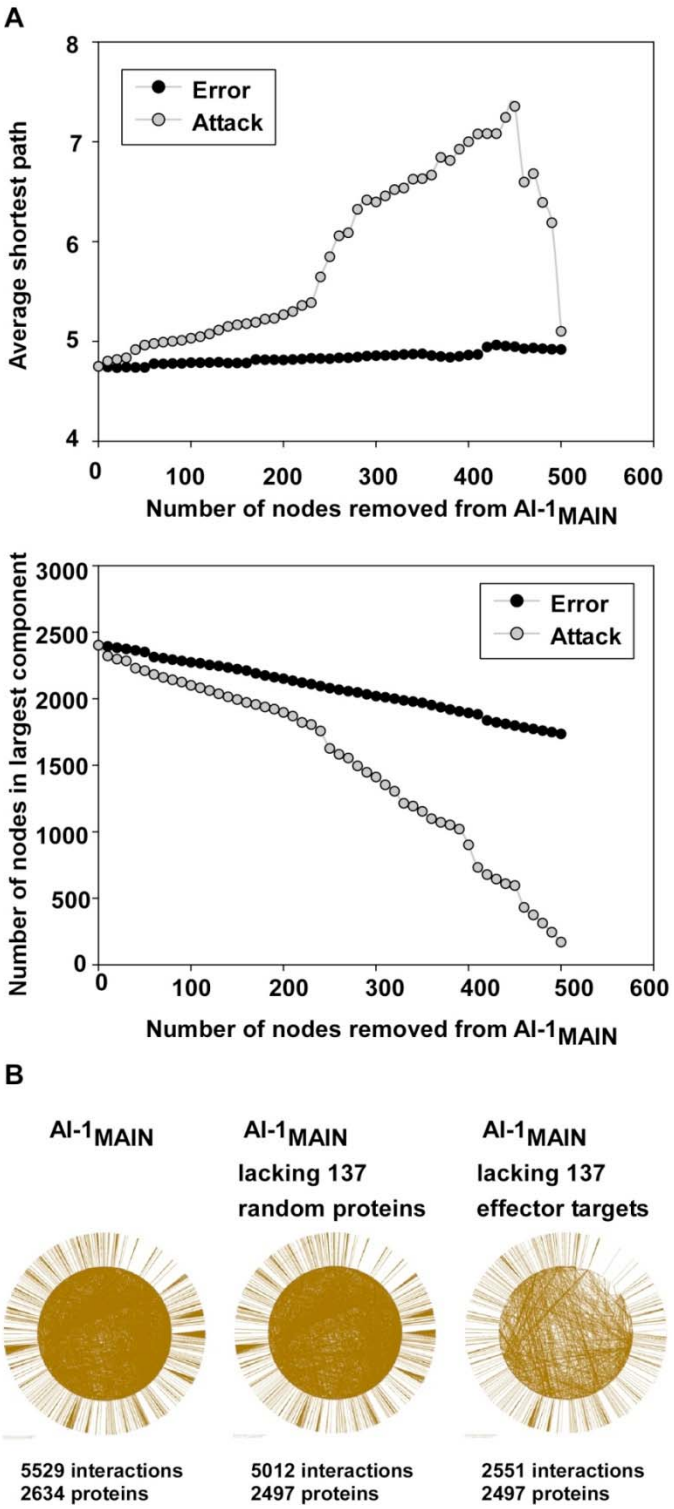


fig. S7 Computational simulations of targeted attacks and random failures on AI-1$_{MAIN}$.

We aimed to determine if AI-1$_{MAIN}$ shared the property of scale-free networks to be resistant to random errors but sensitive to targeted attacks of their hubs (*58*). (**A**) Evolution of the average shortest path (top panel) and the number of nodes in the largest component

(bottom panel) upon random removal of the same number of nodes, either chosen among all nodes ("errors") or only among the nodes of degree >=5 ("attacks"). These simulation show that AI-1_{MAIN} shares the property of scale-free networks to be resistant to random errors but sensitive to targeted attacks of their hubs. (**B**) Circular representation of AI-1_{MAIN} (left), AI-1_{MAIN} after removal of 137 random nodes and the corresponding edges (middle), AI-1_{MAIN} after removal of the 137 effector targets identified in PPIN-1 also present in AI-1_{MAIN} and the corresponding edges (right). All nodes represent proteins and are arranged in a circle; all edges represent interactions; all self loops are eliminated.

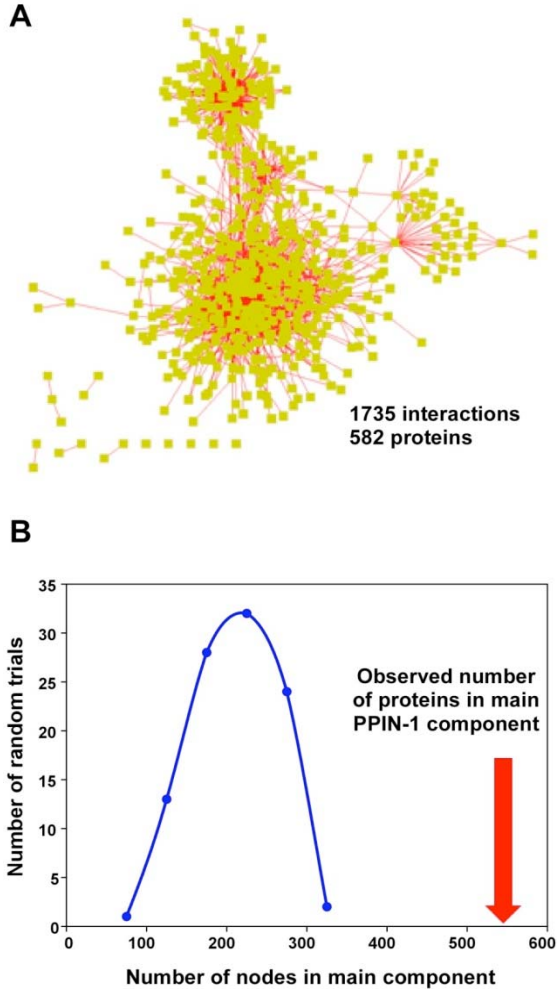


fig. S8 Proteins in PPIN-1 are densely connected in AI-1_{MAIN}.
To evaluate the extent to which the proteins of the plant immune network were connected in the AI-1_{MAIN}, we first calculated that the 632 proteins present in PPIN-1 and also present in AI-1_{MAIN} formed a subnetwork of 582 proteins including 566 forming a single component (**fig. S8**). We then performed 100 random selections of 632 proteins in AI-1_{MAIN} and measured the number of nodes of the largest components of these random controls. This number never reached 566 making the empirical *P*-value for our observation < 0.01. AI-1_{MAIN} (**fig. S8B**). (**A**) A sub network of AI-1_{MAIN} containing 582 PPIN-1 proteins (node; gold) and 1735 Y2H interactions (edges; red) includes 566 nodes and 1723 edges in a single component. (**B**) The graph shows the number of nodes forming the largest component of 100 bootstrapped networks generated by selecting 632 proteins randomly from AI-1_{MAIN} (bottom); the red arrow indicates the observed number of PPIN-1 nodes in the largest component: 566 (SOM).

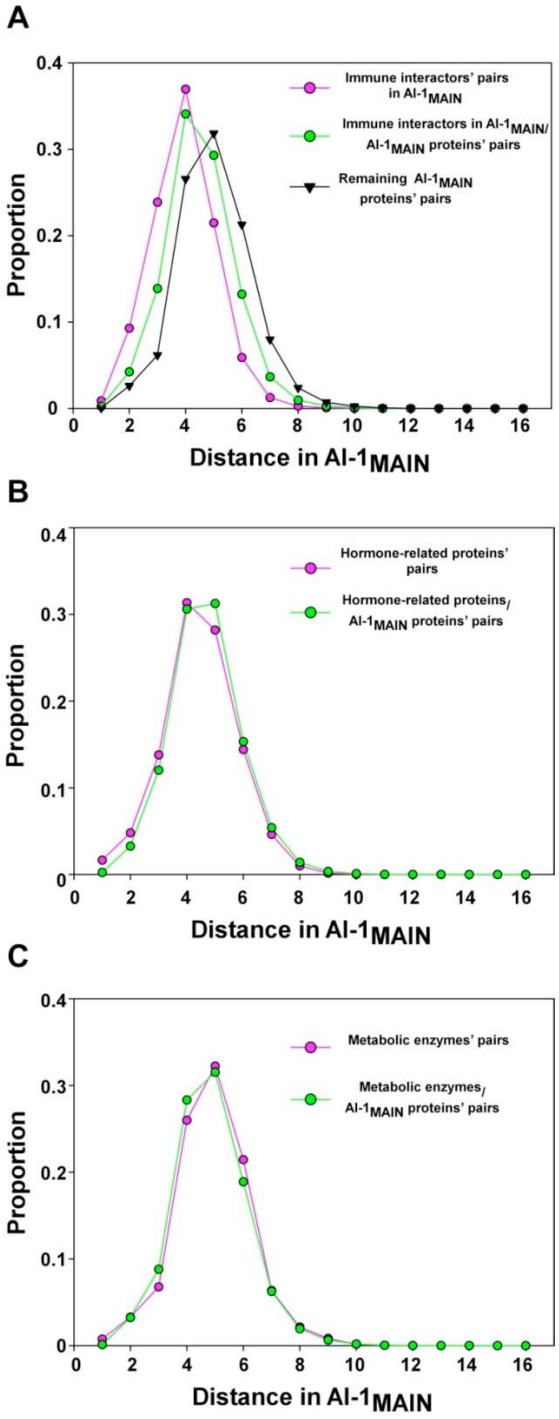


fig. S9 Proteins in PPIN-1 are close to each other in AI-1_{MAIN}
We considered 3 groups of proteins in AI-1_{MAIN}:i) Immune interactors present in both PPIN-1 and AI-1_{MAIN}, (grey nodes in **Fig. 1B**), ii) hormone-related proteins (*25*), and iii) metabolic enzymes (*57*). For each of these groups, we compared the distribution of pairwise shortest paths in AI-1_{MAIN} for pairs of proteins within the group (pink), to pairs consisting of one

protein of the group and one other protein in AI-1_{MAIN} (green) and protein pairs from the remaining of AI-1_{MAIN} (black).

(A) Immune interactors present in AI-1_{MAIN}. The distances between proteins in the first group are significantly shorter than those in the second group according to a Mann Whitney test (p < 2.2e-16). In black is the distribution of pairwise shortest paths in AI-1_{MAIN} for protein pairs that are not present in PPIN-1.

(B) Hormone-related proteins. The distances between proteins in the first group are significantly shorter than those in the second group according to a Mann Whitney test (p < 2.2e-16), but to a lesser extent than in **(A)**.

(C) Metabolic enzymes. The distances between proteins in the first group are significantly longer than those in the second group according to a Mann Whitney test (p < 2.2e-16).

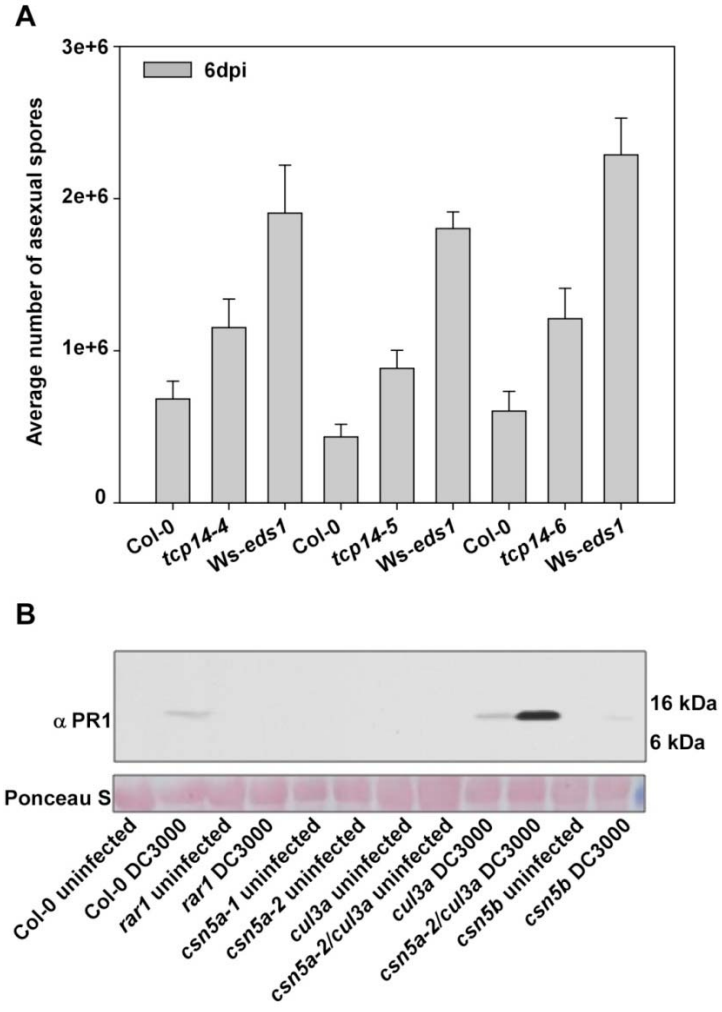


fig. S10: Target validation: Function of TCP14, CSN5 in plant defense.

(A) Loss-of-function *tcp14* mutants (three independent alleles as designated) also exhibit enhanced susceptibility to *Hpa* isolate Noks1. Average number of asexual spores formed 6 dpi in the indicated genotypes was determined Col-0 (susceptible) and Ws-*eds1*

(*e*nhanced *s*usceptibility) genotypes are used as controls. Error bars represent average ± SE of six replicates.

(B) PR1 protein accumulates after infection in the *csn5a cul3* double mutant. Total protein extracts of uninfected tissue or from tissue harvested 2 days after inoculation with *Pto* DC3000 were probed with an anti-PR1 antibody. Ponceau S stain verifies equal loading.

Supporting References

1. S. H. Shiu, A. B. Bleecker, *Sci. STKE* **2001**, re22 (2001).
2. C. Zipfel, J. P. Rathjen, *Curr. Biol.* **18**, R218 (2008).
3. C. Zipfel *et al.*, *Cell* **125**, 749 (2006).
4. C. Zipfel, *Curr. Opin. Plant Biol.* **12**, 414 (2009).
5. P. N. Dodds, J. P. Rathjen, *Nat. Rev. Genet.* **11**, 539 (2010).
6. C. Zipfel *et al.*, *Nature* **428**, 764 (2004).
7. L. Navarro *et al.*, *Plant Physiol.* **135**, 1113 (2004).
8. G. van Ooijen *et al.*, *J. Exp. Bot.* **59**, 1383 (2008).
9. T. K. Eitas, J. L. Dangl, *Curr. Opin. Plant Biol*. **13**, 472 (2010).
10. E. Lukasik, F. L. Takken, *Curr. Opin. Plant Biol.* **12**, 427 (2009).
11. J. W. Mansfield, *Mol. Plant Pathol.* **10**, 721 (2009).
12. M. Thines, S. Kamoun, *Curr. Opin. Plant Biol.* (2010).
13. D. Baltrus *et al.*, *Plos Pathog.*(*in revision*).
14. L. Baxter *et al.*, *Science* (*in press*).
15. M. Dreze *et al.*, (*co-submitted*).
16. J. D. Lewis, D. S. Guttman, D. Desveaux, *Semin. Cell Dev. Biol.* **20**, 1055 (2009).
17. H. Yu *et al.*, *Science* **322**, 104 (2008).
18. K. Venkatesan *et al.*, *Nat. Methods* **6**, 83 (2009).
19. N. Simonis *et al.*, *Nat. Methods* **6**, 47 (2009).
20. P. Braun *et al.*, *Nat. Methods* **6**, 91 (2009).
21. M. Boxem *et al.*, *Cell* **134**, 534 (2008).
22. M. Ashburner *et al.*, *Nat Genet.* **25**, 25 (2000).
23. G. F. Berriz, J. E. Beaver, C. Cenik, M. Tasan, F. P. Roth, *Bioinformatics* **25**, 3043 (2009).
24. A. C. Berglund, E. Sjolund, G. Ostlund, E. L. L. Sonnhammer, Nucleic Acids Res. **36,** D263 (2008)
25. Z. Y. Peng *et al.*, *Nucleic Acids Res.* **37**, D975 (2009).
26. G. Csardi, Nepusz, T., *Inter. Journal* Complex Systems 1695, (2006).
27. P. Shannon *et al.*, *Genome Res.* **13**, 2498 (2003).
28. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
29. M. A. Larkin *et al.*, *Bioinformatics* **23**, 2947 (2007).
30. Z. Yang, R. Nielsen, *Mol. Biol. Evol.* **17**, 32 (2000).
31. Z. Yang, *Mol. Biol. Evol.* **24**, 1586 (2007).
32. M. de Torres-Zabala *et al.*, *Embo J.* **26**, 1434 (2007).
33. T. Eulgem *et al.*, *Plant Physiol.* **135**, 1129 (2004).
34. D. Wang, N. Amornsiripanitch, X. Dong, *PLoS Pathog.* **2**, e123 (2006).
35. C. Denoux *et al.*, *Mol. Plant* **1**, 423 (2008).
36. K. Ramonell *et al.*, *Plant Physiol.* **138**, 1027 (2005).
37. D. Chandran, N. Inada, G. Hather, C. K. Kleindt, M. C. Wildermuth, *Proc. Natl. Acad. Sci. U. S. A.* **107**, 460 (2010).
38. H. Goda *et al.*, *Plant J.* **55**, 526 (2008).

39. R. Thilmony, W. Underwood, S. Y. He, *Plant J.* **46**, 34 (2006).
40. M. G. Rosso *et al.*, *Plant Mol. Biol.* **53**, 247 (2003).
41. J. M. Alonso *et al.*, *Science* **301**, 653 (2003).
42. K. Tatematsu, K. Nakabayashi, Y. Kamiya, E. Nambara, *Plant J.* **53**, 42 (2008).
43. G. Gusmaroli, S. Feng, X. W. Deng, *Plant Cell* **16**, 2984 (2004).
44. G. Gusmaroli, P. Figueroa, G. Serino, X. W. Deng, *Plant Cell* **19**, 564 (2007).
45. Y. Gu *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **105**, 18064 (2008).
46. H. E. Dangl JL, Debener T, Lehnackers H, Ritter C, Crute IR, *Genetic definition of loci involved in Arabidopsis-pathogen interactions*. N. H. C. C. Koncz, and J. Schell, ads., Ed., In Methods in Arabidopsis Research (World Scientific Publishing Co., London, 1992), pp. 393-418.
47. E. B. a. B. Holub, J.L., *Advances in Botanical Research* **24**, 227 (1997).
48. E. Koch, A. Slusarenko, *Plant Cell* **2**, 437 (1990).
49. B. F. Holt, 3rd, Y. Belkhadir, J. L. Dangl, *Science* **309**, 929 (2005).
50. P. C. Stirling, S. F. Bakhoum, A. B. Feigl, M. R. Leroux, *Nat. Struct. Mol. Biol.* **13**, 865 (2006).
51. A. V. Garcia, J. E. Parker, *Trends Plant Sci.* **14**, 479 (2009).
52. S. Robatzek, D. Chinchilla, T. Boller, *Genes Dev.* **20**, 537 (2006).
53. T. K. Eitas, Z. L. Nimchuk, J. L. Dangl, *Proc. Natl. Acad. Sci. U. S. A.* **105**, 6475 2008).
54. D. Swarbreck *et al.*, *Nucleic Acids Res.* **36**, D1009 (2008).
55. B. Aranda *et al.*, *Nucleic Acids Res.* **38**, D525 (2010).
56. C. Stark *et al.*, *Nucleic Acids Res.* **34**, D535 (2006).
57. P. Zhang *et al.*, *Plant Physiol.* **138**, 27 (2005).
58. R. Albert, H. Jeong, A. L. Barabasi, *Nature* **406**, 378 (2000).

DISCUSSION

La biologie systémique, vision holistique des mécanismes moléculaires sous-tendant les relations entre génotypes, environnements et phénotypes, a été l'inspiration au cœur de mes travaux de thèse. Je suis cependant convaincue que, pour permettre de comprendre la Vie, la biologie systémique doit s'intégrer aux quatre autres piliers fondamentaux de la biologie énoncés par Sir Paul Nurse: 1) le gène est la base de l'hérédité, 2) la cellule est l'unité fonctionnelle et structurale de la vie, 3) la vie est basée sur la chimie et 4) l'évolution par sélection naturelle est caractéristique de la vie (Nurse 2003). C'est dans cet esprit que j'ai cherché à observer les phénomènes biologiques, tour à tour et parfois simultanément, à travers le prisme de la biologie systémique et celui de la biologie évolutive.

Malgré mes efforts d'intégration, mes recherches ne prennent pas en compte tous les cinq socles de la biologie. Autre simplification importante, je me suis concentrée sur les systèmes biologiques formés de protéines en interactions physiques, en ignorant temporairement que le mot « interactome» désigne en réalité le réseau moléculaire formé de toutes les relations entre toutes les molécules de la cellule. De plus, les réseaux que j'ai construits et étudiés sont statiques et binaires. Ils ne prennent pas en compte la concentration ou l'état post-traductionnel des protéines, ni la direction, la force, le signe, la régulation ou la dynamique de leurs interactions, alors qu'il est bien établi que ces paramètres donnent leurs propriétés fonctionnelles et dynamiques aux systèmes biologiques (Aldridge, Burke et al. 2006). Bien que réducteurs, je pense néanmoins que ces choix étaient nécessaires étant donné l'état encore largement exploratoire de la cartographie de ces réseaux.

Les milliers d'interactions que nous avons identifiées chez le nématode et *Arabidopsis* nous offrent seulement un aperçu très partiel de la complexité des réseaux d'interactions entre protéines pour ces organismes (<5%, voir chapitre 2). Ce taux de faux négatifs est d'autant plus élevé qu'une grande partie de ma thèse a été consacrée à l'optimisation de la qualité du double hybride en levure en termes de spécificité et de reproductibilité, mais au détriment de sa sensibilité. Malheureusement, mesurer toutes les interactions binaires entre protéines pour un organisme eucaryote est aujourd'hui impossible avec une technologie unique même à saturation, du fait de la notion de sensibilité liée à la technique. Étudier les biais des différentes techniques existantes permettrait sans doute de répartir les paires de protéines à tester entre ces techniques, et ainsi de mieux organiser et réduire le coût de la cartographie des interactomes. Il apparaît également essentiel de développer de nouvelles technologies qui soient adaptables à grande échelle. Étant donné que ce type de cartes représente la base sur laquelle la biologie systémique

de demain va reposer, je m'étonne qu'il n'y ait pas plus d'initiatives internationales pour accélérer le processus de découverte d'interactions physiques entre protéines, comme il y en a eu pour le séquençage du génome humain. On peut espérer que de tels efforts coopératifs seront entrepris dans les années à venir.

La course au séquençage du génome humain a donné l'élan permettant les nombreux projets de séquençage de génomes qui ont suivi. Dans le contexte de l'annotation de ces génomes, il me semble dangereux de négliger les ORFs, ou d'autres unités potentiellement fonctionnelles du génome, seulement parce qu'elles ne sont pas conservées chez d'autres espèce (chapitre 1). L'idée de Jacques Monod selon laquelle « ce qui est vrai du colibacille est vrai pour l'éléphant » pousse avec succès la biologie vers la recherche de principes fondamentaux communs à toutes les espèces vivantes. Cependant, je pense qu'il est aussi important de se pencher sur ce qui rend chaque espèce unique. Ce faisant, j'ai proposé et montré la vraisemblance d'un nouveau mécanisme de naissance de gènes chez la levure *S. cerevisiae* (chapitre 1). Des résultats préliminaires que j'ai obtenus en collaboration avec le laboratoire de F. Ausubel (Massachussetts General Hospital, Boston, MA) semblent indiquer l'existence d'un phénomène similaire chez la bactérie virulente *Pseudomonas aeruginosa*. Ne serait-ce pas ironique si ce mécanisme d'innovation moléculaire qui participe probablement à la spécificité d'une espèce s'avérait à son tour universel ?

Mes propositions techniques et conceptuelles seront, je l'espère, utiles à d'autres scientifiques à la recherche comme moi des nœuds et des liens de l'interactome. Plus fondamentalement, il émerge de mes travaux de thèse l'idée que les systèmes biologiques évoluent, tant par le nombre et la nature de leurs composants (chapitre 1) que par l'organisation de leurs interactions (chapitre 3). Cette notion peut apparaître comme une évidence : bien sûr, puisque le génome encode les composants des réseaux cellulaires, lorsque le génome évolue les réseaux cellulaires font de même. Pourtant, c'est la relation inverse que je souhaite proposer ici : si le génome encode l'information nécessaire à la constitution d'un interactome dans chaque cellule vivante le temps de la vie de cette cellule, l'interactome influencerait « l'écriture » du génome à l'échelle des temps évolutifs.

Cette hypothèse est en partie soutenue par l'observation que, dans le contexte des protéines paralogues, l'identité de séquence de gènes dupliqués reste relativement élevée alors que les profils d'interactions divergent rapidement (Chapitre 3, **Figure 9** et **Document Joint 6**), comme si un petit nombre de mutations sous-tendait ces changements sans perturber les autres contraintes qui pèsent sur les séquences. Il est donc possible que, dans l'interactome, comme peut-être dans les réseaux de transcription bactériens (Mayo, Setty et al. 2006; Lintner, Mishra

et al. 2008), les *liens* (interactions physiques) évoluent plus vite que les *nœuds* (protéines). Dans ce cadre conceptuel, l'unité fonctionnelle des systèmes vivants n'est plus le nœud, mais le lien, sur lequel s'exerceraient les pressions de sélection façonnant l'évolution des séquences.

Pour tester cette hypothèse et déterminer si le re-câblage des interactions physiques influence directement l'évolution des génomes, il faudrait mettre en œuvre un nouveau projet visant à identifier et analyser les mutations fixées sous-tendant des pertes ou gain d'interactions. Cela pourrait être réalisé en comparant les séquences et les profils d'interactions de protéines issues du même ancêtre par duplication avec ceux de leurs orthologues d'une espèce proche demeurés à l'état de copie unique. On pourrait alors utiliser les plateformes expérimentales récemment développées au laboratoire de Marc Vidal (Dreze, Charloteaux et al. 2009; Zhong, Simonis et al. 2009) pour étudier la relation de causalité entre les mutations fixées, les pertes et gains d'interactions et leurs conséquences phénotypiques pour des familles de gènes intéressantes. À plus grande échelle, on apprendrait beaucoup sur l'évolution des systèmes biologiques si l'on pouvait comparer les réseaux interactomes de différentes espèces. Bien qu'il existe de nombreux algorithmes d'alignement de réseaux (Milenkovic, Ng et al. ; Kelley, Yuan et al. 2004; Koyuturk, Kim et al. 2006), ils ne tiennent pas compte des limitations expérimentales de ces réseaux. En conséquence, ils identifient bien les modules conservés, mais manquent de puissance pour analyser les différences entre réseaux. Je pense qu'en couplant le type d'efforts décrits aux chapitres 2 et 3 de ce manuscrit, et en s'inspirant de l'état de l'art existant pour l'analyse de séquences, nous pourrons bientôt concevoir de nouvelles mesures quantitatives permettant de construire des arbres phylogénétiques, calculer les vitesses d'évolution, distinguer les changements neutres des bénéfiques… tout ceci du point de vue des interactomes. Le développement de telles méthodes retraçant l'histoire évolutive des interactions entre protéines avancerait considérablement l'élucidation des principes évolutifs des systèmes biologiques et permettrait en particulier d'estimer dans quelle mesure l'organisation globale des systèmes vivants est universellement conservée (Podani, Oltvai et al. 2001).

Bien que mes travaux de thèse portent sur des questions techniques et fondamentales chez des organismes modèles, j'espère sincèrement qu'ils auront un jour des retombées médicales. Je pense en particulier au cancer, qui peut être vu comme une série d'événements évolutifs à l'intérieur de l'organisme (Marusyk and DeGregori 2008). De nouvelles ORFs apparaissent-elles dans les régions intergéniques des génomes cancéreux? Les protéines dupliquées lors de réarrangements chromosomiques adaptent-elles leurs profils d'interactions

physiques ? Le concept d'attaques ciblées des *hubs* chez les phyto-pathogènes s'applique-t-il aux virus associés au cancer ?

Pour terminer ce manuscrit sur une note plus gaie, je souhaite conclure par un jeu de mot sur la citation de Dobzhansky: *rien n'a de sens en biologie sauf à la lumière de l'évolution,* dont j'ai annoncé en introduction qu'elle était l'inspiration de mes travaux de thèse, et écrire :

Rien n'a de sens en évolution sauf à la lumière de la biologie systémique !

RÉFÉRENCES

Albert, R., H. Jeong, et al. (2000). «Error and attack tolerance of complex networks.» Nature **406**(6794): 378-82.

Aldridge, B. B., J. M. Burke, et al. (2006). "Physicochemical modelling of cell signalling pathways." Nat Cell Biol **8**(11): 1195-203.

Aranda, B., P. Achuthan, et al. "The IntAct molecular interaction database in 2010." Nucleic Acids Res **38**(Database issue): D525-31.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Barabasi, A. L. and R. Albert (1999). "Emergence of scaling in random networks." Science **286**(5439): 509-12.

Beltrao, P. and L. Serrano (2007). "Specificity and evolvability in eukaryotic protein interaction networks." PLoS Comput Biol **3**(2): e25.

Borneman, A. R., P. J. Chambers, et al. (2007). "Yeast systems biology: modelling the winemaker's art." Trends Biotechnol **25**(8): 349-55.

Brachat, S., F. S. Dietrich, et al. (2003). "Reinvestigation of the Saccharomyces cerevisiae genome annotation by comparison to the genome of a related fungus: Ashbya gossypii." Genome Biol **4**(7): R45.

Breitkreutz, B. J., C. Stark, et al. (2008). "The BioGRID Interaction Database: 2008 update." Nucleic Acids Res **36**(Database issue): D637-40.

Calderwood, M. A., K. Venkatesan, et al. (2007). "Epstein-Barr virus and virus human protein interaction maps." Proc Natl Acad Sci U S A **104**(18): 7606-11.

Caldwell, K. S. and R. W. Michelmore (2009). "Arabidopsis thaliana genes encoding defense signaling and recognition proteins exhibit contrasting evolutionary dynamics." Genetics **181**(2): 671-84.

Carvunis, A. R., E. Gomez, et al. (2009). "[Systems biology: from yesterday's concepts to tomorrow's discoveries]." Med Sci (Paris) **25**(6-7): 578-84.

Clamp, M., B. Fry, et al. (2007). "Distinguishing protein-coding and noncoding genes in the human genome." Proc Natl Acad Sci U S A **104**(49): 19428-33.

Clark, A. G., M. B. Eisen, et al. (2007). "Evolution of genes and genomes on the Drosophila phylogeny." Nature **450**(7167): 203-18.

Cliften, P., P. Sudarsanam, et al. (2003). "Finding functional features in Saccharomyces genomes by phylogenetic footprinting." Science **301**(5629): 71-6.

Cusick, M. E., H. Yu, et al. (2009). "Literature-curated protein interaction datasets." Nat Methods **6**(1): 39-46.

Dangl, J. L. and J. D. Jones (2001). "Plant pathogens and integrated defence responses to infection." Nature **411**(6839): 826-33.

de Chassey, B., V. Navratil, et al. (2008). "Hepatitis C virus infection protein network." Mol Syst Biol **4**: 230.

Delbrück, M. (1949). Unités biologiques douées de continuité génétique. Paris, Editions du Centre National de la Recherche Scientifique**:** 33-35.

Dreze, M., B. Charloteaux, et al. (2009). "'Edgetic' perturbation of a C. elegans BCL2 ortholog." Nat Methods **6**(11): 843-9.

Durbin, R., S. Eddy, et al. (1998). Biological sequence analysis. Cambridge, UK, Cambridge University Press.

Dyer, M. D., C. Neff, et al. “The human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis.” PLoS ONE **5**(8): e12089.

Evlampiev, K. and H. Isambert (2008). “Conservation and topology of protein interaction networks under duplication-divergence evolution.” Proc Natl Acad Sci U S A **105**(29): 9863-8.

Fields, S. and O. Song (1989). “A novel genetic system to detect protein-protein interactions.” Nature **340**(6230): 245-6.

Fisk, D. G., C. A. Ball, et al. (2006). “Saccharomyces cerevisiae S288C genome annotation: a working hypothesis.” Yeast **23**(12): 857-65.

Ge, H., A. J. Walhout, et al. (2003). “Integrating ‘omic’ information: a bridge between genomics and systems biology.” Trends Genet **19**(10): 551-60.

Gerstein, M. and D. Zheng (2006). “The real life of pseudogenes.” Sci Am **295**(2): 48-55.

Goffeau, A., B. G. Barrell, et al. (1996). “Life with 6000 genes.” Science **274**(5287): 546, 563-7.

Gunsalus, K. C., H. Ge, et al. (2005). “Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis.” Nature **436**(7052): 861-5.

Halfmann, R., S. Alberti, et al. “Prions, protein homeostasis, and phenotypic diversity.” Trends Cell Biol.

Han, J. D., D. Dupuy, et al. (2005). “Effect of sampling on topology predictions of protein-protein interaction networks.” Nat Biotechnol **23**(7): 839-44.

Hanada, K., T. Kuromori, et al. (2009). “Increased expression and protein divergence in duplicate genes is associated with morphological diversification.” PLoS Genet **5**(12): e1000781.

Hart, G. T., A. K. Ramani, et al. (2006). “How complete are current yeast and human protein-interaction networks?” Genome Biol **7**(11): 120.

Hartl, D. L. and A. G. Clark (1997). Principles of Population Genetics. Sunderland, Massachussetts, Sinauer Associates.

Hasty, J., D. McMillen, et al. (2002). “Engineered gene circuits.” Nature **420**(6912): 224-30.

Hedges, S. B., J. Dudley, et al. (2006). “TimeTree: a public knowledge-base of divergence times among organisms.” Bioinformatics **22**(23): 2971-2.

Ideker, T. and D. Lauffenburger (2003). “Building with a scaffold: emerging strategies for high- to low-level cellular modeling.” Trends Biotechnol **21**(6): 255-62.

Ingolia, N. T., S. Ghaemmaghami, et al. (2009). “Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.” Science **324**(5924): 218-23.

Innan, H. and F. Kondrashov “The evolution of gene duplications: classifying and distinguishing between models.” Nat Rev Genet **11**(2): 97-108.

Ispolatov, I., P. L. Krapivsky, et al. (2005). “Duplication-divergence model of protein interaction network.” Phys Rev E Stat Nonlin Soft Matter Phys **71**(6 Pt 1): 061911.

Ito, T., T. Chiba, et al. (2001). “A comprehensive two-hybrid analysis to explore the yeast protein interactome.” Proc Natl Acad Sci U S A **98**(8): 4569-74.

Jacq, C., J. R. Miller, et al. (1977). “A pseudogene structure in 5S DNA of Xenopus laevis.” Cell **12**(1): 109-20.

Jeong, H., S. P. Mason, et al. (2001). “Lethality and centrality in protein networks.” Nature **411**(6833): 41-2.

Kaessmann, H. “Origins, evolution, and phenotypic impact of new genes.” Genome Res **20**(10): 1313-26.

Keefe, A. D. and J. W. Szostak (2001). "Functional proteins from a random-sequence library." Nature **410**(6829): 715-8.
Kelley, B. P., B. Yuan, et al. (2004). "PathBLAST: a tool for alignment of protein interaction networks." Nucleic Acids Res **32**(Web Server issue): W83-8.
Kellis, M., N. Patterson, et al. (2003). "Sequencing and comparison of yeast species to identify genes and regulatory elements." Nature **423**(6937): 241-54.
Kimura, M. (1968). "Evolutionary rate at the molecular level." Nature **217**(5129): 624-6.
Kondrashov, F. A., I. B. Rogozin, et al. (2002). "Selection in the evolution of gene duplications." Genome Biol **3**(2): RESEARCH0008.
Koyuturk, M., Y. Kim, et al. (2006). "Pairwise alignment of protein interaction networks." J Comput Biol **13**(2): 182-99.
Kumar, S. (2005). "Molecular clocks: four decades of evolution." Nat Rev Genet **6**(8): 654-62.
Leitch, A. R. and I. J. Leitch (2008). "Genomic plasticity and the diversity of polyploid plants." Science **320**(5875): 481-3.
Levy, E. D. and J. B. Pereira-Leal (2008). "Evolution and dynamics of protein interactions and networks." Curr Opin Struct Biol **18**(3): 349-57.
Li, Q. R., A. R. Carvunis, et al. (2008). "Revisiting the Saccharomyces cerevisiae predicted ORFeome." Genome Res **18**(8): 1294-303.
Lintner, R. E., P. K. Mishra, et al. (2008). "Limited functional conservation of a global regulator among related bacterial genera: Lrp in Escherichia, Proteus and Vibrio." BMC Microbiol **8**: 60.
Long, M., E. Betran, et al. (2003). "The origin of new genes: glimpses from the young and old." Nat Rev Genet **4**(11): 865-75.
Lynch, M. and J. S. Conery (2000). "The evolutionary fate and consequences of duplicate genes." Science **290**(5494): 1151-5.
Marusyk, A. and J. DeGregori (2008). "Declining cellular fitness with age promotes cancer initiation by selecting for adaptive oncogenic mutations." Biochim Biophys Acta **1785**(1): 1-11.
Maslov, S., K. Sneppen, et al. (2004). "Upstream plasticity and downstream robustness in evolution of molecular networks." BMC Evol Biol **4**: 9.
Mayo, A. E., Y. Setty, et al. (2006). "Plasticity of the cis-regulatory input function of a gene." PLoS Biol **4**(4): e45.
Milenkovic, T., W. L. Ng, et al. "Optimal network alignment with graphlet degree vectors." Cancer Inform **9**: 121-37.
Milo, R., S. Shen-Orr, et al. (2002). "Network motifs: simple building blocks of complex networks." Science **298**(5594): 824-7.
Monod, J. and F. Jacob (1961). "Teleonomic mechanisms in cellular metabolism, growth, and differentiation." Cold Spring Harb Symp Quant Biol **26**: 389-401.
Nagalakshmi, U., Z. Wang, et al. (2008). "The transcriptional landscape of the yeast genome defined by RNA sequencing." Science **320**(5881): 1344-9.
Nilsson, T., M. Mann, et al. "Mass spectrometry in high-throughput proteomics: ready for the big time." Nat Methods **7**(9): 681-5.
Nishimura, M. T. and J. L. Dangl "Arabidopsis and the plant immune system." Plant J **61**(6): 1053-66.
Novick, A. and M. Weiner (1957). "Enzyme Induction as an All-or-None Phenomenon." Proc Natl Acad Sci U S A **43**(7): 553-66.

Nurse, P. (2003). "The great ideas of biology." Clin Med **3**(6): 560-8.
Ohno, S. (1970). Evolution by Gene Duplication. New York, Springer-Verlag.
Pastor-Satorras, R., E. Smith, et al. (2003). "Evolving protein interaction networks through gene duplication." J Theor Biol **222**(2): 199-210.
Pena-Castillo, L. and T. R. Hughes (2007). "Why are there still over 1000 uncharacterized yeast genes?" Genetics **176**(1): 7-14.
Plewczynski, D. and K. Ginalski (2009). "The interactome: predicting the protein-protein interactions in cells." Cell Mol Biol Lett **14**(1): 1-22.
Podani, J., Z. N. Oltvai, et al. (2001). "Comparable system-level organization of Archaea and Eukaryotes." Nat Genet **29**(1): 54-6.
Presser, A., M. B. Elowitz, et al. (2008). "The evolutionary dynamics of the Saccharomyces cerevisiae protein interaction network after duplication." Proc Natl Acad Sci U S A **105**(3): 950-4.
Rigaut, G., A. Shevchenko, et al. (1999). "A generic protein purification method for protein complex characterization and proteome exploration." Nat Biotechnol **17**(10): 1030-2.
Rual, J. F., K. Venkatesan, et al. (2005). "Towards a proteome-scale map of the human protein-protein interaction network." Nature **437**(7062): 1173-8.
Scannell, D. R. and K. H. Wolfe (2008). "A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast." Genome Res **18**(1): 137-47.
Schoener, T. W. "The newest synthesis: understanding the interplay of evolutionary and ecological dynamics." Science **331**(6016): 426-9.
Simonis, N., J. F. Rual, et al. (2009). "Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network." Nat Methods **6**(1): 47-54.
Stein, L. D., Z. Bao, et al. (2003). "The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics." PLoS Biol **1**(2): E45.
Swarbreck, D., C. Wilks, et al. (2008). "The Arabidopsis Information Resource (TAIR): gene structure and function annotation." Nucleic Acids Res **36**(Database issue): D1009-14.
Thomas, R. (1978). "[Feedback loops and their biological significance (proceedings)]." Arch Int Physiol Biochim **86**(4): 902-3.
Uetz, P., Y. A. Dong, et al. (2006). "Herpesviral protein networks and their interaction with the human proteome." Science **311**(5758): 239-42.
Uetz, P., L. Giot, et al. (2000). "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae." Nature **403**(6770): 623-7.
Vazquez, A. (2003). "Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations." Phys Rev E Stat Nonlin Soft Matter Phys **67**(5 Pt 2): 056104.
Venkatesan, K., J. F. Rual, et al. (2009). "An empirical framework for binary interactome mapping." Nat Methods **6**(1): 83-90.
Vilella, A. J., J. Severin, et al. (2009). "EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates." Genome Res **19**(2): 327-35.
Waddington, C. H. (1957). The Strategy of the Genes. London, Geoge Allen & Unwin.
Wagner, A. (2001). "The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes." Mol Biol Evol **18**(7): 1283-92.
Wagner, A. (2003). "How the global structure of protein interaction networks evolves." Proc Biol Sci **270**(1514): 457-66.
Walhout, A. J. and M. Vidal (2001). "High-throughput yeast two-hybrid assays for large-

scale protein interaction mapping." Methods **24**(3): 297-306.
Yu, H., P. Braun, et al. (2008). "High-quality binary protein interaction map of the yeast interactome network." Science **322**(5898): 104-10.
Zhong, Q., N. Simonis, et al. (2009). "Edgetic perturbation models of human inherited disorders." Mol Syst Biol **5**: 321.

FIN