Study of differential allelic expression in the breast cancer intermediate-risk susceptibility genes CHEK2, ATM and TP53

Binh Thieu Tu Nguyen Nguyen-Dumont

N° d'ordre: 344-2010 Année 2010

THESE DE L'UNIVERSITE DE LYON

délivrée par

L'Université Claude Bernard Lyon I

Ecole Doctorale BMIC

Diplôme de Doctorat

(arrêté du 7 août 2006)

soutenue publiquement le 15 décembre 2010

par

Tú NGUYEN-DUMONT

Study of differential allelic expression in the breast cancer intermediate-risk susceptibility genes *CHEK2*, *ATM* and *TP53*

Jury :

Dr	Sean V. TAVTIGIAN	Directeur de thèse
Dr	Francine DUROCHER	Rapporteur
Dr	Ana Teresa MAIA	Rapporteur
Dr	Janet HALL	Examinateur
Dr	Sylvie MAZOYER	Examinateur

During my Ph.D, I have received support from many persons, both in my professional and personal life. I would like to thank all my colleagues, friends and family who supported me and contributed to the completion of this work.

I especially want to express my gratitude to ***Sean Tavtigian*** *for giving me that exceptional opportunity to work and learn with him. Thank you for accepting me in the GCS group and for proposing me such an interesting project. Thank you for trusting and encouraging me all these years. Lastly, than k you for your support, despite the distance (and LaTeXediting).*

I am also grateful to ***Fabienne Lesueur*** *for her constant support during my Ph.D work. Thank you for your commitment to this project, your permanent availability for discussions and your help in correcting the drafts to this mémoire.*

I would like to thank ***Francine Durocher*** *and* ***AnaTeresa Maia*** *for accepting without any hesitation to review this work.*

I would like to thank ***Janet Hall*** *and* ***Sylvie Mazoyer****, my comité de suivi de thèse, for their valuable guidance and advices throughout the carrying out of this project. Thank you also for providing the last pieces of material needed to complete this project.*

Many thanks to all the persons I worked with in the ***GCS Group*** *every day, for their help and encouragements.*

A special acknowledgment to ***Lars Jordheim****. Although not my supervisor on the paper, you have been acting in this role at the beginning of the DAE project. Thank you for your advices to the inexperienced Masters student in the first years. Thank you for always being supportive, open to answer all the questions and solve the issues encountered along the road. I have learnt a lot from you.*

I would like to thank ***Melissa Southey****. Your mentorship during my Masters studies encouraged me to pursue a scientific career.*

I also want to acknowledge the financial support I received from ***Fondation de France*** *and* ***IARC****.*

Je souhaite maintenant remercier mes amis. Ils savent tout ce que je leur dois.

En particulier, ***Mathilde*** *(Pub'Math!) et* ***Karima****, malgré notre parcours parsemé d'obstacles (la rentrée 2005...), nous savons maintenant que nous avons eu raison de persévérer. Merci à toutes les deux pour tout, depuis tout ce temps.*

Je remercie bien sûr ***Amélie****,* ***Maxime*** *et* ***Camille****,* ***Véra*** *et* ***Stéphane****,* ***Sandrine*** *et* ***James****,* ***Thomas****,* ***Bertrand****, le meilleur soutien moral tous les jours, c'est le gang du CIRC.*

Enfin, je remercie ma famille et mes proches. Cette thèse est en particulier dédiée à mes parents. Cám on ***Ba*** *và* ***Má****.*

Je tiens aussi à remercier ***ma soeur****,* ***mes beau-parents*** *et* ***mes belle-soeurs****, pour votre soutien moral et aussi logistique! Merci aussi à tous les autres babysitters (Tom!). Sans vous, la fin aurait été (encore plus) difficile.*

Merci à ***Simplice*** *d'avoir cru en moi depuis si longtemps.*

Ong Nôi*, mes cousins* ***Nguyên****, gia dình* ***Pham****, et mes cousins* ***Chanfreau****: merci pour votre présence et votre soutien.*

Pour finir, j'aimerais te remercier, ***Arnaud****, le meilleur assistant-éditeur, assistant-illustrateur, assistant-technique dont on puisse rêver. C'est la fin d'une première étape, allons voir ce que l'avenir nous réserve à Melbourne, avec* ***Nam Son*** *dans notre valise.*

Contents

List of abbreviations

A : Adenosin

A-T : Ataxia-Telangiectasia

ATM : *Ataxia Telangiectasia Mutated*

BRCA1 : *Breast Cancer 1*

BRCA2 : *Breast Cancer 2*

C : Cytosin

cDNA : Complementary DNA

CI : Confidence Interval

CIN : Chromosomal instability

CHEK2 : *Protein kinase CHK2*

DAE : Differential allelic expression

DNA : Deoxyribonucleic acid

dsDNA : Double-stranded DNA

EBV : Epstein-Barr Virus

eQTL : Expression Quantitative Trait Loci

G : Guanin

GWA : Genome wide association

HRM : High-resolution melting curve

LCL : Lymphoblastoid cell line

mRNA : messenger RNA

miRNA : micro RNA

MSI : Microsatellite instability

NMD : Non-sense mediated mRNA decay

OMIM : Online Mendelian Inheritance in man

PBS : Phosphate buffered saline

PCR : Polymerase chain reaction

RFLP : Restriction fragment length polymorphism

RNA : Ribonucleic acid

rs : RefSNP, Single Nucleotide Polymorphism reference number

RT-PCR : Reverse Transcription-PCR

SNP : Single Nucleotide Polymorphism

siRNA : Small interfering RNA

ssDNA : Single-stranded DNA

T : Thymin

Tm : Melting temperature

TP53 : *Tumor Protein p53*

U : Unit

UTR : Untranslated Region

List of Figures

List of Tables

Introduction

A key feature of all higher eukaryotes is the defined life span of the organism. This is a property that extends to the individual somatic cells, whose growth and division are highly regulated. A notable exception is provided by cancer cells, which arise as variants have lost their usual growth control. Cancer cells acquire the ability to grow in inappropriate locations and to propagate indefinitely via the acquisition of many non-lethal genetic and epigenetic alterations in various genes including those that control cellular proliferation and apoptosis. These acquired cellular traits are often ultimately lethal to the individual organism in which they occur. Many genes can be the target of genetic damage (or mutation) and the loss or changed function of these mutated genes can contribute in several different ways to the unregulated growth advantage of the cancer cells [Pharoah and Caldas, 1999].

Although the vast majority of cancers are sporadic (not associated with a familial cancer syndrome) and due to the acquisition of somatic mutations during the life of a stem cell in a given organ, some are linked to a constitutional mutation that is found in every cell of the body. This second type of mutation is usually inherited, located within a tumor suppressor gene, transmissible by the gametes (leading to the term germline mutation), and results in the individual carrying the mutation being genetically predisposed to the development of certain types of cancer. In many cases, a single copy of a mutated gene transmitted by the father or mother, gives carriers a more than 50% greater likelihood of developing cancer during their lifetime than the general population. These “hereditary” forms are thought to represent 2% to 5% of all cancers. This type of predisposition is reflected clinically in familial aggregations of cancers in a single branch of the family, bilateral cancers in twin organs, and cancers diagnosed at a young age.

Recent, rapid advances in fundamental genetics have led to tremendous developments in the understanding of the molecular biology of cancer: the genes involved in the initiation and progression of tumorigenesis, and both the function of their products and how disruptions in these molecules can contribute to cancer. Indeed, from the late 1980s through the mid 1990s, linkage analysis and positional cloning provided a successful approach to the discovery of high-risk susceptibility genes for many diseases. For example, screening of the coding region of *BRCA1* and *BRCA2* has identified some significant structural alterations that give rise to protein truncation, protein shortening and protein deletion in some multiple-case breast cancer families and in some young women diagnosed with breast cancer without a family history of breast cancer. Similarly, screening of mismatch repair genes (*hMLH1*, *hMSH2*, *hMSH6*, *hPMS2*) has identified similar structural changes in some families with Hereditary Non-Polyposis Colorectal Cancer (HNPCC) syndrome and in some young people diagnosed with colorectal cancer without a family history of colorectal cancer. However, at least for the common cancers, the success rate of these approaches to identify additional susceptibility genes (and/or the deleterious sequence variants that they harbor) has decreased, despite the proportion of unexplained multiple-case cancer families remaining significant. While genotype information from well characterized susceptibility genes such as *APC*, *BRCA1*, *BRCA2*, *MLH1*, *MSH2* and *PTEN* have obvious clinical utility, these cancer susceptibility genes only account for approximately 20% to 25% of the heritable risk of these diseases.

Therefore, it has become necessary to consider the possibility of additional mechanisms being involved in inherited genetic susceptibility and development of disease. In this perspective, it can be hypothesized that genetic variation in transcriptional regulatory elements could also influence risk of disease. Indeed, several studies have shown that differences in gene expression levels account for a major part of the variation within and among species [King and Wilson, 1975, Johnson and Porter, 2000, Levine, 2002]. It might well be expected that variations in disease phenotype would frequently be explained by changes in transcript expression levels rather than by structural alteration of genes. Consequently, gene expression regulation provides a potential mechanism for generating cellular variation and may be the underlying explanation for a proportion of cancer syndromes that have not been resolved by germline coding region mutation screening in currently known cancer predisposition genes.

Bibliographical review

I

The genetic bases of cancer development

I.1 Initiation and control of tumorigenesis

Cancer is a multistage process in which initiation of a tumor requires several steps, which may be followed by further changes that advance the tumorigenic state. Six essential cellular functions governing cell proliferation and homeostasis have to be acquired by cells to become malignant. They are often referred to as the "hallmarks of cancer" [Hanahan and Weinberg, 2000]. These acquired capabilities are insensitivity to growth-inhibitory signals, self-sufficiency in growth signals, evasion of programmed cell-death (apoptosis), unlimited replicative potential, sustained angiogenesis, and tissue invasion and metastasis (Figure I.1). The sequence of acquisition of these capabilities varies widely among tumors of different types, but also among tumors of the same type [Hanahan and Weinberg, 2000].

Transformation of normal cells may occur spontaneously, but a variety of agents, or carcinogens, may also increase the frequency with which cells are converted into neoplastic forms. Such carcinogens belong to diverse categories ranging from lifestyle factors (ex: tobacco smoking), occupational exposures (ex: asbestos),

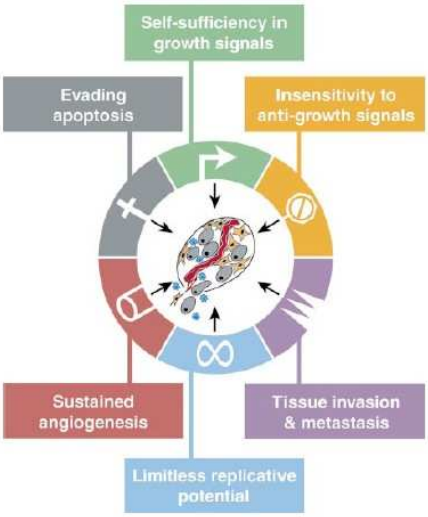

Self-sufficiency in growth signals
Evading apoptosis
Insensitivity to anti-growth signals
Sustained angiogenesis
Tissue invasion & metastasis
Limitless replicative potential

Figure I.1: The hallmarks of cancer. To become malignant, a cell has to acquire six essential capabilities that collectively interact to modify the normal cellular pattern of growth and development. After [Hanahan and Weinberg, 2000].

dietary habits (ex: aflatoxins) to environmental exposures (ex: radiation). Transformation may as well result from infection by DNA or RNA viruses, bacteria or parasites, the most significant ones being hepatitis B and C viruses, human papillomavirus and *Helicobacter pylori.* Tumorigenesis is characterized by the accumulation of genetic changes in somatic cells through random events and/or by the action of these carcinogens [Pharoah and Caldas, 1999, Stewart and Kleihues, 2003].

Most genes targeted by somatic mutation are either proto-oncogenes or tumor suppressor genes. Mutations in these genes give cells some kind of growth or survival advantage over neighboring normal cells of the same type. Proto-oncogenes are positive regulators of the cell cycle progression. They fall into several groups ranging from transmembrane proteins, kinases and their receptors, to transcription factors. Oncogenic mutations in these genes usually result in a gain-of-function in which their activity is inappropriately activated. In contrast, tumor suppressor genes normally impose some kind of constraint on the cell cycle or cell growth, and release of this constraint by a loss-of-function mutation is tumorigenic.

Tumor suppressor genes have been subclassified into two groups. The first subclass is gatekeeper genes, which are negative regulators of the cell cycle, controlling the pathways of cell division and proliferation. The second subclass is caretaker genes. Their primary function is to control the accuracy of cell division [Stewart and Kleihues, 2003]. Caretaker genes are mainly involved in DNA repair. After radiation, chemical or spontaneous damage, repair systems can recognize mispaired, altered or missing bases in DNA or other structural distortions of the double helix. Mutations in caretaker genes alter the mechanisms that repair damaged DNA, thereby increasing the probability that mutations will go unrepaired and be transmitted to daughter cells after replication and cell division [Morgan et al., 1998]. Caretaker genes are also involved in the control of genomic instability, by ensuring correct chromosomal segregation during mitosis [Hanahan and Weinberg, 2000].

The development of a malignant tumor from a normal cell is a long and complex accumulation of changes, which can take years to decades. Tumor progression is driven by accumulation of mutations and/ or epigenetic changes to oncogenes and tumor suppressors. However, since mutations are normally infrequent, the normal rate of somatic mutation is unlikely to be sufficient to account for all the accumulation of mutations required for a tumor to develop [Simpson, 1997, Hanahan and Weinberg, 2000, Stewart and Kleihues, 2003]. Thus, one means by which cancer cells increase the number of mutations in their genome is by inactivating some of their repair systems, so that spontaneous mutations accumulate instead of being removed [Hanahan and Weinberg, 2000, Stewart and Kleihues, 2003, Auranen et al., 2005].

I.2 The genetic determinants of cancer susceptibility

"Cancer genetics" can refer both to somatic cell genetics and genetic susceptibility. Somatic cell genetics focuses on mutations that are acquired by an individual's cells during their lifetime and the role that those mutations play during tumor initiation and progression. In contrast, genetic susceptibility focuses on inherited genetic variation in cancer susceptibility genes and the effects of that inherited variation on an individual's lifetime cancer risk. This section describes the latter phenomenon.

Alterations leading to the six hallmarks of cancer previously mentioned have both genetic and epigenetic origins.

I.2.1 Genetic changes

The population genetics definition of a polymorphism is a naturally occurring sequence variant that has a frequency of greater than 1% in a population. In the human gene pool, their estimated frequency is about one every 1000 base pairs. Approximately 90% of DNA polymorphisms are single nucleotide polymorphisms (SNPs) [Collins et al., 1998, Lewin, 2004]. SNPs are distributed throughout the human genome, in coding and non-coding regions and, as used here "polymorphism" does not imply anything about function. Although the majority of DNA polymorphisms are functionally neutral, a proportion of them are likely to exert effects on the function of the encoded protein or the regulation of gene expression. These "minor variations" among individuals can result in inter-individual differences to environmental factors, disease susceptibility and responsiveness to therapy [Webb, 2002]. In the past few years, many research projects have investigated the possible association between disease risk and the inheritance of specific genetic variants, including SNPs.

Mutations are rarer events than polymorphisms: their population genetics definition is a sequence variant with a frequency of less than 1%, and, as used here, the word does not necessarily imply anything about effect on gene function [Lewin, 2004]. A variety of genetic mutations can alter the function of individual genes. These include point mutations, i.e. affecting a single base pair, small insertions/deletions or larger genetic changes such as chromosomal rearrangements and gene amplification. Mutations can be further categorized functionally. Within the coding sequence, point mutations may take the form of missense variants that affect protein function, or nonsense or frameshift mutations that lead to loss of protein function. Mutations outside the coding sequence can affect transcription, translation and mRNA splicing and processing. In contrast, mutations can also be neutral, *i.e.* without any effect on protein function, and take the form of silent substitutions or neutral intronic variants. To date, most mutation screening projects have focused on coding sequences. Interesting sequence variants have been found in genes whose products play roles in a variety of biochemical pathways, including DNA replication, recombination and repair, hormone synthesis and degradation, hormonal signal transduction, cell cycle progression and checkpoint control, as well as transcriptional regulation.

Inherited mutations, when affecting caretaker genes, result in affected cells being genetically unstable and extremely prone to acquire further genetic changes that favor cancer development. Indeed, genetic instability is one major feature of cancer cells [Pharoah and Caldas, 1999, Hanahan and Weinberg, 2000]. Genetic instability can occur both at the chromosome and nucleotide levels. Chromosomal instability (CIN) is caused by systems that act on partitioning at mitosis or recombination during cell division. CIN may include excess or loss of one or more chromosomes, as well as breakage of two chromosomes, with transfer and fusion of parts of the broken fragments onto each other (translocation). For instance, aneuploidy is considered a feature of many cancer cells and is thought to develop as a result of CIN.

Genetic instability at the level of the nucleotide leads to subtle sequence alterations within minisatellites and microsatellites, which are repetitions of a short DNA sequence motif occurring abundantly and randomly throughout the human genome (*e.g.* CACACACACA) [Lewin, 2004]. Minisatellites are generally 0.1 to 20 kb long whereas microsatellites are less than 0.1 kb. Microsatellite instability (MSI) occurs due to faulty mismatch DNA repair pathways, which induces variation in the number of tandem repeats of DNA sequence and happens preferentially at di-nucleotide repeat sequences. This type of polymorphisms are multi-allelic when generally SNPs are bi-allelic. The high variability of microsatellites make them especially useful for genomic mapping, because there is a high probability that individuals will present allelic variation at such a locus.

I.2.2 Epigenetic changes

Until recently, tumor initiation and progression has mostly been considered a genetic process in which cells in the developing tumor acquire successive genetic lesions that provide the cells with a growth or survival advantage. The focus on genetic alterations in cancer research has perhaps initially led to an underestimation (or at least under-investigation) of the contribution of epigenetic mechanisms, which are alterations in gene function that are mediated by factors other than changes in primary DNA sequence. Epigenetic phenomena follow an inheritance process that is independent from the classical Mendelian inheritance.

It has become increasingly apparent that multiple changes in cancer cells, i.e. activation of oncogenes, silencing of tumor suppressor genes, inactivation of DNA repair systems, and thus CIN and MSI, are also caused by epigenetic abnormalities [Pharoah and Caldas, 1999, Jaenisch and Bird, 2003, Secko, 2005, Perera and Bapat, 2007]. DNA methylation and histone acetylation are the most common non-mutational mechanisms that disrupt gene function and expression [Jaenisch and Bird, 2003]. DNA methylation is the covalent modification of the C-5 position of cytosine (C) residues and occurs primarily at CpG dinucleotides.

In contrast, Cs in the enhancers and promoters of active genes are not/less methylated [Lewin, 2004]. Histones can be modified by acetylation, methylation, phosphorylation, ubiquitination and Poly-ADP ribosylation, which ultimately influence the protein-DNA interaction and can modulate the recruitment of cellular machinery that alter the chromatin state.

Parental imprinting for instance is an epigenetic phenomenon. An imprint is a reversible modification of DNA that causes differential expression of maternally and paternally inherited homologous genes. A particular gene is expressed only from one of the two alleles, depending on which parent it was inherited from. The specific pattern of methyl groups in the parental chromosomes is mainly responsible for achieving monoallelic gene expression without altering the genetic sequence [Lewin, 2004]. New research suggests that variation in the imprint left on a genome by a parent can influence tumor development [Webb, 2002, Feinberg and Tycko, 2004]. Such variation may take the form of loss of imprinting (LOI). LOI involves the activation of the normally silent copy of growth-promoting genes, or silencing of the normally transcribed copy of tumor suppressor genes. Thus mutations in parental imprinting can influence cell differentiation and may as a consequence increase cancer risk [Secko, 2005].

I.3 Evidence for an effect of heredity in common cancers

There is general agreement that environmental factors and somatic events are the predominant contributors to the causation of cancer [Lichtenstein et al., 2000]. Natural selection acts to eliminate genetic mutations in the germline that contribute to disease formation. However there are some instances where disease is due to inherited genetic alterations in germ cells, which provide the carrier with an increased lifetime disease risk. Germline variations themselves do not necessarily cause cancer. They only affect an individual's consequent cancer risk, after exposure to carcinogens [Stewart and Kleihues, 2003].

A familial component to various cancers has been recognized for many years. Indeed, over 140 years ago, the French physician Paul Broca characterized the pattern of breast and other cancers over four generations of his wife's family. Since then, collection and description of pedigrees from such "cancer-prone" families have provided empirical evidence for an heritability component to cancer susceptibility. Dissecting heritable genetic from non-genetic variation in disease risk and identifying the proportion of susceptibility to cancer that can be accounted for by inherited genetic factors has been a great challenge for epidemiologists and geneticists studying familial clustering of cancer.

Usually, family aggregation is assessed by studying relatives of affected subjects and comparing their rates of illness with those of controls (unaffected individuals) and their relatives. However, familial aggregation of a trait is a necessary but not sufficient condition to infer the importance of genetic susceptibility. This is due to environmental and cultural influences also aggregating in families, leading to family clustering and excessive familial risk. Incidentally, an under-utilized design in the search for the effects of shared environmental risk factors is comparison of cancer incidences among spouses, which has provided meaningful results in the study of passive smoking and anogenital infections [Hemminki and Dong, 2000]. Studying spouses of cancer patients brings out the increased risk for unrelated but cohabiting individuals.

Several approaches for discriminating genetic from environmental influences are available in studies of human diseases, although practical difficulties often limit their use [Risch, 2001]. The most powerful design examines risks in relatives of affected versus control adoptees because adoption creates a separation between individuals' biological and environmental effects. Since it is difficult to access information on biological relatives of adoptees, adoption studies usually focus on common diseases or trait outcomes only.

Another study design involves twins [Ahlbom et al., 1997, Lichtenstein et al., 2000]. Identical (monozygote) twins derive from the fission of a single fertilized egg and thus, inherit identical genetic material. By contrast, dizygote twins are derived from two distinct fertilized eggs. Consequently, those twins have the same biological links as full siblings, although they may be more physiologically related on account of the sharing of the same prenatal intra-uterine experience. Comparing the similarity of monozygote twins with same-sex dizygote twins is a common approach for determining the magnitude of genetic influence on a disease and this technique has been applied to a broad range of disorders, including cancer. These studies rely on the assumption that monozygote and dizygote twins display a comparable degree of similarity because of the sharing of the same environmental factors, so the difference in concordance rates between monozygote and dizygote twins is a reflection of genetic factors. The proportion of twins who have both the illness and an affected twin is called probandwise concordance.

In a Swedish population-based study, Ahlbom et al. linked the Twin Registry to the Cancer Registry in order to identify cases of cancer in twins [Ahlbom et al., 1997]. Although the estimates were very low, they found increased probandwise concordance in monozygote versus dizygote twins for colorectal, breast, cervical and prostate cancers, suggesting the importance of genetic factors for these sites. In their study, monozygote and dizygote concordances were comparable for stomach and lung cancers, indicating a weaker genetic contribution in these cancers. The authors suggest that smoking habits are most likely an important source of familial effects for lung cancer. Overall, the authors found genetic effects to influence cancer risk.

Similarly, in a large, population-based twin study of cancer in Sweden, Denmark and Finland, Lichtenstein et al. observed that generally the twin of a person with cancer had an increased risk of having the same cancer [Lichtenstein et al., 2000]. There were increased concordance rates in monozygote versus dizygote twins for common cancers, in particular for cancers of the colorectum, breast and prostate. In contrast to the previous study, the authors also found increased concordance for stomach and lung cancers.

Table I.1: The three components of phenotypic variance. Adapted from [Lichtenstein et al., 2000].

Effect	Definition	Indication of effect*	Examples
Hereditary	The proportion of phenotypic variance accounted for by inherited genetic differences among persons (heritability).	Similarity greater in MZ twins than in DZ twins.	Additive and dominant genetic effects.
Shared environmental	The proportion of phenotypic variance accounted for by environmental factors shared by both twins, thus contributing to similarity between them.	Similarity among both MZ and DZ twins greater than would be expected from genetic effects alone.	Passive smoking during childhood (lung cancer) or similar dietary habits (stomach cancer).
Nonshared environmental	The proportion of phenotypic variance accounted for by environmental factors causing differences between twins.	Lack of similarity in both MZ and DZ twins.	Sporadic mutations, occupational exposure, or viral infections.

* MZ, monozygotic and DZ, dizygotic

In their analysis, Lichtenstein et al. divided phenotypic variation into three components (Table I.1) and estimated the magnitude of the contributions of genetic factors and both shared and nonshared environmental factors to the development of cancer at various sites (Table I.2). The highest contribution to disease risk was observed for non-shared environmental factors, which include any unique environmental cause of cancer that is not inherited and not shared between twins, e.g. occupational exposure. Using the particular multifactorial model chosen by Lichtenstein et al, their contribution to risk of cancer was found to range from 58% to 82% for cancers of the prostate and uterus respectively. Then, the estimates for contribution of shared environmental factors such as smoking, diet or human papillomavirus infection indicated an increased susceptibility but they did not reach statistical significance. The authors suggested that twin studies may have limited power to detect such effects. Lastly, the authors found statistically significant effects of heritable factors, ranging from 27 to 42% for breast, colorectal and prostate cancer. Their work also suggest evidence of limited heritability of leukemia and of cancer of the stomach, lung, pancreas, ovary and bladder but the estimates did not reach statistical significance.

Table I.2: Effects of the phenotypic variance components in cancers at various sites. Adapted from [Lichtenstein et al., 2000].

	Proportion of variance [95% CI]			
Site or type	Heritable factors	Shared environmental factors	Nonshared environmental factors	Fit of model p-value
Stomach	0.28 [0 – 0.51]	0.10 [0 – 0.34]	0.62 [0.49 – 0.76]	1.0
Colorectum	0.35 [0.10 – 0.48]	0.05 [0 – 0.23]	0.60 [0.52 – 0.70]	0.93
Pancreas	0.36 [0 – 0.53]	0.00 [0 – 0.35]	0.64 [0.47– 0.86]	0.92
Lung	0.26 [0 – 0.49]	0.12 [0 – 0.34]	0.62 [0.51 – 0.73]	0.88
Breast	0.27 [0.04 – 0.41]	0.06 [0 – 0.22]	0.67 [0.59 – 0.76]	0.93
Cervix uteri	0.00 [0 – 0.42]	0.20 [0 – 0.35]	0.80 [0.57 – 0.97]	0.96
Corpus uteri	0.00 [0 – 0.35]	0.17 [0 – 0.31]	0.82 [0.64 – 0.98]	0.99
Ovary	0.22 [0 – 0.41]	0.00 [0 – 0.24]	0.78 [0.59 – 0.99]	1.0
Prostate	0.42 [0.29 – 0.50]	0.00 [0 – 0.09]	0.58 [0.50 – 0.67]	0.09
Bladder	0.31 [0 – 0.45]	0.00 [0 – 0.28]	0.69 [0.53 – 0.86]	0.64
Leukemia	0.21 [0 – 0.54]	0.12 [0 – 0.41]	0.66 [0.45 – 0.88]	0.99

For colorectal, breast and prostate cancer, the estimated hereditary components were slightly higher in the younger than in the older groups. These findings were in accordance with observations that hereditary effects are stronger in early-onset cancers [Stewart and Kleihues, 2003]. Nevertheless, although an hereditary component was found, this study reinforced the hypothesis that, at the population level, environmental exposures are responsible for the largest single component of cancer incidence.

Family history and twin concordance are the only pieces of evidence informative of a possible heritable etiology, with an important caveat that environmental causes of cancer may also cause familial aggregation. However, assessment of the contribution of inherited and environmental factors to the causation of cancer in twin studies have had a relatively small impact on research and clinical practice because twins are rare and only a few registries go back far enough in time to provide enough cases of cancer for reliable conclusions to be drawn [Lichtenstein et al., 2000].

I.4 Inherited cancer syndromes

Numerous studies have addressed the degree to which site-specific cancers run in families, in particular those of the breast, colon and prostate. Such studies are useful to derive a global view of the familiality of cancer. In order to identify the genetic explanations of these familial cancer disorders, various molecular genetic approaches have been used successfully in the past decades.

The first major susceptibility genes for the common cancers were identified in the early 1990's and since then, a considerable amount of knowledge about genetic cancer susceptibility and the underlying susceptibility genes have been gathered. Table I.3 shows some susceptibility genes that have been associated with common cancers such as breast and colon, as well as others linked to rarer inherited cancers syndromes [Stewart and Kleihues, 2003, Nagy et al., 2004]. Many of these syndromes show almost complete penetrance, *i.e.* a very large fraction of individuals affected by age 70.

Nevertheless, to date, known cancer syndromes with identified gene defects only explain 5–10% of all cancers [Peto et al., 1999, Nagy et al., 2004]. Inherited mutations in susceptibility genes are relatively rare, except in some populations, which have arisen from a small numbers of founders and remained genetically isolated. In that particular case, mutations can achieve higher frequencies and therefore account for a larger fraction of cancer in the population. Molecular genetic discoveries that have resulted from the study of families with heritable cancer have changed the way these families are counseled, managed and provided with appropriate medical care.

Table I.3: Highly penetrant hereditary cancer syndromes. Adapted from [Nagy et al., 2004].

Syndrome	Gene(s)	Population incidence	Penetrance*
Ataxia-telangiectasia	*ATM*	1/30,000 to 1/100,000	100%
Cowden syndrome	*PTEN*	1/200,000	90–95%
Familial adenomatous polyposis	*APC*	1/5000 to 1/10,000	∼ 100%
Familial malignant melanoma	*CDKN2A, CDK4*	Unknown	∼ 100%
Fanconi anaemia	*FANCA, FANCB, FANCC, FANCD, FANCE, FANCF, FANCG, FANCL*	1/360,000	100%
Hereditary breast–ovarian cancer syndrome	*BRCA1* and *BRCA2*	1/500 to 1/1000	Up to 85%
Hereditary diffuse gastric cancer	*CDH1*	Unknown, rare	90%
Hereditary nonpolyposis colon cancer	*MLH1, MSH2, MSH6, PMS1, PMS2*	1 in 400	90%
Hereditary papillary renal cell carcinoma	*MET*	Unknown	Unknown, but reduced
Li–Fraumeni syndrome	*TP53*	Rare	90–95%
Multiple endocrine neoplasia II (MEN2)	*RET*	1/30,000	70–100%
Neurofibromatosis type I	*NF1*	1/3000	100%
Neurofibromatosis type II	*NF2*	1/40,000	100%
Peutz–Jeghers syndrome (PJS)	*LKB1 (STK11)*	1/200,000	95–100%
Retinoblastoma, hereditary (RB)	*RB*	1/13 500 to 1/25,000	90%
von Hippel–Lindau (VHL)	*VHL*	1/36,000	90–95%
Xeroderma pigmentosum	*XPA, ERCC3, XPC, ERCC2, XPE, ERCC4, ERCC5*	1/1,000,000◇	100%

* Penetrance estimates are up until age 70 years, include both malignant and benign features and with the exception of MEN2, describe clinical penetrance. For MEN2, biochemical testing is 95–100% by age 70.

◇ Incidence of XP in Japan is 1/40,000

II

Inherited breast cancer susceptibility

II.1 Background

Today, breast cancer is the most commonly occurring cancer among women, accounting for 22% of all female cancers and with an estimated annual worldwide incidence of about one million cases [Oldenburg et al., 2007]. Well established determinants known to increase breast cancer risk are summarized in Table II.1. These can be from either endogenous or exogenous sources, such as early age at menarche, late age at menopause, late pregnancy or nulliparity, as well as use of hormone replacement therapy. Other behavioral or environmental risk factors include diet, alcohol intake, overweight and obesity, tobacco use and radiation exposure [Stewart and Kleihues, 2003, Narod, 2006, Oldenburg et al., 2007].

Family history is also a well established risk factor for breast cancer. The risk increases with the number of relatives affected at young age or past history of disease. In Western countries, the overall lifetime risks for women who have no affected relative, one affected relative or two affected relatives are 7.8%, 13.3% and 21.1% respectively [Oldenburg et al., 2007].

Table II.1: Breast cancer risk factors. Adapted from [Oldenburg et al., 2007].

Genetic factors	Positive family history of breast cancer; any first or second degree family member with breast cancer; carrier of a know breast cancer susceptibility gene.
Demographic factors	Geographical region (Western countries); female sex; increasing age; low socio-economical status.
Endogenous hormonal factors	Older age at menopause (> 54); early age of menarche (< 12); nulliparity and late pregnancy; no breastfeeding; low physical activity.
Exogenous hormonal factors	Usage of oral contraceptives; usage of hormone replacement therapy.
Physical characteristics	Obesity in postmenopausal women; tall stature; high insulin-like growth factor I (IGF-I) levels; history of atypical proliferative benign breast disease; history of breast cancer; dense tissue at mammography; high bone density in postmenopausal women.
Environmental factors	Exposure to ionizing radiation, in particular at young age.
Behavioral factors	Alcohol intake; tobacco smoking; low folate intake; high intake of unsaturated fat and well-done meat.

II.2 The genetic epidemiology of hereditary breast cancer

II.2.1 Strategies for identifying breast cancer susceptibility genes and variants

Genetic epidemiology progresses through different study designs aiming at answering different questions, in order to address the role of genetic factors in determining susceptibility to disease:

- familial aggregation studies: is there a genetic component to the disease?
- segregation studies: what is the pattern of inheritance? (i.e. dominant or recessive)

- linkage studies: on which region of the chromosome is the disease-related gene/allele located?
- association studies/ mutation-screening studies: which allele of which gene is associated with the disease?

Thus, once a genetic component to the disease has been implicated, mapping methods are used to identify genes and genetic variants influencing susceptibility to disease, without prior knowledge of which or how abnormally functioning proteins are involved in pathogenesis. Once the gene is mapped, then its product can be characterized and its contribution to etiology defined.

There have been multiple large-scale searches for genes involved in susceptibility to breast cancer using both linkage and association studies, with the ultimate aim of discovering new genes or variants allowing for better risk prediction.

Linkage studies

Linkage studies have been the mainstay of geneticists and epidemiologists for localizing susceptibility genes for breast cancer for a long time. Linkage analysis examines the cosegregation of a marker and a trait in large pedigrees at high-risk. Essentially, linkage analysis relies on the fact that if two or more genetic loci are in very close physical proximity, they are likely to segregate together during meiosis.

For gene mapping, linkage analysis uses known polymorphic markers, which are scattered throughout the genome, and analyzes their segregation with disease phenotypes in related individuals. Alternatively, one can examine marker allele sharing between pairs of affected relatives, for example using the sib-pair method. If relative pairs share marker alleles more often than would be expected by chance, this suggests that a susceptibility locus may be linked to the marker.

The statistical measure of linkage is the "logarithm of the odds", or LOD score. The LOD score is the log_{10} of the odds in favor of finding the observed combination of alleles at the loci studied if they are linked. Positive LOD scores favor the presence of linkage, whereas negative LOD scores indicate that linkage is less likely. A LOD score of +3 or greater is considered to be strong evidence of linkage (1000:1 odds for linkage).

Linkage analysis helps identifying a candidate region, then positional cloning is used to narrow the candidate region until the gene and its mutations are found. This approach has been successful in identifying the high-risk genes *BRCA1* and *BRCA2*. However, these investigations to map the site of breast cancer susceptibility genes require recruitment of large families with multiple affected relatives, hence creating a limitation for the use of this methodology. For lower risk variants, association studies provide a more powerful approach.

Association studies

Association studies compare the frequency of genetic variants in breast cancer cases and controls, and are convenient because they do not require high-risk families, as does linkage analysis. The power to detect alleles of modest effect is much larger for association than linkage studies.

The candidate gene approach Until a few years ago, almost all association studies focused on candidate genes selected by the investigators based on their potential role in tumorigenesis. Such candidate genes encode proteins involved in apoptosis, cell cycle control or DNA repair for instance. Association studies aim to detect alleles in these candidate genes, which influence susceptibility to disease themselves or which are in linkage disequilibrium with the disease causing variant.

Some issues that have hampered association studies of candidate genes are small study size, a limited number of markers used to characterize the gene, failure to adjust for multiple testing and lack of replication of findings. Another issue to take into consideration when looking at results from these studies is the potential bias towards publishing significant findings. The more extreme a finding is, the more likely it is to be published (publication bias). Further, researchers may not even submit negative findings for publication (selective reporting bias). These factors largely affect the power of this kind of study.

Genome-wide association studies The human gene pool harbors an estimated 10 million common SNPs. Groups of SNPs in close physical proximity to each other are often in linkage disequilibrium; these tend to be transmitted together across generations, resulting in so-called haplotype blocks. Because of the disequilibrium, it only takes a few tag SNPs to capture the great majority of SNP variation within each block [Gabriel et al., 2002]. The ability of SNPs to tag DNA haplotypes underlies the rationale for genome wide association (GWA) study. The SNPs contained in the human genome may either directly cause changes in phenotype or tag nearby mutations containing the causal variant that influence individual variation and susceptibility to disease.

Although GWA does not differ from candidate gene SNP association studies technically, the scale of genotyping with hundreds of thousands of SNPs in large series of patients has proven to be successful. Surprisingly, most of the SNP variations associated with disease have not been found in the coding region of DNA. Instead, they were usually located in the large non-coding DNA regions or in intronic sequences.

The major difference between a GWA study and a candidate gene study is that GWA studies do not make any prior assumptions about genes and their functions. The associated genetic variations are considered as indicators of the region of the human genome where the causal variant is likely to reside. Most genetic variations are associated with the geographical and historical populations in which the mutations first arose. Thus, studies must take account of the geographical and racial background of the individual enrolled in such studies, controlling for what is called population stratification.

Case-control mutation screening studies

Case-control mutation screening is an approach that is designed to address the challenge of identifying genes that harbor uncommon or rare intermediate risk variants. If there are strong a priori reasons to suspect that a particular gene may influence a trait then this gene may be screened for functional variants even before there is any mapping data to implicate it.

Mutation screening can be split into two fundamental processes. A primary screen is used to detect the presence of a sequence variation in a particular DNA fragment. A secondary screen (usually by Sanger sequencing) is used to confirm the results of the primary screen.

Primary screen can be achieved through a variety of methods, from conformational analyses (SSCP: single strand conformation polymorphism), heteroduplex analyses (DGGE: denaturing gradient gel electrophoresis, DHPLC: denaturing high performance liquid chromatography, HRM: high-resolution melting curve analysis) or protein truncation tests (PTT) [Isaacs and Rebbeck, 2008]. High density microarrays have also made it possible to perform mutation screening using chips, either through sequencing by hybridization (SBH) or through arrayed primer extension (APEX).

II.2.2 Genetic risk, a continuous variable.

What is the risk spectrum of breast cancer susceptibility genes and their variants?

Susceptibility genes and their pathogenic sequence variants fall into a spectrum from high-risk through intermediate-risk to modest-risk. High-risk refers to sequence variants with odds ratios (OR) $\geq$ 5.0, intermediate-risk refers to odds ratios in the range of 5.0> OR >2.0, and modest-risk refers to OR $\leq$ 2.0.

Carrier frequency can also be divided into 3 strata. Common refers to sequence variant with allele frequencies ≥10%, uncommon refers to variants with frequencies in the range of 1% to 10% and rare refers to variants with frequencies of <1%.

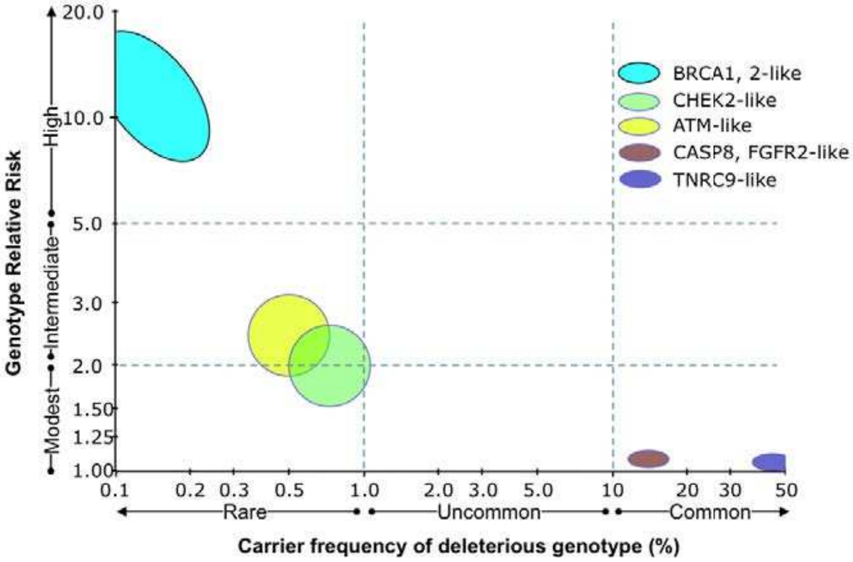


Figure II.1: Risk spectrum of breast cancer susceptibility gene and their genetic variants carrier frequency. After [Boyle and Levin, 2008]

Thus, in terms of relative risks and allele frequencies, nine categories of deleterious sequence variants/ cancer susceptibility genes are defined by the 3x3 stratifications (Figure II.1).

What fraction of the risk of common cancer is attributable to each of these categories of genes/sequence variants?

High-risk genes and variants. For common cancers, no common high-risk variants have been identified for common cancers. Given the constraints on incidence and observed familial risk, it seems that this category of variants does not exist. Uncommon high-risk variants are sometimes found as founder mutations in specific population but appear not to exist in the general population [Boyle and Levin, 2008]. Lastly, linkage analysis followed by positional cloning led to the discovery of susceptibility genes, such as *BRCA1* and *BRCA2* that harbor rare, high-risk variants.

High-risk of breast cancer also involves rare syndromes caused by germline mutations in *TP53* and *PTEN*. These mutations are very rare and hence account for a smaller proportion of the familial risk. These 4 genes and associated syndromes will be detailed in the next section. *LKB1/STK11* (Peutz-Jeghers syndrome) or *CDH1* (hereditary diffuse gastric cancer syndrome) are also associated with elevated risks of breast cancer, although the risks and prevalence of mutations in these genes are not well defined [Oldenburg et al., 2007, Stratton and Rahman, 2008]. Studies of the role that these six genes together play in the risk of breast cancer are not consistent with their accounting for more than 20% of the familial risk of the disease [Thompson and Easton, 2004, Antoniou and Easton, 2006].

Genome-wide analyses using large numbers of families without mutations in *BRCA1* or *BRCA2* have not mapped additional high-risk susceptibility loci [Smith et al., 2006]. Although this does not exclude the existence of as yet unidentified mutations associated with high-risk of breast cancer, it strongly suggests that, if they exist, they are not likely to account for a fraction of the familial aggregation of breast cancer as large as that attributed to the established high-risk genes.

This suggests that the remaining $\sim 80\%$ of the familial risk of breast cancer must be explained by the other categories of genes and variants.

Intermediate-risk genes and variants. The two best-understood intermediate-risk genes for breast cancer susceptibility are *CHEK2* and *ATM* [Meijers-Heijboer et al., 2002, Ahmed and Rahman, 2006]. As described more thoroughly in the next section, *CHEK2* and *ATM* both encode checkpoint kinases involved in DNA repair. *BRIP1* and *PALB2* have been more recently described as intermediate risk genes [Seal et al., 2006, Rahman et al., 2007]. BRIP1 was discovered as a binding partner of BRCA1 and is implicated in BRCA1 activities related to DNA repair. PALB2 was discovered as a protein associated with BRCA2 and is also involved in DNA repair activities.

Although there is currently some imprecision in the risk estimates, it is clear that mutations in *ATM*, *CHEK2*, *BRIP1* and *PALB2* confer an approximately two to threefold risk of breast cancer. In each of these four genes, there are multiple different pathogenic mutations, each of which is generally uncommon or rare (Figure II.1). Current estimates suggest that mutations in these four genes together account for 2.3% of the familial risk of breast cancer [Stratton and Rahman, 2008].

Modest-risk genes and variants. As shown on the graph of Figure II.1, breast cancer familial aggregation is also explained by common variants that confer very modest increases in risk. The currently known susceptibility alleles of this type have been discovered through association studies, either through the candidate gene approach or, more recently, through GWA studies. Progress in this area has been enabled by pooling of the data and resources from very large numbers of cases and controls from many different locations and ethnic group, in order to reach substantial power to detect small effects [Stratton and Rahman, 2008].

GWA studies have successfully identified a number of loci associated with modest increases of breast cancer risk [Easton et al., 2007, Stratton and Rahman, 2008, Isaacs and Rebbeck, 2008]. These include *CASP8*, which encodes caspase 8, a member of the cysteine-aspartic acid protease family involved in apoptosis, and *FGFR2*, which encodes the fibroblast growth factor receptor. Susceptibility loci have also been associated to regions with no known protein-coding genes (8q and 2q) [Easton et al., 2007].

The population prevalence of each risk allele is high, ranging from 28% to 87% [Stratton and Rahman, 2008]. However, the increased risks of breast cancer conferred by these susceptibility alleles are low. This set of variants confer risks of 1.3-fold or less, with the FGFR2 and 2q susceptibility alleles at the high end of this spectrum [Stratton and Rahman, 2008, Boyle and Levin, 2008]. Thus, although these predisposing alleles are common, their contribution to the familial risk of breast cancer is relatively small and support a polygenic model for breast cancer susceptibility.

II.2.3 A polygenic model for breast cancer susceptibility

Germline mutations in well identified susceptibility genes account for a relatively small proportion of the total breast cancer incidence (approximately 5-10%) [Peto et al., 1999, Nagy et al., 2004]. However, lifestyle and environmental factors that cluster in families are unlikely to explain all of the residual familial clustering, so the obvious implication is that additional susceptibility genes do exist [Struewing, 2004, Antoniou and Easton, 2006, Oldenburg et al., 2007]. Some argue that there must still be some unknown, rare, highly penetrant mutations accounting for breast cancer cases in such high- risk families [Walsh and King, 2007].

However, others have argued that the polygenic model is the best fitting model to account for the residual familial aggregation of breast cancer after excluding the known high-penetrance mutations [Houlston and Peto, 2004, Antoniou and Easton, 2006]. Under this model, susceptibility to breast cancer is conferred by a large number of genetic variants which combine additively or multiplicatively, resulting in a range of susceptibilities in the population. The risk associated with each one of these is small, but a woman with several susceptibility alleles is at a higher risk.

As mentioned earlier, there have been multiple large-scale searches for genes involved in the susceptibility to breast cancer using association studies. Technical advances have established public health (in terms of attributable fraction) importance of modest-risk SNPs, but clinical relevance has yet to be established for either modest-risk SNPs or mutations in the intermediate-risk genes such as *ATM* and *CHEK2*. The fraction of missing heritability that is going to be explained by as yet unknown genetic contributions from modest-risk or intermediate-risk susceptibility genes is not known. However, incorporation of these genes into polygenic models should allow for better risk prediction. Then, whether this would translate to cost-effective improvement in patient management remains an open question.

II.3 Known breast cancer susceptibility genes and their associated hereditary syndromes

To date, along with colon cancer, genetic susceptibility to breast cancer is better understood than genetic susceptibility to any other common cancers. As mentioned briefly in the previous section, a number of susceptibility genes have been identified and associated with inherited breast cancer syndromes.

An estimated 20-25% of familial aggregation is explained by known inherited breast cancer genes that are divided into "high-risk" and "modest to intermediate risk" breast cancer susceptibility genes [Thompson and Easton, 2004, Oldenburg et al., 2007]. The firsts confer a relative lifetime risk higher than 4 and generally much higher at young ages. The seconds are associated with a doubling of breast cancer risk [Walsh and King, 2007, Oldenburg et al., 2007].

II.3.1 High-risk genes

BRCA1 and *BRCA2*, the hereditary breast ovarian cancer syndrome

The two most important breast cancer susceptibility genes, *BRCA1* and *BRCA2*, were identified by linkage analysis and positional cloning in the mid 1990's [Miki et al., 1994, Wooster et al., 1995]. The *BRCA1* gene is located on chromosome 17q21 and the *BRCA2* gene is located on chromosome 13q12. *BRCA1* has 24 exons and encodes a protein of 1863 amino-acid (AA), while *BRCA2* has 27 exons and encodes a protein of 3418 AA. Both are ubiquitously expressed in humans with the highest levels in testis, ovaries and thymus. These two genes belong to the caretaker category of tumor suppressor genes.

Germline mutations in *BRCA1* and *BRCA2* are rare but confer high risks of hereditary breast ovarian cancer (HBOC) syndrome. Indeed, the associated relative risk is approximately 10- to 20-fold. An increased relative risk of male breast cancer has been found in *BRCA2* mutation carriers. It was initially hypothesized that the vast majority of multiple-case breast cancer families and families with HBOC would be caused by mutations in *BRCA1* or *BRCA2* genes [Honrado et al., 2006].

The frequency of *BRCA1* and *BRCA2* mutations has been estimated to be 0.1% for either genes, in most populations [Antoniou et al., 2002]. However, the occurrence can be higher for founder mutations. For instance, the frequency of the 6174delT mutation in *BRCA2* in the Ashkenazi Jewish population is 1.5% (although this mutation seems not to be restricted to this population) [Berman et al., 1996]. A recent meta-analysis of twenty-two population-based and hospital-based studies reported average risks by age 70 years of 65% and 45% in *BRCA1* and *BRCA2*-mutation carriers respectively [Antoniou et al., 2003].

The most significant cancer, other than breast cancer, in *BRCA1* and *BRCA2* mutation carriers is ovarian cancer, as is suggested by the name of the syndrome.

The average cumulative risk of ovarian cancer by age 70 years were 39% and 11% in *BRCA1* and *BRCA2* mutation carriers respectively [Antoniou et al., 2003]. The figures were higher in HBOC families with early-onset index cases (51% and 32% respectively). *BRCA1* and *BRCA2* mutation carriers are also at elevated risks of other cancers, such as malignant melanoma as well as colorectal, gastric, pancreatic, uterine and prostate cancers [Stewart and Kleihues, 2003, Isaacs and Rebbeck, 2008].

Genetic testing for mutations in these genes in high-risk families is now well established. In addition to *BRCA1* and *BRCA2*, other genes can be considered well established breast cancer susceptibility genes and account for the residual risk of familial breast cancer [Walsh and King, 2007].

TP53 **and Li-Fraumeni syndrome**

The *TP53* gene is located on chromosome 17p13.1, and encodes a protein involved in many cellular pathways that control cell proliferation and homeostasis, such as cell cycle, apoptosis and DNA-repair. The expression of the *TP53* gene is activated in response to various stress signals, including DNA damage. However, the proportion of early-onset breast cancer in the general population explained by *TP53* mutations is small as mutations are rarer than *BRCA1* and *BRCA2* mutations [Isaacs and Rebbeck, 2008]. Fewer than 400 families with germline mutations have been reported worldwide. These germline mutations have been implicated in most families with Li–Fraumeni (LF) syndrome [Oldenburg et al., 2007, Isaacs and Rebbeck, 2008].

LF syndrome is characterized by childhood sarcoma, early-onset breast cancer, brain tumors and a variety of other cancers. Mutations in the *TP53* gene account for about 70% of families fulfilling the classical criteria for LF syndrome (e.g. one patient with a sarcoma diagnosed $<$ 45 years, with a first-degree relative with any cancer diagnosed $<$ 45 years, and an additional first or second degree relative diagnosed with cancer $<$ 45 years or a sarcoma at any age), whereas they are less common in breast cancer/sarcoma families selected solely on the occurrence of breast and/or ovarian cancer [Oldenburg et al., 2007].

Susceptibility to cancer in LF families follows an autosomal dominant pattern of inheritance and among families with a known germline *TP53* mutation, the probability of developing an invasive cancer (excluding carcinomas of the skin) approaches 50% by the age of 30, as compared to an age-adjusted population incidence of cancer of 1%. It is estimated that more than 90% of *TP53* mutation carriers will develop cancer by the age of 70.

One of the most frequently occurring cancers in LF families is breast cancer with an estimated penetrance in *TP53* mutation carriers of 28–56% by the age of 45 years. The peak incidence for breast cancer is between 20 and 40 years, in contrast to the other frequent occurring neoplasms, which mainly develop in young children, suggesting that hormonal stimulation of the mammary glands in puberty is an important cofactor. Germline mutations in *TP53* are found at very low prevalence (<0.5%) among unselected, early-onset cases of breast cancer.

PTEN and Cowden's syndrome

Cowden's syndrome (CS) is an autosomal dominant disorder, characterized by carcinomas of the breast, thyroid and endometrium, multiple hamartomas and mucocutaneous lesions. Published estimates of CS incidence are 1/200,000 (see Table I.3). This syndrome is caused by germline mutations in the *PTEN* tumor suppressor gene [Liaw et al., 1997]. *PTEN* is located on chromosome 10q23.3 and codes for a phosphatase with both phospholipid and protein phosphatase activities, and plays a crucial role in controlling cell growth and migration, mediating apoptosis and cell cycle arrest. Mutations in *PTEN* have been involved in sporadic cancers but 80% of CS families present a germline mutation in this gene.

The most common malignancy seen in CS is adenocarcinoma of the breast, with lifetime risks in female patients estimated to be 25 to 50%, compared to 12 to 13% in the general population [Isaacs and Rebbeck, 2008]. The average onset of breast cancer in such patients is lower (30-35 years) than in sporadic cases, as commonly described in other inherited cancer syndromes. Fackenthal et al. have also suggested an increased risk for male breast cancer in CS patients carrying a germline mutation in *PTEN* [Fackenthal et al., 2001].

II.3.2 Intermediate-risk genes

ATM

The *ATM* gene is located on chromosome 11q22–23. It codes for a protein kinase whose many substrates include the products of *TP53*, *BRCA1* and *CHEK2*. ATM plays a central role in sensing and signaling the presence of DNA double-strand breaks. Mutations in the *ATM* gene cause the rare recessive disorder Ataxia-Telangiectasia (AT) (Table I.3) [Savitsky et al., 1995]. AT is characterized by cerebellar degeneration (ataxia), dilated blood vessels in the eyes and skin (telangiectasia), immunodeficiency, chromosomal instability, increased sensitivity to ionizing radiation and a highly increased susceptibility to cancer, in particular leukaemia's and lymphomas, as reviewed in [Chun and Gatti, 2004, Taylor and Byrd, 2005].

The estimated incidence of AT is 1:30,000 to 1:100,000 with an *ATM* mutation carrier frequency of approximately 0.5% [Nagy et al., 2004, Ahmed and Rahman, 2006, Stratton and Rahman, 2008]. Studies based on relatives of AT and breast cancer case-control studies have estimated that the relative risk of breast cancer in heterozygous carriers of *ATM* mutations is in the order of 2 [Thompson et al., 2005*b*], with some evidence of higher relative risk under the age of 50 years. Specific mutations, notably 7271T>G, may confer higher breast cancer risks. The results have not been replicated in subsequent studies [Thompson et al., 2005*a*, Thompson et al., 2005*b*, Oldenburg et al., 2007]. However, Tavtigian and colleagues did show that missense substitutions in the FAT and Kinase domains, including 7271T>G, confer greater risk than do truncating variants [Tavtigian et al., 2009]. The role of missense substitutions uncovered in this paper also somewhat increases the best estimate for the population carrier frequencies of variants in *ATM* that are pathogenic for breast cancer.

CHEK2

The *CHEK2* gene is located on chromosome 22q12.1 and there are several *CHEK2* pseudogenes scattered throughout the genome. CHEK2 is involved in the maintenance of genomic stability and functions downstream of ATM to phosphorylate several substrates, including p53, Cdc25C and BRCA1, leading to cell cycle arrest, activation of DNA repair or apoptosis in response to DNA double-stranded breaks. Since *CHEK2* plays a key role in the DNA damage pathway, loss of function of the protein may allow cells to evade normal cell cycle checkpoints, ultimately leading to tumor initiation or progression.

The CHEK2*1100delC deletion, falling in the kinase domain of the protein, has been widely studied for its contribution to inherited breast cancer susceptibility [Oldenburg et al., 2003]. The frequency of CHEK2*1100delC differs between ethnic populations, and is higher in the North of Europe and low or absent in other countries [Honrado et al., 2006]. The CHEK2-Breast Cancer Consortium reported a frequency of 5.1% for the CHEK2*1100delC variant in familial breast cancer cases who tested negative for *BRCA1* and *BRCA2* mutations, as opposed to 1.1% of carriers in the control population [Meijers-Heijboer et al., 2002]. This intermediate-risk breast cancer susceptibility allele almost triples the risk of developing the disease in unselected breast cancer cases [CHEK2 Breast Cancer Case-Control Consortium, 2004].

Whereas no other pathogenic variants occur at a significant frequency, other founder mutations in *CHEK2* have been associated with an increased risk of cancer. As reviewed in [Antoni et al., 2007], the I157T mutation in exon 3 also shows some population specificity but it confers a more modest risk than the CHEK2*1100delC allele. This missense mutation is reported to increase the risk of breast cancer by 1.5 fold. Further, the splicing mutation IVS2+1G>A, which creates a premature termination codon, has a frequency of 0.3% in the Polish population and is much rarer in other studied populations. This allele is associated with a 2- to 4-fold increased risk of breast cancer. In addition, S428F, a substitution specific to the Ashkenazi population has been identified in exon 11, increasing breast cancer risk by approximately 2-fold [Shaag et al., 2005]. Lastly, a 5 395 bp genomic deletion that leads to the loss of exons 9 and 10, resulting in a truncated protein, has been identified in families of central Europe ancestry [Cybulski et al., 2007]. The deletion was present in 1% of unselected breast cancer cases and in 0.9% of the early-onset cases.

Though first discovered in breast cancer patients, *CHEK2* mutations have since been reported to predispose to a range of cancer types, including ovarian, prostate, kidney and colorectal cancers [Nevanlinna and Bartek, 2006], supporting the hypothesis that *CHEK2* is a multiorgan cancer susceptibility gene [Antoni et al., 2007].

II.4 Outlook on the search for the "missing heritability" of breast cancer

Despite the remarkable progress made in the past decades, most of the familial risk of breast cancer remains unexplained, highlighting the need for ongoing efforts to identify the "missing heritability" of breast cancer. The remaining familial aggregation of breast cancer will likely be explained by three different categories of variants: rare high-risk variants, rare intermediate-risk variants and common low-penetrance variants [Stratton and Rahman, 2008].

The identification of variants underlying each of these categories requires different strategies and technologies. GWA studies have successfully identified a number of common genetic variants associated with very modest increases of breast cancer risk. There will certainly be additional common genetic variants identified through this approach [Stratton and Rahman, 2008]. However, such new variants will not contribute significantly to the remaining unexplained familial aggregation of breast cancer since they will be associated with very small effects.

The remaining familial clustering of breast cancer will likely be explained by rare genetic mutations in genes that convey an intermediate to high-risk of breast cancer. Massive parallel sequencing (MPS), or "next-generation sequencing", offers a new strategy to discover such category of genes and variants. MPS enables whole-exome mutation screening in pedigrees, with considerable throughput advantages over older sequencing and mutation screening techniques. One can expect that growth in the power of MPS techniques will help case-control mutation screening to evolve from candidate gene through whole pathway to exome and then genome as did single marker SNP association studies. The power of such an approach relies on a careful selection of the pedigrees to be included in the study, in order to maximize the likelihood to find a single, high-penetrance, autosomal dominant mutation segregating in each pedigree, that would account for the breast cancer cases in that family.

III

Gene expression regulation

Understanding the genetic and molecular mechanisms that give rise to phenotypic differences in humans and other complex genomes remains a major challenge. Such differences can arise from a single DNA sequence variant affecting a protein coding sequence. Historically, earlier studies have focused almost exclusively on such variants, in coding sequences and regions immediately surrounding candidate genes.

However, recent technological developments have enabled whole genome scans that interrogated most of the human genome, including non-coding DNA regions that had not been studied previously. These whole genome association (WGA) studies found some of the strongest signals of association in non-coding regions, either in large introns or far way from any annotated loci. The mechanisms connecting the identified sequence variants to the etiology of diseases are still unclear but regulation of gene expression remains a foremost candidate.

III.1 The determinants of gene expression regulation

Studies of gene regulation have classified regulatory interactions based on their effect in *cis* or *trans*. The terms *cis* and *trans* were introduced by Haldane to describe differences in the configuration of mutant alleles in heterozygotes, in analogy to *cis* and *trans* isomers in chemistry. In the *cis* configuration, two mutations were inherited together, whereas in *trans*, they would be on different members of a pair of homologous chromosomes. Nowadays, these terms are used to describe particular types of regulatory interactions.

However, in gene expression regulation literature, a certain confusion exists on the usage of the terms *cis* and *trans*. Some use *cis* and *trans* in accordance to the original definitions, with *cis* regulatory variation having an allele-specific impact on gene expression and *trans* regulatory variation affecting both alleles. The underlying molecular nature of the variation, *i.e.* whether the mutation acts at the level of DNA or RNA, or alters a protein is not specified.

Sequence variants are also often said to be in *cis* or *trans* on the basis of their physical distance from the regulated target gene. Regulatory variation mapped near the gene is classified as *cis*. This distance-based classification may lead to misclassification of long distance *cis* elements as *trans* (e.g. a regulatory element located far upstream its target transcription start site), as well as *trans* regulatory elements existing near their gene target (e.g. a transcription factor that regulates an adjacent gene).

Thus, as pinpointed by Rockman et al, using the same terms *cis* and *trans*, some describe the pattern of co-inheritance of a trait and a locus (-linking), whereas others describe the mechanism of action of a locus to a trait (-acting) [Rockman and Kruglyak, 2006]. In their review, the authors suggest to employ the terms "local" and "distant" to classify regulatory variations, underlining that both local and distant regulations can comprise sequence variants in *cis* and in *trans*-acting factors. In accordance with Rockman et al., we will use the term "*cis* " to refer to the variation located on the same gene and the same chromosome as the causal variant. Thus, an SNP falling in a regulatory region 500,000 bp upstream the target gene will be considered to have a *cis*-effect.

III.1.1 Local regulatory variation

Sequence variants can affect the target gene itself by altering classic *cis*-regulatory DNA sequences. These include promoter regions, enhancers, silencers and insulators, which regulate transcription initiation. Enhancers and silencers act over distance to potentiate or repress transcription. Insulator sequences prevent enhancers and silencers from inappropriately regulating a neighboring gene. Sequence variants in these *cis*-regulatory elements have an allele-specific effect on gene regulation [Lewin, 2004, Maston et al., 2006].

Although gene expression regulation occurs mostly at the level of transcription [Wray et al., 2003, Williams et al., 2007], mutations or polymorphism can also contribute to variation in gene expression at the post-transcriptional level. Genetic variation may occur outside the promoter regions and may involve introns and 5' or 3' untranslated regions (UTRs). Gene expression variation can then result from alterations of binding sites for molecular complexes involved in transcription initiation, mRNA stability, processing efficiency, or splicing, leading to differential recruitment of transcription factors and Small interfering RNAs (SiRNAs), differential concentrations of nuclear and cytoplasmic mRNA or differential mRNA isoform expression [Pastinen et al., 2004].

Sequence variants occurring in the coding region of a gene can also affect its expression. For instance, in the case of transcripts bearing a mutation that creates a premature stop codon, the nonsense-mediated mRNA decay (NMD) mechanism is triggered and specifically degrades the transcripts bearing the mutation [Conti and Izaurralde, 2005]. These are all potential important contributors to variation in gene expression in an allele-specific manner (Figure III.1).

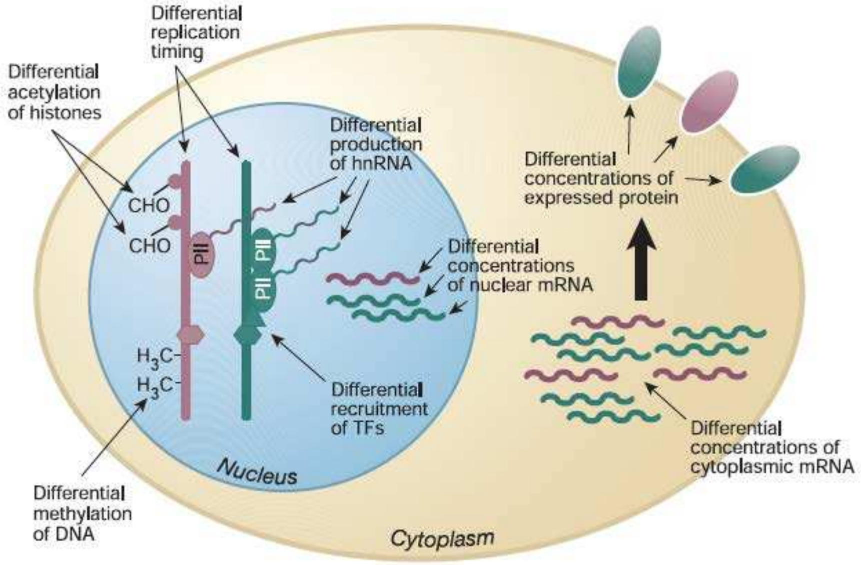

Differential replication timing
Differential acetylation of histones
CHO
CHO
PII
PII
PII
H$_3$C-
H$_3$C-
Differential production of hnRNA
Differential concentrations of expressed protein
Differential concentrations of nuclear mRNA
Differential recruitment of TFs
Differential concentrations of cytoplasmic mRNA
Nucleus
Differential methylation of DNA
Cytoplasm

Figure III.1: Cellular phenomena associated with *cis*-acting regulation.
Sequence variants in *cis*-regulatory elements have an allele-specific effect on gene regulation. PII, RNA polymerase II. After [Pastinen and Hudson, 2004]

Besides these allele-specific effect variations, Rockman et al. describe three other situations of local regulatory variation [Rockman and Kruglyak, 2006]. First, the case of autoregulatory genes, where sequence variants in the gene itself act in *trans* (*i.e.* affect both alleles) to modify its expression. This category of variation can as well refer to sequence variants in genes regulated by feedback loops. Finally, local regulatory variation can be due to a sequence variant located in a neighboring gene that regulates the expression of the gene of interest[1].

[1] Regulatory variation affecting gene expression in an allele-specific manner is *cis*-variation strictly speaking. However, in some instances, variation may affect both alleles although mapping of the gene will disclose regulatory loci coinciding with the position of the source gene, leading some reports in the literature to misuse the term "*cis*-acting regulation".

III.1.2 Distant regulatory variation

Distant regulatory variation is typically associated with transcription factors, which regulate transcription initiation. For instance, polymorphisms in their DNA-binding domain or in a protein-protein interaction domain are likely to account for differential recruitment of transcription factors, hence underlying *trans*-effects, which can be highly pleiotropic because of the large number of downstream genes that could be affected.

However variation in transcript abundance is not necessarily equivalent to variation in transcription *per se*. Although the most common point of control lies in transcription initiation, genetic variations influencing gene expression may also reside within several other sites of action. These mechanisms include polyadenylation and splicing, intracellular trafficking, mRNA decay, translational controls, post-translational modification, and protein decay. Sequence variants affecting any gene involved in one of these cellular machineries could act in *trans* on the regulation of gene expression.

Distant regulatory variation can as well display *cis*-effects in the case of cellular components involved in an allele-specific manner in post-transcriptional processes accounting for mRNA stability, processing and degradation (Figure III.1). Distant *cis*-acting regulatory elements also include enhancers, which are a type of regulatory sequences involved in transcription initiation. As mentioned above, enhancers elements can participate to local regulatory variation but can also be located a considerable distance upstream or downstream of their target gene, and modulate expression independently of their orientation. Enhancers are often targets of tissue-specific or temporal regulation [Lewin, 2004, Pennacchio et al., 2006, Visel et al., 2009]. This implicates that SNPs affecting binding of transcription factors in those enhancers may be tissue specific, and this further means that disease-specific studies may be limited by access to normal tissues of the relevant cell type from appropriate subjects.

III.1.3 Epigenetic factors

DNA methylation

Gene transcription can be affected by epigenetic mechanisms. Indeed, transcriptional repression of a gene can occur without altering its DNA sequence but rather through DNA methylation, a post-replication modification that is predominantly found in cytosines of the dinucleotide sequence CpG. An epigenetic inheritance is established as long as the maintenance methylase DNMT1 acts constitutively to copy the methylation state from the parent DNA strand to the daughter strand after each cycle of replication. Thus allele-specific expression can be associated with differential methylation of genomic loci [Jaenisch and Bird, 2003].

Histone modification

In addition, chromatin structure is an integral part of controlling gene expression. DNA is not directly packaged in the final structure of chromatin. There are several levels of organization. The nucleosome provides the fundamental building block of chromatin. Its component and structure are well characterized: it consists of about 200 bp of DNA spooled around an octamer of proteins called histones. The histone octamer itself has a kernel that consists of two copies each of H2A, H2B, H3 and H4 core histones, with the N-terminal tails extending out of the nucleosome. These tails have sites for modifications that are important for chromatin function and hence, for gene regulation. Modifications of histone tails that are triggers for chromatin reorganization occur on specific serine, lysine and arginine residues and include acetylation, which is usually associated with gene activation, methylation, which is associated either with gene activation or inactivation and lastly, phosphorylation and ubiquitination. Once established, changes in chromatin may persist through cell divisions, creating an epigenetic state in which the properties of the gene are determined by the self perpetuating structure of the chromatin. Post-translational histone modification is hence another epigenetic phenomenon that can be associated with allele-specific expression [Jaenisch and Bird, 2003].

The microRNA pathway

The discovery of microRNAs (miRNAs) added a new layer to the complexity of gene regulation. The miRNA pathway is an essential part of genetic regulation of ancient origin. miRNA are short RNA sequences abundant in many genomes, including worms, flies, plants and mammals [He and Hannon, 2004]. They interact with the Argonaute proteins to join an effector complex that targets the 3'UTR of a mRNA in order to induce silencing of the target gene. Lim et al. have shown that a single miRNA can downregulate expression of hundreds of its target genes when overexpressed in HeLa cell lines [Lim et al., 2005].

III.2 The study of gene expression phenotypes

Under the term "genetical genomics", Jansen and Nap were among the first to suggest combining expression profiling and mapping methods with the aim to further study the gene expression network as a whole. [Jansen and Nap, 2001]. Others have followed, and numerous studies, meticulously reviewed by Stamatoyannopoulos, have provided insight into our understanding of the role of gene expression regulation in organizing complex genomes [Stamatoyannopoulos, 2004].

III.2.1 What is the extent of natural variation in gene expression?

One question of central importance is the biological variability of gene expression in the context of naturally occurring populations. It was hypothesized more than three decades ago that much of phenotypic variation among closely related organisms is due to variation in gene expression rather than to alterations in protein sequences [King and Wilson, 1975].

Inter-species variation

Actually, it has been shown that the evolution of regulatory genetic pathways has played an important role in speciation [Johnson and Porter, 2000, Levine, 2002]. In a broad variety of organisms, changes in gene expression have led to morphological evolution and other adaptative phenotypes, such as beak morphology in Darwin's finches [Abzhanov et al., 2004], wing pigmentation patterns in flies [Gompel et al., 2005], branching structure in maize [Clark et al., 2006] and even parental care in rodents [Hammock and Young, 2005].

Another interesting example is provided by a class of regulatory genes, the Hox genes, which encode DNA-binding proteins and control early development in arthropods. The genetics of this phylum have been extensively analyzed in the last century, and the genes responsible for segmentation and limb development have been identified. Levine described some of the mechanisms of limb evolution identified to date [Levine, 2002]. Changes in Hox gene expression patterns are, for instance, responsible for the conversion of swimming limbs in branchiopods into feeding appendages in isopods, two species of crustaceans. Evolution also acted through changes in Hox target genes in different insects via evolution of Hox protein-binding sites. This second mechanism accounts for the dipterans (such as Drosophila) having rudimentary wings, called halteres, in place of hindwings in lepidopterans (such as moths).

Inter-population and inter-individual variations

Recently, several studies in humans, other mammals, yeast and plants have reported differences in gene expression levels accounting for a major part of inter-strain or inter-population variation. Based on the genetical genomics approach, studies were conducted for instance in yeasts and showed extensive genetic variation for gene expression [Brem et al., 2002, Yvert et al., 2003]. Brem et al. carried out investigation of inter-strain variation in expression levels of 6215 genes in *Saccharomyces cerevisiae.* They found that 1528 (25%) of the genes in the yeast genome were differentially expressed at $p<0.005$.

In humans, Spielman and colleagues first reported the analysis of HapMap samples to study the difference in gene expression that could be accounted for by genetic variants [Spielman et al., 2007]. They found a role for regulatory polymorphisms in the prevalence of complex diseases. Stranger and colleagues measured gene expression in 270 lymphoblastoid cell lines (LCLs) derived from unrelated individuals from the four HapMap populations, *i.e.* Caucasian from European ancestry (CEU), Chinese (CHB), Japanese (JPT), and Yoruba from Nigeria (YRI) [Stranger et al., 2007]. The authors tested population differences in gene expression and found that 17-29% of genes presented with significant differences in mean expression levels between pairs of HapMap populations. Storey et al. also used a subset of samples from the HapMap project to characterize patterns of natural gene-expression variation [Storey et al., 2007]. They studied 16 CEU and YRI unrelated individuals and found that about 17% of genes are differentially expressed among these two populations.

The same study by Storey and al. estimated that 83% of genes are differentially expressed among individuals. Several other studies have looked into inter-individual variability in gene expression [Oleksiak et al., 2002, Schadt et al., 2003, Whitney et al., 2003]. Whitney et al. have identified inter-individual variation while surveying variation in gene expression patterns in peripheral blood from 75 healthy donors.

Overall, these studies agree that most expression variation is due to variations among individuals rather than among populations. Nevertheless, the experimental design of these studies does not permit to distinguish the proportion of *cis* versus *trans*-acting factors. In addition to advance our understanding of the general mechanisms of gene regulation, determining the proportion of genes regulated by *cis*-acting factors will be of major importance for further fine-scale characterization of functional genetic variation [Vasemägi and Primmer, 2005].

III.2.2 How heritable are patterns of gene expression?

Considerable effort has been made to address this other fundamental question in the last decade and several studies describing the genetic basis of transcriptional variation have convincingly provided evidence that it is a heritable trait.

In their study of the yeast, Brem et al. concluded that the established expression phenotypes were highly heritable traits, with 84% of the median proportion of expression difference between strains explained by genetic variation [Brem et al., 2002].

More recent studies have shown that gene expression phenotypes are heritable in family pedigrees. Schadt et al. analyzed 40 descendants of 16 pedigree founders and observed that 29% of 2726 differentially expressed genes exhibited heritable expression phenotypes [Schadt et al., 2003]. In another study, Cheung et al. found clear evidence for familial aggregation of expression phenotypes by studying five genes previously found to present high inter-individual variability in 49 unrelated individuals, 41 siblings and 10 pairs of monozygotic twins. The greatest variability was found between unrelated individuals, intermediate variability among siblings and the least variability between twins [Cheung et al., 2003]. Monks et al. also used LCLs from the CEPH to perform a large survey of the heritability of gene-expression traits in segregating human populations [Monks et al., 2004]. They measured expression for 23 499 genes in LCL of 15 CEPH families members. Of the total set of genes, 2340 were found to be differentially expressed, of which 31% had significant heritability.

The above-mentioned studies have provided evidence for a significant heritable component of individual variation in gene expression.

III.3 Strategies for determining the architecture of gene expression regulation

Although accumulating evidence shows that regulatory variations contribute to many important phenotypes, the genetic architecture of gene expression regulation is still elusive. Since they can directly modify transcript abundance, sequence variants in regulatory elements have been proposed to be the major determinant of gene expression variation. Yet, unlike coding sequence variants where the consequences of non-synonymous variation may be resolved at the level of the protein phenotype, defining how variation at the DNA sequence level will induce differences in transcript abundance has proven problematic. Indeed, characterization of the effect of *cis*-acting sequence variants in regulatory regions is a great challenge due to the difficulty to locate these regions. In addition, regulatory variants are not robustly detected by sequence analysis since SNP identification by screening regulatory regions does not consistently allow prediction of the effect of observed SNPs on gene expression [Wang and Sadée, 2006, Gilad et al., 2008]. Thus, knowledge of the effect of genetic variants affecting mRNA transcription is very limited. Currently, two strategies are most commonly used for assaying expression levels for the purpose of uncovering the nature of their genetic bases.

III.3.1 Expression quantitative trait loci (eQTL) mapping

What are eQTLs?

Recently, mRNA transcript abundance has been considered as a quantitative trait (QT), *i.e.* a continuously variable phenotype, that can be used with considerable power as a surrogate to study gene expression regulation [Cookson et al., 2009]. Using the well-established linkage and association mapping approaches, expression quantitative trait loci (eQTL) mapping has become a widespread tool for identifying genetic variants that affect gene regulation. It is important to note that the term eQTL refers to the mapped locus that regulates the mRNA level and not the mRNA expression trait (the QT) itself.

The contribution of microarray analysis

Although the importance of gene regulation for controlling biological processes has been recognized for over 30 years, it is only in the last decade that the tools necessary to study changes in gene expression on a large-scale have become available, especially with the advent of microarrays. RNA abundance is measured by exploiting hybridization of RNA fragments to short sequences of complementary oligonucleotides, or probes, fixed to the array substrate. The measurement of the phenotype is correlated to the brightness of an array spot or probe. DNA microarrays allow to measure the expression phenotypes of many genes in a genome simultaneously. These phenotypes are then mapped to specific genomic regions using genome-wide genetic markers [Morley et al., 2004, Gilad et al., 2006].

Several groups have taken advantage of the microarray technology to perform global analyses of the variability of gene expression in order to further understand the role of transcriptional regulation. Such studies have provided insight into transcriptional regulation in yeast, mice, maize and humans [Schadt et al., 2003]. In particular, they have consistently helped to address two of the questions mentioned in the previous section, regarding the extent of natural variation of gene expression [Brem et al., 2002, Oleksiak et al., 2002, Whitney et al., 2003, Stranger et al., 2007] and its heritability [Cheung et al., 2003, Monks et al., 2004, Morley et al., 2004].

By allowing simultaneous capture of many regulatory interactions, DNA microarrays have enabled genome-wide mapping studies of eQTLs, thus enhancing our understanding of the genetic architecture of gene expression.

Linkage and association mapping

The identification of eQTLs relies on the principle that expression levels can be analyzed with the same study designs and statistical methods traditionally used for mapping any other quantitative trait phenotype. Thus, mapping methods for eQTL can be classified in linkage and association methods. As described in the previous chapter, linkage mapping uses a study design that is based on tracking the transmission of chromosomes in families. This approach aims to identify markers, or chromosomal segments, whose transmission pattern is correlated with the phenotype. By contrast, association mapping uses samples of unrelated individuals with the aim to identify markers whose genotype is correlated with the phenotype at the population level.

The limits to eQTL studies

Although they provide a global view, studies investigating inter-individual variation in gene expression using microarrays are limited by the accuracy of hybridization-based gene expression profiling. In addition, measurement of gene expression variation levels by microarrays may be affected by many non-genetic factors such as environmental effects, epigenetic modifications, as well as experimental exposures such as differences in establishing and culturing cell lines, or, for experiments from primary tissues, subtle differences in tissue acquisition conditions between subjects.

III.3.2 Differential allelic expression (DAE) assays

Differential allelic expression (DAE) studies represent a different approach than eQTL studies to discovering factors that might affect gene expression levels.

Allelic variation in gene expression

Mendelian inheritance assumes that genes from maternal and paternal chromosomes contribute equally to human development. However, DAE, *i.e.* the preferential expression between alleles, appears to be a common feature of the human genome. DAE has traditionally been associated with X-chromosome inactivation, *i.e.* the silencing of one of the X-chromosomes, and imprinting, *i.e.* the expression of certain genes in a parent-of-origin-specific manner.

Cowles et al. have examined DAE in 69 mouse genes [Cowles et al., 2002]. They screened spleen, liver and brain tissues of two F1 hybrid mice from five strains' combinations and thus five genetic backgrounds. With an average ratio of 1.5 as a threshold for detection, they identified 4 genes clearly displaying *cis*-variation in expression.

Apart from this one study conducted in mice, allelic variation in gene expression has more frequently been analyzed in human populations. Recently, several studies have investigated DAE in autosomal non-imprinted genes and found that allelic differences in gene expression are relatively common across the human genome. Indeed, allelic variation may affect up to 50% of human genes [Yan et al., 2002*b*, Bray et al., 2003, Lo et al., 2003, Pastinen et al., 2004, Pant et al., 2006, Serre et al., 2008]. Furthermore, Yan et al. used a pedigree analysis of families of individuals showing DAE and demonstrated that allele specific differences in expression were transmitted by Mendelian inheritance [Yan et al., 2002*b*].

An alternate approach to microarray expression profiling

Allele-specific expression assays represent a fundamentally different approach to investigating factors affecting gene expression levels. In such studies, disruption or alteration of gene expression levels is examined through a careful survey of whether the two alleles of a gene are equally expressed. This approach relies on relative quantification of allelic transcripts within heterozygous individuals, using a transcribed SNP as marker. It has major advantages over more conventional methods for investigating gene expression variation based on a comparison between individuals, as discussed elsewhere [Bray et al., 2003, Buckland, 2004, Pastinen and Hudson, 2004, Jordheim et al., 2008]. Since they come from the same tissue sample and have therefore been subjected to the same environmental influences (such as genetic *trans*-acting factors and experimental exposures, including mRNA degradation) both alleles should be equally expressed in the absence of *cis*-acting sequence variation or allele-specific epigenetic effects affecting the expression of the target mRNA. Thus, the strength of this approach is that each allele acts as an internal control for confounding factors, disclosing *cis*-variation effects without being confounded by any *trans*-variation effects.

DAE analysis has the potential to enhance our ability to identify regulatory genetic variation by revealing the existence of regulatory variations without directly identifying or requiring prior knowledge of specific *cis*-regulatory SNPs. In some cases, observation of DAE will be explained by genetic variants in coding regions, such as truncating mutations resulting in NMD or splice junction mutations resulting in an unstable transcript. DAE can also be the signature of a heterozygous carriage of a regulatory variant. In addition, DAE assays can highlight the existence of epigenetic factors controlling gene expression, which would not have been detected by standard eQTL approaches.

Experimental approaches to allele-specific expression studies

Thus far, several techniques have been reported for quantitative analysis of DAE. Reporter gene assays are one approach to assess allele-specific transcript abundance. The limit of these systems is that they produce results outside of the normal chromosomal context and do not permit the elucidation of epigenetic effects [Wang and Sadée, 2006]. In contrast, direct assessment of the relative abundance of allelic-transcripts allows investigation of DAE in a normal chromosomal environment. In such studies, measurements of allelic expression require prior amplification of the region surrounding an intragenic marker SNP in cDNA, as well as in genomic DNA, where both copies of the gene are assumed to be present in equal proportions [Pastinen and Hudson, 2004]. Detection of DAE is then based on the detection of deviation of the allelic ratio in cDNA from the expected equimolar allelic ratio, in samples from subjects who are heterozygous for the marker SNP, and ideally with probe signals indicative of equimolarity provided by reference to genomic DNA from the same subjects and departures from equimolarity calibrated by reference to controlled mixing experiments.

For instance, single-base extension (SBE) assays starts with prior amplification of the region surrounding the intragenic marker SNP, by PCR for genomic DNA or RT-PCR for cDNA, then SBE of a primer adjacent to the polymorphic site in the presence of fluorescently labeled dideoxynucleotides, and finally detection on a DNA sequencer. Peak heights correlate with the relative transcripts levels for the two alleles and the allelic ratio is inferred by comparison with known mixtures used as a reference standard [Yan et al., 2002*b*, Cowles et al., 2002, Pastinen et al., 2004].

Allele-specific quantitative real-time PCR (qRT-PCR) is another assay commonly used for DAE assessment. qRT-PCR often serves to further confirm findings in studies using other gene expression assessment methodology in the first place [Lo et al., 2003, Storey et al., 2007]. This method has also been used by others as a stand-alone method [Chen et al., 2008, Maia et al., 2009, Bellini et al., 2010], usually using Taqman® technology. With this approach, each allele is detected by a complementary probe labeled with a different fluorochrome (usually VIC and FAM) generating two distinct signals during the qRT-PCR. Relative transcript levels are extrapolated from the linear relation between Log2[allelic ratio] and ΔCT.

High-throughput analysis was performed by Lo et al., who screened human fetal liver and kidney tissues for DAE using a microarray platform (Lo et al. 2003). The authors used a genotyping technology, the Affymetrix HuSNP chip system, where the alleles are distinguished on the basis of their hybridization specificity to probes matching the two allelic forms of the marker SNP. These data suggest that their high-throughput method is quantitative enough to confidently detect genes with a greater than twofold differences in allelic expression. Other studies have used the same method of allele-specific arrays [Ronald et al., 2005, Pant et al., 2006, Milani et al., 2007, Serre et al., 2008]

DAE can also be determined by polymerase loading assay, which is based on isolating transcriptionally active DNA fragments by immunoprecipitating DNA bound to the RNA polymerase II enzyme. The isolated DNA fragments can be assessed for SNPs in heterozygous samples to determine relative allelic transcriptional activity as a surrogate for relative allelic expression [Knight et al., 2003]. This technique abrogates the need for a transcribed polymorphism in the gene of interest, which is one of the limitations of the other approaches to DAE assessment. Any polymorphism in coding or non-coding DNA within 1kb of the transcriptional start site or the 3'UTR can be used as marker with this method.

III.4 The relevance of expression studies to disease susceptibility

Altered gene expression of multiple genes could represent a common mechanism for inherited susceptibility to complex diseases or variation in drug responsiveness. Thus the study of variation affecting gene expression, mRNA processing and translation may lead to the identifications of biomarkers and optimized therapy.

III.4.1 eQTL studies

Genome-wide association (GWA) studies of complex human diseases have been spectacularly successful in identifying new loci in the past few years. However, there is a substantial gap between SNP association from a GWA study and understanding the contribution of the locus to disease. Current evidence suggests that only a small fraction of the causal loci consists of variants directly affecting the protein amino-acid sequence. Thus, a large fraction of the identified loci are expected to have a regulatory role on gene expression via effects on transcription, mRNA stability and splicing. Combining eQTL and GWA studies allows to characterize those loci that are identified outside coding regions and thus, likely to be involved in gene regulation.

For instance, the utility of combining eQTL and disease mapping studies is illustrated by a GWA study of asthma that identified a series of SNPs strongly associated with the risk of disease [Moffatt et al., 2007]. The region of association contained 19 genes, none of which were obvious candidates for disease susceptibility. eQTL data on the same families provided evidence for a consistent and strong association ($p<10^{-22}$) of asthma-associated SNPs and *cis*-effect on transcript levels for one of the gene: *ORMDL3* (ORM1-like 3).

Simultaneous assessment of gene expression and genetic variation on a genome-wide basis, in a large number of individuals, can provide substantial information for dissecting the genetics of complex diseases. eQTL studies can allow to assign a regulatory role to a polymorphism located in a genomic region identified in GWA studies of disease, for which there might not be prior evidence of potential effect on a particular phenotype or disease susceptibility [Meyer et al., 2008, Fransen et al., 2010].

Hence combining eQTL mapping with results from traditional linkage or association studies has been recognized as a very promising strategy for identifying sequence variants underlying complex traits. eQTL studies may also allow to link genes and genetic variants to cellular phenotypes, such as sensitivity and response to chemotherapy. Lastly, eQTL studies may provide insight on whole networks of genes associated with complex traits and diseases [Gilad et al., 2006, Cookson et al., 2009].

III.4.2 DAE studies

Several studies of genes involved in known disease-related pathways have addressed the question of whether small genetic changes in these genes can give rise to alterations in transcript abundance sufficient to predispose an individual to disease.

Excluding results due to nonsense mediated decay, the first example of DAE in germline cells was described by Yan et al. in individuals with familial adenomatous polyposis (FAP), for whom no mutation in the *APC* gene had been identified [Yan et al., 2002*a*]. The authors found that decreased expression of one of the *APC* alleles was associated with risk of FAP. They also report that subtle changes in the expression of *APC* could contribute to attenuated forms of polyposis.

Interleukin-10 (IL10) plays a key role in the regulation of immune response and thus is involved in the pathogenesis and outcomes of various diseases. Kurreeman et al. conducted a study of genetic variations in the *IL10* gene aiming at analyzing the large inter-individual differences in the production of IL10 [Kurreeman et al., 2004]. They found that allele-specific regulation at the mRNA level were responsible for the differential production of IL10 among individuals.

The *DAPK1* gene, which has a pro-apoptotic role in the programmed cell death pathway and is predominantly expressed in the brain and lung, has been found to be strongly associated with an elevated risk of Alzheimer's disease (AD) [Li et al., 2006]. As *DAPK1* transcripts are differentially expressed, the authors suggested that *DAPK1* genotype or activity may influence risk of AD by influencing the cell number in the hippocampus and/ or by influencing the response to environmental stimuli such as amyloid beta.

More recently, Chen et al. have shown that the *BRCA1* and *BRCA2* genes displayed DAE and that these differences in expression contribute to risk of breast cancer [Chen et al., 2008]. The authors found DAE to be more significant for *BRCA1* than *BRCA2* for both familial and non-familial breast cancer patients, as compared to cancer-free women.

In conclusion, recent literature on the genetic predisposition to various diseases reports growing evidence for the role of DAE, confirming that alternate mechanisms than deleterious coding mutations are likely to contribute to disease susceptibility.

Aims of the thesis

Considerable evidence supports the hypothesis that sequence variation in a large number of genes contribute to the risk of common cancers such as breast, colon and prostate cancer. In the mid 1990's, linkage analysis provided a successful approach to the discovery of high-risk susceptibility genes for these diseases. More recently, large case-control genotyping studies have identified common modest-risk SNPs and case-control mutation screening has emerged as a useful strategy for identifying and characterizing intermediate-risk susceptibility genes. While some breast cancer susceptibility alleles have been clearly defined in these three well-established classes of genetic variants, they do not account for more than 30-35% of the relative risk of breast cancer (high and intermediate-risk genes account for about 25%, and common SNPs from GWAS may account for 10%).

The Genetic Cancer Susceptibility group at the International Agency for Research on Cancer focuses on the evaluation of inherited genetic factors in the etiology and outcome of cancer. In particular, the group is carrying out an international breast cancer genetics study aiming to identify new potentially deleterious genetic variants in candidate susceptibility genes conferring an intermediate-risk of breast cancer. This is performed by mutation screening coding exons and proximal intronic splice consensus sequences of the candidate genes, in large series of cases and controls. Mutation screening of the *ATM* [Tavtigian et al., 2009] and *CHEK2* genes [Le Calvez-Kelm et al., submitted, see Annex II] have recently been completed in our laboratory and work on additional genes including RAD51, BARD1, and RAD50 is nearing completion.

Although mutation scanning projects have focused for many years on variations in coding sequences, structural alterations caused by genetic variants are not the only possible explanation for variations in disease phenotype. Gene expression regulation provides an alternate mechanism for generating cellular variation and may be the underlying explanation for a proportion of cancer syndromes that have not been resolved by germline coding region variants in currently known cancer predisposition genes.

This Ph.D work is an integral component of the candidate gene sequence variant discovery project of wide scope undertaken at our laboratory. In order to fully assess the contribution of the *CHEK2*, *ATM* and *TP53* genes to breast cancer susceptibility, we used the differential allelic expression (DAE) approach to screen these genes for the signature of *cis*-regulation. A diagram of the workflow is presented on the next page (Figure 1). The strength of our approach is to combine mutation-screening, which provides a list of candidate genetic variants and/or genotypes at polymorphic "probe" nucleotides, and assessment of DAE . This combination can be a powerful tool for identifying *cis*-acting variation affecting gene expression at the mRNA level. Assessing DAE whenever possible, in each of the genes included in the mutation screening project, will allow to focus regulatory sequence variant discovery efforts on the subset of genes that are most likely to harbor regulatory variants altering gene expression.

Specific aims

Aim 1 Assemble a discovery panel of lymphoblastoid cell lines derived from high-risk breast cancer patients.

Aim 2 Develop an appropriate assay for the detection of differential allelic expression, from the experimental aspects to the development of bioinformatics tools specifically dedicated to the analysis.

Aim 3 Test each gene for differential allelic expression.

Aim 4 Characterize the sequence variation that may contribute to the observed differential allelic expression.

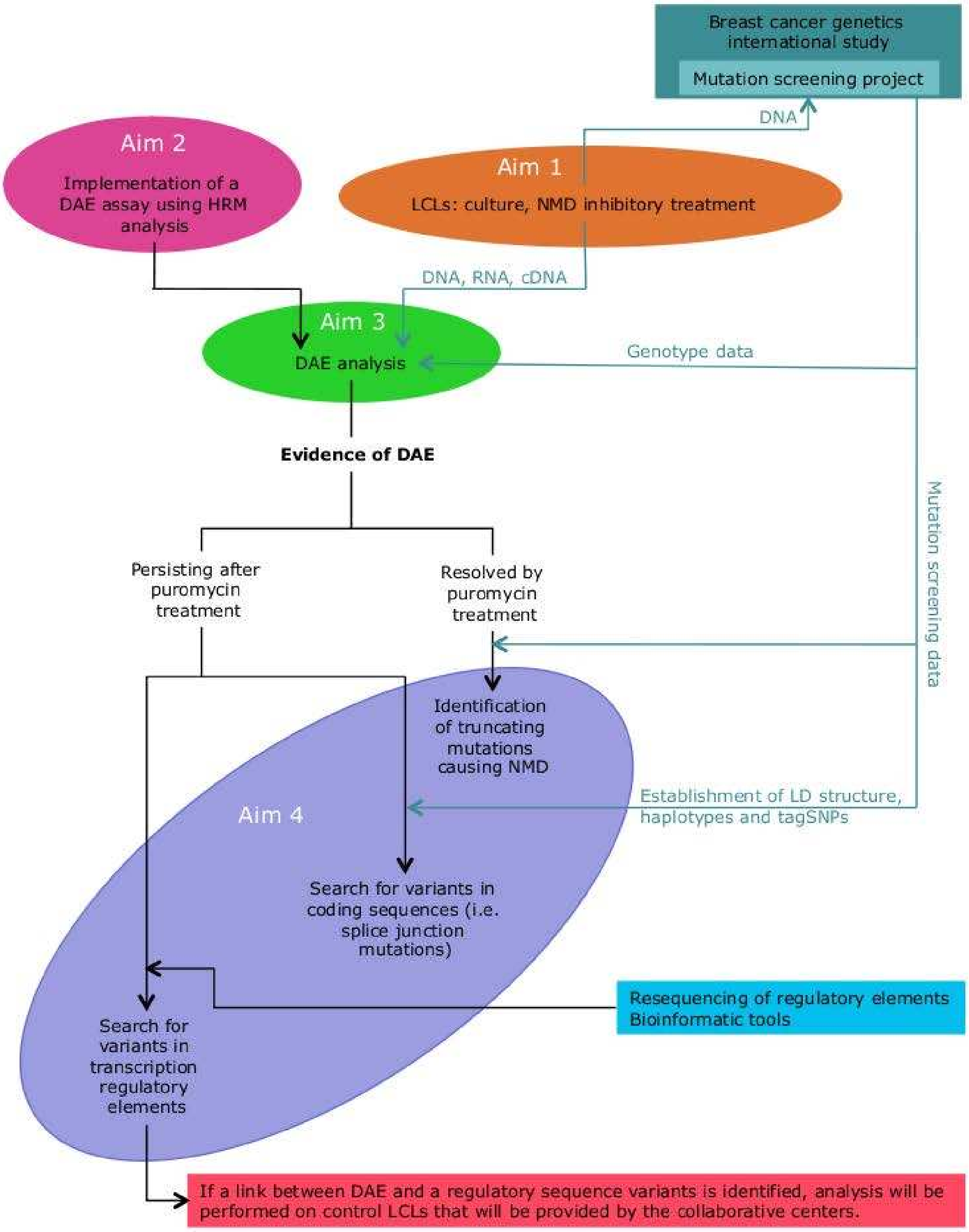

Breast cancer genetics international study
Mutation screening project
DNA
Aim 2
Implementation of a DAE assay using HRM analysis
Aim 1
LCLs: culture, NMD inhibitory treatment
DNA, RNA, cDNA
Aim 3
DAE analysis
Genotype data
Evidence of DAE
Persisting after puromycin treatment
Resolved by puromycin treatment
Mutation screening data
Identification of truncating mutations causing NMD
Aim 4
Establishment of LD structure, haplotypes and tagSNPs
Search for variants in coding sequences (i.e. splice junction mutations)
Resequencing of regulatory elements
Bioinformatic tools
Search for variants in transcription regulatory elements
If a link between DAE and a regulatory sequence variants is identified, analysis will be performed on control LCLs that will be provided by the collaborative centers.

Figure III.2: Diagram describing the aims of the Ph.D project.
DAE stands for differential allelic expression and NMD for non-sense mediated mRNA decay. Puromycin treatment inhibits NMD.

We chose to perform DAE assessment by high-resolution melting (HRM) curve analysis. In the next chapter, as a preamble, an overview of the methods and applications related to HRM technology will be provided. Then, the first section of the Result chapter will describe our general experimental approach for DAE assessment by HRM, and address Aim 1 and Aim 2. This will be followed by results from the studies of the *CHEK2*, *ATM* and *TP53* genes, addressing Aim 3 and Aim 4 for each of these genes.

The components of the diagram will be further deciphered as we progress through theses sections.

Introducing high-resolution melting curve analysis

I

The basic underlying principles to HRM analysis

Melting temperature (Tm) is one of the characteristics of DNA and cDNA. A double-stranded DNA (dsDNA) molecule melts into two molecules of single-stranded DNA (ssDNA) under conditions that overcome the interacting forces between bases. The Tm of a dsDNA fragment is defined as the temperature at which 50% of the DNA melts, i.e becomes single stranded. Each dsDNA fragment has its own distinct Tm.

Melting methods exist, that can be used to distinguish between alleles and detect sequence variation within a given PCR amplicon. After PCR amplification, PCR products are subject to a denaturation and renaturation phase, then melting of dsDNA by continual increase of temperature is easily monitored with the aid of instruments that allow for highly controlled temperature transitions and appropriate, usually fluorescence based, data acquisition. Fluorescence instrumentation has recently been introduced with the HR-1™ and LightScanner®, but high resolution methods have also been adopted by real-time PCR instruments, such as Roche's LightCycler480 [Herrmann et al., 2006, Herrmann et al., 2007].

II

Amplicon melting analysis

Amplicon melting analysis relies on the use of intercalating dsDNA-binding dyes in the PCR reaction, which fluorescence intensity varies during the melting analysis (Figure II.1) [Lipsky et al., 2001, Wittwer et al., 2003].

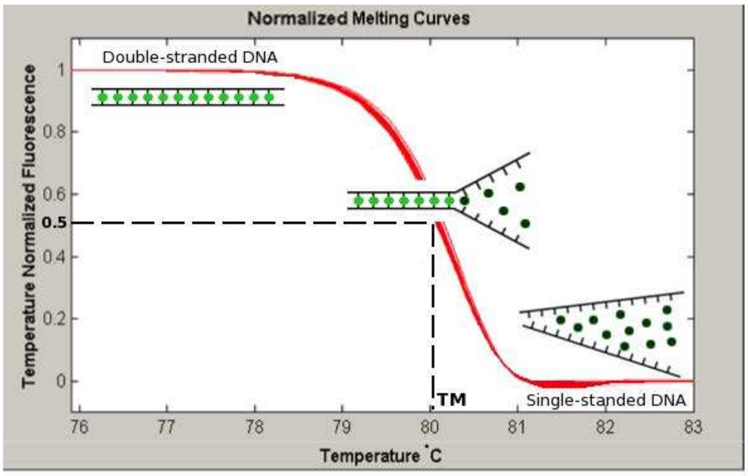


Figure II.1: High-resolution melting uses intercalating dyes that fluoresce only in the presence of double stranded DNA. At the beginning of the HRM analysis, after PCR amplification, there is a high level of fluorescence. As the temperature increases, the two strands of DNA denaturate and the dye is released, thus leading to a decrease of fluorescence .

LC Green® was the first saturating dye available. Other dyes have since been commercialized, including LC Green®Plus (Idaho Technology), Syto9® (Invitrogen), EvaGreen® (Biotum) and LightCycler® 480 ResoLight Dye (Roche).

During high-resolution melting (HRM) curve analysis, as the temperature increases, the specific sequence of the amplicon (mainly, its GC content and its length) determines the melting behavior. When the fluorescence signal is plotted against the temperature, the fluorescence intensity decreases as the dsDNA becomes single stranded and the dye is released (Figure II.2-A). Opposite homozygous samples are distinguished by difference in Tm. Homozygotes are in green and red on Figure II.2. Heterozygous samples are distinguished from homozygotes by altered curve shape (blue on Figure II.2). Heterozygous samples produce four different kinds of duplexes after denaturation and reannealing: two completely matched duplexes (homoduplexes) and two duplexes mismatching at the polymorphic position (heteroduplexes) [Montgomery et al., 2007]. Since heteroduplexes melt at a lower temperature than homoduplexes, they each have a characteristic melting pattern. The sum of all transitions are small but can reliably be detected by melting curve analysis, resulting in a skewed composite melting curve [Erali et al., 2008].

Melting curves are usually converted into negative first-derivative melting plots to reveal melting transitions of the probe-target duplexes as peaks (Figure II.2-B). The pattern of the difference plot may also be used for amplicon analysis (Figure II.2-C).

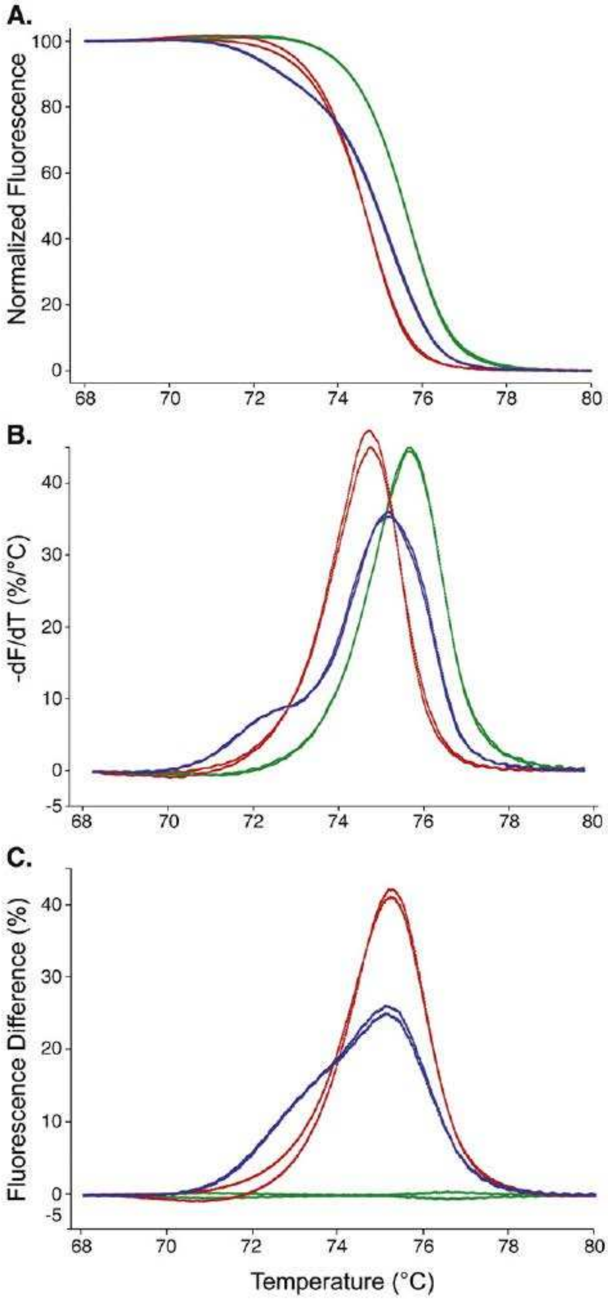


Figure II.2: Amplicon melting analysis for duplicate samples of factor V (Leiden) 1691 G>A homozygous common (green), homozygous rare (red) and heterozygous (blue) samples. (A) Normalized melting curves, (B) derivative plots, and (C) difference plots. From [Erali et al., 2008].

III

Probe melting analysis

For known mutation or SNP detection, HRM analysis may also be applied to labeled or unlabeled probes. These probes are specifically designed to anneal either to the wild-type or the variant sequence. In this case, the fluorescence changes as a result of the probe denaturing away from the amplicon [Zhou et al., 2004, Liew et al., 2006].

A single nucleotide mismatch between the probe and its target amplicon can significantly reduce the Tm of the probe-amplicon heteroduplex, offering a straightforward means to genotyping target sequences [Liew et al., 2004]. Since the Tm may be approximated by taking the derivative of the melting curve, samples with single peaks indicate a homozygous genotype, and those with two peaks indicate a heterozygous genotype. The higher temperature peak corresponds to the allele perfectly matched to the probe whereas the lower temperature peak corresponds to the mismatched allele.

In presence of a probe, an asymmetric PCR is required because it leads to the production of more copies of the strand to which the probe will bind to. This reduces competitive binding and favors the annealing of the probe. The fluorescence then varies during the melting phase, increasing or decreasing according to the type of probe that is used.

III.1 Fluorescein labeled probes

A fluorescein probe is designed to bind to a sequence variant of interest, in the vicinity of guanosines (Gs) on the opposite target strand. At the beginning of the melting phase, the probe hybridizes to the excess strand produced by the asymmetric PCR. As long as probe and target form a duplex, the fluorescence is quenched by the Gs. These nucleotides are especially good quenchers [Crockett and Wittwer, 2001].

Once the Tm is reached, the probe melts away and the fluorescence increases sharply (Figure III.1-A). All three possible genotypes at the polymorphic position targeted by the probe can be distinguished on the derivative curve (Figure III.1-B).

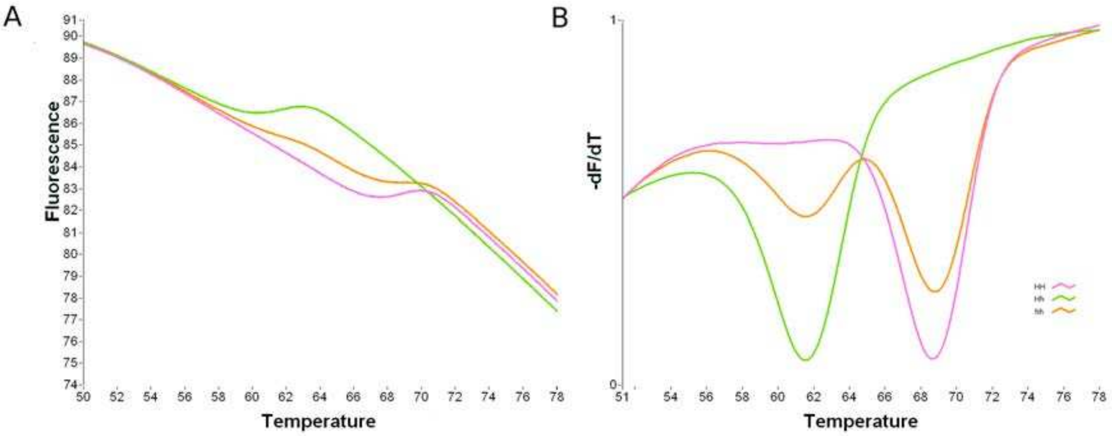


Figure III.1: Genotyping of the SNP R72P in the *TP53* gene, by use of a fluorescein probe melting analysis on the HR-1$^{\text{TM}}$ instrument. (A) Normalized melting curves, (B) derivative plots. Genotypes are homozygous common (HH), homozygous rare (hh) and heterozygous (Hh).

III.2 SimpleProbes

A SimpleProbe is another type of olignonucleotide probe, labeled with a proprietary linked green dye at either the 5' or the 3' end. In contrast to a fluorescein probe, which fluoresces most when single-stranded, the physical characteristics of the linker molecule of a SimpleProbe allow the dye to fluoresce when hybridized to its target DNA strand. Once the Tm is reached, the probe melts away, the linker then quenches the dye and the fluorescence decreases [Gameau et al., 2005].

This phenomenon translates into a sharp drop on the melting curve (Figure III.2-A) and upward peaks on the derivative curve (Figure III.2-B). The sequence surrounding the labeled end of the SimpleProbe is important and should avoid guanosines on the opposite strand, as it may affect the fluorescence signal.

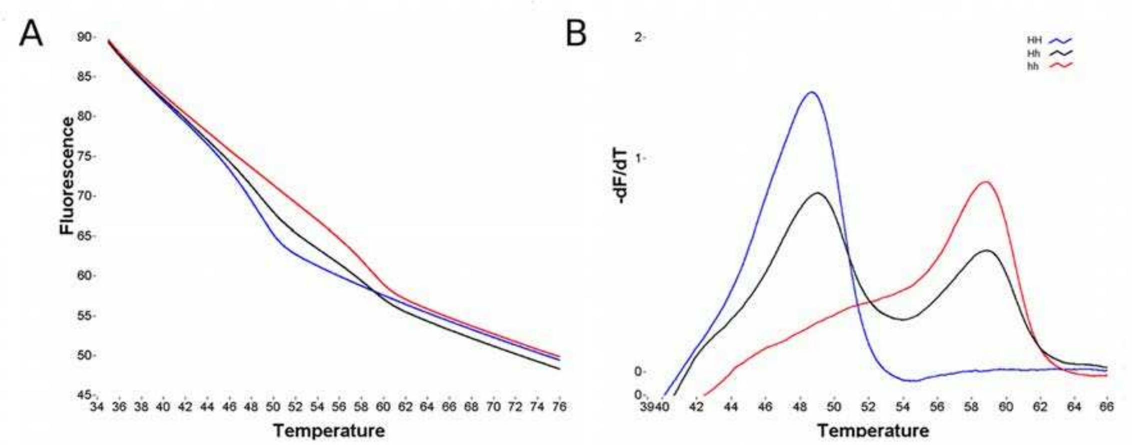


Figure III.2: Genotyping of the SNP rs2236142 in the *CHEK2* gene, by use of a SimpleProbe on the HR-1™ instrument. (A) Normalized melting curves, (B) derivative plots. Genotypes are homozygous common (HH), homozygous rare (hh) and heterozygous (Hh).

III.3 Unlabeled probes

Saturating dsDNA dyes such as LC Green® also allow genotyping with unlabeled probes that have no fluorescent labels [Liew et al., 2007]. Direct amplicon genotyping does not require probes, as described above. However, whereas heterozygotes are easily detected by a change in melting curve shape, differentiating opposite homozygotes may be more challenging due to their small Tm difference. The use of unlabeled probe thus increases genotyping accuracy in amplicon melting assays (Figure III.3).

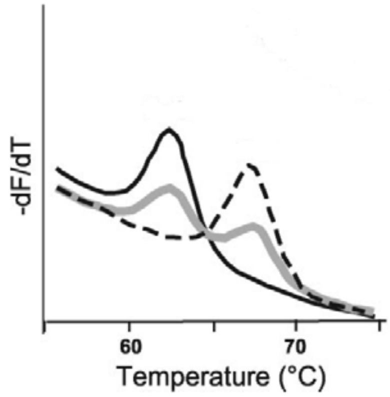


Figure III.3: Genotyping of the SNP G542X in the *CFTR* gene, by use of an unlabeled probe on the HR-1™ instrument. Genotypes are homozygous wild type (thin black line), heterozygous (thick gray line), and homozygous mutant (dashed line). From [Zhou et al., 2004].

Using the *TP53* gene, Garritano et al. have assessed the performance of a variety of available HRM-based genotyping assays, including unlabeled probe analysis, and described a series of solutions to handle the difficulties that may arise in large-scale application of HRM to mutation screening and genotyping at the *TP53* locus. In particular, the authors report specific HRM assays that render possible genotyping of 2 or more, sometimes closely spaced, polymorphisms within the same amplicon, using unlabeled probes, and showed that multiplex PCR reaction is feasible [Garritano et al., 2009 - see Annex I].

IV

Applications of HRM analysis

The major reported advantages of HRM are high sensitivity but also minimal post-PCR sample manipulation, thus ease of use, throughput and cost-effectiveness. Furthermore, HRM analysis is a non-destructive method, and subsequent analysis by sequencing can still be performed after melting analysis [Isaacs and Rebbeck, 2008, Wittwer, 2009, Vossen et al., 2009].

As recently reviewed by [Erali et al., 2008] and [Vossen et al., 2009], HRM allows to perform a panoply of molecular genetic and epigenetic analyses, such as genotyping [Wittwer et al., 2003, Liew et al., 2004, Zhou et al., 2004, Graham et al., 2005, Palais et al., 2005], mutation screening [Reed and Wittwer, 2004, Chou et al., 2005, Margraf et al., 2006, Takano et al., 2008] and methylation profiling [Worm et al., 2001, Wojdacz and Dobrovic, 2007].

HRM analysis is now recognized as a robust multi-purpose analytical tool for research, as well as molecular diagnostic and clinical ends.

IV.1 Clinical purposes

One of many examples of clinical application of HRM analysis was reported by Margraf et al. [Margraf et al., 2006]. The authors developed a screen for *RET* proto-oncogene mutations associated with multiple endocrine neoplasia type 2 (MEN2) syndromes. Genetic testing of *RET* mutations can identify patients at risk of thyroid cancer before disease progression. The assay was designed to amplify 6 *RET* exons and to include all known pathogenic mutations in a total of 20 codons. The assay can be used more specifically to detect a mutation that is known to be present in a family. In the course of a blinded study, 100% concordance was observed in comparison with sequencing, the gold standard approach, which is a time-consuming and expensive open-tube method. HRM was validated as a fast and accurate method for detecting or genotyping *RET* mutations.

Also in diagnostic settings, van der Stoep et al. designed a screening test to cover the *BRCA1* gene. Their validation study included a large panel of 170 *BRCA1* variants and 197 controls [van der Stoep et al., 2009]. They described an HRM assay that allowed mutation screening of all of the exons of the gene and included unlabeled probes to identify nine commonly occurring polymorphisms of the *BRCA1* gene, thus avoiding unnecessary sequence analysis upon detection of these non-pathogenic variants. The authors also aimed at performing a thorough interlaboratory evaluation and validation of HRM analysis, to confirm its accuracy and robustness, and provided a list of guidelines for setting up and implementing HRM as a scanning technique for new genes in diagnostic.

IV.2 Research purposes

The special issue of Human Mutation: Focus on High-Resolution Melting Technology [Volume 30, Issue 6, June 2009] contains several papers describing methods application of HRM for sequence variant detection, demonstrating its current state of the art. For instance, Rouleau et al. described a quantitative

PCR and HRM analysis in one instrument to scan for both quantitative (deletions/duplications) and qualitative nucleotide changes in the *MLH1* gene [Rouleau et al., 2009]. Dobrowolski et al. described the use of HRM to scan the entire 16.6kb human mitochondrial genome (mtDNA) for sequence variants in less than two hours [Dobrowolski et al., 2009].

This special issue of Human mutation, which also highlighted the limits of HRM and challenges that may be encountered, included a paper of ours that addressed the need for cost-effective solutions to some of these challenges with a methodological improvement to the basic HRM mutation screening strategy.

HRM analysis is the technology chosen by our laboratory to perform large case-control mutation screening studies, aiming at identifying candidate genes and variants conferring an intermediate-risk of breast cancer [Tavtigian et al., 2009]. A particularly demanding application of HRM is analysis of candidate intermediate-risk susceptibility genes by case-control mutation screening, which requires complete mutation screening of >1000 cases and controls to achieve reasonable statistical power. We have actually screened the coding exons of *ATM* and *CHEK2* (published[1]), and *RAD51*, *BARD*1 and *RAD50* (unpublished) in >1000 subjects, and did not meet any technical difficulties with the vast majority of exons/amplicons when using standard HRM mutation scanning.

However, detection sensitivity for rare unknown variants may be problematic during mutation scanning of fragments containing common SNPs using HRM. This may happen in one of the following situations: the melting profile of rare unknown variants can be masked by one from a common homozygote, by the extra noise present in large scale melt curve studies, or buried within the melt curve data of an amplicon whose data complexity overcomes the standard software's ability to form groups.

[1] [Tavtigian et al, 2009] and [Le Calvez-Kelm et al, 2011- see Annex II] respectively.

In the following paper, we demonstrated that simultaneous scanning and genotyping using unlabeled probes allows better differentiation of multiple variants. An unlabeled probe was included in asymmetric PCR such that both probe/product and full length product duplexes were produced. From a single melting curve, both genotyping data (at low temperature) and scanning data (at high temperature) were extracted. Two exons of *ATM*, each including a common variant that interferes with standard scanning, were analyzed by high-resolution melting on 384-well plates. Simultaneous scanning and genotyping of 1356 subjects was performed. For analysis, the curves were grouped by probe/target melting (by the genotype of the common variant) and amplicon scanning was performed on each group. Up to 9 different genotype combinations were distinguished and the curve clusters were completely concordant to sequencing. Furthermore, the sequencing burden from common variants can be dramatically reduced.

Article I

Description and Validation of High-Throughput Simultaneous Genotyping and Mutation Scanning by High-Resolution Melting Curve Analysis

Nguyen-Dumont T, Le Calvez-Kelm F, Forey N, McKay-Chopin S, Garritano S, Gioia-Patricola L, De Silva D, Weigel R, Breast CFR, kConFab, Sangrajrang S, Lesueur F, Tavtigian SV.

Hum Mutat, 2009 ; (6) :884-90.

Human Mutation
OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Description and Validation of High-Throughput Simultaneous Genotyping and Mutation Scanning by High-Resolution Melting Curve Analysis

Tú Nguyen-Dumont,[1] Florence Le Calvez-Kelm,[1] Nathalie Forey,[1] Sandrine McKay-Chopin,[1] Sonia Garritano,[1] Lydie Gioia-Patricola,[1] Deepika De Silva,[2] Ron Weigel,[2] Breast Cancer Family Registries (BCFR), Kathleen Cuningham Foundation Consortium for research into Familial Breast cancer (kConFab), Suleeporn Sangrajrang,[3] Fabienne Lesueur,[1] and Sean V. Tavtigian[1]*

[1]International Agency for Research on Cancer (IARC), Lyon, France; [2]Idaho Technology, Inc., Salt Lake City, Utah; [3]Research Division, National Cancer Institute, Bangkok, Thailand

For the Focus Section on HRMA Technology

Received 4 September 2008; accepted revised manuscript 11 November 2008.
Published online 4 April 2009 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.20949

ABSTRACT: Mutation scanning using high-resolution melting curve analysis (HR-melt) is an effective and sensitive method to detect sequence variations. However, the presence of a common SNP within a mutation scanning amplicon may considerably complicate the interpretation of results and increase the number of samples flagged for sequencing by interfering with the clustering of samples according to melting profiles. A protocol describing simultaneous high-resolution gene scanning and genotyping has been reported. Here, we show that it can improve the sensitivity and the efficiency of large-scale case–control mutation screening. Two exons of *ATM*, both containing an SNP interfering with standard mutation scanning, were selected for screening of 1,356 subjects from an international breast cancer genetics study. Asymmetric PCR was performed in the presence of an SNP-specific unlabeled probe. Stratification of the samples according to their probe-target melting was aided by customized HR-melt software. This approach improved identification of rare known and unknown variants, while dramatically reducing the sequencing effort. It even allowed genotyping of tandem SNPs using a single probe. Hence, HR-melt is a rapid, efficient, and cost-effective tool that can be used for high-throughput mutation screening for research, as well as for molecular diagnostic and clinical purposes.
Hum Mutat 30, 884–890, 2009. © 2009 Wiley-Liss, Inc.

KEY WORDS: high-resolution melting curve analysis; HR-melt; high throughput mutation scanning; genotyping; *ATM*

Introduction

A key step in the search for potentially pathogenic genetic variants in disease susceptibility genes is mutation screening of coding exons and proximal intronic splice consensus sequences of the entire gene in large subject series. Mutation scanning using high-resolution melting curve analysis (HR-melt) prior to sequencing has been described as an effective, sensitive, and economical method to detect genetic variations and to reduce sequencing efforts [De Leeneer et al., 2008; Reed and Wittwer, 2004; Takano et al., 2008]. HR-melt analysis using unlabeled probes can also be used for genotyping [Liew et al., 2004; Seipp et al., 2007; Zhou et al., 2004]. HR-melt relies on the use of the double-stranded DNA fluorescent dye LCGreen® Plus (Idaho Technology, Inc., Salt Lake City, Utah) and specifically designed instruments for data collection, such as the LightScanner® (Idaho Technology), which can be used for high-throughput analyses. HR-melt offers several obvious advantages as compared to traditional mutation scanning methods. Not only efficient, this method is secure due to its closed-tube nature, and is amenable to automation for high-throughput mutation discovery. It provides simultaneous acquisition of up to 384 fluorescent melting signals in about 5 min and also fits seamlessly into a resequencing workflow because of its nondestructive nature. A particularly demanding application of HR-melt is analysis of candidate intermediate-risk susceptibility genes by case–control mutation screening, which will often require complete mutation screening of >1,000 cases and >1,000 controls to achieve reasonable statistical power. However, the presence of a common SNP within a mutation scanning amplicon may considerably complicate the interpretation of results and increase the number of samples flagged for sequencing by interfering with the clustering of melt curve groups according to melting profiles.

Recently, a protocol for simultaneous mutation scanning and genotyping using HR-melt analysis has been described [Montgomery et al., 2007; Zhou et al., 2005]. The method combines both LCGreen Plus dye and unlabeled oligonucleotide probes in an asymmetric PCR, leading to simultaneous production of probe-target and whole amplicon double-stranded DNA duplexes that can be analyzed from the same HR-melt run.

In this study, we aimed to apply the method to improve sensitivity and efficiency of large-scale case–control mutation scanning of the *ATM* gene (GenBank reference sequence NM_000051; MIM 607585) in some specific situations (Fig. 1). We have actually screened the 62 coding exons of *ATM* in >1,000 subjects, and did not meet any technical difficulties with the vast majority of exons/amplicons when using standard HR-melt mutation scanning (Tavtigian et al., unpublished results).

*Correspondence to: Sean V. Tavtigian, Genetic Susceptibility Group, International Agency for Research on Cancer, 150 Cours Albert-Thomas, 69372 Lyon Cedex 08, France. E-mail: tavtigian@iarc.fr

However, screening of some *ATM* amplicons illustrates challenges that may be encountered and cost-effective solutions to these challenges. We chose the 36th and 59th coding exons of *ATM*, both containing a common SNP that interferes with standard HR-melt mutation scanning (Table 1). Mutation screening of the 36th coding exon was challenging because the amplicon contains a common missense SNP (c.5557G>A, allele frequency of 23% in the Caucasian population) adjacent to a less common SNP (c.5558A>T, 1% in the Caucasian population), as well as two other rare SNPs (c.5497–15G>C and c.5497–8T>C, 0.5% and 1% in the Caucasian population, respectively). The 59th coding exon amplicon contains an A>C substitution (c.8786+8, 1.7% in the Caucasian population), located downstream of the splice donor site, disturbing standard HR-melt analysis.

For each amplicon, an unlabeled probe was designed to anneal to the region surrounding the common SNP of interest. Stratification of the samples according to their probe-target melting profiles was facilitated by customized LightScanner software (Idaho Technology). The conceptual idea is that common SNP genotypes are called from the probe-target melting data. Analysis of the whole amplicon melt curve data (e.g., mutation scanning) is then performed separately on heterozygous and homozygous sample subsets to distinguish curve shape differences due to presence of other unknown variants. This approach is particularly valuable for large-scale mutation screening studies, where systematic resequencing of the whole gene in all samples is too laborious and expensive.

Materials and Methods

Origin of Samples

Mutation screening was performed on 697 early onset and/or familial breast cancer cases and 659 controls enrolled in an international breast cancer genetics study. These include subjects collected through the Northern California, Australian, and Ontario sites of the Breast Cancer Family Registry (BCFR), subjects collected through the Kathleen Cuningham Foundation Consortium for research into Familial Breast cancer (kConFab, Melbourne, Australia), and subjects enrolled in a Thai case–-control study. The mutation screening included in this project had approval by the IARC Institutional Review Board (IRB) and the local IRBs of each of the centers from which we received samples. All DNAs were extracted from lymphocyte samples or lympho-

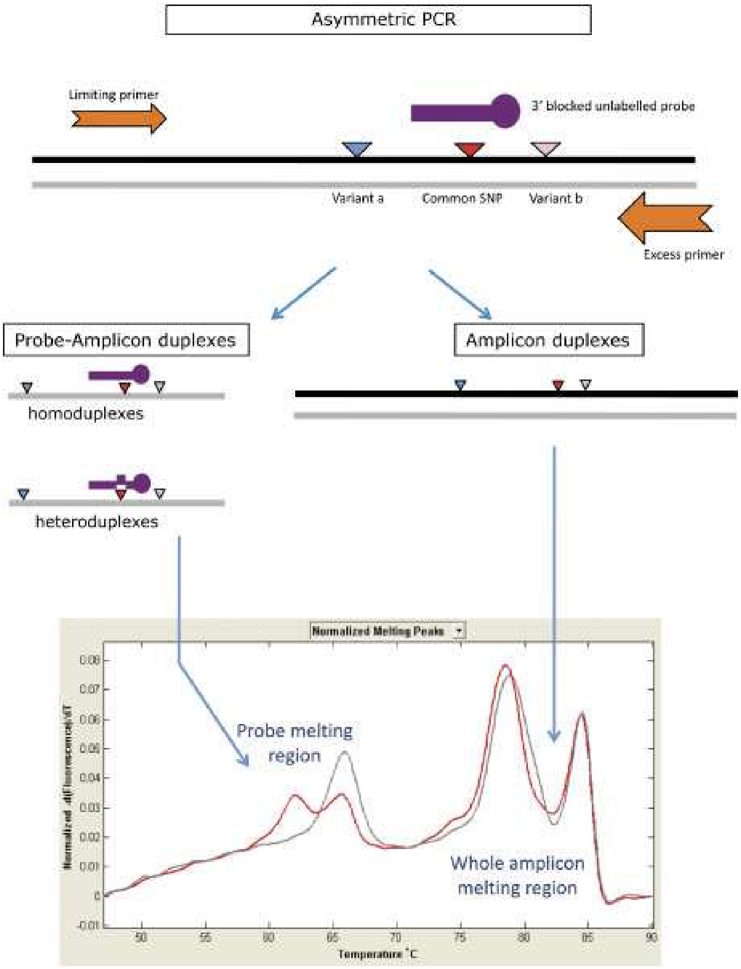


Figure 1. Principle of simultaneous genotyping and mutation scanning using high-resolution melting analysis. For DNA amplicons containing a common SNP, an unlabelled probe is designed to target the SNP. The probe is blocked at the 3′ end to prevent extension during amplification. In presence of a probe, an asymmetric PCR reaction is required so that more copies of the strand to which the probe anneals are produced. This favors probe-target annealing and reduces probe competition with the complementary DNA strand. Both probe-amplicon duplexes and whole amplicon duplexes melting regions can be observed from the same melting run, in two separate temperature windows, allowing genotyping and mutation scanning analyses to be performed simultaneously.

blastoid cell lines using standard procedures, then normalized at a concentration of 15 ng/µl and arrayed in 384-well plates. Each plate included negative controls (with no DNA), and a DNA sample from Chimpanzee was added. This sample is used as quality control to assess the efficiency of the HR-melt assay for rare sequence variant detection, as the Chimpanzee genome is evolutionarily close enough to the Human genome that almost all amplicons work, but different enough from human that most amplicons will harbor a few sequence variations.

Table 1. Nomenclature of All Sequence Variations Detected in the 36th and 59th Coding Exons of *ATM*, in the Breast Cancer Genetics Study

Amplicon	HGVS[a]	Protein	rs Number
36th Coding exon	c.5497–15G>C	—	rs3092828
	c.5497–8T>C	—	rs3092829
	c.5557G>A	p.Asp1853Asn	rs1801516
	c.5558A>T	p.Asp1853Val	rs1801673
	c.5633C>T	p.Ser1878Leu	—
59th Coding exon	c.8672–43T>C	—	—
	c.8672–22T>G	—	rs56172540
	c.8741T>C	p.Ile2914Thr	—
	c.8786+8A>C	—	rs4986839
	c.8786+11T>C	—	—

[a]Number based on transcript sequence (NM_000051), +1 as A of ATG start codon.

PCR Amplification

Nested primer pairs were designed to amplify specifically the 36th and 59th coding exons of *ATM* (NM_000051), including intron–exon junctions, and are available on request. Unlabeled probes were designed to anneal to the SNP of interest, following Idaho Technology's recommendations. All probes were blocked at their 3′ end by C3-blocker to prevent extension during PCR.

For the 36th coding exon, the unlabeled probe (5′-CCAA**G**A-TACAAATGAATCAT-3′) was designed to target the major allele of the common SNP c.5557G>A. This probe also annealed to the major allele of the adjacent SNP c.5558A>T (Fig. 2A). For the 59th codon exon, the unlabeled probe (5′-CAGAAGGTAAGTG**A**-TATGAAGTAAAGGAGG-3′) was designed to target the major allele of SNP c.8786+8A>C.

Primary PCR (PCR_1) was performed in an 8-µl reaction volume containing 30 ng of template DNA, 1.5 mM $MgCl_2$, 200 µM dNTP, 200 nM forward and reverse primers, 0.04 U/µl of Platinum® Taq Polymerase (Invitrogen, Paisley, Scotland), and 1 × PCR buffer supplied by the manufacturer. The amplification protocol consisted of 25 cycles with amplification steps at 94°C, 60°C, and 72°C for 30 s each.

Asymmetric nested PCR (PCR_2) was then performed in a 6-µl reaction volume containing 2 µl of 1:100 diluted PCR_1 product, 1.5 mM $MgCl_2$, 132 µM dNTP, 100 nM limiting primer, 500 nM excess primer (primer asymmetry ratio of 1:5), 500 nM unlabeled

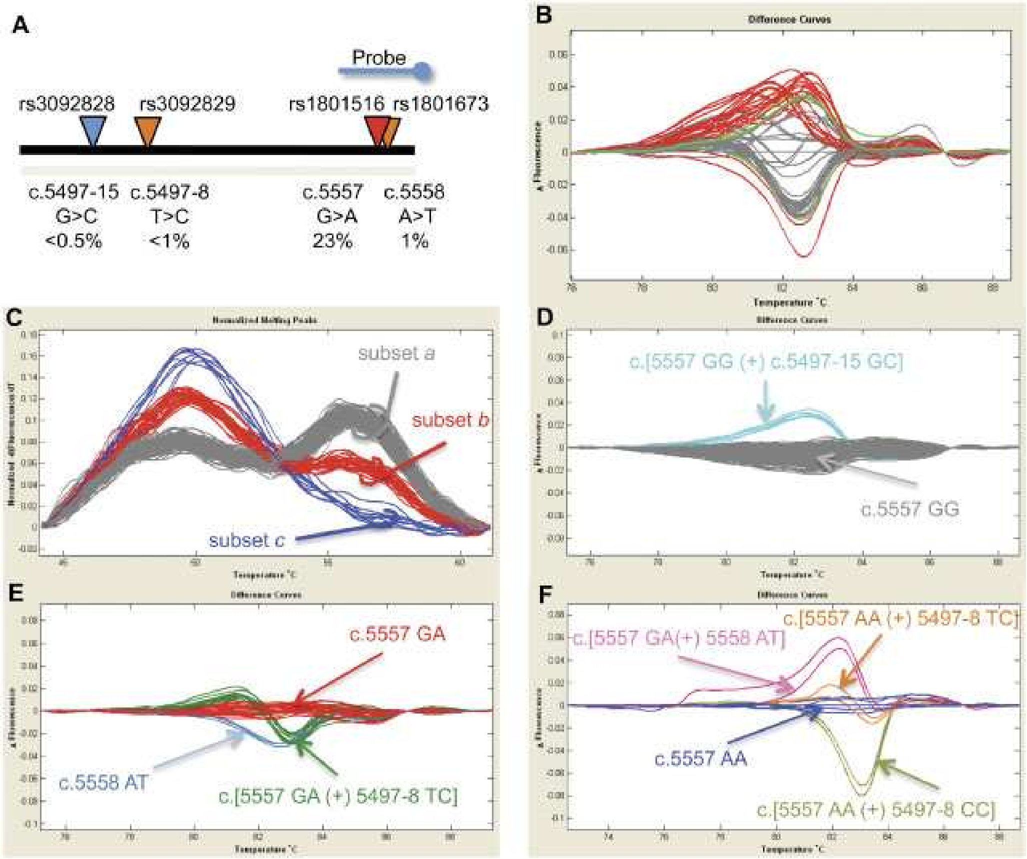


Figure 2. High-resolution melting analysis of the 36th coding exon of *ATM*: results for a set of 384 samples. **A:** Relative positions of SNPs and of the unlabeled probe. The probe was designed to anneal to the common SNP c.5557G>A and adjacent c.5558A>T. **B:** Standard mutation scanning. The melting temperature (Tm) difference plot failed to stratify samples by genotype effectively. **C:** Simultaneous genotyping and mutation scanning allowed classification of the samples into three subsets. **D–F:** Difference plots of the amplicon melting profiles of subset *a* (D), *b* (E), and *c* (F) allowed identification of a total of nine combinations of genotypes. Genotypes are indicated on the figures.

probe, 0.48 × LCGreen Plus, 0.04 U/μl of Platinum Taq Polymerase, and 1 × PCR buffer. The amplification protocol consisted of 55 cycles with amplification steps of 94°C for 30 s, 60°C for 30 s, and 72°C for 40 s each. For an optimal efficiency of HR-melt, PCR_2 primers were designed to amplify amplicons with a maximum length of 350 bp.

HR-melt Analysis

Prior to LightScanner analysis, PCR_2 products were heated to 94°C, then slowly cooled to 20°C to promote heteroduplex formation and detection. Melting was monitored from 35°C to 94°C on a LightScanner instrument. PCR amplification led to simultaneous production of probe-target and whole amplicon duplexes that were analyzed from the same HR-melt run. Since probe-target duplexes are shorter than whole amplicon double stranded DNA duplexes, they melt at a lower temperature. Short probe-target duplexes and larger whole amplicon double-stranded DNA duplexes can therefore be analyzed in two distinct temperature windows (Fig. 1).

Genotyping and mutation scanning analyses were carried out using the LightScanner software. The region of the probe melting was analyzed using the "Genotyping" mode and the region of DNA melting was analyzed using the "Scanning" mode [Montgomery et al., 2007]. Stratification of the samples according to their probe-target melting profile was facilitated by a customization of the commercial LightScanner software. This new version provides the option to export probe-target melting groups as subsets, which are subsequently used for independent scanning of the amplicons according to probe-target melting profile results.

Sequencing

PCR_2 products showing different melting curves from the reference group were sequenced using the BigDye Terminator, version 1.1 (Applied Biosystems, Foster City, CA) and run on a 96-capillary Spectrumedix Sequencer (Transgenomics, Glasgow, UK) according to the manufacturers' recommendations.

Results

Many research groups have published studies on mutation screening of the *ATM* gene and have reported many different rare sequence variants detected with various methodologies. Some have used SSCP with a relatively low sensitivity, others have used DHPLC with a relatively higher sensitivity. Recently, we have completed a large-scale mutation screening of *ATM* using a relatively new procedure, HR-melt. Although the object of this report is to describe a methodological improvement to the basic HR-melt protocol that enhances the effectiveness of mutation screening, we would like to emphasize that basic HR-melt mutation screening strategy provides good sensitivity for detection of sequence variants. For instance, we assessed the results of 13 studies where different mutation scanning methodologies had been used for mutation detection in *ATM* [Angèle et al., 2003; Atencio et al., 2001; Broeks et al., 2008; Brunet et al., 2008; Buchholz et al., 2004; Dörk et al., 2001; Gonzalez-Hormazabal et al., 2008; Izatt et al., 1999; Livingston et al., 2004; Maillet et al., 2002; Renwick et al., 2006; Sommer et al., 2003; Teraoka et al., 2001; Thorstenson et al., 2001]. When considering only Caucasian subjects for this pooled analysis, 142 carriers of a rare missense variant (i.e., with carrier frequency <1%) were observed among 2,661 subjects (mutation detection rate: $142/(2661*3056) = 0.000017$ variants*subjects^{-1}*codons^{-1}). We compared those results to our data. We identified 101 carriers of a rare missense variant when performing the analysis by HR-melt on 1,356 subjects of non-African origin (mutation detection rate: $101/(1356*3056) = 0.000024$ variants*subjects^{-1}*codons^{-1}). Hence, the rate ratio between our results and the pooled analysis is 1.4 (P value = 0.008) confirming the higher sensitivity of the HR-melt approach. Nonetheless, we observed that for the two amplicons containing the 36th and 59th coding exons, classic mutation scanning by HR-melt analysis failed to provide clear variant clustering due to the presence of a frequent SNP within the amplicons studied (Figs. 2B and 3A).

Analysis of the 36th coding exon of *ATM* was hampered by the presence of several SNPs reported in this amplicon: a common SNP immediately adjacent to a rare SNP (c.5557G>A and c.5558A>T, respectively) and two other rare SNPs (c.5497–8C>T and c.5497–15G>C) (Fig. 2A). We assessed an unlabeled oligonucleotide probe designed to anneal to the common G allele at the polymorphic site c.5557G>A and to the common A allele at the adjacent rarer SNP position c.5558A>T. The LightScanner software identified three different groups in the "Genotyping" mode (Fig. 2C), which were analyzed independently in the "Scanning" mode. Subset *a* was scanned for mutations and the automatic call identified two melting profiles. Sequencing analysis revealed that all samples from this subset carried the common G allele for c.5557 and some were also c.5497–15 GC heterozygotes (Fig. 2D). Mutation scanning analysis of subset *b* identified heterozygous samples that were either c.5557GA or c.5558AT. A third group emerging from subset *b* analysis corresponded to double heterozygotes c.[5557GA(+)5497CT] (Fig. 2E). Subset *c* revealed four different groups in mutation scanning: c.5557AA, c.[5557AA(+)5497–8CT], c.[5557AA(+)5497–8CC], and last, c.[5557GA(+)5558AT] (Fig. 2F). Analysis of the whole sample set (1,356 subjects) identified one additional rare missense variant (c.5633C>T).

Another example illustrating the difficulty of interpreting HR-melt using the standard mutation scanning mode in the presence of a common SNP is provided by the exon 59 amplicon (Fig. 3A). Within the reference group, sequencing of a few samples with melt curves near the edge of the "normal" melt curve distribution revealed that some were AC heterozygous for the SNP c.8786+8. Since their melting pattern was hardly distinguishable from the one produced by the wild-type group, other samples carrying the same SNP could have been missed, even though our standard practice is to sequence a small fraction of samples from the edge of the HR-melt normal grouping. Moreover, other variants in the vicinity of this SNP might also produce a melting pattern similar to that of the reference group. Thus, to improve the detection of SNP c.8786+8A>C and the detection of new rare nearby variants, an unlabeled probe was designed to hybridize to the common A allele of the SNP. As expected, genotyping allowed distinction of homozygous c.8786+8AA samples from heterozygous c.8786+8AC samples. Analysis of the probe-target melting region in the same 384 samples also identified two samples presenting a third distinct profile. One of the samples was c.8786+8CC. The second sample, from the Chimpanzee DNA used as quality control, was homozygous CC for a new variant (c.8786+11) located downstream SNP c.8786+8, and therefore in the probe-target region (Fig. 3B). Both had been missed in the standard mutation scanning analysis initially performed. Groups corresponding to the three probe-target melting profiles were further analyzed as individual subsets using the mutation scanning mode. No novel variant was identified in the c.8786+8AA subset (Fig. 3C) nor in the c.8786+8AC subset (Fig. 3D). However, analysis of the whole sample set (1,356 subjects) using the

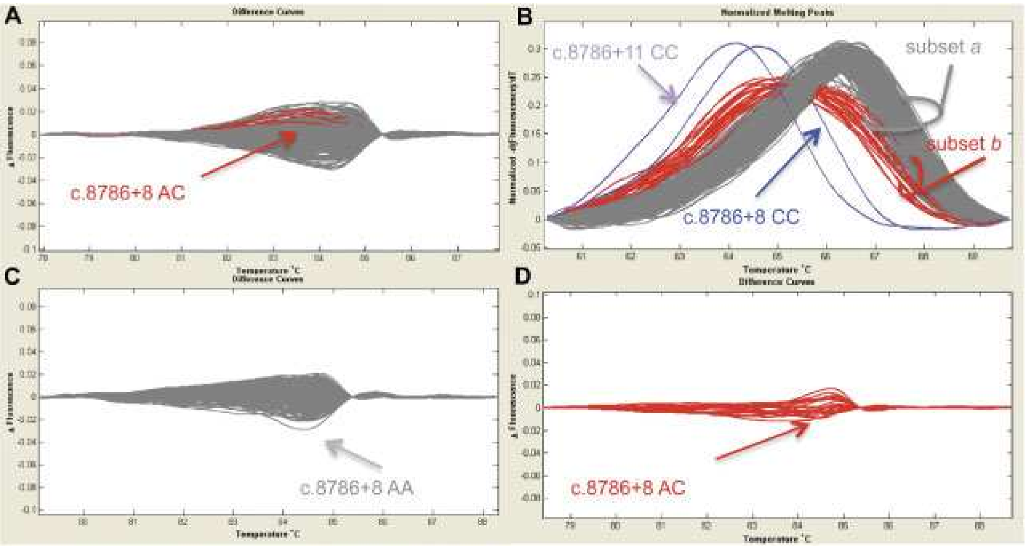


Figure 3. High-resolution melting analysis of the 59th coding exon of *ATM*: results for a set of 384 samples. **A:** Standard mutation scanning did not reveal any variant. Random sequencing from the edge of the "normal" distribution revealed the presence of several c.8786+8AC heterozygotes. **B:** An unlabeled probe was designed to anneal to this SNP. Simultaneous genotyping and mutation scanning allowed classification of the samples into the two principal subsets (c.8786+8 major allele homozygotes and heterozygotes) plus two additional melt curves corresponding to the genotypes indicated on the figure. **C,D:** Difference plots of the amplicon melting profiles of subset *a* (C) and *b* (D) did not identify novel rare variants.

simultaneous genotyping and mutation scanning approach succeeded in identifying a total of four rare variants (c.8672–43T>C, c.8672–22T>G, c.8741T>C, and the chimpanzee variant c.8786+11T>C), which had not been detected or were hardly distinguishable using the standard mutation scanning mode (data not shown). Thus, simultaneous genotyping and mutation scanning substantially improved the characterization of the samples where standard mutation scanning provided ambiguous and not fully reliable results.

Discussion

Here, we discuss the usefulness of simultaneous genotyping and mutation scanning in the context of large-scale mutation screening projects. The search for new potentially deleterious genetic variants in candidate susceptibility genes requires the screening of coding sequences and splice junctions of entire genes in large sets of cases and controls. HR-melt analysis has repeatedly been reported as an efficient method for mutation scanning. Although it has been reported that different heterozygotes within the same amplicon could be distinguished from each other based on their curve shape differences [Graham et al., 2005; Garritano et al., 2009], mutation screening performed on a large number of samples renders the analysis more complex. Moreover, screening of entire genes often requires the screening of genomic regions containing common SNPs that can interfere with the mutation scan and complicate the interpretation of the results [De Leeneer et al., 2008]. Systematic resequencing of all variant samples is the most common approach to this issue. However, when applied on large series, this latter approach is expensive, laborious, and time-consuming [Sevilla et al., 2002].

Simultaneous genotyping and mutation scanning by HR-melt analysis represents an attractive alternative for high-throughput analysis. The genotyping method was chosen because it could be easily integrated in our existing mutation scanning workflow. Other genotyping methods could have been chosen, but they would have added extra steps to our mutation screening protocol, and would also have required the use of another laboratory instrument. By performing genotyping and mutation scanning simultaneously using HR-melt, we avoided multiple manipulations, and waste of biological material and reagents. Laboratory contamination issues were also reduced. For amplicons that contain a common SNP, we postulated that stratification of HR-melt data by common SNP genotype prior to mutation scanning analysis would increase the detection sensitivity for those rare variants, whose melting patterns may be either: 1) essentially the same as, and consequently masked by, the melt curve of a common SNP heterozygote; 2) masked by the extra noise present in a large-scale melt curve analysis that contains two common genotypes; or 3) buried within the melt curve data of an amplicon whose data complexity overcomes the standard software's ability to group.

Here, we showed that simultaneous genotyping and mutation scanning is suitable to easily distinguish up to nine different genotype combinations, in the case of the 36th coding exon of *ATM*. Automatic clustering by the analysis software showed complete concordance with sequencing results. In addition, this approach offers the advantage of directly queuing asymmetric PCR products for sequencing. We validated on a series of 90 samples that sequencing reactions from asymmetric PCR products and standard sequencing reactions performed equally.

Study of the 59th coding exon pointed out that the position of a variant within the amplicon and/or the nature of the sequence surrounding the variant are likely to play a critical role on the accuracy of mutation detection by standard HR-melt analysis. Our study provides evidence that in some sequence contexts, some sequence variants may be missed by the classical HR-melt approach, especially when mutation scanning is performed in a 384-well format. This issue has been discussed in a technical assessment of the HR-melt protocol by the UK National Genetics Reference Laboratory (www.ngrl.org.uk/Wessex/downloads.htm), and the authors concluded that there were sequence variations "intrinsically difficult" to detect by HR-melt.

Table 2. Unlabeled Probes Used to Perform the Simultaneous Genotyping and Mutation Scanning of ATM, in the Breast Cancer Genetics Study

Amplicon	HGVS[a]	rs Number	Allele frequency[b] (%)	Probe sequence[c]
2nd Coding exon	c.146C>G	rs1800054	3.3	5′-GGCATTCAGATT**C**CAAACAAGGAAA-3′
6th Coding exon	c.735C>T	rs3218674	2.5	5′-GACTTTGGCTGT**C**AACTTTCGAA-3′
12th Coding exon	c.2119T>C	rs4986761	9.9	5′-AATTACTCA**T**CTGAGGTGAGATTTTTTA-3′
16th Coding exon	c.2572T>C	rs1800056	2.5	5′-CATCCATGAATCTA**T**TTAACGAT-3′
21st Coding exon	c.3161C>G	rs1800057	2.5	5′-CGTAGGCTGATC**C**TTATTCAAAATGGGC-3′
22nd Coding exon	c.3284–4delT	rs1799757	6.7	5′-GTTTGTTTGTTTGC**T**TGCTTGTTTT-3′
28th Coding exon	c.4258C>T	rs1800058	5.8	5′-ATTCTT**C**TTGCCATATGTGAGC-3′
29th Coding exon	c.4578C>T	rs1800889	5	5′-ATACC**C**CTTGTGTATGAGCA-3′
36th Coding exon	c.5557G>A; c.5558A>T	rs1801516; rs1801673	17.5; 0.8	5′-CCAA**GA**TACAAATGAATCAT-3′
38th Coding exon	c.5793T>C	rs3092910	0.5[d]	5′-TTTTAATGATGC**T**TTCTGGCTGGATTT-3′
43rd Coding exon	c.6348–54T>C	Unreported	[d]	5′-GCTATTTATACATG**T**ATATCTTAGGGTTCTGTTT-3′
59th Coding exon	c.8786+8A>C	rs4986839	1.7	5′-CAGAAGGTAAGTG**A**TATGAAGTAAAGGAGG-3′

[a] Number based on transcript sequence (NM_000051), +1 as A of ATG start codon.
[b] Frequency reported in European population in dbSNP.
[c] Polymorphic position is indicated in bold.
[d] Found to have a frequency >1% in our sample set.

Simultaneous genotyping and mutation scanning represents therefore a valuable asset since it can easily be integrated in large-scale, high-throughput, mutation scanning workflows. Although the risk of missing a rare variant might still remain, this method showed better sensitivity for the identification of novel rare variants, and better accuracy for distinguishing different genotype groups, than the standard HR-melt mutation scanning method. Having validated this approach to screen the 36th and 59th coding exons of *ATM* efficiently, eight additional probes were designed to improve the mutation screening of the whole gene in our sample sets. Our general experimental strategy was to design an unlabeled probe for each variant reported to have a frequency >1% in the dbSNP database or in our sample series, in the regions of interest.

Thus, 10 out of 66 *ATM* amplicons could have been predicted beforehand to require simultaneous genotyping and mutation scanning (15%). We also applied this approach to two additional amplicons during the course of the study to facilitate their mutation screening. The first one contained an SNP found to be common in our sample series (rs3092910:T>C), and the second one contained a novel SNP 54 bp upstream of the 43rd coding exon of *ATM*, that we initially identified using the standard mutation scanning approach (Table 2). For all studied amplicons, cycling conditions (annealing temperature and number of cycles) were optimized in presence of LCGreen for mutation screening. Our experience showed that after PCR optimization, none of the *ATM* amplicons failed to amplify in the presence of LCGreen. For amplicons requiring simultaneous genotyping and mutation scanning, we reoptimized the PCR conditions in presence of the probe. We also verified that the 1:5 primer concentrations ratios would not impair the HR-melt analysis. Initial protocols had to be modified in some cases, especially by adjusting $MgCl_2$ concentration.

Using our strategy, a higher level of confidence in mutation scanning results can be reached when simultaneously proceeding to genotyping using unlabeled probes. We have shown that this approach can dramatically reduce the amount of sequencing required, compared to sequencing all variants that have a melt curve indicative of the presence of a sequence variant, and recommend the method whenever one of three criteria is met: 1) the cost of excess sequencing due to the presence of a known common variant in an amplicon will exceed the ~$50 to $75 setup cost of the unlabeled probe assay; 2) there is great concern that the presence of a known common SNP will mask the presence of an unknown rare SNP; or 3) it is important, within the mutation screening context, to detect all of the minor allele homozygotes of a common SNP located within an amplicon of interest.

The potential of HR-melt for cost-effective and sensitive high-throughput genotyping and mutation scanning has been reported in numerous studies. For example, Takano et al. [2008] and De Leeneer et al. [2008] described HR-melt as an economical screening method to detect mutations in *BRCA1* and *BRCA2*. In their work, the authors emphasized the advantages, both in time and cost, offered by the use of HR-melt. Cost-effective and rapid methods for screening are indeed highly needed for mutation screening and testing, particularly for molecular diagnostic purposes in medium and low-resources countries. For mutation discovery studies, this technique would also be beneficial since it enables large-scale case–control or population studies at low cost, but with a sensitivity and an accuracy higher than the current mutation scanning gold-standard, DHPLC [Chou et al., 2005].

In conclusion, simultaneous genotyping and mutation scanning is another methodology that confirms that HR-melt is a rapid, efficient, and cost-effective tool that can be used for high-throughput mutation screening for research, as well as for molecular diagnostic and clinical purposes.

Acknowledgments

We thank the many families who have contributed to the Breast Cancer Family registries (BCFR) and Kathleen Cuningham Foundation Consortium for research into Familial Breast cancer (kConFab). We also thank Heather Thorne, Eveline Niedermayr, all the kConFab research nurses and staff, the heads and staff of the Family Cancer Clinics, and the Clinical Follow Up Study (funded by NHMRC grants 145684, 288704, and 454508) for their contributions to this resource. T.N.-D. is the recipient of a fellowship from the Fondation de France and S.G. is the recipient of a fellowship from University of Pisa and a Special Training Award from IARC. The content of this manuscript does not necessarily reflect the views or policies of the NCI or any of the collaborating centers in the CFR, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government or the CFR. This work was supported by the National Cancer Institute, National Institutes of Health under RFA-CA-06-503 and R01-CA121245-01A2, and through cooperative agreements with members of the Breast Cancer Family Registry and P.I.s, including Cancer Care Ontario (U01 CA69467), Northern California Cancer Center (U01 CA69417), and University of Melbourne (U01 CA69638). kConFab is supported by grants from the National Breast Cancer Foundation, the National Health and Medical Research Council (NHMRC), and by the Queensland Cancer Fund, the Cancer Councils of New South Wales, Victoria, Tasmania, and South Australia, and the Cancer

Foundation of Western Australia. Authored by members of the BCFR group: specifically Cancer Care Ontario, Fred A. Litwin Center for Cancer Genetics, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada; the Northern California Cancer Center, Fremont, CA; and the Centre for Genetic Epidemiology, University of Melbourne, Carlton, Victoria, Australia; and members of the kConFab group: specifically the Peter MacCallum Cancer Centre, Melbourne, Australia.

References

Angèle S, Romestaing P, Moullan N, Vuillaume M, Chapot B, Friesen M, Jongmans W, Cox DG, Pisani P, Gérard JP, Hall J. 2003. ATM haplotypes and cellular response to DNA damage: association with breast cancer risk and clinical radiosensitivity. Cancer Res 63:8717–8725.

Atencio DP, Iannuzzi CM, Green S, Stock RG, Bernstein JL, Rosenstein BS. 2001. Screening breast cancer patients for ATM mutations and polymorphisms by using denaturing high-performance liquid chromatography. Environ Mol Mutagen 38:200–208.

Broeks A, Braaf LM, Huseinovic A, Schmidt MK, Russell NS, van Leeuwen FE, Hogervorst FB, Van 't Veer LJ. 2008. The spectrum of ATM missense variants and their contribution to contralateral breast cancer. Breast Cancer Res Treat 107:243–248.

Brunet J, Gutierrez-Enriquez S, Torres A, Berez V, Sanjose S, Galceran J, Izquierdo A, Menendez JA, Guma J, Borras J. 2008. ATM germline mutations in Spanish early-onset breast cancer patients negative for BRCA1/BRCA2 mutations. Clin Genet 73:465–473.

Buchholz TA, Weil MM, Ashorn CL, Strom EA, Sigurdson A, Bondy M, Chakraborty R, Cox JD, McNeese MD, Story MD. 2004. A Ser49Cys variant in the ataxia telangiectasia, mutated, gene that is more common in patients with breast carcinoma compared with population controls. Cancer 100:1345–1351.

Chou LS, Lyon E, Wittwer CT. 2005. A comparison of high-resolution melting analysis with denaturing high-performance liquid chromatography for mutation scanning: cystic fibrosis transmembrane conductance regulator gene as a model. Am J Clin Pathol 124:330–338.

De Leeneer K, Coene I, Poppe B, De Paepe A, Claes K. 2008. Rapid and sensitive detection of BRCA1/2 mutations in a diagnostic setting: comparison of two high-resolution melting platforms. Clin Chem 54:982–989.

Dörk T, Bendix R, Bremer M, Rades D, Klöpper K, Nicke M, Skawran B, Hector A, Yamini P, Steinmann D, Weise S, Stuhrmann M, Karstens JH. 2001. Spectrum of ATM gene mutations in a hospital-based series of unselected breast cancer patients. Cancer Res 61:7608–7615.

Garritano S, Gemignani F, Voegele C, Nguyen-Dumont T, Le Calvez-Kelm F, De Silva D, Lesueur F, Landi S, Tavtigian SV. 2009. Determining the effectiveness of high resolution melting analysis for SNP genotyping and mutation scanning at the TP53 locus. BMC Genetics 10:5 PMID: 19222838 [PubMed-in process].

Gonzalez-Hormazabal P, Bravo T, Blanco R, Valenzuela CY, Gomez F, Waugh E, Peralta O, Ortuzar W, Reyes JM, Jara L. 2008. Association of common ATM variants with familial breast cancer in a South American population. BMC Cancer 8:117.

Graham R, Liew M, Meadows C, Lyon E, Wittwer CT. 2005. Distinguishing different DNA heterozygotes by high-resolution melting. Clin Chem 51:1295–1298.

Izatt L, Greenman J, Hodgson S, Ellis D, Watts S, Scott G, Jacobs C, Liebmann R, Zvelebil MJ, Mathew C, Solomon E. 1999. Identification of germline missense mutations and rare allelic variants in the ATM gene in early-onset breast cancer. Genes Chromosomes Cancer 26:286–294.

Liew M, Pryor R, Palais R, Meadows C, Erali M, Lyon E, Wittwer C. 2004. Genotyping of single-nucleotide polymorphisms by high-resolution melting of small amplicons. Clin Chem 50:1156–1164.

Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA. 2004. Pattern of sequence variation across 213 environmental response genes. Genome Res 14:1821–1831.

Maillet P, Bonnefoi H, Vaudan-Vutskits G, Pajk B, Cufer T, Foulkes WD, Chappuis PO, Sappino AP. 2002. Constitutional alterations of the ATM gene in early onset sporadic breast cancer. J Med Genet 39:751–753.

Montgomery J, Wittwer CT, Palais R, Zhou L. 2007. Simultaneous mutation scanning and genotyping by high-resolution DNA melting analysis. Nat Protoc 2:59–66.

Reed GH, Wittwer CT. 2004. Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. Clin Chem 50:1748–1754.

Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, North B, Jayatilake H, Barfoot R, Spanova K, McGuffog L, Evans DG, Eccles D, Breast Cancer Susceptibility Collaboration (UK), Easton DF, Stratton MR, Rahman N. 2006. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet 38:873–875.

Seipp MT, Durtschi JD, Liew MA, Williams J, Damjanovich K, Pont-Kingdon G, Lyon E, Voelkerding KV, Wittwer CT. 2007. Unlabeled oligonucleotides as internal temperature controls for genotyping by amplicon melting. J Mol Diagn 9:284–289.

Sevilla C, Moatti JP, Julian-Reynier C, Eisinger F, Stoppa-Lyonnet D, Bressac-de Paillerets B, Sobol H. 2002. Testing for BRCA1 mutations: a cost-effectiveness analysis. Eur J Hum Genet 10:599–606.

Sommer SS, Jiang Z, Feng J, Buzin CH, Zheng J, Longmate J, Jung M, Moulds J, Dritschilo A. 2003. ATM missense mutations are frequent in patients with breast cancer. Cancer Genet Cytogenet 145:115–120.

Takano EA, Mitchell G, Fox SB, Dobrovic A. 2008. Rapid detection of carriers with BRCA1 and BRCA2 mutations using high resolution melting analysis. BMC Cancer 8:59.

Teraoka SN, Malone KE, Doody DR, Suter NM, Ostrander EA, Daling JR, Concannon P. 2001. Increased frequency of ATM mutations in breast carcinoma patients with early onset disease and positive family history. Cancer 92:479–487.

Thorstenson YR, Shen P, Tusher VG, Wayne TL, Davis RW, Chu G, Oefner PJ. 2001. Global analysis of ATM polymorphism reveals significant functional constraint. Am J Hum Genet 69:396–412.

Zhou L, Myers AN, Vandersteen JG, Wang L, Wittwer CT. 2004. Closed-tube genotyping with unlabeled oligonucleotide probes and a saturating DNA dye. Clin Chem 50:1328–1335.

Zhou L, Wang L, Palais R, Pryor R, Wittwer CT. 2005. High-resolution DNA melting analysis for simultaneous mutation scanning and genotyping in solution. Clin Chem 51:1770–1777.

Results

I

Implementation of an HRM assay for DAE analysis

I.1 DAE assessment by HRM: the general experimental approach

Differential allelic expression (DAE) is observed when the two alleles of a given gene produce different levels of transcript. Here, we aimed to demonstrate that HRM technology has potential to detect alteration in allele-specific transcript abundance.

I.1.1 General description of the approach

In order to make a differential measurement of the level of expression of the two alleles of a gene from a patient sample, one must be able to distinguish between the alleles. As in other DAE approaches described previously in the bibliographical review, the HRM approach requires a polymorphism within the mRNA sequence as a copy-specific tag.

The HRM approach for DAE analysis relies on the use of a fluorescently labeled oligonucleotide probe designed to anneal to the sequence surrounding the selected marker SNP, with standard PCR reagents, in an nested asymmetric PCR reaction[1]. To validate our approach, we tested both fluorescein probes and SimpleProbes.

Similarly to genotyping analysis, data acquisition in HRM instruments for DAE assessment consists in monitoring changes in the fluorescence properties of the probe, as it dissociates from the two allelic templates, while the probe-target duplexes are continuously heated. Data analysis displays melting curves which are plots of the temperature versus fluorescence for each sample. Since the probe has a complete complementarity to one variant while mismatching the other variant at the polymorphic position, thus resulting in a melting temperature difference between the two alleles, this technique allows the two alleles in heterozygous individuals to be distinguished by distinct peaks on the derivative curve. Then, the relative allelic abundance of the analyzed transcript is inferred from the ratio of the peak heights (Figure I.1).

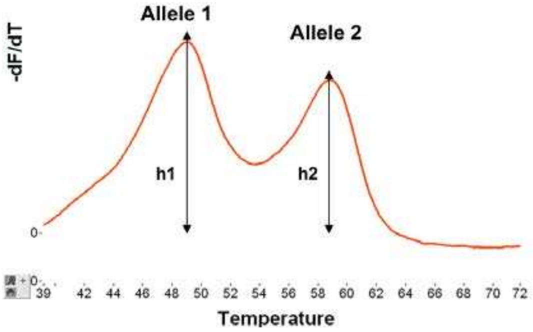


Figure I.1: Derivative melting profile of a SimpleProbe obtained from a heterozygous individual for the SNP rs2236142. Peak heights are measured and Allele 1/ Allele 2 ratio is calculated as h1/h2.

[1] Primers, probes and cycling conditions are detailed in the section of the memoir reporting the results for each of the assessed genes. Here, we aimed to present an overview of the experimental approach.

I.1.2 The hardware

The premises of the application of melting curve analysis to DAE assessment have been described by our collaborators in a study investigating mRNA degradation due to the non-sense mediated mRNA decay (NMD) mechanism in the *BRCA2* gene [Ware et al., 2006]. Ware et al. performed their analyses on the LightCycler 2.0 instrument, which is a carousel-based capillary system developed by Roche, that can analyze up to 32 samples. The LightCycler was the instrument by which melting analysis was first introduced [Wittwer et al., 1997]. Some years later, several suppliers have launched different devices and systems that allowed for high-resolution analysis [Wittwer et al., 2003, Herrmann et al., 2006].

Although they permit melting analysis, real-time PCR instruments such as the LightCycler or the Rotor-Gene 3000 (Corbett) are not primarily intended for melting analysis. Yet, the power of DNA melting analysis depends directly on the resolution of the melting instrument [Herrmann et al., 2006].

In this study, we chose the two platforms that performed the best in a comparative assessment of different HRM platforms, namely the HR-1™ and LightScanner® instruments (Idaho Technology) (Figure I.2) [Herrmann et al., 2006, Herrmann et al., 2007]. Both are dedicated instruments for HRM analysis. The HR-1™ is the most sensitive HRM system currently available. It uses glass capillaries and can analyze only one sample at a time, when the LightScanner® is a 96 or 384 plate-based system, with 5 times the resolution of other instruments of the same throughput.

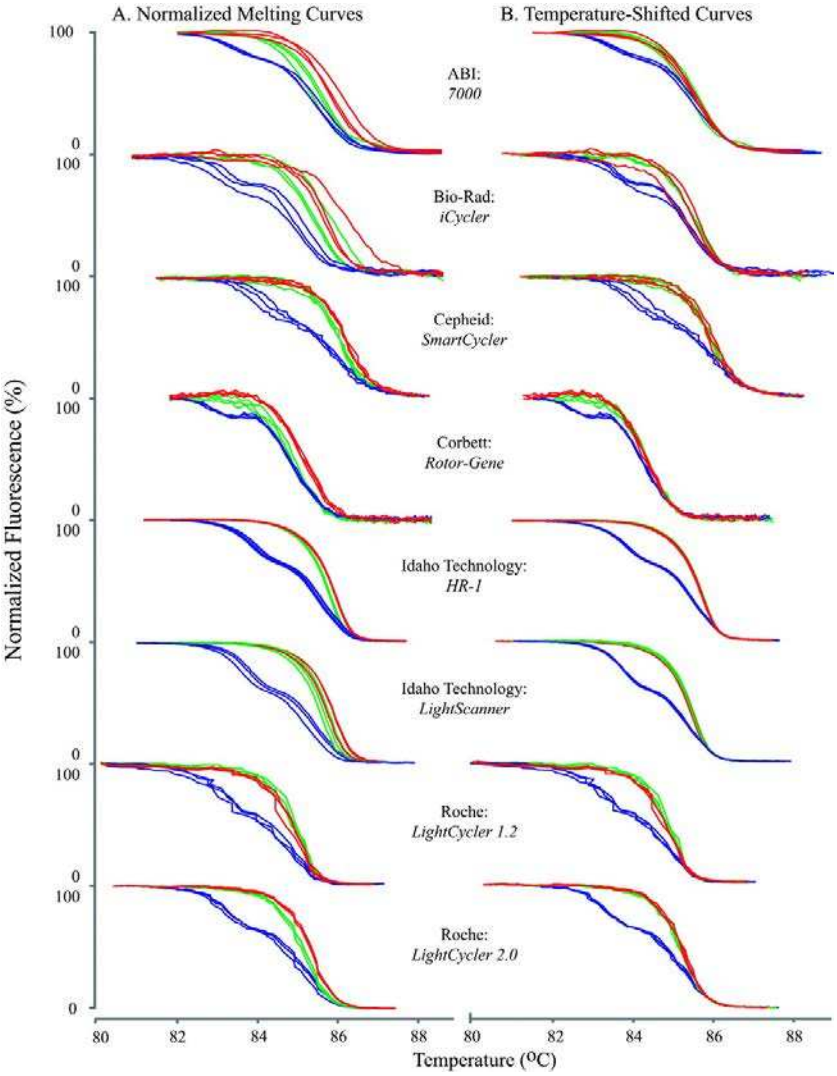


Figure I.2: Cross-platform comparison of melting instruments.

Melting curves of a 110-bp amplicon including the sickle cell SNP in the presence of LC Green®Plus. Each genotype was melted and displayed in triplicate on 8 different instruments. Wild-type samples are shown in green, heterozygotes in blue, and the homozygous mutants in red. (A) normalized melting curves for genotyping, (B) temperature-shifted curves for heterozygote scanning. From [Herrmann et al., 2006].

I.1.3 Assembling informative samples

Selection of LCLs

As described on the diagram of the Aims section, the case-control mutation screening project on breast cancer susceptibility genes ongoing in our laboratory provided genotyping information for common SNPs in the genes of interest. For each gene assessed for DAE, we were therefore able to select a set of breast cancer patients, heterozygous for a selection of marker SNPs and for which lymphoblastoid cell lines (LCLs) were available for further DAE assessment[2]. The LCLs included in our study were derived from subjects, who were considered to be at high risk of carrying a genetic predisposition to breast cancer due to an early age at onset and/or family history, and for whom no mutation in *BRCA1* or *BRCA2* genes had been identified.

Samples preparation

In some of the genes that we aimed to analyze, DAE could result from NMD, a cellular mechanism responsible for the specific degradation of an allele bearing a premature stop codon. In order to address this issue, RNA was prepared from each LCL under two culture conditions. One condition was a standard LCL culture condition, which was also the source of the genomic DNAs. The second condition involved cells that had been treated with puromycin, a translation inhibitor frequently used to stabilize transcripts containing a premature stop codon subject to NMD, without stabilizing either the wild-type transcript or transcripts containing a premature stop codon that are nonetheless insensitive to NMD. [Ware et al., 2006]. Complementary DNA (cDNA) was prepared from these two sources of RNA and will be hereafter referred to as "cDNA" and "puro-cDNA", respectively.

[2] Again, a more thorough description of the marker SNP selected and LCLs enrolled in each DAE study has been included in the section of the memoir reporting the corresponding results.

Before cDNA preparation, RNA integrity was controlled using the BioAnalyzer and RNA NanoChip II kit. Good quality RNAs, producing an RNA integrity number (RIN) ≥ 8, were selected for further analysis [Schroeder et al., 2006]. Whenever the quality threshold was not reached for a sample, RNA extraction was repeated so that all the RNAs used in this study had a minimum RIN of 8.

The NMD inhibitory treatment was really an anticipation to address Aim 4. As shown in the Aims diagram, in some cases, DAE could be explained by sequence variants identified during the mutation screening process, which investigated the coding exons and proximal intronic splice consensus sequences of candidate genes. These sequence variants include truncating mutations that induce NMD.

Observed DAE could also result from splice junction variants that lead to an unstable transcript. After eliminating the possibility of DAE linked to sequence variants found in the coding region, one can reorient the search towards sequence variation in non-coding regions, using bioinformatics methods, which allow for instance to identify SNPs in putative transcription factor binding sites and to study the evolutionary conservation of the surrounding sequences [Wasserman and Sandelin, 2004, Jordheim et al., 2008].

I.1.4 Statistical criteria for evidence of DAE

In the end, a test for DAE by HRM analysis is very much like biallelic marker genotyping, with two differences. First, the sample is cDNA, not genomic DNA. Second, the DAE test is done only on heterozygous samples and departure from the expected 1:1 ratio of a perfect heterozygote is indicative of differential expression. Although not the template of interest, genomic DNA must be assessed to provide the expected peak heights ratio value for a 1:1 allelic ratio. Thus, genomic DNA serves as internal control to control for any bias in the binding of the fluorescent probe to the two alleles. The level of allelic imbalance of an individual is calculated by dividing the signal ratio of the cDNA by its reference, the corresponding ratio of genomic DNA.

To allow for statistical calculations, each LCL is assessed in several PCR_1 replicates of genomic DNA, cDNA and puro-cDNA. Statistical significance for allelic imbalance is calculated using Student's t-test. Criteria for DAE are the following: i) the point estimate of the difference between genomic DNA and cDNA ratios should be greater than 20%; ii) at p-value ≤ 0.05, and iii) with the 95% confidence interval of the point estimate not including a null difference.

I.2 The early stages of the method: the HR-1™ instrument

I.2.1 Preliminary study: assessment of the *TP53* gene

Assay design for the *TP53* gene

Selection of a polymorphic cDNA marker and informative LCLs For assay implementation, we chose to study the *TP53* gene, which contains a number of common polymorphisms and rare mutations [Szymańska and Hainaut, 2003]. Among the SNPs of the coding region, we selected SNP c.215C>G (*p53R72P*, rs1042522). This SNP is responsible for a proline (C<u>C</u>G: P, ancestral allele) to an arginine (C<u>G</u>G: R) substitution at codon 72 of exon 4 of *TP53*. It will be hereafter referred to as the R72P polymorphism.

TP53 was not included in the case-control mutation screening project. However, previous work by Gemignani et al suggested the existence of a common mechanism leading to the disruption of the allelic expression balance for that gene. The authors reported that individuals homozygous for the C variant of R72P had a reduced expression of *TP53* compared to GG homozygotes. Heterozygous individuals had an intermediate level of expression [Gemignani et al., 2004]. These findings guided our choice to assay this gene during our preliminary work.

For this preliminary study, a panel of 74 LCLs, derived from breast cancer patients, were genotyped for the R72P polymorphism, by Taqman technology as described later. Twenty-five heterozygotes, for which RNA was immediately available, were used to test the assay. Two GG homozygotes and two CC homozygotes were also selected for further tests.

The HRM assay In their study of the *BRCA2* gene, Ware et al. used a SimpleProbe approach. We attempted to use the same kind of probe but were unsuccessful in obtaining a clear fluorescent signal. This was probably due to the presence of many Gs in the area surrounding the marker SNP [Gameau et al., 2005]. We thus turned to a fluorescein-labeled probe.

The fluorescein probe (5'-GGCTGCTCCCCGCGTGGC-3') was incorporated in the PCR_2 reaction. PCR_2 products were overlaid with clear oil before transfer into a glass capillary and analysis with the HR-1™ instrument. Melting curves were obtained by continuous measurement of the fluorescence during heating from 40 to 80°C. Although one sample is analyzed at a time, the turnaround was fast enough (2 minutes) to allow reasonable throughput.

Relative abundance of each allele was obtained from the ratio of the peak heights calculated on the first negative derivative plot of the fluorescence. Peak heights were measured manually. Then measurements were included in an Excel sheet to calculate peaks ratios and to determine the statistical significance of the observed allelic imbalance, using Student's *t*-test.

Mixing experiment A standard curve was generated by mixing genomic DNA from individuals homozygous for the common (G) and rare (C) variant, in the following G:C ratios: 9:1, 8:2, 7:3, 6:4, 5:5, 4:6, 3:7, 2:8 and 1:9 (data not shown). Linearity of the method was validated by calculating the determination coefficient. An R^2 of 0.9918 was obtained, establishing that fluorescein probe melting analysis could accurately allow detection of changes in the relative abundance of the two alleles in a heterozygous sample, by assessment the heights of the peaks corresponding to each allelic variant.

Assessment of genomic DNA Next, we aimed to determine the average ratio observed in genomic DNA, representing a perfect 1:1 ratio of the two alleles, in our collection of 25 heterozygotes for R72P. Despite the fact that the copy number of each allele is theoretically equal in LCLs, we found that the two peak heights ratio differed from 1. Indeed, we actually observed an average ratio of 0.42 [95% CI = 0.385-0.448] across the 25 heterozygous samples. The assay was replicated and the same average ratio, and similarly low standard deviation and CI values (95% CI = 0.386- 0.448), were obtained from the second assay. These data generated from known heterozygotes demonstrated the between samples reproducibility of the assay and the normal variation in a situation where both alleles are theoretically equally abundant.

First published application of the *TP53* assay

Implementation of the assay was conducted in collaboration with Dr Lars P. Jordheim, who was studying acute myeloid leukemia (AML). Our results for DAE assessment of *TP53* will be reported in Chapter III of this section.

Jordheim et al. previously reported that mRNA expression levels of certain genes have shown predictive value for the outcome of AML-patients treated with cytarabine [Jordheim and Dumontet, 2007]. In the following article, the DAE-approach was used to i) to investigate whether interindividual variations in mRNA expression levels of genes involved in the cellular response to cytarabine could, at least in part, be due to genetic polymorphisms and ii) to identify the regulatory SNPs responsible for these differences.

Using leucoblasts from AML patients treated with cytarabine, the HRM approach was used to assess genes with a key role in the metabolism and mechanism of action of cytarabine, and with strong evidence of clinical relevance. These were the genes for the equilibrative nucleoside transporter 1, *SLC29A1*, deoxycytidine kinase, *DCK*, cytidine deaminase, *CDA*, cytosolic 5'-nucleotidases II and III, *NT5C2* and *NT5C3*, and the tumor suppressor gene *TP53*. Different extents of DAE were observed but the causative variants could not be identified through a bioinformatics approach.

Article II

Differential Allelic Expression in Leukoblast from Patients with Acute Myeloid Leukemia Suggests Genetic Regulation of CDA, DCK, NT5C2, NT5C3, and TP53

Jordheim LP, **Nguyen-Dumont T**, Thomas X, Dumontet C, Tavtigian SV.

Drug Metab Dispos, 2008; 36(12):2419-23.

Supplemental Material can be found at:
http://dmd.aspetjournals.org/cgi/content/full/dmd.108.023184/DC1

0090-9556/08/3612-2419–2423$20.00
DRUG METABOLISM AND DISPOSITION
Copyright © 2008 by The American Society for Pharmacology and Experimental Therapeutics
DMD 36:2419–2423, 2008

Vol. 36, No. 12
23184/3405592
Printed in U.S.A.

Short Communication

Differential Allelic Expression in Leukoblast from Patients with Acute Myeloid Leukemia Suggests Genetic Regulation of *CDA*, *DCK*, *NT5C2*, *NT5C3*, and *TP53* S

Received July 18, 2008; accepted September 4, 2008

ABSTRACT:

mRNA expression levels of certain genes have shown predictive value for the outcome of cytarabine-treated AML-patients. We hypothesized that genetic variants play a role in the regulation of the transcription of these genes. We studied leukoblasts from 82 patients with acute myeloid leukemia and observed various extent and frequency of differential allelic expression in the *CDA*, *DCK*, *NT5C2*, *NT5C3*, and *TP53* genes. Our attempts to identify the causative regulatory single nucleotide polymorphisms by a bioinformatics approach did not succeed. However, our results indicate that genetic variations are at least in part responsible for the differences in overall expression levels of these genes.

The deoxynucleoside analog cytarabine (1-β-D-arabinofuranosylcytosine) is a major component of the chemotherapeutic treatment of patients with acute myeloid leukemia (AML). The most important limitation for the use of deoxynucleoside analogs in the clinic is the presence of primary or acquired resistance. Several studies have identified clinically relevant mechanisms of resistance in patients with leukemia or other malignant diseases (Jordheim and Dumontet, 2007). In particular, the mRNA expression level of several genes has been correlated to the outcome of the treatment with cytarabine or gemcitabine (2′-2′-difluorodeoxycytidine), another analog of deoxycytidine. In all of these studies, large variations in the expression levels of genes involved in cytarabine metabolism have been observed between patients, suggesting the presence of important regulatory mechanisms. In addition to differences in levels and activities of transcription factors and stability of mRNA, variations in the genomic sequence of the gene and its regulatory elements can influence the mRNA level. In fact, at least 25 to 35% of interindividual differences in gene expression are supposed to be caused by *cis*-acting variations (Pastinen and Hudson, 2004).

When a heterozygous genetic variation induces a difference in mRNA expression level, the two corresponding alleles are expressed at different levels. This is called differential allelic expression (DAE) or allelic expression imbalance (Pastinen and Hudson, 2004). Currently, DAE is studied in samples heterozygous for an exonic variation (exonic single nucleotide polymorphisms; cSNP) used as a marker to determine the relative amount of transcripts from the two alleles. This method allows the distinction between *cis*- and *trans*-acting effects because the cellular environment and mRNA extraction are exactly the same (Stamatoyannopoulos, 2004). The cSNP used for the assessment of DAE is not necessarily responsible for the allelic expression imbalance, and additional investigations are needed to identify the functional regulatory variant (regulatory SNP; rSNP) or the underlying epigenetic modification (Milani et al., 2007).

Specific mRNA expression levels can be used to predict the outcome of cancer patients treated with chemotherapy. Because genetic variants are partially responsible for variations in gene expression, these could potentially be used as more precise markers for this prediction (Stamatoyannopoulos, 2004; Abraham et al., 2006). For AML, the use of reliable predictive markers would substantially increase the treatment success rate and the overall management of the cancer patients. Our main goal in this study was to investigate whether interindividual variations in mRNA expression levels of genes involved in the cellular response to cytarabine could, at least in part, be due to genetic polymorphisms. As a secondary goal, we tried to identify the rSNPs responsible for these differences. We focused on genes with a key role in the metabolism and mechanism of action of cytarabine and with strong evidence of clinical relevance. We chose the genes for the equilibrative nucleoside transporter 1, *SLC29A1*, deoxycytidine kinase, *DCK*, cytidine deaminase, *CDA*, cytosolic 5′-nucleotidases II and III, *NT5C2* and *NT5C3*, and the tumor suppressor gene *TP53*. We used a method based on high-resolution melting-curve analysis to assess the DAE of these genes, determined their relative

L.P.J. received a post-doctoral grant from Ligue Contre le Cancer-Comité du Rhône, and T.N.-D. received a PhD grant from Fondation de France. Work in our laboratories was partially funded by Oddrun Mjålands Stiftelse for Kreftforskning and Medeco's Stiftelse.

Parts of this work were previously presented as a poster as follows: Jordheim LP, Nguyen T, Thomas X, Dumontet C, and Tavtigian SV (2007) Differential allelic expression in genes involved in the cellular response to the nucleoside analogue cytarabine. *Gordon Research Conference on Nucleosides, Nucleotides & Oligonucleotides;* 2007 July 1–6; Newport, RI. Gordon Research Conferences, West Kingston, RI.

Article, publication date, and citation information can be found at http://dmd.aspetjournals.org.

doi:10.1124/dmd.108.023184.

S The online version of this article (available at http://dmd.aspetjournals.org) contains supplemental material.

ABBREVIATIONS: AML, acute myeloid leukemia; DAE, differential allelic expression; cSNP, exonic single nucleotide polymorphism; rSNP, regulatory SNP; PCR, polymerase chain reaction; gDNA, genomic DNA; RT, reverse transcription.

expression level, and tried to identify causative rSNPs in leukoblasts from 82 patients with AML.

Materials and Methods

Biological samples were obtained from 82 patients with AML at diagnosis before initiation of therapy, all followed in the Hematology Department of Edouard Herriot Hospital in Lyon, France. Approval was obtained from Lyon Protocol Review Board, and written informed consent was provided according to the Declaration of Helsinki. Mononuclear cells including leukemic cells were isolated by Ficoll-Hypaque sedimentation from peripheral blood ($n = 33$) and bone marrow ($n = 49$). Median percentages of blast cells in peripheral blood and in bone marrow were 61 (range, 13–99%) and 69% (range, 20–95%), respectively.

Total RNA and genomic DNA were extracted with TRIzol Reagent (Invitrogen, Cergy Pontoise, France), and cDNA synthesis was performed with 1 μg of total RNA using SuperScript III Reverse Transcriptase (Invitrogen) and oligo(dT) primers.

SNPs were genotyped by high-resolution melting-curve analysis of polymerase chain reaction (PCR) products in the presence of LCGreen Plus+ (Idaho Technologies, Salt Lake City, Utah) (Reed and Wittwer, 2004) or a fluorescent probe (Crockett and Wittwer, 2001), using a LightScanner Instrument (Idaho Technologies). All PCR primers and cycling conditions are described in supplemental tables.

DAE was assessed with the same technique as the genotyping, using cDNA and genomic DNA based on a previously described method (Ware et al., 2006). All of the samples were amplified five times. Melting curves were analyzed with the HR-1 Instrument Analysis Software (Idaho Technologies), peak heights were measured manually, and the ratio of the two peaks corresponding to the two alleles was calculated for cDNA and genomic DNA. Allelic expression imbalance of each sample was calculated as the ratio (mean ratio for cDNA)/(mean ratio for gDNA). A sample was considered to have DAE if this ratio was <0.8 or >1.2 and if the difference between the mean ratios was statistically significant at $p < 0.01$ as calculated with the Student's *t* test. The linearity of the method for the assessment of DAE was validated using mix of genomic DNA from common and rare homozygous samples containing 20 to 80% of each allele. R^2 values were between 0.9918 and 0.9979 for *CDA*, *NT5C2*, *NT5C3*, and *TP53*. For *NT5C3*, PCR products of PCR1 were digested with PvuII to avoid interference with mRNA from pseudogene *NT5C3P1* as described earlier (Marinaki et al., 2001).

SNPs situated in putative transcription factor binding sites, or rSNPs, were identified in the region situated upstream of the start codon of genes of interest using the public databases RAVEN (http://www.cisreg.ca/cgi-bin/RAVEN/a), CONSITE (http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite/), and TESS (http://www.cbil.upenn.edu/cgi-bin/tess/tess). Their evolutionary conservation was verified by comparative sequence analysis with mouse, dog, and cow using seqcom comparisons in FamilyRelations II (http://family.caltech.edu/) with a window size of 20 base pairs and a threshold of 0.8. Putative rSNPs for *DCK* and *SLC29A1* were selected from the literature.

mRNA expression levels were determined by relative quantitative reverse transcription (RT)-PCR on an ABI PRISM 7900 sequence detection system (Applied Biosystems, Courtaboeuf, France) using *GAPDH* as internal standard and the comparative threshold cycle method as described in the user's guide. TaqMan gene expression assays were Hs00156401_m1 (CDA), Hs00176127_m1 (DCK), Hs00366992_m1 (NT5C2), Hs00826433_m1 (NT5C3), Hs00191940_m1 (SLC29A1), and Hs00153349_m1 (TP53).

The Student's *t* test was used for statistical analysis of DAE (differences in mean of ratios between cDNA and gDNA) and of the association between genotype and mRNA expression (differences in mean mRNA expression between genotype groups).

Results and Discussion

The DAE protocol can only be applied to samples that are heterozygous for a marker cSNP in the target gene. Genotyping *CDA*, *DCK*, *NT5C2*, *NT5C3*, and *TP53* in our series of 82 patients revealed that 13 to 41 (15.9–51.3%) were heterozygous for selected cSNPs with theoretically high frequency, and thereby suitable for the DAE experiments (Table 1). For *NT5C2*, two highly frequent cSNPs were genotyped, but only rs3740387 in exon 18 was retained for DAE-assessment. *SLC29A1* was excluded from the analysis because only one heterozygote for the cSNP was found. This is consistent with the reported low frequencies of cSNPs in *SLC29A1* in Europeans (Osato et al., 2003). Statistically significant DAE was observed in 57.7, 50.0, 8.7, 38.5, and 16.7% of positive samples for *CDA*, *DCK*, *NT5C2*, *NT5C3*, and *TP53*, respectively (Fig. 1; Table 1). The extent of imbalanced allelic expression varied from 20% up to monoallelic expression of *NT5C3* in four samples. The percentage of DAE-positive samples for each gene was not different between samples from peripheral blood and bone marrow. DAE has been reported to be tissue-dependent (Wilkins et al., 2007), but here we studied the same cells (leukoblasts) in two different environments (peripheral blood and bone marrow). Detection of DAE provides strong evidence that *cis*-genetic variation is involved in the determination of the expression level of these genes. We observed DAE in leukoblasts that are the target cells for the cytarabine-based treatment of AML. Therefore, *cis*-regulation of these genes could have a direct impact on the efficiency of the chemotherapeutic drug used for treatment of AML. Because DAE was assessed in samples heterozygous for a marker cSNP only and the linkage disequilibrium with the causative SNP is unknown, it is difficult to estimate the rate of DAE in the whole population of 82 AML patients. The observed DAE could be due to genetic or epigenetic variants in transcription factor binding sites or by nonsense-mediated mRNA decay. We did not have biological material to study nonsense-mediated mRNA decay, but we continued our research on putative rSNPs.

Eighteen putative rSNPs in the 5′-region of our genes of interest were identified by bioinformatics tools or selected from the literature (Shi et al., 2004; Fitzgerald et al., 2006; Gilbert et al., 2006; Joerger et al., 2006; Myers et al., 2006; Sugiyama et al., 2007). Compared with sequencing of large upstream regions of genes, this method allows screening of thousands of kb in silico to make a selection of potential rSNPs in either proximal promoters or more distant candidate enhancers (Wasserman and Sandelin, 2004). The different databases identified various SNPs situated in potential transcription factor binding sites within the regulatory elements and for which the two alleles potentially did not have the same affinity for transcription factors (data not shown). Sequence conservation through mouse, dog, and cow was more or less constant. After genotyping of these SNPs, their comparison with the DAE status did not show any correlation (Table 1). This result might be explained by the limited power of the study of some SNPs (few heterozygote samples) or reflect the fact that the studied SNPs do not intervene in the regulation of the expression of these genes. Functional rSNPs in our genes of interest can be as follows: SNPs that we did not chose to genotype; situated in sequences not reported on the publicly available databases; or situated elsewhere than upstream of the start codon. In addition, the analysis might have been complicated by the presence of several rSNPs in the same gene.

Median values (and ranges) for relative mRNA expression in leukoblasts from 67 patients were 8.1 (0–215.3) for *CDA*, 12.6 (2.5–162.6) for *DCK*, 8.0 (0–88.7) for *NT5C2*, 0.8 (0.1–6.0) for *NT5C3*, 1.0 (0–18.2) for *SLC29A1* ($n = 65$), and 1.0 (0–6.6) for *TP53*. This result confirmed our previous publications reporting large interindividual variations in gene expression between AML patients (reviewed in Jordheim and Dumontet, 2007). If the role of the functional rSNP is important compared with other regulating parameters, genotypes should be correlated to the mRNA expression of the regulated gene, with heterozygous samples between the two groups of homozygotes. Comparison between the genotype groups of cSNPs and rSNPs did

TABLE 1

Genotype, DAE, and quantitative RT-PCR data for all studied cSNPs and rSNPs

Two to six SNPs were genotyped in each gene and correlated to mRNA expression level as determined by quantitative RT-PCR and the DAE status.

Gene	SNP Information				Genotyping Results				Quantitative RT-PCR Results			DAE Results				
	SNP ID	cSNP/rSNP	Nucleotide Position^a	Nucleotide Change	HH^b (%)	Hh^b (%)	Hh^b (%)	Negative	mRNA HH^c (*n*)	mRNA Hh^c (*n*)	mRNA hh^c (*n*)	Hh DAE+^{d}	Hh DAE−^{d}	HH/hh DAE+^{d}	HH/hh DAE−^{d}	Negative Genotype DAE+ or DAE−^{d}
CDA	rs2072671	cSNP	+79	A>C	28 (35.00)	41 (51.25)	11 (13.75)	2	19.33 ± 21.68 (19)	19.16 ± 7.67 (36)	5.75 ± 3.14 (11)	15	11			
CDA	rs6690069	rSNP	−1172	G>A	62 (86.11)	7 (9.72)	3 (4.71)	10	17.54 ± 9.11 (51)	35.46 ± 40.15 (4)	6.95 ± 8.08 (3)	1	2	12	8	3
CDA	rs10916823	rSNP	−897	C>A	71 (100.00)	0 (0.00)	0 (0.00)	11	18.23 ± 8.47 (58)			0	0	13	10	3
CDA	rs532545	rSNP	−451	C>T	23 (32.86)	35 (50.00)	12 (17.14)	12	21.25 ± 25.61 (16)	20.28 ± 9.42 (29)	8.38 ± 3.92 (12)	12	8	1	2	3
CDA	rs12095662	rSNP	−378	T>C	71 (100.00)	0 (0.00)	0 (0.00)	11	18.23 ± 8.47 (58)			0	0	13	10	3
CDA	rs602950	rSNP	−92	A>G	25 (35.21)	33 (46.48)	13 (18.31)	11	21.67 ± 24.07 (17)	19.64 ± 9.86 (28)	8.32 ± 3.96 (12)	12	7	1	2	4
DCK	rs11544786	cSNP	+28624	C>T	69 (85.15)	13 (15.85)	0 (0.00)	0	19.18 ± 6.02 (56)	21.58 ± 8.14 (11)		5	5			
DCK	SNP-360	rSNP	−360	C>G	79 (97.53)	2 (2.47)	0 (0.00)	1	19.15 ± 5.35 (64)	30.54 ± 34.90 (2)		0	0	5	5	0
DCK	SNP-243	rSNP	−243	G>T	81 (100.00)	0 (0.00)	0 (0.00)	1	19.49 ± 5.27 (66)			0	0	5	5	0
DCK	rs2306744	rSNP	−201	C>T	80 (100.00)	0 (0.00)	0 (0.00)	2	19.74 ± 5.32 (65)			0	0	5	5	0
NT5C2	rs10883841	cSNP	+7	T>C	61 (79.22)	16 (20.78)	0 (0.00)	5	15.71 ± 4.56 (49)	9.40 ± 4.27 (15)		0	7	2	13	1
NT5C2	rs3740387	cSNP	+85248	C>T	30 (38.46)	27 (34.62)	21 (26.92)	4	12.16 ± 4.56 (26)	15.34 ± 7.53 (22)	14.32 ± 6.66 (19)	2	21			
NT5C2	rs12781668	rSNP	−19360	T>A	81 (100.00)	0 (0.00)	0 (0.00)	1	13.82 ± 3.54 (67)			0	0	2	21	0
NT5C2	rs7917650	rSNP	−2486	C>G	51 (66.23)	21 (27.27)	5 (6.49)	5	12.09 ± 3.38 (39)	18.15 ± 9.80 (19)	5.02 ± 1.64 (4)	1	9	1	10	2
NT5C2	rs12261294	rSNP	−238	G>A	35 (45.45)	34 (44.16)	8 (10.39)	5	13.04 ± 4.49 (26)	13.39 ± 4.50 (30)	5.00 ± 2.06 (6)	1	15	1	3	3
NT5C3	rs3750117	cSNP	+14,603	C>T	38 (46.34)	32 (39.02)	12 (14.63)	0	1.31 ± 0.44 (33)	1.12 ± 0.37 (25)	1.10 ± 0.96 (9)	10	16			
NT5C3	rs13228639	rSNP	−26881	A>G	36 (45.57)	36 (45.57)	7 (8.86)	9	1.27 ± 0.42 (33)	1.14 ± 0.47 (21)	0.88 ± 0.40 (4)	10	13	0	0	3
NT5C3	rs7778958	rSNP	−6937	G>A	37 (45.68)	34 (41.98)	10 (12.35)	1	1.27 ± 0.43 (33)	1.16 ± 0.40 (25)	1.15 ± 1.08 (8)	9	14	0	2	1
NT5C3	rs4723239	rSNP	−6441	A>G	76 (95.00)	4 (5.00)	0 (0.00)	2	1.26 ± 0.31 (61)	0.65 ± 0.62 (4)		0	4	9	12	1
NT5C3	rs4316067	rSNP	−5933	T>C	35 (43.75)	32 (40.00)	13 (16.25)	2	1.10 ± 0.42 (25)	1.25 ± 0.47 (29)	1.51 ± 0.74 (11)	5	9	5	6	1
SLC29A1	rs8187641	cSNP	+3312	T>C	78 (98.73)	1 (1.27)	0 (0.00)	3	1.56 ± 0.62 (66)	1.12 ± 0.00 (1)						
SLC29A1	rs747199	rSNP	−706	G>C	54 (67.50)	24 (30.00)	2 (2.50)	2	1.71 ± 0.92 (43)	1.24 ± 0.54 (20)	1.64 ± 1.44 (2)					
TP53	rs1042522	cSNP	+441	G>C	49 (60.49)	23 (28.40)	9 (11.11)	1	1.29 ± 0.35 (40)	1.57 ± 0.58 (21)	1.17 ± 0.61 (6)	3	15			
TP53	rs17885803	rSNP	−12565	G>A	64 (84.21)	11 (14.47)	1 (1.32)	6	1.29 ± 0.27 (53)	1.77 ± 1.24 (9)	0.81 ± 0.00 (1)	2	3	1	11	1
TP53	rs17883670	rSNP	−11805	C>G	75 (100.00)	0 (0.00)	0 (0.00)	7	1.36 ± 0.29 (62)			0	0	3	13	2

HH, frequent homozygote; Hh, heterozygote; hh, rare homozygote.

[a] Position with respect to ATG as +1.

[b] Genotyping results.

[c] Mean values for mRNA expression level ±95% confidence interval.

[d] Data only concerning samples with determined DAE status.

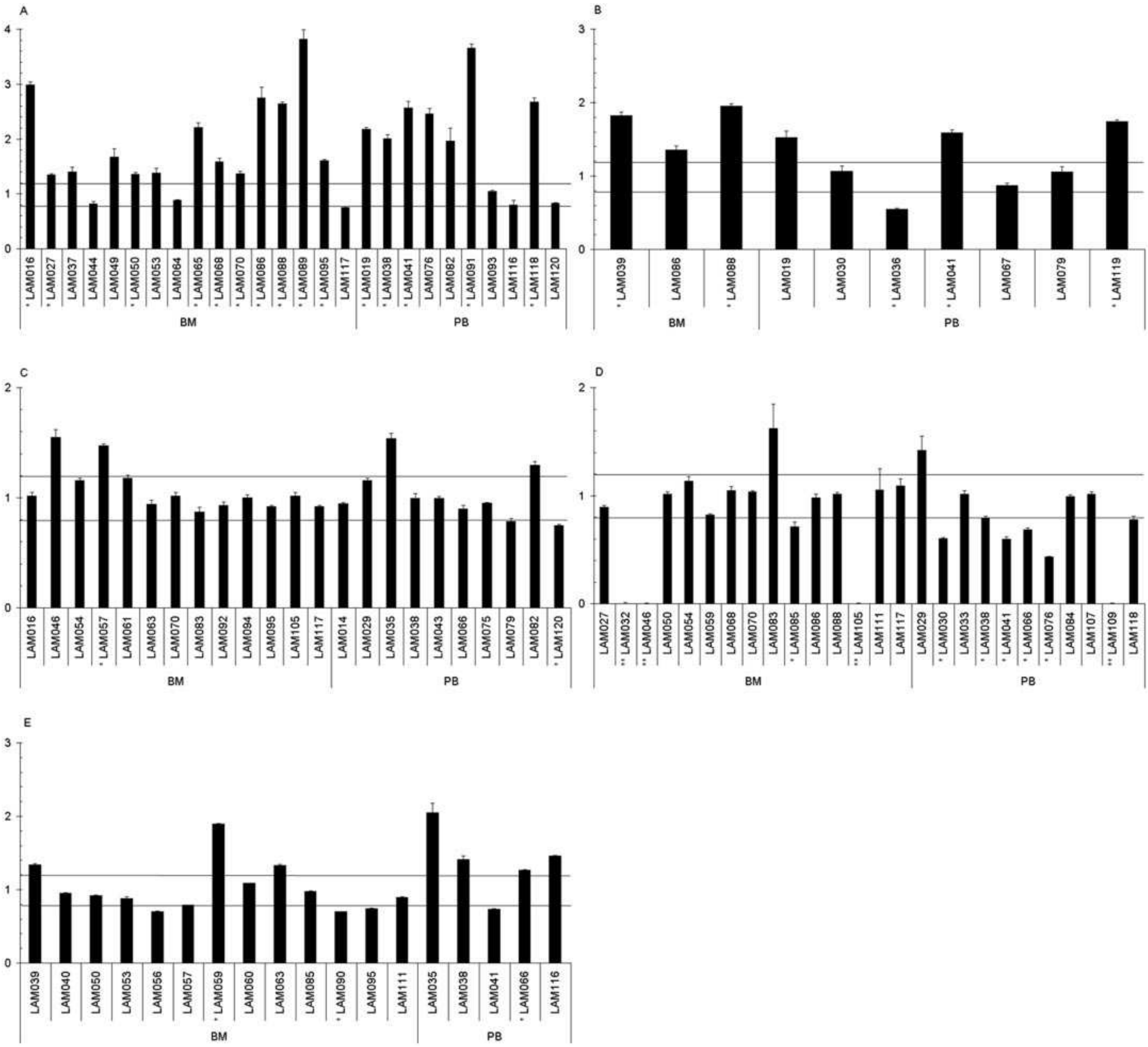


FIG. 1. Allelic-specific expression of CDA (A), DCK (B), NT5C2 (C), NT5C3 (D), and TP53 (E) in leukoblasts from AML patients. All samples heterozygous for cSNPs in *CDA* (rs2072671), *DCK* (rs11544786), *NT5C2* (rs10883841), *NT5C3* (rs3750117), and *TP53* (rs1042522) were assessed for DAE (5 replicates), and statistical significance was calculated using the Student's *t* test for comparison of the mean of the ratios of cDNA versus genomic DNA as explained under *Materials and Methods*. The *y*-axis shows the value of the ratio of mean cDNA ratios over mean gDNA ratios. Samples without DAE have a ratio of 1. Horizontal bars indicate the lower (0.8) and upper (1.2) limits of the ±20% zone. The *x*-axis shows sample IDs. *, samples with a ratio lower than 0.8 or higher than 1.2 and with $p < 0.01$ using the Student's *t* test; **, samples with monoallelic expression of NT5C3. BM, samples from blood marrow; PB, samples from peripheral blood.

not show statistically significant differences (Table 1). However, trends were seen for rs2072671 in *CDA* [$p = 0.07$ for comparison between heterozygote (19.16, $n = 36$) and rare homozygous (5.75, $n = 11$) samples] and for rs12261294 in *NT5C2* [$p = 0.10$ for comparison between the pool of common homozygous and heterozygote (13.23, $n = 56$) samples and the rare homozygous (4.96, $n = 6$) samples]. We did not observe higher expression for *SLC29A1* in samples with C alleles for rs747199 as previously reported (Myers et al., 2006). Lower promoter activity has been shown for T alleles of rs532545 in *CDA*, which is consistent with our mRNA data in TT-samples for this variant (Fitzgerald et al., 2006; Gilbert et al., 2006). When subgroups of samples were compared for their correlation between genotypes and mRNA expression, statistically significant differences were observed in some cases. This was the case, for example, for rs4316067 in *NT5C3* [$p = 0.005$ for comparison between heterozygote ($n = 11$) and rare homozygous ($n = 4$) samples from peripheral blood only] and rs1042522 in *TP53* [$p = 0.001$ for comparison between common homozygous ($n = 22$) and heterozygotes ($n = 14$) samples from blood marrow only]. Comparison of gene expression levels between samples with or without DAE showed no differences (data not shown), thus eliminating low expression of the target gene as a bias of DAE assessment (Pastinen and Hudson, 2004).

This work provides proof that genes involved in the cellular response to cytarabine are subject to genetic or epigenetic regulation in leukemic blasts. The fact that a patient shows differential allelic expression in a cytarabine-related gene would not have an affect on the response to the treatment. However, this clearly indicates that interindividual differences in gene expression with predictive power in cohorts of AML patients treated with cytarabine can at least partially be explained by genetic variations. In addition to providing an explanation to previous data available in this field, our results strongly encourage the search of causative variants for the differences in expression levels.

International Agency for Research on Cancer,
Lyon, France (L.P.J., T.N.-D., S.V.T.);
Université de Lyon,
Institut National de la Santé et de la Recherche Médicale U590,
Lyon, France (L.P.J., C.D.); and
Department of Hematology,
Hôpital Edouard Herriot,
Lyon, France (X.T.)

L. P. Jordheim
T. Nguyen-Dumont
X. Thomas
C. Dumontet
S. V. Tavtigian

References

Abraham J, Earl HM, Pharoah PD, and Caldas C (2006) Pharmacogenetics of cancer chemotherapy. *Biochim Biophys Acta* **1766:**168–183.

Crockett AO and Wittwer CT (2001) Fluorescein-labeled oligonucleotides for real-time pcr: using the inherent quenching of deoxyguanosine nucleotides. *Anal Biochem* **290:**89–97.

Fitzgerald SM, Goyal RK, Osborne WR, Roy JD, Wilson JW, and Ferrell RE (2006) Identification of functional single nucleotide polymorphism haplotypes in the cytidine deaminase promoter. *Hum Genet* **119:**276–283.

Gilbert JA, Salavaggione OE, Ji Y, Pelleymounter LL, Eckloff BW, Wieben ED, Ames MM, and Weinshilboum RM (2006) Gemcitabine pharmacogenomics: cytidine deaminase and deoxycytidylate deaminase gene resequencing and functional genomics. *Clin Cancer Res* **12:**1794–1803.

Joerger M, Bosch TM, Doodeman VD, Beijnen JH, Smits PH, and Schellens JH (2006) Novel deoxycytidine kinase gene polymorphisms: a population screening study in Caucasian healthy volunteers. *Eur J Clin Pharmacol* **62:**681–684.

Jordheim LP and Dumontet C (2007) Review of recent studies on resistance to cytotoxic deoxynucleoside analogues. *Biochim Biophys Acta* **1776:**138–159.

Marinaki AM, Escuredo E, Duley JA, Simmonds HA, Amici A, Naponelli V, Magni G, Seip M, Ben-Bassat I, Harley EH, et al. (2001) Genetic basis of hemolytic anemia caused by pyrimidine 5′ nucleotidase deficiency. *Blood* **97:**3327–3332.

Milani L, Gupta M, Andersen M, Dhar S, Fryknäs M, Isaksson A, Larsson R, and Syvänen AC (2007) Allelic imbalance in gene expression as a guide to cis-acting regulatory single nucleotide polymorphisms in cancer cells. *Nucleic Acids Res* **35:**e34.

Myers SN, Goyal RK, Roy JD, Fairfull LD, Wilson JW, and Ferrell RE (2006) Functional single nucleotide polymorphism haplotypes in the human equilibrative nucleoside transporter 1. *Pharmacogenet Genomics* **16:**315–320.

Osato DH, Huang CC, Kawamoto M, Johns SJ, Stryke D, Wang J, Ferrin TE, Herskowitz I, and Giacomini KM (2003) Functional characterization in yeast of genetic variants in the human equilibrative nucleoside transporter, ENT1. *Pharmacogenetics* **13:**297–301.

Pastinen T and Hudson TJ (2004) Cis-acting regulatory variation in the human genome. *Science* **306:**647–650.

Reed GH and Wittwer CT (2004) Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. *Clin Chem* **50:**1748–1754.

Shi JY, Shi ZZ, Zhang SJ, Zhu YM, Gu BW, Li G, Bai XT, Gao XD, Hu J, Jin W, et al. (2004) Association between single nucleotide polymorphisms in deoxycytidine kinase and treatment response among acute myeloid leukaemia patients. *Pharmacogenetics* **14:**759–768.

Stamatoyannopoulos JA (2004) The genomics of gene expression. *Genomics* **84:**449–457.

Sugiyama E, Kaniwa N, Kim SR, Kikura-Hanajiri R, Hasegawa R, Maekawa K, Saito Y, Ozawa S, Sawada J, Kamatani N, et al. (2007) Pharmacokinetics of gemcitabine in Japanese cancer patients: the impact of a cytidine deaminase polymorphism. *J Clin Oncol* **25:**32–42.

Ware MD, DeSilva D, Sinilnikova OM, Stoppa-Lyonnet D, Tavtigian SV, and Mazoyer S (2006) Does nonsense-mediated mRNA decay explain the ovarian cancer cluster region of the BRCA2 gene? *Oncogene* **25:**323–328.

Wasserman WW and Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5:**276–287.

Wilkins JM, Southam L, Price AJ, Mustafa Z, Carr A, and Loughlin J (2007) Extreme context specificity in differential allelic expression. *Hum Mol Genet* **16:**537–546.

Address correspondence to: Dr. Lars P. Jordheim, Laboratoire de Cytologie Analytique, INSERM U590, Faculté de Médecine Rockefeller, 8, Avenue Rockefeller, 69008 Lyon, France. E-mail: jordheim@yahoo.com

I.2.2 Preliminary assessment of the *CHEK2* gene

Assay design for the *CHEK2* gene

Selection of polymorphic cDNA markers and informative samples Based on the genotyping data provided by the mutation screening project, we selected as markers the SNPs with the highest frequencies so it would allow us to maximize the number of available informative LCLs. These were the SNPs g.4879C>G (rs2236142) and g.4953G>A (rs2236141) located in the 5'UTR region of the gene.

Thirty-two were heterozygotes for rs2236142 and seventeen were heterozygotes for rs2236141. Eight individuals were double heterozygotes. RNA was prepared from the selected LCLs, under the two cell culture conditions mentioned previously, in order to be able to evaluate the effect of NMD inhibition on allelic expression.

Specificities to *CHEK2* assay We employed specific primer pairs in the PCR_1 to amplify genomic DNA and cDNA. Both markers were assayed by SimpleProbe melting analysis.

PCR product melting curves were obtained from the HR-1$^{™}$ by melting from 35 to 75°C. Given the large umber of samples to assess, we sought to reduce the work load by automatizing peak height measurements and ratio calculations. Therefore, each analysis file generated by the HR-1$^{™}$ software was exported and processed with Excel macro that I wrote. For each sample, the HR-1$^{™}$ provided the fluorescence estimate at each temperature point. Based on this information, the macro was written to draw the derivative curve, to measure the peak heights, to calculate the peaks ratios and to perform the Student's *t*-test.

Mixing experiment

We created a range of melting curves associated with known allelic imbalance. Using SNP rs2236142 as marker, we produced bi-allelic templates with the following G:C proportions: 1:9, 2:8, 3:7, 4:6; 5:5, 6:4, 7:3; 8:2, 9:1. As shown in Figure I.3, the melting profile of the mixtures of opposite homozygotes reflected the relative contribution of each amplicon to the total mixture. Peak height ratios were calculated with the Excel macro. Following regression analysis, the coefficient of determination used to assess the relationship between allelic imbalance and peak height ratios measurement using our method was $R^2 = 0.974$.

For the second marker SNP rs2236141, regression analysis provided an $R^2 = 0.973$ (data not shown).

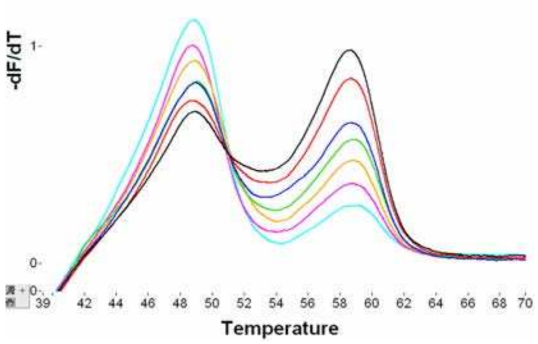


Figure I.3: Mixing experiment to assess the efficiency of HRM analysis for detection of DEA with a SimpleProbe using the HR-1™ instrument. Bi-allelic templates were artificially created with genomic DNA of opposite homozygous for SNP rs2236142. The determination coefficient (R^2) between the expected and the observed allelic ratios was 0.974 in this experiment.

Preliminary DAE results

The preliminary study of *CHEK2* was performed on the largest set of available samples, *i.e.* the heterozygotes for SNP rs2236142.

Out of the 32 available LCLs, four met our statistical criteria for presence of DAE. Figure I.4- A and B show respectively the genomic DNA and the cDNA melting profiles of one of these samples. Following NMD inhibitory treatment, we observed that the cDNA profile changed towards a genomic DNA profile, suggesting that the NMD pathway plays a role in the DAE observed in that sample (Figure I.4-C). We found the same pattern in all four samples with statistically significant DAE.

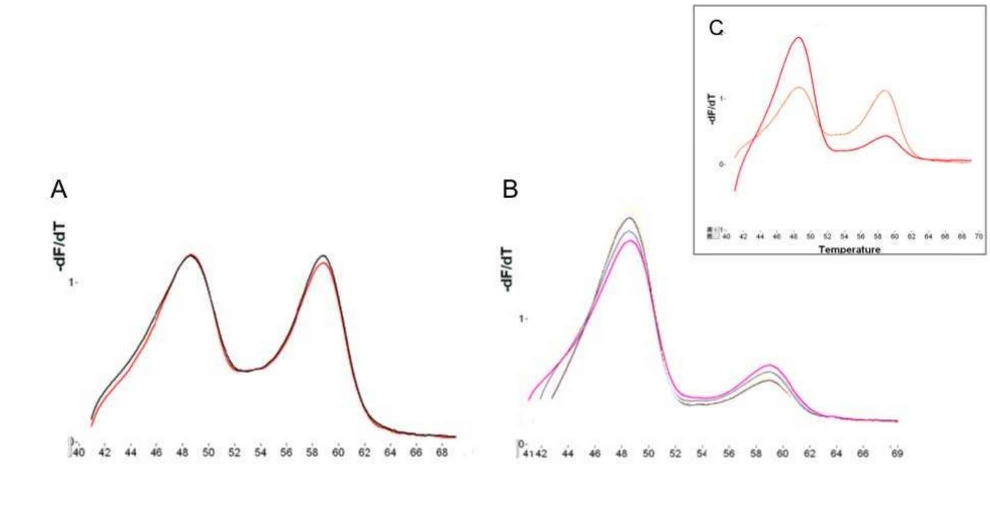


Figure I.4: Evidence for DAE of the *CHEK2* gene in an heterozygous sample for SNP rs2236142. Melting curves of four replicates of genomic DNA (A) and four replicates of cDNA (B) suggest the existence of DAE in the assessed individual. Insert (C) shows the melting profile of one replicate of cDNA (in red) and one replicate of puro-cDNA (orange). The differences in peak height ratios between the two support a role for NMD in the observed DAE. Melting curves were obtained from the HR-1$^{\text{TM}}$ instrument.

From the mutation screening project data, we were able to determine that the four LCLs evincing statistically significant DAE were heterozygous carriers of the truncating mutation CHEK2*1100delC. The CHEK2*1100delC deletion, falling in the kinase domain of the protein, has been widely studied for its contribution to inherited breast cancer susceptibility [Oldenburg et al., 2003]. This mutation induces a premature stop codon in exon 10, and causes the truncation of the protein at codon 381 thus abrogating its kinase activity. Our results support previous findings that the NMD pathway specifically targets mRNA bearing the 1100delC mutation, thus inducing alterations in the expression of these transcripts.

I.3 Upscaling the method

I.3.1 A higher throughput with the LightScanner® instrument

Our preliminary work showed that HRM analysis with the HR-1™ instrument was an accurate approach to assess DAE. However, our protocol requires that DNA and cDNA analyses be performed in four PCR_1 replicates, for each assessed LCL. To complete the study *CHEK2* and *TP53*, as well as for *ATM*, for which we were expecting to gather at least 50 LCLs, throughput was becoming an important issue. Consequently, despite its very high accuracy, the single capillary throughput format of the HR-1™ instrument was an important limitation for our analysis. It was time-consuming to read and to process the data.

Some months after initiation of the DAE project, our laboratory acquired a 384-LightScanner®. Since the LightScanner® was reported to produce almost as good scanning specificity as the HR-1™ [Herrmann et al., 2006, Herrmann et al., 2007], we aimed to adapt our protocol to this instrument. The different sources of template DNA (genomic DNA, cDNA and puro-cDNA) and their replicates would thus be assayed in parallel, which would allow for a more consistent comparison of their melting profiles.

To verify the ability of the LightScanner® to detect small allelic variations, we performed a mixing experiment as we did with the HR-1™. We confirmed that the LightScanner® had high enough resolution to detect small variation in allelic balance, using both SimpleProbes and fluorescein probes (Figure I.5).

In our experiment, each HRM run was performed in a 96-well format. Each plate analyzed a batch of 8 LCLs. Four primary PCR replicates were performed for each of theses LCLs, with genomic DNA, puro-cDNA and cDNA. Thus, each row of the plate corresponded to one assayed LCL and proper organization of the samples on the plate was very important to respect, in order to be able to use the analysis tool described hereafter.

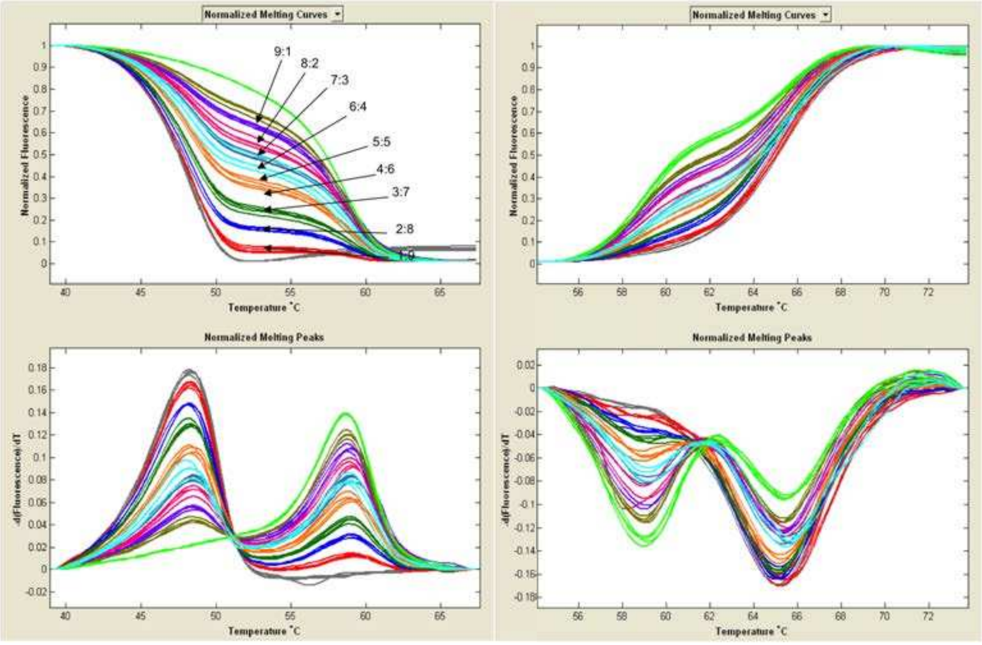


Figure I.5: Mixing experiment to assess the efficiency of HRM analysis for detection of DEA with a SimpleProbe and a fluorescein probe using the LightScanner® instrument. Bi-allelic templates were artificially created with genomic DNA of opposite homozygous for SNP rs2236142. The determination coefficient (R^2) between the expected and the observed allelic ratios were 0.963 in this experiment.

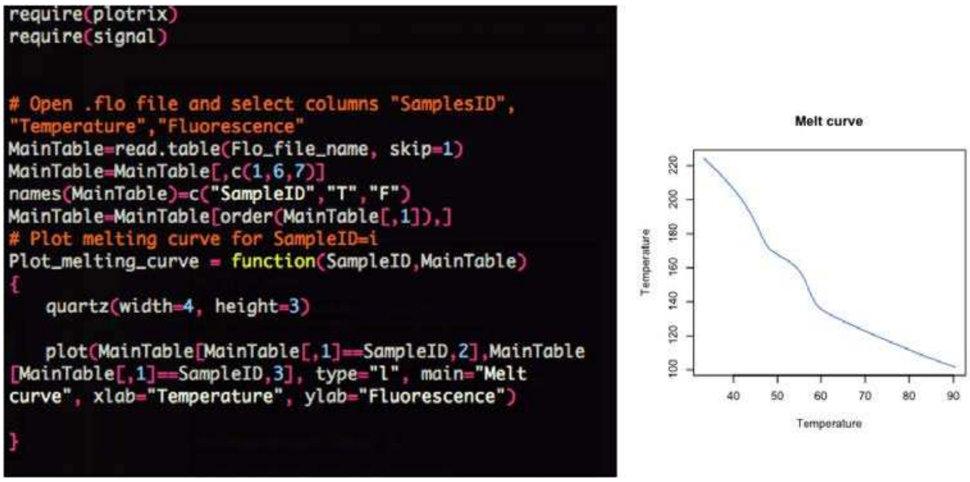
I.3.2 Automation of the measurements with an R-script

To improve the analysis of allelic expression, an analysis tool was developed using R statistical computing software in order to process data acquired with HRM commercialized software. The script provides allelic imbalance estimates and subsequent statistical calculations that are required to assess DAE.

We observed that the LightScanner® software was not able to detect peak heights when these were too small. We decided to write our own software package to perform DAE analysis. We aimed to create a tool able to use the output format of the LightScanner® instrument, plot the melt curves and their derivatives, then calculate the peak heights, deduce the allelic ratios, and lastly, perform statistical analyses on the observed allelic imbalance.

R provides a wide variety of mathematical, statistical and graphical tools, which were flexible enough to address theses specific needs. The script is available from http://sourceforge.net/projects/hrmdae. Installation instructions for R are available from http://www.R-project.org.

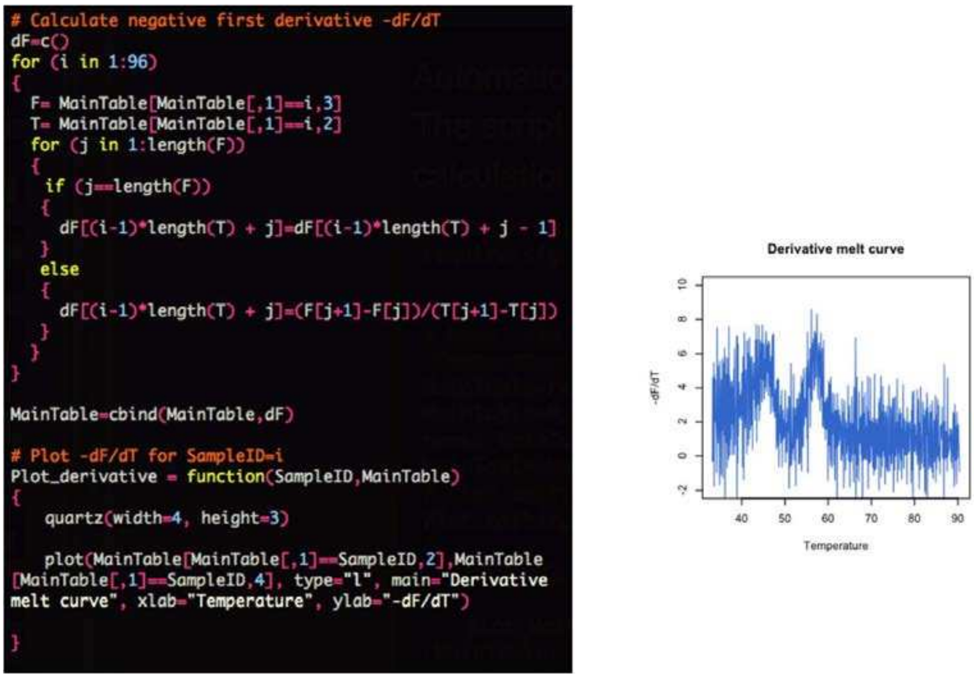
Following the HRM run, a variety of data files are generated by the LightScanner® software, including a .flo file that provides the estimates of the fluorescence for each temperature point, for the whole set of samples. Prior to executing the R-script, the user must export this .flo file. The user must specify a number of parameters, such as the sample ID, the approximate Tms observed during the HRM run and the cut-offs to be used in the statistical calculations. Once the appropriate information has been entered, the user can execute the script to launch the DAE analysis.



```
require(plotrix)
require(signal)


# Open .flo file and select columns "SamplesID",
"Temperature","Fluorescence"
MainTable=read.table(Flo_file_name, skip=1)
MainTable=MainTable[,c(1,6,7)]
names(MainTable)=c("SampleID","T","F")
MainTable=MainTable[order(MainTable[,1]),]
# Plot melting curve for SampleID=i
Plot_melting_curve = function(SampleID,MainTable)
{
   quartz(width=4, height=3)

   plot(MainTable[MainTable[,1]==SampleID,2],MainTable
[MainTable[,1]==SampleID,3], type="l", main="Melt
curve", xlab="Temperature", ylab="Fluorescence")

}
```

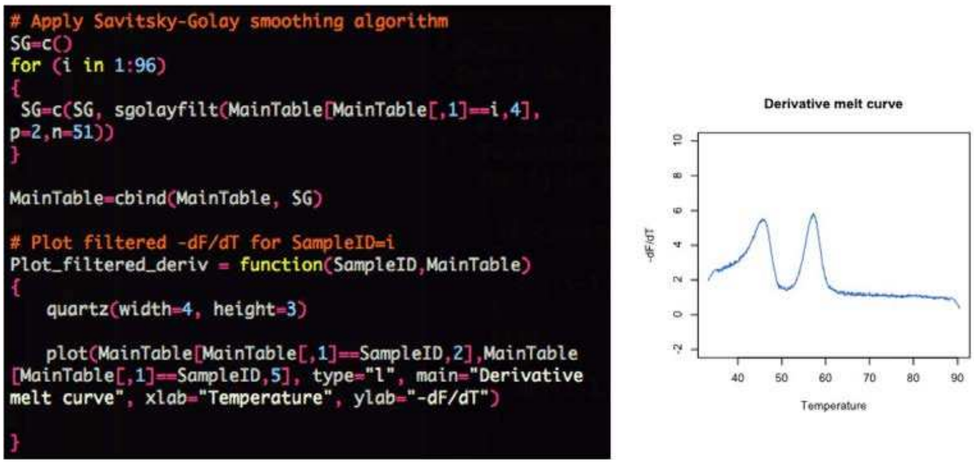
Figure I.6: R-script: plot of the melting curve.

For each of the 96 samples, the R-script can plot the melting curve (Figure I.6) and calculate the derivative curve (Figure I.7) . The resolution of the LightScanner® is so high that the raw signal seems messy. We used the Savitsky-Golay algorithm to smooth the signal and to obtain a neat smoothed curve(Figure I.8). In order to subtract noise due to the amplification, the curve is then normalized to a baseline provided by a no-template control (Figures I.9 and I.10). The R-script measures the peak heights for the 96 samples (Figure I.11), calculates the peak height ratios,



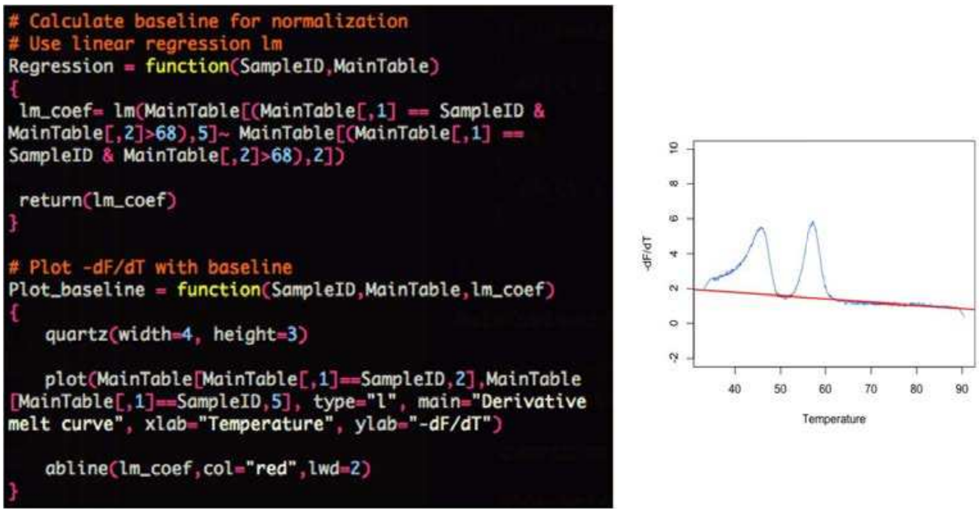
```
# Calculate negative first derivative -dF/dT
dF=c()
for (i in 1:96)
{
  F= MainTable[MainTable[,1]==i,3]
  T= MainTable[MainTable[,1]==i,2]
  for (j in 1:length(F))
  {
   if (j==length(F))
   {
    dF[(i-1)*length(T) + j]=dF[(i-1)*length(T) + j - 1]
   }
   else
   {
    dF[(i-1)*length(T) + j]=(F[j+1]-F[j])/(T[j+1]-T[j])
   }
  }
}

MainTable=cbind(MainTable,dF)

# Plot -dF/dT for SampleID=i
Plot_derivative = function(SampleID,MainTable)
{
   quartz(width=4, height=3)

   plot(MainTable[MainTable[,1]==SampleID,2],MainTable
[MainTable[,1]==SampleID,4], type="l", main="Derivative
melt curve", xlab="Temperature", ylab="-dF/dT")

}
```

Figure I.7: R-script: plot of the derivative melting curve.



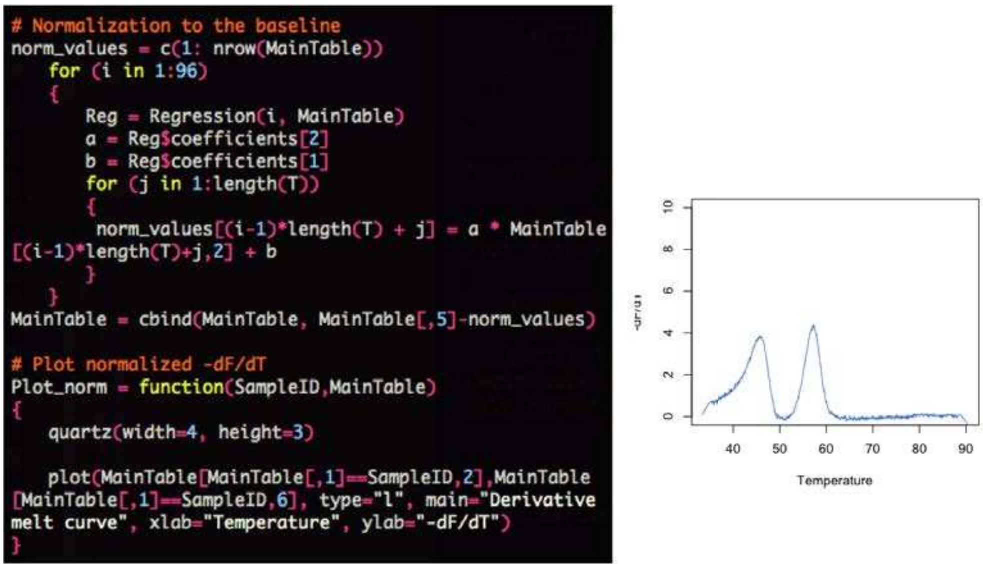
```
# Apply Savitsky-Golay smoothing algorithm
SG=c()
for (i in 1:96)
{
 SG=c(SG, sgolayfilt(MainTable[MainTable[,1]==i,4],
p=2,n=51))
}

MainTable=cbind(MainTable, SG)

# Plot filtered -dF/dT for SampleID=i
Plot_filtered_deriv = function(SampleID,MainTable)
{
   quartz(width=4, height=3)

   plot(MainTable[MainTable[,1]==SampleID,2],MainTable
[MainTable[,1]==SampleID,5], type="l", main="Derivative
melt curve", xlab="Temperature", ylab="-dF/dT")

}
```

Figure I.8: R-script: smoothing of the derivative melting curve using Savitsky-Golay algorithm.



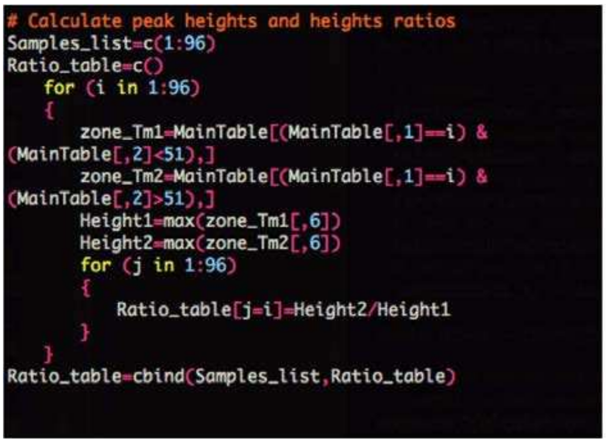
```
# Calculate baseline for normalization
# Use linear regression lm
Regression = function(SampleID,MainTable)
{
 lm_coef= lm(MainTable[(MainTable[,1] == SampleID &
MainTable[,2]>68),5]~ MainTable[(MainTable[,1] ==
SampleID & MainTable[,2]>68),2])

 return(lm_coef)
}

# Plot -dF/dT with baseline
Plot_baseline = function(SampleID,MainTable,lm_coef)
{
    quartz(width=4, height=3)

    plot(MainTable[MainTable[,1]==SampleID,2],MainTable
[MainTable[,1]==SampleID,5], type="l", main="Derivative
melt curve", xlab="Temperature", ylab="-dF/dT")

    abline(lm_coef,col="red",lwd=2)
}
```

Figure I.9: R-script: calculation of the baseline.



```
# Normalization to the baseline
norm_values = c(1: nrow(MainTable))
    for (i in 1:96)
    {
        Reg = Regression(i, MainTable)
        a = Reg$coefficients[2]
        b = Reg$coefficients[1]
        for (j in 1:length(T))
        {
         norm_values[(i-1)*length(T) + j] = a * MainTable
[(i-1)*length(T)+j,2] + b
        }
    }
MainTable = cbind(MainTable, MainTable[,5]-norm_values)

# Plot normalized -dF/dT
Plot_norm = function(SampleID,MainTable)
{
    quartz(width=4, height=3)

    plot(MainTable[MainTable[,1]==SampleID,2],MainTable
[MainTable[,1]==SampleID,6], type="l", main="Derivative
melt curve", xlab="Temperature", ylab="-dF/dT")
}
```

Figure I.10: R-script: normalization to the baseline.



```
# Calculate peak heights and heights ratios
Samples_list=c(1:96)
Ratio_table=c()
    for (i in 1:96)
    {
        zone_Tm1=MainTable[(MainTable[,1]==i) &
(MainTable[,2]<51),]
        zone_Tm2=MainTable[(MainTable[,1]==i) &
(MainTable[,2]>51),]
        Height1=max(zone_Tm1[,6])
        Height2=max(zone_Tm2[,6])
        for (j in 1:96)
        {
            Ratio_table[j=i]=Height2/Height1
        }
    }
Ratio_table=cbind(Samples_list,Ratio_table)
```

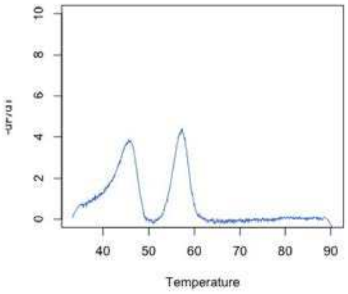
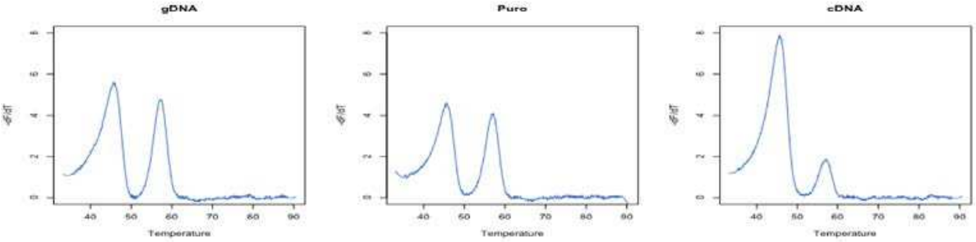
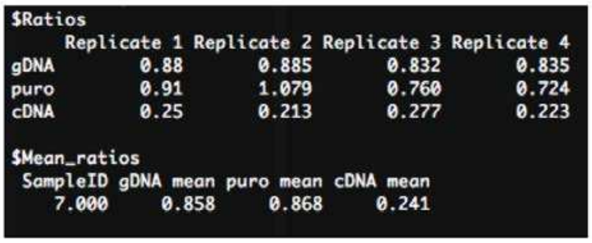


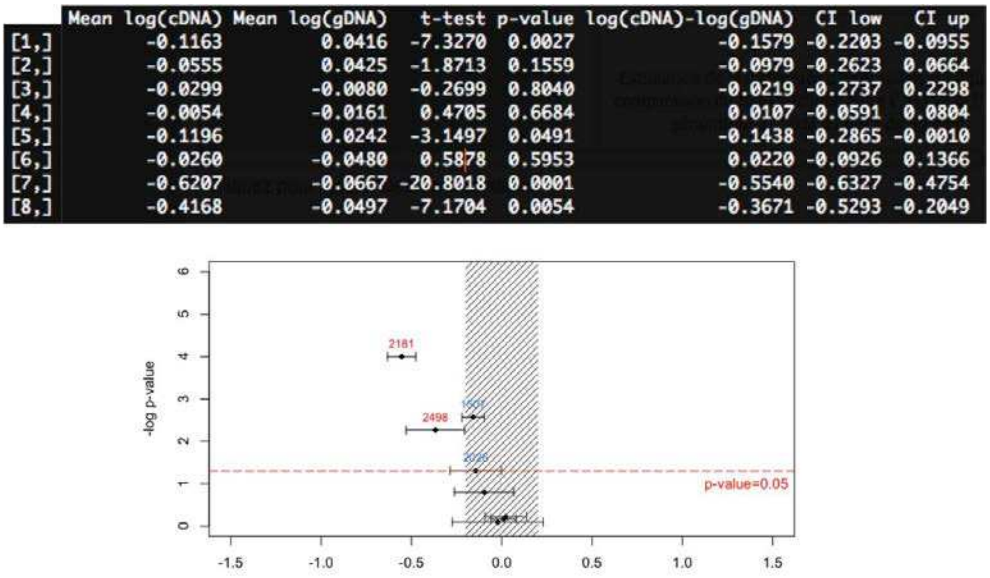
Figure I.11: R-script: measurement of the peak heights.





```
$Ratios
      Replicate 1 Replicate 2 Replicate 3 Replicate 4
gDNA         0.88       0.885       0.832       0.835
puro         0.91       1.079       0.760       0.724
cDNA         0.25       0.213       0.277       0.223

$Mean_ratios
 SampleID gDNA mean puro mean cDNA mean
    7.000     0.858     0.868     0.241
```

Figure I.12: R-script: calculation of the peak height ratios. Example of the 7th sample from the samples series.



```
     Mean log(cDNA) Mean log(gDNA)   t-test p-value log(cDNA)-log(gDNA)  CI low   CI up
[1,]        -0.1163         0.0416  -7.3270  0.0027             -0.1579 -0.2203 -0.0955
[2,]        -0.0555         0.0425  -1.8713  0.1559             -0.0979 -0.2623  0.0664
[3,]        -0.0299        -0.0080  -0.2699  0.8040             -0.0219 -0.2737  0.2298
[4,]        -0.0054        -0.0161   0.4705  0.6684              0.0107 -0.0591  0.0804
[5,]        -0.1196         0.0242  -3.1497  0.0491             -0.1438 -0.2865 -0.0010
[6,]        -0.0260        -0.0480   0.5878  0.5953              0.0220 -0.0926  0.1366
[7,]        -0.6207        -0.0667 -20.8018  0.0001             -0.5540 -0.6327 -0.4754
[8,]        -0.4168        -0.0497  -7.1704  0.0054             -0.3671 -0.5293 -0.2049
```

Figure I.13: R-script: statistical test and result plot of 8 individuals from the samples series. The script adds the ID of the samples showing interesting results in red.

groups the genomic DNA, cDNA and puromycin-cDNA replicates and averages them for each of the 8 individuals on the plate (Figure I.12). A statistical analysis is performed using a Student's *t*-test to compare the genomic DNA to the cDNAs. Eventually, R returns a summary of the measurements and a plot, where the x-axis represents the level of allelic imbalance, and the y-axis represents the p-value for the Student's *t*-test. The hatched area and the red line represent a level of DAE $\leq$ 20% and the threshold for a p-value $=$ 0.05, respectively (Figure I.13). Results of DAE analyses can be saved as image files or excel files for further reporting.

Optional plots and tables are also available for further checks. For instance, the user can plot the melting curve of each replicate sample or the averaged curve for the replicates for each individual. As can be seen on Figure I.13, in this series of 8 individuals, two are clear outliers, suggesting they both carry statistically significant DAE. Thus we aimed to look at the individual plots of each outlier. We were able to verify visually that the melting profile of the genomic DNA of the individual was normal. We were also able to observe that the DAE is visible only in the cDNA melting curve and not in the puromycin-treated cDNA.

Our package allows users to set options to perform both DAE assessment with a fluorescein-labeled probe or with a SimpleProbe. Preliminary results obtained with the HR-1™ instrument and the Excel macro have been counterchecked and confirmed by writing another script specific to the data generated by HR-1™. Table I.1 shows a comparison of the duration of the DAE analysis between the HR-1™ and the LightScanner® instruments, for 96 samples, using their respective R scripts.

DAE step	HR-1™ instrument	LightScanner® instrument
PCRs	Same duration	Same duration
Data acquisition	2 days	12 minutes
Data analysis	1 full day	15 minutes

Table I.1: Comparison of the duration of the DAE analysis between the HR-1™ and the LightScanner® instruments, for 96 samples.

In conclusion, we found that analysis on the LightScanner® instrument combined with the R script is of greater practical efficiency, and yields accuracy comparable to the HR-1™. Following successful accomplishment of Aim 2, i.e. the implementation of an appropriate assay for DAE assessment, the chapters below will present the results from the study of the *CHEK2*, *ATM* and *TP53* genes.

II

DAE assessment of the *CHEK2* gene

The purpose of the following paper was to present our novel approach based on high-throughput HRM analysis and the R-script we have developed. We describe the assessment of the breast cancer susceptibility gene *CHEK2*, using HRM analysis of two SimpleProbes, designed for marker SNPs rs2236141 and rs2236142. We were able to test a total of 41 LCLs. We observed statistically significant DAE in 4 LCLs. The fact that our sample set included LCLs from patients carrying the 1100delC mutation, known to induce NMD, provided us a positive control to validate our assay.

Article III

Detecting differential allelic expression using high-resolution melting curve analysis: assessment of the breast cancer susceptibility gene *CHEK2*

Nguyen-Dumont T, Jordheim LP, Michelon J, McKay-Chopin S, Forey N,
kConFab, Sinilnikova O, Le Calvez-Kelm F, Southey MC,
Tavtigian SV, Lesueur F.

(accepted)

Detecting differential allelic expression using high-resolution melting curve analysis: application to the breast cancer susceptibility gene CHEK2

Tú Nguyen-Dumont[1], Lars P. Jordheim[2] , Jocelyne Michelon[1] , Nathalie Forey[1] , Sandrine McKay-Chopin[1] , kConFab[3] , Olga Sinilnikova[4] , Florence Le Calvez-Kelm[1] , Melissa C. Southey[5] , Sean V. Tavtigian[6] and Fabienne Lesueur*[1]

[1]Genetic Cancer Susceptibility Group, IARC, 69372 Lyon, France
[2]INSERM U590, Université Lyon 1, Lyon, France
[3]Peter MacCallum Cancer Center, East Melbourne 2, VIC 3002, Australia
[4]Unité Mixte de Génétique Constitutionnelle des Cancers Fréquents, Hospices Civils de Lyon, Centre Léon Bérard, 69373 Lyon, France
[5]Department of Pathology, The University of Melbourne, VIC 3010, Australia
[6]Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT 84112, USA

Email: Tú Nguyen-Dumont - nguyent@students.iarc.fr; Lars P. Jordheim - lars-petter.jordheim@univ-lyon1.fr; Jocelyne Michelon - michelon@iarc.fr; Nathalie Forey - forey@iarc.fr; Sandrine McKay-Chopin - chopin@iarc.fr; kConFab - heather.thorne@petermac.org; Olga Sinilnikova - sinilnik@lyon.fnclcc.fr; Florence Le Calvez-Kelm - lecalvez@iarc.fr; Melissa C. Southey - msouthey@unimelb.edu.au; Sean V. Tavtigian - sean.tavtigian@hci.utah.edu; Fabienne Lesueur*- lesueurf@iarc.fr;

*Corresponding author

Abstract

Background: The gene *CHEK2* encodes a checkpoint kinase playing a key role in the DNA damage pathway. Though *CHEK2* has been identified as an intermediate breast cancer susceptibility gene, only a small proportion of high-risk families have been explained by genetic variants located in its coding region. Alteration in gene expression regulation provides a potential mechanism for generating disease susceptibility. The detection of differential allelic expression (DAE) represents a sensitive assay to direct the search for a functional sequence variant within the transcriptional regulatory elements of a candidate gene. We aimed to assess whether *CHEK2* was subject to DAE in lymphoblastoid cell lines (LCLs) from high-risk breast cancer patients for whom no mutation in BRCA1 or BRCA2 had been identified.

Methods: We implemented an assay based on high-resolution melting (HRM) curve analysis and developed an analysis tool for DAE assessment.

Results: We observed allelic expression imbalance in 4 of the 41 LCLs examined. All four were carriers of the truncating mutation 1100delC. We confirmed previous findings that this mutation induces non-sense mediated mRNA decay. In our series, we ruled out the possibility of a functional sequence variant located in the promoter region or in a regulatory element of *CHEK2* that would lead to DAE in the transcriptional regulatory milieu of freely proliferating LCLs.

Conclusions: Our results support that HRM is a sensitive and accurate method for DAE assessment. This approach would be of great interest for high-throughput mutation screening projects aiming to identify genes carrying functional regulatory polymorphisms.

Background

The *CHEK2* gene (cell cycle checkpoint kinase 2) is a multiorgan tumour susceptibility gene involved in the maintenance of genomic stability. CHEK2 functions downstream of ATM to phosphorylate several substrates, including p53, Cdc25C and BRCA1, leading to cell cycle arrest, activation of DNA repair or apoptosis in response to DNA double-stranded breaks. Since *CHEK2* plays a key role in the DNA damage pathway, loss of function of the protein may allow cells to evade normal cell cycle checkpoints, ultimately leading to tumour initiation or progression. The CHEK2*1100delC deletion, falling in the kinase domain of the protein, has been widely studied for its contribution to inherited breast cancer susceptibility [1]. This mutation induces a premature stop codon in exon 10, and causes the truncation of the protein at codon 381 thus abrogating its kinase activity. The frequency of CHEK2*1100delC differs between ethnic populations, and is higher in the North of Europe and low or absent in other countries [2]. The CHEK2-Breast Cancer Consortium reported a frequency of 5.1% for the CHEK2*1100delC variant in familial breast cancer cases who tested negative for *BRCA1* and *BRCA2* mutations, as opposed to 1.1% of carriers in the control population [3]. This intermediate-risk breast cancer susceptibility allele almost triples the risk of developing the disease in unselected breast cancer cases (OR= 2.34; 95% CI[1.72 - 3.20]) [4]. Other founder mutations in *CHEK2* have been associated with an increased risk of cancer [5]. Though first discovered in breast cancer patients, *CHEK2* mutations have since been reported to predispose to a range of cancer types, including ovarian, prostate, kidney and colorectal cancers [6], supporting the hypothesis that *CHEK2* is a multiorgan cancer susceptibility gene [5].

As part of an international breast cancer genetics study aiming to investigate candidate genes conferring an intermediate-risk of breast cancer, we mutation screened the coding exons and the adjacent proximal introns of *CHEK2* in 1415 cases and 1204 controls. The main goal of this study was to evaluate and to compare the role of truncating mutations, splice junction mutations and rare missense substitutions in breast cancer susceptibility (Le Calvez-Kelm et al., manuscript submitted). In order to fully assess the contribution of *CHEK2* in breast cancer susceptibility, we aimed to test whether the gene was subject to differential allelic expression (DAE). In such a case, it would be worth extending variant discovery efforts from the coding sequence of the gene to known or predicted regulatory regions to search for causal variants. Indeed, phenotypic variation may be influenced by sequence variations in genes by alterations in the quality or in the quantity of the encoded proteins [7]. These changes are transmitted from the gene to the protein in the guise of modifications of the sequence or the abundance of mRNA. From this perspective, it can be hypothesized that gene expression regulation may be the underlying explanation for a proportion of cancer that have not been resolved yet by mutation screening of coding region in currently known cancer predisposition genes.

Allelic imbalance was first described in parental imprinting and X-chromosome inactivation but it is becoming clear that *cis*-acting variations in gene expression occur commonly in the human genome, playing a key-role in human phenotypic variability [8–10]. Characterization of the effect of *cis*-acting polymorphisms in regulatory regions is a great challenge due to the difficulty to locate these regions. In addition, regulatory variants are not robustly detected by sequence analysis since SNP identification by screening regulatory regions does not consistently allow prediction of the effect of observed SNPs on gene expression. Thus, knowledge of the effect of genetic variants affecting mRNA transcription is very limited. One possible approach to address this issue is the examination of disruption/alteration of gene expression level. The most sensitive test for this phenomenon is to carry a careful survey of whether two alleles of a gene are equally expressed. This approach has been used in studies aiming at identifying functional cis-variants that can have a role in susceptibility to breast [11, 12] and colorectal cancer [13, 14] . In some cases, observation of DAE will be explained by a truncating mutation resulting in non-sense mediated mRNA decay (NMD) or by a splice junction mutation resulting in an unstable transcript. However, DAE can also be the signature of a heterozygous carriage of a regulatory variant [15] or of an epigenetic event (methylation) [16].

In this study, we used a high-resolution melting (HRM) analysis approach to perform allele-specific expression measurement in *CHEK2*. As in currently used methods for investigating DAE, this approach is applied to individual subjects who are heterozygous for an exonic marker SNP, specifically targeted by a labelled probe, called SimpleProbe [17, 18]. Data acquisition on HRM instruments consists of monitoring changes in the fluorescence intensity of the probe, as it dissociates from the two allelic templates, while the probe-target duplexes are continuously heated. We have already reported the use of this methodology to compare the relative abundance of allelic transcripts in a study investigating mRNA degradation due to NMD in *BRCA2* [18], and in a group of selected genes involved in the cellular response to the cytotoxic agent cytarabine [19]. In these studies, DAE analysis was limited by the single-capillary throughput of the HRM device used, the HR-1$^{\text{TM}}$ instrument, and allelic imbalance was quantified manually. Here, we report additional experiments and testing, as well as up-scaling possibilities with a high-throughput HRM device, the LightScanner® instrument that uses a 384-well plate format. To improve the analysis of allelic expression, an analysis tool was developed using R in order to process data acquired with HRM commercialized software. Our script provides allelic imbalance estimates and subsequent statistical calculations that are required to assess DAE.

Methods

Lymphoblastoid cell lines

We used a total of 89 lymphoblastoid cell lines (LCLs) derived from breast cancer patients, who were considered to be at high risk of carrying a genetic predisposition to cancer due to an early age at onset and/or family history, and for whom no mutation in *BRCA1* or *BRCA2* genes had been identified. Biological samples were obtained from Creighton University School of Medicine (Omaha, NE, USA, 33 familial cases), Centre Léon Bérard (CLB, Lyon, France, 21 patients diagnosed below age 50) and the Kathleen Cuningham Consortium for Research into Familial Breast Cancer (kConFab, Melbourne, Australia, 35 familial cases). LCLs were established by Epstein-Barr virus immortalization of patients' blood lymphocytes. Cells were maintained in RPMI 1640 medium (Invitrogen, Cergy-Pontoise, France) supplemented with 20% fetal calf serum (VWR, Fontenay-sous-bois, France), 0.4% fungizon (Qiagen, Courtaboeuf, France) and 1% penicilin-streptomycin (Invitrogen), in 5% CO_2 incubator at 37°C with 95% humidity. For NMD inhibition, LCLs were treated for 6 hours with 100 μM puromycin (Sigma Aldrich, St Quentin Fallavier, France).

DNA samples

Genomic DNAs and total RNAs were extracted from LCLs using Puregene DNA isolation kit (Qiagen) and NucleoSpin RNA II kit (Machery Nagel, Hoerdt, France), respectively. Integrity of RNA was controlled using the BioAnalyzer and RNA NanoChip II kit (Agilent, Massy, France) according to the manufacturer's instructions. RNAs harbouring an RNA integrity number (RIN) ≥ 8 were selected for further analysis [20]. Whenever the quality threshold was not reached, a new RNA extraction was performed so that all the RNAs used in this study had a minimum RIN of 8. Complementary DNA (cDNA) synthesis was performed from 1 μg total RNA using SuperScript^{TM} III First Strand Synthesis System for RT-PCR (Invitrogen) with oligo(dT) primers, according to the manufacturer's instructions.

Mutation screening

The 89 subjects included in this study were drawn from a large-scale case-control mutation screening study involving 1415 cases and 1204 controls, that has been described elsewhere [21,22]. CHEK2*1100delC carriers were all confirmed by direct sequencing on genomic DNA (For mutation screening results, see Additional file 1).

PCR amplification for DAE assessment

DAE was assessed in four replicates of primary PCR (PCR1), both with cDNA, cDNA from puromycin-treated LCL, and genomic DNA (For primers and probes, see Additional file 2). PCR1 contained 2 μl template DNA in 1X PCR Buffer, 1.5 mM MgCl2, 0.13 mM dNTP, 0.2 μM forward and reverse primers specific to genomic DNA or cDNA, and 0.05 Units Platinum Taq Polymerase (Invitrogen), in a final volume of 8 μl. The temperature cycling protocol was: 94°C for 3 minutes; 30 cycles at 94°C for 30 seconds, 62°C for 45 seconds and 72°C for 30 seconds; and finally 72°C for 5 minutes. To reduce competitive binding of the probe and the complementary strand during the melting curve analysis, the secondary PCR (PCR2) was carried out asymmetrically, with the primer generating the target strand at a 5-fold higher concentration (0.5 μM) than the primer for the other strand (0.1 μM). In addition, PCR2 contained 2 μl of 1:15 diluted in TE^{-4} PCR1 products combined with 0.9X Buffer, 1.38 mM MgCl2, 0.12 mM each dNTP, 0.5 μM SimpleProbe (Tib Molbiol, Berlin, Germany) and 0.4 Units of Taq Platinum Polymerase in a final volume of 6 μl. Clear oil (Avatech) saturated with Tween 80 (Sigma Aldrich) was used to overlay PCR reactions. The temperature cycling protocol was the same as above, except that 45 cycles were performed. DAE analyses were performed in batches of 96 samples, corresponding to 4 replicates of genomic DNA, 4 replicates of cDNA and 4 replicates of puromycin-treated cDNA derived from 8 different LCLs.

High-resolution melting analysis

PCR product melting curves were obtained from the HR-1$^{\mathrm{TM}}$ and the LightScanner® instruments by melting from 35 °C to 75 °C. Data were obtained with the supplied software (HR-1$^{\mathrm{TM}}$ v1.5 and LightScanner® Software v2.0, respectively), and then exported to an analysis tool that we developed in R, a programming language and software environment for statistical computing and graphics (http://cran.r-project.org). R scripts were developed in order to retrieve the data, to apply the Savitsky-Golay filter to smooth the derivative melting curves and to calculate the peak heights. For each sample, ratios were measured from 4 PCR-replicates and the mean ratio was calculated across all replicate samples. The R scripts are available on http://sourceforge.net/projects/hrmdae. The level of allelic imbalance for each individual was determined from the difference between the log of the signal ratio in cDNA and the corresponding log ratio in genomic DNA. Statistical significance for the allelic imbalance was calculated using Student's t-test. Criteria for DAE were the following: i) the point estimate of the difference between genomic DNA and cDNA ratios should be greater than 20%; ii) the Student's t-test p-value should be ≤ 0.05, and iii) the 95% confidence interval of the point estimate should not include 0 [13,18,19,23].

Results

Genotyping of CHEK2 exonic SNPs

The main goal of the initial case-control mutation-screening project was to identify rare, potentially pathogenic genetic variants within the coding sequence and the proximal intronic splice consensus sequences of *CHEK2*. This mutation screening simultaneously provided the genotype of all common coding SNPs for every subject enrolled in the molecular epidemiology study. For 89 of the breast cancer patients investigated, LCLs were available to conduct DAE analysis.

In order to make a differential measurement of the level of expression of the two alleles of a gene for a given patient, one must be able to distinguish between the alleles. We used the two most common exonic SNPs that were identified during the mutation screening process, namely rs2236142 and rs2236141, and only the cell lines that are heterozygote for at least one of the two SNPs were selected for further analysis. These two markers are reported to be common in the dbSNP database (Minor allele frequency of 49.2% and

25.4% in European populations, respectively). Thirty-two out of 89 cell lines were heterozygotes for rs2236142 and 17 out of 89 were heterozygotes for rs2236141 (AdditionalTable 1). Eight individuals were double heterozygotes.

Evaluation of the HRM method to detect DAE

This technique relies on the distinction between the two alleles in heterozygous individuals using differences in melting temperature (Tm) with a derivative fluorescent signal correlated to the relative abundance of each transcript. We first verified that HRM could distinguish between the two alleles of each SNP in our experimental conditions, by assaying genomic DNA and cDNA from all three genotypes. Analysis of the melting curves of the homozygous samples showed a transition at a Tm specific to each allele (Figure 1). Melting transitions were converted into peaks on the derivative plot. Heterozygous samples presented transitions and peaks corresponding to each allele at both Tm. A no-template control was taken as baseline to subtract local background value to the fluorescence intensity of the samples.

To examine the feasibility of detecting DAE by melting curve analysis, we created a range of melting curves associated with known allelic imbalance. Using homozygous genomic DNAs, we produced bi-allelic templates with increasing minor allele:major allele proportions (9:1, 8:2, 7:3, 6:4, 5:5, 4:6, 3:7, 2:8, 1:9). Allelic (im)balance was observed as the ratio of the peak heights of the fluorescence signal. As expected, the melting profiles of the mixtures of opposite homozygotes reflected the relative contribution of each allele to the total mixture (Figure 2A). For both SNPs we observed good correlations between allelic imbalance and peak height ratio measurements. For rs2236142, R^2=0.974 on HR-1^{TM} and R^2=0.963 on LightScanner® (Figure 2B); for rs2236141, R^2=0.973 on HR-1^{TM} and R^2=0.963 on LightScanner® (data not shown). The mixing experiments showed that the measured allelic ratios varied in a linear relationship with the dilution ratios. Altogether, these results show that HRM is able to accurately detect different extents of DAE.

Assessment of DAE for CHEK2 in LCLs from breast cancer patients

Mutation screening of the 89 LCLs identified four carriers of the CHEK2*1100delC mutation (see Additional file 1). This mutation induces a premature termination codon and has been reported to trigger the NMD pathway, which leads to the specific degradation of mRNAs bearing such deleterious mutation [24]. In order to distinguish DAE that would be caused by NMD from DAE that would be caused by a regulatory variant altering the level of expression of the transcript, cDNA was derived from LCLs treated and untreated with puromycin, from each individual.

We performed quantitative measurements on genomic DNA and on both types of cDNA. Genomic DNA served as an internal control and provided the expected peak heights ratio value for a 1:1 allelic ratio, thereby controlling for any bias in the binding of the fluorescent probe to the two alleles. Because of differences in fluorescence yield, measured peak heights ratios differed from unity when genomic DNAs were assessed. However, the melting profiles of genomic DNA were in accordance with what was expected from the mixing experiment.

The first series of analysis using the SNP rs2236142 as marker included 32 heterozygous individuals for this coding SNP. The statistical threshold for DAE was reached in four individuals (Figure 3). Mutation screening results indicated that these four patients carried the CHEK2*1100delC mutation (see Additional file 1). Observed levels of DAE varied from 37% to 60%, revealing a substantial expression imbalance of an order likely to have biological importance. NMD inhibitory treatment on these four LCLs showed melting curves profiles tending towards the genomic curve profile, which is the reference for a 1:1 allelic ratio (Figure 4A). This confirms previous findings that the CHEK2*1100delC mutation leads to allele-specific degradation by triggering the NMD pathway [3]. None of the 28 other individuals of this first series showed allelic imbalance, according to our statistical criteria (Figure 4-B). The second series of analysis used SNP rs2236141 as marker and included 17 heterozygous individuals for this coding SNP. Eight of them were also heterozygous for SNP rs2236142 and had already tested negative for DAE with the first marker. The statistical threshold for DAE was not reached in any of the remaining 9 samples (Figure 3).

Discussion

Our work supports the high sensitivity of HRM for the detection and quantification of DAE. We have shown that HRM is able to detect DAE associated with NMD in LCLs carrying a non-sense mutation in *CHEK2*. Although no DAE was observed in the patients who do not carry the 1100delC mutation, the series investigated here was limited, and we cannot rule out that *cis*-regulatory variants in *CHEK2* may lead to DAE in a tissue specific manner [23]. However, this later hypothesis could not be tested since no breast tissue was available from these patients.

The approach used in our study relies on subjects who are heterozygous for a coding SNP and allows relative quantification of allelic transcripts. This methodology has major advantages over more conventional methods for investigating DAE based on the comparison of gene expression between individuals as discussed elsewhere [7, 9, 19] . Since they come from the same tissue sample and have therefore been subjected to the same environmental influences (such as genetic *trans*-acting factors and experimental exposures, including mRNA degradation) both alleles should be equally expressed in the absence of *cis*-acting sequence variation or epigenetic effects affecting the expression of the target mRNA. Thus, the strength of this approach is that each allele acts as an internal control for confounding factors, disclosing *cis*-variation effects without being confounded by any *trans*-variation effects.

Here, we report a complete solution for HRM analysis that can be used on both the HR-1$^{\mathrm{TM}}$ (1 single capillary) and LightScanner® (384-well plate format) instruments, with the format depending on the required throughput. Access to DAE assessment technology can be cost prohibitive for many laboratories. HRM provides a good alternative when compared to methodologies based for instance, on the use of capillary electrophoresis for single-base extension assays, such as SnapShot assays [10], allele-specific quantitative real-time PCR [11] and microarray platform [8]. Advantages offered by HRM analysis include its rapidity, cost-effectiveness and security due to its closed-tube nature. Though the HR-1$^{\mathrm{TM}}$ is reported to provide a better accuracy [25], both instruments performed well to identify the 4 carriers of the CHEK2*1100delC variant showing DAE in the absence of puromycin treatment in our study. However, given the number of samples to test, analysis with the HR-1$^{\mathrm{TM}}$ instrument ended up being much more time consuming (Table 1). The results obtained with the LightScanner® instrument showed that this methodology can be applied in larger-scale studies, provided that LCL material is available, while maintaining high accuracy and remaining cost-effective. Indeed, the protocol is relatively inexpensive since it only requires standard PCR reagents and a small amount of fluorescent probe.

The script we developed using R computing software was made compatible with both instruments and greatly reduces the time of analysis. Once HRM data are acquired, the normalization of the curves, peak heights measurements, ratios calculations and statistical analysis are performed automatically within less than 15 minutes for a set of 96 samples when using the LightScanner® instrument. The output consists in a summary table of the peak heights, relative allelic ratios, and the Student's t-test values, as well as a plot on which DAE carriers are highlighted. The script can also display other information on demand, such as melting curve profiles which can be displayed for each replicate or by average of 4 replicates for each individual (see examples in Figure 3 and 4).

In DAE analysis by HRM, the peak heights obtained from the melting curve reflect the relative abundance of each allele's transcript. The reproducibility and precision of the assay are reasonable as seen in the small standard deviations associated with the calculations. The accuracy of the method was illustrated by the consistency of the allelic expression estimates across multiple replicates assay within the same individual sample. Genomic DNA ratios varied within a very narrow range, showing the excellent reproducibility and precision of the assay on DNA derived from LCL. The intra-sample variation in replicate analysis was higher for mRNA ratios than for DNA ratios, possibly owing to RNA stability. In addition, at low copy numbers of mRNA, the stochastic distribution of the RNA templates may be a major source of variation and hence affect the accuracy of DAE analysis, by generating disagreeing replicate results for instance [26]. In a DAE study, the main optimization issue is the ability to select a subset of 2-3 marker SNPs so that as many individuals as possible are heterozygous for at least one of the markers. Subsets of individuals giving the most heterozygotes at 2 loci should be chosen in order to maximize redundancy, and to self-check for error reduction. Unfortunately, in the present study, no individual heterozygous for both SNPs showed evidence of DAE. Detection of DAE in a candidate gene may be indicative of the presence of a coding or regulatory variant altering expression of the gene product. However, DAE-based approaches can point out the presence of a regulatory causative variant only if the subjects are heterozygous for the causative variant (and of course for the coding SNP serving as marker). In some situations, the coding SNP used to distinguish both alleles may be itself responsible for the observed DAE, or it can be on linkage disequilibrium (LD) with it, *i.e.* on the same haplotype. In the case of no LD between the marker and the dysfunctional variant, it is still possible to map the variant, as previously reported by others [27, 28].

Conclusions

Allele-specific expression assays can be applied to identify genetic variants located in regions essential for gene expression regulation or splicing. Thus, identification of a list of genes for which DAE has been detected would yield a considerable reduction of the amount of work, by focusing discovery effort on the subset of genes that are most likely to harbour coding or regulatory variants that may alter gene expression. The approach reported here allows revealing the existence of regulatory variations without directly identifying or requiring prior knowledge of specific *cis*-regulatory SNPs. DAE assays can also highlight the existence of epigenetic factors controlling gene expression [29].

Analysis of the relative allelic ratios of marker SNPs circumvents the issue of confounding *trans*-acting factors. Any significant differences in these ratios support the existence of DAE and hence, *cis*-acting polymorphisms determining gene expression. The primary goal of this type of study is to identify sequence variants that are likely to alter gene expression and gene product function, and thereby influence susceptibility to breast cancer. However, to demonstrate that some of these variants actually show disease association, large-scale epidemiological studies are required and may ultimately lead towards the identification of causal genetic factors responsible for susceptibility to disease. In the context of such high-throughput studies, instead of LCLs, one can use blood samples, a tissue that is easier to collect than breast tissues. Identification and elucidation of rare intermediate-risk genetic variants associated with susceptibility to cancer will contribute to a better understanding of the aetiology of the disease.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TN-D carried out the DAE experiments and drafted the manuscript; LPJ participated in the study design and helped to draft the manuscript; JM carried out the cell culture; NF and SM-C carried out the mutation screening; kConFab and OS provided the cell lines; FL-K participated in the development of the laboratory workflow and helped to draft the manuscript; MCS participated in the experiment design; SVT conceived the study, participated in its design and coordination and helped to draft the manuscript. FL participated in the study coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We wish to thank Heather Thorne, Eveline Niedermayr, all the kConFab research nurses and staff, the heads and staff of the Family Cancer Clinics, and the Clinical Follow Up Study (funded by NHMRC grants 145684, 288704 and 454508) for their contributions to this resource, and the many families who contribute to kConFab. kConFab is supported by grants from the National Breast Cancer Foundation, the National Health and Medical Research Council (NHMRC) and by the Queensland Cancer Fund, the Cancer Councils of New South Wales, Victoria, Tasmania and South Australia, and the Cancer Foundation of Western Australia. We gratefully acknowledge Arnaud Dumont for his help on developing the R script. TN-D was the recipient of a fellowship from Fondation de France and a Special Trainee Award from the International Agency for Research on Cancer. LPJ was the recipient of a fellowship from Ligue Contre le Cancer – Comité du Rhône. MCS is a National Health and Medical Research Council Senior Research Fellow and Victorian Breast cancer Research Consortium (VBCRC) Group Leader. This work was also supported by the Canadian Institutes of Health Research team grant CRN-87521-IC0898832 and by the National Cancer Institute, National Institutes of Health (NIH) grant R01 CA121245.

References

1. Oldenburg RA, Kroeze-Jansema K, Kraan J, Morreau H, Klijn JGM, Hoogerbrugge N, Ligtenberg MJL, van Asperen CJ, Vasen HFA, Meijers C, Meijers-Heijboer H, de Bock TH, Cornelisse CJ, Devilee P: **The CHEK2*1100delC variant acts as a breast cancer risk modifier in non-BRCA1/BRCA2 multiple-case families.** *Cancer Res* 2003, **63**(23):8153–8157.

2. Honrado E, Osorio A, Palacios J, Benitez J: **Pathology and gene expression of hereditary breast tumors associated with BRCA1, BRCA2 and CHEK2 gene mutations.** *Oncogene* 2006, **25**(43):5837–5845.

3. Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M, Elstrodt F, van Duijn C, Bartels C, Meijers C, Schutte M, McGuffog L, Thompson D, Easton D, Sodha N, Seal S, Barfoot R, Mangion J, Chang-Claude J, Eccles D, Eeles R, Evans DG, Houlston R, Murday V, Narod S, Peretz T, Peto J, Phelan C, Zhang HX, Szabo C, Devilee P, Goldgar D, Futreal PA, Nathanson KL, Weber B, Rahman N, Stratton MR: **Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations.** *Nat Genet* 2002, **31**:55–59.

4. CHEK2 Breast Cancer Case-Control Consortium: **CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies**. *Am J Hum Genet* 2004, **74**(6):1175–82.

5. Antoni L, Sodha N, Collins I, Garrett MD: **CHK2 kinase: cancer susceptibility and cancer therapy - two sides of the same coin?** *Nat Rev Cancer* 2007, **7**(12):925–36.

6. Nevanlinna H, Bartek J: **The CHEK2 gene and inherited breast cancer susceptibility.** *Oncogene* 2006, **25**(43):5912–5919.

7. Buckland PR: **Allele-specific gene expression differences in humans.** *Hum Mol Genet* 2004 Oct 1, **13 Spec No 2**:R255–60.

8. Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, Lee MP: **Allelic variation in gene expression is common in the human genome.** *Genome Res* 2003 Aug, **13**(8):1855–1862.

9. Bray NJ, Buckland PR, Owen MJ, O'Donovan MC: **Cis-acting variation in the expression of a high proportion of genes in human brain.** *Hum Genet* 2003 Jul, **113**(2):149–153.

10. Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW: **Allelic variation in human gene expression.** *Science* 2002 Aug 16, **297**(5584):1143.

11. Chen X, Weaver J, Bove BA, Vanderveer LA, Weil SC, Miron A, Daly MB, Godwin AK: **Allelic imbalance in BRCA1 and BRCA2 gene expression is associated with an increased breast cancer risk.** *Hum Mol Genet* 2008, **17**(9):1336–1348.

12. Azzato EM, Lee AJX, Teschendorff A, Ponder BAJ, Pharoah P, Caldas C, Maia AT: **Common germ-line polymorphism of C1QA and breast cancer survival**. *Br J Cancer* 2010, **102**(8):1294–9.

13. Yan H, Dobbie Z, Gruber SB, Markowitz S, Romans K, Giardiello FM, Kinzler KW, Vogelstein B: **Small changes in expression affect predisposition to tumorigenesis.** *Nat Genet* 2002 Jan, **30**:25–26.

14. Valle L, Serena-Acedo T, Liyanarachchi S, Hampel H, Comeras I, Li Z, Zeng Q, Zhang HT, Pennison MJ, Sadim M, Pasche B, Tanner SM, de la Chapelle A: **Germline allele-specific expression of TGFBR1 confers an increased risk of colorectal cancer**. *Science* 2008, **321**(5894):1361–5.

15. Milani L, Gupta M, Andersen M, Dhar S, Fryknäs M, Isaksson A, Larsson R, Syvänen AC: **Allelic imbalance in gene expression as a guide to cis-acting regulatory single nucleotide polymorphisms in cancer cells**. *Nucleic Acids Res* 2007, **35**(5):e34.

16. Feinberg AP, Tycko B: **The history of cancer epigenetics**. *Nat Rev Cancer* 2004, **4**(2):143–53.

17. Gameau LJ, Brown LD, Moore MA, E DJ, DemYan WB: **Optimization of LightTyper genotyping assays**. *Biochemica* 2005, **3**.

18. Ware MD, DeSilva D, Sinilnikova OM, Stoppa-Lyonnet D, Tavtigian SV, Mazoyer S: **Does nonsense-mediated mRNA decay explain the ovarian cancer cluster region of the BRCA2 gene?** *Oncogene* 2006 Jan 12, **25**(2):323–328.

19. Jordheim LP, Nguyen-Dumont T, Thomas X, Dumontet C, Tavtigian SV: **Differential allelic expression in leukoblast from patients with acute myeloid leukemia suggests genetic regulation of CDA, DCK, NT5C2, NT5C3, and TP53.** *Drug Metab Dispos* 2008 Dec, **36**(12):2419–2423.

20. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T: **The RIN: an RNA integrity number for assigning integrity values to RNA measurements**. *BMC Mol Biol* 2006, **7**:3.

21. Nguyen-Dumont T, Calvez-Kelm FL, Forey N, McKay-Chopin S, Garritano S, Gioia-Patricola L, De Silva D, Weigel R, Sangrajrang S, Lesueur F, Tavtigian SV, Breast Cancer Family Registries (BCFR), Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer (kConFab): **Description and validation of high-throughput simultaneous genotyping and mutation scanning by high-resolution melting curve analysis**. *Hum Mutat* 2009, **30**(6):884–90.

22. Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F, Byrnes GB, Chuang SC, Forey N, Feuchtinger C, Gioia L, Hall J, Hashibe M, Herte B, McKay-Chopin S, Thomas A, Vallée MP, Voegele C, Webb PM, Whiteman DC, Australian Cancer Study, Breast Cancer Family Registries (BCFR), Kathleen Cuningham Foundation Consortium for Research into Familial Aspects of Breast Cancer (kConFab), Sangrajrang S, Hopper JL, Southey MC, Andrulis IL, John EM, Chenevix-Trench G: **Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer**. *Am J Hum Genet* 2009, **85**(4):427–46.

23. Wilkins JM, Southam L, Price AJ, Mustafa Z, Carr A, Loughlin J: **Extreme context specificity in differential allelic expression**. *Hum Mol Genet* 2007, **16**(5):537–46.

24. Conti E, Izaurralde E: **Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species**. *Curr Opin Cell Biol* 2005, **17**(3):316–25.

25. Herrmann MG, Durtschi JD, Bromley LK, Wittwer CT, Voelkerding KV: **Amplicon DNA melting analysis for mutation scanning and genotyping: cross-platform comparison of instruments and dyes**. *Clin Chem* 2006, **52**(3):494–503.

26. Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, Lepage P, Lavergne K, Villeneuve A, Gaudin T, Brandstrom H, Beck A, Verner A, Kingsley J, Harmsen E, Labuda D, Morgan K, Vohl MC, Naumova AK, Sinnett D, Hudson TJ: **A survey of genetic and epigenetic variation affecting human gene expression.** *Physiol Genomics* 2004 Jan 15, **16**(2):184–193.

27. Pastinen T, Ge B, Gurd S, Gaudin T, Dore C, Lemire M, Lepage P, Harmsen E, Hudson TJ: **Mapping common regulatory variants to human haplotypes**. *Hum Mol Genet* 2005, **14**(24):3963–71.

28. Fogarty MP, Xiao R, Prokunina-Olsson L, Scott LJ, Mohlke KL: **Allelic expression imbalance at high-density lipoprotein cholesterol locus MMAB-MVK**. *Hum Mol Genet* 2010, **19**(10):1921–9.

29. Pinsonneault JK, Papp AC, Sadee W: **Allelic mRNA expression of X-linked monoamine oxidase a (MAOA) in human brain: dissection of epigenetic and genetic factors.** *Hum Mol Genet* 2006 Sep 1, **15**(17):2636–2649.

Figures

Figure 1 - Principle of high-resolution melting curve analysis (HRM) for detection of allelic expression imbalance

A single labelled fluorescent probe is designed with complete complementarity to one allele of the exonic SNP chosen as marker, while mismatching the other allele. Following an asymmetric PCR reaction in presence of the probe, HRM analysis allows the alleles in heterozygous individuals to be distinguished by differences in their melting temperatures (Tm), with a fluorescent signal correlated to the relative abundance of each transcript. The Allele 2/Allele 1 ratio is calculated as h2/h1.

Figure 2 - Mixing experiment to assess efficiency of HRM for detection of differential allelic expression

(A) SimpleProbe® melting curves generated on the LightScanner® instrument from mixing series of opposite homozygous genomic DNAs for the marker SNP rs2236142 in CHEK2. Mixing ratios are indicated on the figure (G allele: C allele ratio). (B) The determination coefficient (R2) between the expected and the observed allelic ratios was 0.963. Each value corresponds to the mean value of 4 replicate measurements.

Figure 3 - R plot showing the DAE assay results for the 41 heterozygous individuals enrolled in the study

The level of DAE is calculated by dividing the allelic ratio in cDNA by the corresponding ratio in genomic DNA (log cDNA-log gDNA). Statistical significance for DAE is evaluated using Student's t-test. Evidence for DAE is reached when i) the point estimate of the level of DAE (plotted on the horizontal axis) is greater than 20%, ii) the Student's t-test p-value (plotted on the vertical axis) is ≤ 0.05, and iii) the 95% confidence interval of the point estimate (based on 4 replicate assays) does not include 0. Samples above the horizontal line and outside the hatched area reached the statistical threshold for DAE. In our experiment, four samples met all criteria (Samples 2181, 2498, 2500 and 2666).

Figure 4 - Non-sense mediated mRNA decay causes differential allelic expression in CHEK2*1100delC carriers

Allelic ratio measurements were performed on genomic DNA (gDNA), cDNA derived from LCLs in standard cell culture condition, and cDNA from LCLs treated with puromycin, an NMD inhibiting agent. (A) For a carrier of the mutation, comparison of gDNA and cDNA melting profiles supports the existence of DAE. Puromycin-cDNA profile resembles gDNA, supporting the role of NMD in the DAE observed in this individual. (B) The wild-type sample shows similar profiles in all three situations. HRM profiles were generated with the R script.

Tables

Table 1 - Comparison of the duration of the DAE analysis between the HR-1^{TM} and the LightScanner® instruments, for 96 samples

The HR-1^{TM} instrument can only analyze a single sample per run making data analysis time consuming. The LightScanner® instrument, with its 384-well plate format, is of greater practical efficiency. Data analysis was performed using an analysis tool that we developed.

DAE step	HR-1$^{\text{TM}}$ instrument	LightScanner® instrument
PCRs	Same duration	Same duration
Data acquisition	2 days	12 minutes
Data analysis	1 full day	15 minutes

Additional Files

Additional file 1 — Mutation screening results

File name: additional file1.pdf

File format: pdf

Title of dataset: Mutation screening results for the 41 breast cancer samples enrolled in the DAE study.

Description: Additional table showing the mutation screening results for the 41 breast cancer samples enrolled in the DAE study.

Additional file 2 — Primers and probes

File name: additional file2.doc

File format: word document

Title of dataset: Primers and probes used in the DAE study on the *CHEK2* gene.

Description: Additional table showing the primers and probes sequences used to perform the DAE study.

Software availability

A copy of our R script code has been made available on Sourceforge.net.

Project name: HRMdae project;

Project home page: http://sourceforge.net/projects/hrmdae;

Operating system(s): Platform independent, R environment;

Programming language: R v2, or above;

Licence: GPL v3;

Any restrictions to use by non-academics: None.

Figure 1

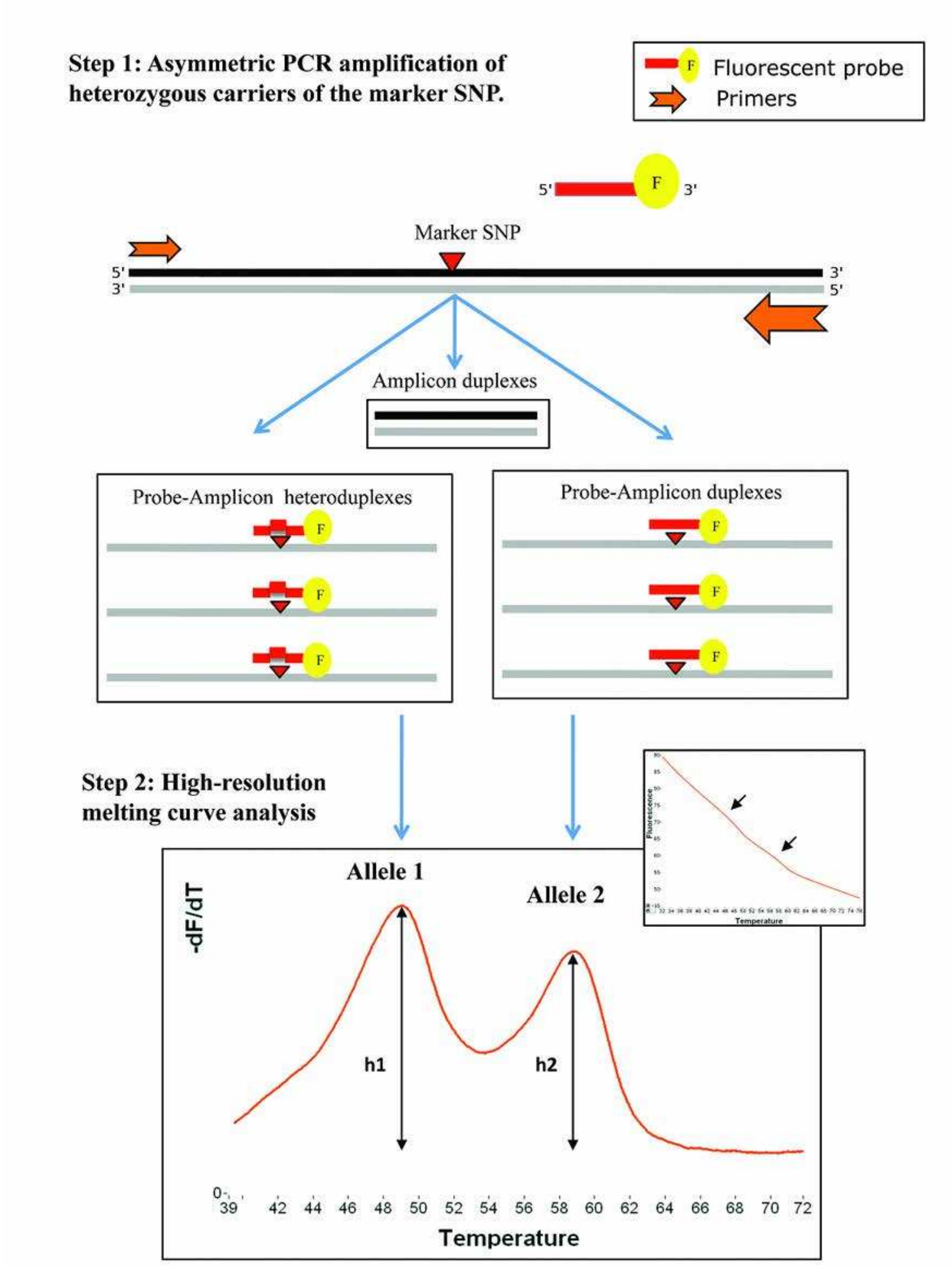

Step 1: Asymmetric PCR amplification of heterozygous carriers of the marker SNP.
F
Fluorescent probe
Primers
5'
F
3'
Marker SNP
5'
3'
3'
5'
Amplicon duplexes
Probe-Amplicon heteroduplexes
F
F
F
Probe-Amplicon duplexes
F
F
F
Step 2: High-resolution melting curve analysis
Fluorescence
Temperature
-dF/dT
Allele 1
Allele 2
h1
h2
0
39
42
44
46
48
50
52
54
56
58
60
62
64
66
68
70
72
Temperature

Figure 2

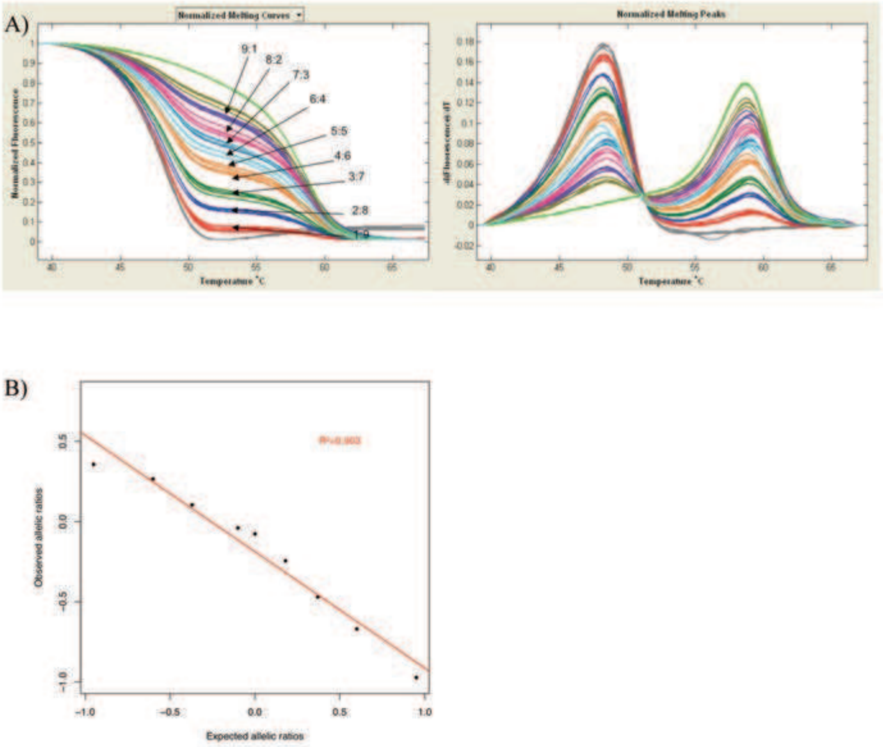

A)
Normalized Melting Curves
9:1
8:2
7:3
6:4
5:5
4:6
3:7
2:8
1:9
Normalized Fluorescence
Temperature °C
Normalized Melting Peaks
Temperature °C
B)
Observed allelic ratios
Expected allelic ratios
-1.0
-0.5
0.0
0.5
1.0

Figure3

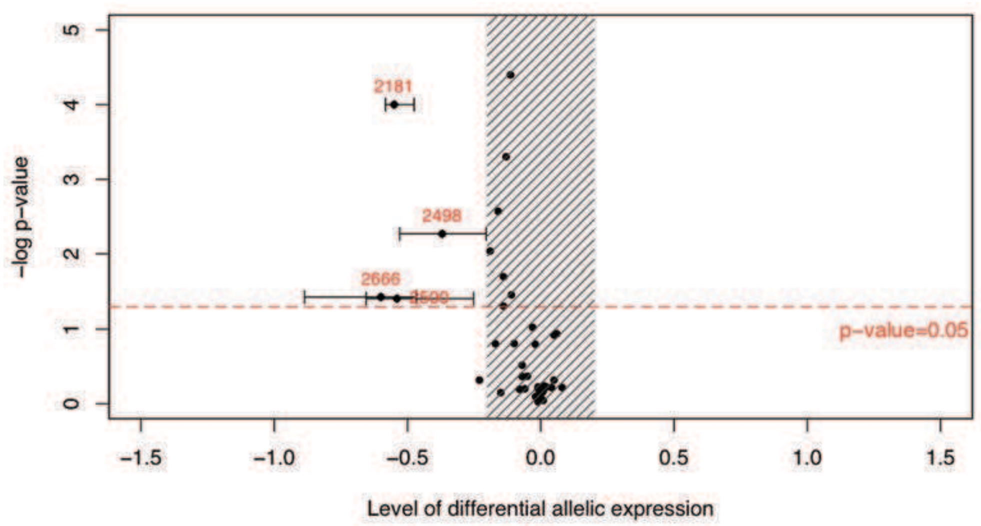
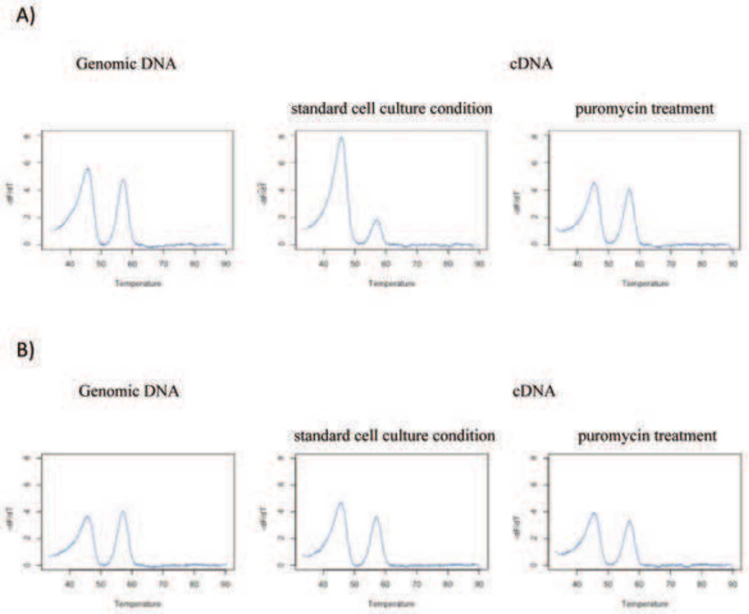


Figure 4



Additional file 1 - Mutation screening results for the 41 breast cancer samples enrolled in the DAE study

Sample	Genotype	Genotype	1100delC	Other variants	DAE results[d]	
ID	at rs2236142[a]	at rs2236141[b]	carrier	(HGVS nomenclature)[c]	DAE value [95%CI]	p-value
1507	Het	CC		c.319+38_319+39insA	-0.16 [-0.22; -0.10]	0.0027
1526	Het	Het			-0.10 [-0.26; 0.07]	0.16
1802	CC	Het			-0.01 [-0.06; 0.04]	0.59
1928	Het	CC			-0.02 [-0.27; 0.23]	0.80
1967	Het	CC		c.319+38_319+39insA	0.01 [-0,06; 0.08]	0.47
2026	Het	Het			-0.14 [-0.29; 0.00]	0.049
2166	Het	CC		c.319+38_319+39insA	0.02 [-0.09; 0.14]	0.60
2181	Het	CC	Yes	c.319+38_319+39insA	-0.55 [-0.58; -0.48]	1.10-4
2193	Het	Het		c.252A>G	-0.07 [-0.25; 0.11]	0.30
2212	Het	CC		c.252A>G+c.319+38_319+39insA	-0.01 [-0.51; 0.49]	0.94
2234	GG	Het			-0.11 [-0.14; -0.09]	4.10-5
2247	Het	CC		c.319+38_319+39insA	0.01 [-0.33; 0.36]	0.90
2443	Het	CC			-0.03 [-0.07; 0.01]	0.095
2472	Het	CC		c.319+38_319+39insA	-0.01 [-0.08; 0.05]	0.62
2497	GG	Het			-0.02 [-0.05; 0.01]	0.16
2498	Het	CC	Yes	c.319+38_319+39insA	-0.37 [-0.53; -0.20]	0.0054
2499	GG	Het			-0.01 [-0.09; 0.07]	0.77
2500	Het	CC	Yes		-0.54 [-0.66; -0.25]	0.039
2526	Het	CC			0.04 [-0.16; 0.24]	0.61
2529	Het	CC			-0.14 [-0.25; -0.04]	0.020
2534	CC	Het			0.01 [-0.06; 0.07]	0.058
2536	Het	CC			0.06 [-0.03; 0.15]	0.12
2539	Het	CC			-0.05 [-0.23; 0.13]	0.43
2541	Het	CC			-0.07 [-0.31; 0.17]	0.43
2542	Het	CC			0.08 [-0.36; 0.52]	0.60
2557	GG	Het			-0.13 [-0.16; -0.10]	5.10-4
2570	GG	Het			-0.01 [0.07; 0.06]	0.82
2574	GG	Het			0.01 [-0.03; 0.04]	0.70
2665	Het	CC			-0.23 [-1.14; 0.68]	0.48
2666	Het	CC	Yes	c.319+38_319+39insA	-0.60 [-0.89; -0.47]	0.037
2667	Het	CC			-0.15 [-1.33; 1.03]	0.71
2668	Het	Het		c.444+24C>T	-0.08 [-0.55; 0.39]	0.64
2669	Het	Het			-0.11 [-0.21; -0.01]	0.035
2670	Het	Het			-0.19 [-0.33; -0.09]	0.0092
2671	Het	Het			-0.17 [-0.41; 0.30]	0.16
2674	Het	CC			0.05 [-0.02; 0.12]	0.12
2677	Het	Het			-0.06 [-0.41; 0.30]	0.63
2678	Het	CC			-0.01 [-0.15; 0.13]	0.85
2679	Het	CC		c.319+38_319+39insA	0.00 [-0.04; 0.03]	0.84
2680	Het	Het			0.05 [-0.14; 0.03]	0.84
2691	GG	Het			0.01 [-0.05; 0.07]	0.58

[a] GG, frequent homozygote; Het, heterozygote; CC, rare homozygote.

[b] CC, frequent homozygote; Het, heterozygote.

[c] Number based on transcript sequence (NM_007194), +1 as A of ATG start codon.

[d] DAE is expressed as the difference between the log of the signal ratio of the cDNA from the patient's LCL and the corresponding log ratio of genomic DNA.

Additional file 2**:** Primers and probes used in the DAE study of *CHEK2*.

Specific Primer´ or Probe		Oligonucleotide sequence
Primary PCR on genomic DNA	Forward primer	5ÕGCAAAGAGAGCGTCTAACCAG-3Õ
	Reverse primer	5ÕGCAGAGTGGCGCTAAACCT-3Õ
Primary PCR on cDNA	Forward primer	5ÕATCTAGCCGTGGTCACTCGT-3Õ
	Reverse primer	5ÕTAGGACCCACTTCCCTGAAA-3Õ
Secondary PCR	Forward primer	5ÕCAAAGAGAGCGTCTAACCAGACTAAT-3Õ
	Reverse primer	5ÕCAGATACAAACTCCACCCTCAGC-3Õ
Simpleprobe for rs2236142*		5ÕTAAGTTCCGCTCT**C**CCTTCTAAA-3Õ
Simpleprobe for rs2236141*		5ÕTCCTCATTGGTCC**G**GCGG-3Õ

*Polymorphic position is indicated in bold.

´Marker SNPs rs2236141 and rs2236142 were located in the same amplicons.

III

DAE assessment of the *TP53* and *ATM* genes

Following successful application of the HRM assay to the *CHEK2* gene, we aimed to assess DAE of the *TP53* and *ATM* genes. Our preliminary study of *TP53* included 25 informative LCLs. The case-control mutation screening project ongoing in the laboratory identified 14 heterozygotes for the marker SNP in *ATM* from samples for which we already had LCLs on hand. However, the lab's mutation screening studies identified many other subjects that were heterozygous at the probe SNPs. Accordingly, to increase the sample size and reach at least 50 heterozygotes for each gene, we requested LCLs from collaborators involved in the breast cancer genetics study, based on their genotype data obtained through the mutation screening project. After receipt, these LCLs were grown under both standard and NMD inhibitory conditions, and then used to prepare RNA, cDNA, and genomic DNA.

TP53 presents a complex pattern of alternative splicing. Recent studies report 10 different p53 isoforms, which differential production involves regulatory mechanisms at the level of transcription, RNA processing and translation [Marcel and Hainaut, 2009]. The proximal promoter p53P1 initiates the synthesis of the fully spliced variant FSp53 and the p53I2 variant, which retains the entire intron 2. A second promoter, p53P2, has been described in intron 1 and generates a 1 125 bp transcript. A third internal promoter p53P3 has been described, which generates the p53I4 mRNA. A part of intron 4 is retained, followed by all of the gene's downstream exons correctly spliced. In the present study, we chose to focus on the canonical form FSp53, as well as the p53I2 transcript since it is generated from the same proximal promoter as FSp53.

ATM is less complicated, however one must be careful when selecting heterozygotes for c.5557 so that the individuals are not double heterozygotes at the c5558 position because it will interfere with probe binding and decrease the accuracy of the fluorescent signal.

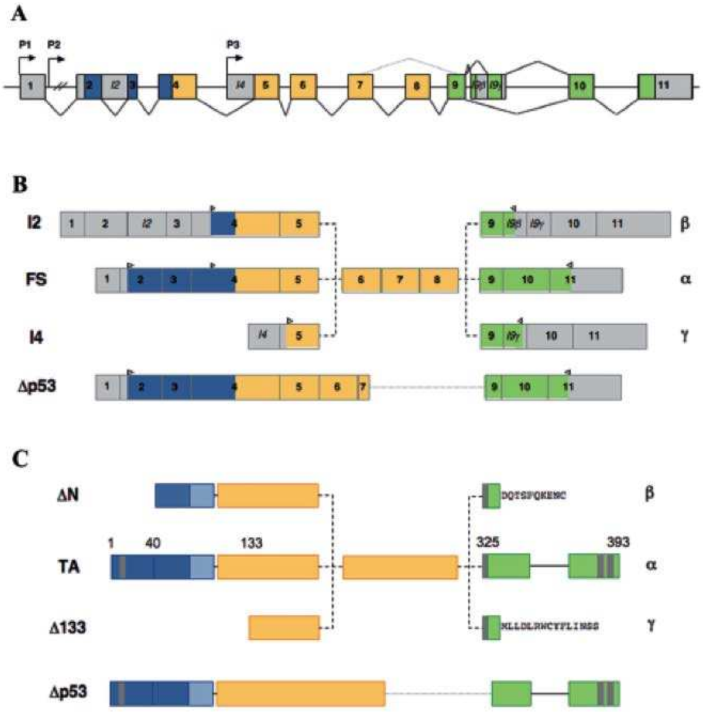


Figure III.1: The *TP53* gene is subject to alternative splicing. From [Marcel and Hainaut, 2009].

Article IV

Differential allelic expression assessment of the breast cancer susceptibility genes ***TP53*** and ***ATM***

(in preparation)

Introduction

The first major susceptibility genes for breast cancer were identified in the early 1990's by linkage analysis and since then, a considerable amount of knowledge about genetic cancer susceptibility and the underlying susceptibility genes have been gathered. More recently, large case-control genotyping studies have identified common modest-risk SNPs and case-control mutation screening has emerged as a useful strategy for identifying and characterizing intermediate-risk susceptibility genes. While some breast cancer susceptibility alleles have been clearly defined in these three well-established classes of genetic variants, they are estimated to account for 30-35% of the relative risk of breast cancer (high and intermediate-risk genes account for about 25%, and common SNPs from GWAS may account for 10%).

The *TP53* gene (NM_000546) belongs to the high-risk susceptibility genes category. *TP53* is located on chromosome 17p13.1, and encodes a protein involved in many cellular pathways that control cell proliferation and homeostasis, such as cell cycle, apoptosis and DNA-repair. The expression of the *TP53* gene is activated in response to various stress signals, including DNA damage [Oldenburg et al., 2007]. Indeed, germline mutations in *TP53* are associated with increased risk of developing breast cancer, in particular in Li-Fraumeni patients [Olivier et al., 2003].

TP53 presents a complex pattern of alternative splicing. Recent studies report 10 different p53 isoforms, which differential production involves regulatory mechanisms at the level of transcription, RNA processing and translation [Marcel and Hainaut, 2009]. The proximal promoter p53P1 initiates the synthesis of the fully spliced variant FSp53 and the p53I2 variant, which retains the entire intron 2. A second promoter, p53P2, has been described in intron 1 and generates a 1 125 bp transcript. A third internal promoter p53P3 has been described, which generates the p53I4 mRNA. A part of intron 4 is retained, followed by all exons correctly spliced. In the present study, we chose to focus on the canonical form

FSp53, as well as the p53I2 transcript since it is generated from the same proximal promoter as FSp53.

The *ATM* gene (NM_000051) is one of the best-understood intermediate-risk genes for breast cancer susceptibility. *ATM* is located on chromosome 11q22-23 and codes for a protein kinase whose many substrates include the products of *TP53*, *BRCA1* and *CHEK2*. ATM plays a central role in sensing and signaling the presence of DNA double-strand breaks. Mutations in the *ATM* gene cause the rare recessive disorder Ataxia-Telangiectasia (AT) in biallelic carriers. In monoallelic carriers, *ATM* mutations have been reported to induce breast cancer susceptibility [Ahmed and Rahman, 2006, Renwick et al., 2006].

Although mutation scanning projects have focused for many years on variations in the coding sequences of such susceptibility genes, structural alterations caused by genetic variants are not the only possible explanation for variations in disease phenotype. Gene expression regulation provides an alternate mechanism for generating cellular variation and may be the underlying explanation for a proportion of cancer syndromes that have not been resolved by germline coding region variants in currently known cancer predisposition genes. Accumulating evidence shows that regulatory variations contribute to many important phenotypes.

Yet, unlike coding sequence variants where the consequences of non-synonymous variation may be resolved at the level of the protein phenotype, defining how variation at the DNA sequence level will induce differences in transcript abundance has proven problematic. Indeed, characterization of the effect of *cis*-acting sequence variants in regulatory regions is a great challenge due to the difficulty to locate these regions. In addition, regulatory variants are not robustly detected by sequence analysis since SNP identification by screening regulatory regions does not consistently allow prediction of the effect of observed SNPs on gene expression [Wang and Sadée, 2006, Gilad et al., 2008]. Thus, knowledge of the effect of genetic variants affecting mRNA transcription is very limited.

Currently, two strategies are most commonly used for assaying gene expression levels. In combination to the well-established linkage and association mapping approaches, expression quantitative trait loci (eQTL) mapping has become a widespread tool for identifying genetic variants that affect gene regulation [Gilad et al., 2006, Cookson et al., 2009]. Differential allelic expression (DAE) assays represent a fundamentally different approach to investigating factors affecting gene expression levels [Yan et al., 2002*b*, Bray et al., 2003, Lo et al., 2003, Pastinen et al., 2004, Pant et al., 2006, Serre et al., 2008, Jordheim et al., 2008, Maia et al., 2009, Azzato et al., 2010].

In such studies, disruption or alteration of gene expression levels is examined through a careful survey of whether the two alleles of a gene are equally expressed. This approach relies on relative quantification of allelic transcripts within heterozygous individuals, using a transcribed SNP as marker. Since they come from the same tissue sample and have therefore been subjected to the same environmental influences (such as genetic *trans*-acting factors and experimental exposures, including mRNA degradation) both alleles should be equally expressed in the absence of *cis*-acting sequence variation or allele-specific epigenetic effects affecting the expression of the target mRNA. Thus, the strength of this approach is that each allele acts as an internal control for confounding factors, displaying *cis*-variation effects without being confounded by any *trans*-variation effects.

We recently described a novel approach based on high-throughput HRM analysis for DAE assessment [Nguyen-Dumont et al. *submitted*]. A test for DAE by HRM analysis consists in a quantitative genotyping experiment, using fluorescent probes. We report here DAE assessment of the breast cancer susceptibility genes *ATM* and *TP53*, using HRM analysis of fluorescein probes [Crockett and Wittwer, 2001].

Materials and methods

Origin of samples

The LCLs included in our study were derived from subjects, who were considered to be at high risk of carrying a genetic predisposition to breast cancer due to an early age at onset and/or family history, and for whom no mutation in *BRCA1* or *BRCA2* genes had been identified. Biological samples were obtained from different sources: Creighton University School of Medicine (Omaha, NE, USA), Centre Léon Bérard (CLB, Lyon, France), the Kathleen Cuningham Consortium for Research into Familial Breast Cancer (kConFab, Melbourne, Australia) and Samuel Lunenfeld Research Institute (Toronto, Ontario, Canada).

Cell culture

In the genes that we wish to analyze, DAE could result from non-sense mediated mRNA decay (NMD) induced by the specific degradation of the transcript from an allele bearing a premature stop codon [Conti and Izaurralde, 2005]. In order to address this issue, RNA was prepared from each LCL under two culture conditions. One was standard LCL culture conditions (this was also the source of the DNAs). The second condition was cells that have been treated with puromycin, a translation inhibitor frequently used to prevent the effect of NMD [Ware et al., 2006].

Cells were maintained in RPMI 1640 medium (Invitrogen, Cergy-Pontoise, France) supplemented with 20% fetal calf serum (VWR, Fontenay-sous-bois, France), 0.4% fungizon (Qiagen, Courtaboeuf, France) and 1% penicilin-streptomycin (Invitrogen), in 5% CO_2 incubator at 37° with 95% humidity. When sufficient growth had occurred, viable cells were counted using Burker cell counting chamber and Trypan blue. Cell suspension was then dispensed to yield approximately four million cells per flask, and volume was adjusted to 10 ml. Cells were incubated overnight in an upright position. The following day, one flask was maintained under standard culture conditions while the other was treated with 100 μM puromycin (Sigma Aldrich, St Quentin Fallavier, France) for 6 hours, in an horizontal position.

Genomic DNA and RNA extraction, cDNA preparation.

Genomic DNA and total RNA were respectively extracted from LCL using Puregene DNA isolation kit (Qiagen) and NucleoSpin RNA II kit (Machery Nagel). Before cDNA preparation, RNA integrity was controlled using the BioAnalyzer and RNA NanoChip II kit. Good quality RNAs, harboring an RNA integrity number (RIN) ≥ 8, were selected for further analysis [Schroeder et al., 2006].
Complementary DNA (cDNA) synthesis was performed from 1 μg total RNA using SuperScriptTM III First Strand Synthesis System for RT- (Invitrogen) with oligo(dT) primers, according to manufacturer's instructions.

Selection of polymorphic markers, primers and probes

Differential measurement of the level of expression of the two alleles of a gene are performed in subjects heterozygous at a coding polymorphism specifically targeted by a fluorescein probe. Primers and probe sequences are listed in Table III.1 and Table III.2 for *TP53* and *ATM* respectively.

The *TP53* gene contains a number of common polymorphisms and rare mutations. Among the common polymorphisms of the coding region, we selected the SNP c.215C>G (*p53R72P*, rs1042522). This SNP is responsible for a proline (C$\underline{\text{C}}$G: P, ancestral allele) to an arginine (C$\underline{\text{G}}$G: R) substitution at codon 72 of exon 4 of *TP53*. Of the currently know *TP53* transcripts variants, we chose to assess DAE of FSp53, the canonical form, and of p53I2, which differ by alternative splicing of intron 2. Primer pairs for PCR$_1$ were designed to ensure specific amplification of the genomic DNA and cDNAs variants. The forward primer for FSp53 overlapped exons 2 and 3; the forward primer for p53I2 hybridized in intron 2; for both transcript variants, the reverse primer overlapped exons 3 and 4.

For the study of the *ATM* gene, we selected the SNP c.5557G>A (*D1853N*, rs1801516). This SNP is responsible for an aspartic acid (<u>G</u>AT: D, ancestral allele) to an asparagine (<u>A</u>AT: N) substitution at codon 1853 of the 36th coding exon of *ATM*. Primer pairs for primary PCR (PCR$_1$) were designed to specifically amplify genomic DNA and cDNA.

Genotyping for these markers SNPs was performed by HRM analysis as described elsewhere [Nguyen-Dumont et al., 2009, Garritano et al., 2009]. For the *ATM* gene, careful selection of subjects was conducted in order to avoid samples also heterozygous at the variable adjacent position (rs1801673), as double heterozygotes may interfere with efficiency of probe binding.

DAE assessment

DAE assays were performed as described elsewhere [Nguyen-Dumont et al., *submitted*]. Briefly, HRM for DAE assessment requires a fluorescent oligonucleotide probe designed to anneal to the sequence surrounding the marker SNP, with standard PCR reagents, in a nested asymmetric PCR reaction. For secondary PCR (PCR$_2$), both genomic DNA and cDNA were amplified using the same set of primers. PCR product melting curves were obtained from the LightScanner® instrument, by melting from 40 to 85°. Data were acquired with LightScanner® Software v2.0, and then analyzed using `R` (http://cran.r-project.org).

The relative abundance of each allele was obtained from the ratio of the peak heights calculated from the derivative of the melting curve. For each sample, ratios were measured from 4 PCR$_1$-replicates and averaged. The level of allelic imbalance for each individual was determined as [log (ratio cDNA) - log (ratio gDNA)]. Statistical significance for the allelic imbalance was calculated using a Student's *t*-test. Criteria for statistically significant DAE were: i) the point estimate of the difference between cDNA and genomic DNA ratios should be greater than 20%; ii) the Student's *t*-test p-value should be ≤ 0.05, and iii) the 95% confidence interval of the point estimate should not include 0.

Results

We first verified that the fluorescein probes designed for both marker SNPs in *ATM* and *TP53* could detect small variations in allelic imbalance. For both genes, artificial bi-allelic ratios were generated by mixing genomic DNA from individuals homozygous for the common and rare variant, in the following ratios: 9:1, 8:2, 7:3, 6:4, 5:5, 4:6, 3:7, 2:8 and 1:9. Figure III.1 shows the melting curves obtained for *ATM*. Linearity of the method was validated by calculating the determination coefficient. An R^2 of 0.891 was obtained, establishing that fluorescein probe melting analysis could accurately allow detection of changes in the relative abundance of the two alleles in a heterozygous sample, by assessment the heights of the peaks corresponding to each allelic variant (Figure III.2).

We were able to gather a total of 50 and 52 LCLs derived from heterozygous individuals for the marker SNP D1853N in *ATM* and R72P in *TP53*, respectively. The statistical threshold for DAE was not reached in any of the two studies (Figure III.3 and III.4). Data from LCLs pre-treated with puromycin, a translation inhibitor used to block the NMD mechanism, suggest the absence of a genetic variant creating a premature stop codon targeted by NMD, in the tested samples. In our study set, difference between cDNA and genomic DNA was too weak to call DAE.

Discussion

The subjects enrolled in this study were expected to be at high-risk of carrying a genetic predisposition to breast cancer. However, our results suggest null or weak *cis*-variation effects in our set of LCLs. No statistically significant DAE was observed during our assessment of the *TP53* and *ATM* genes by HRM analysis. Application of other criteria or use of another statistical test may have identified samples with DAE. However, results from our previous study of the *CHEK2* gene using the HRM approach showed that the +/-20% cut-off allows detection of DAE due to a biological mechanism such as NMD. Borderline sample are hypothesized to carry an imbalanced allelic expression too weak to justify further investigation at this stage. Thus, although it might not be the most appropriate strategy, our approach to identify possible outliers was very conservative.

In particular, our data did not confirm previous observations of allelic imbalance for the *TP53* gene, in LCLs. Gemignani et al. have suggested the existence of a common mechanism leading to the disruption of the allelic expression balance for that gene. The authors reported that homozygous individuals for the C variant of R72P had a reduced expression of *TP53* compared to G homozygotes, and heterozygous individuals had an intermediate level of expression [Gemignani et al., 2004]. Bellini et al. also recently reported DAE of *TP53* [Bellini et al., 2010]. However, in the latter study, the authors investigated the Δ133 form generated from p53I4 mRNA, which is never amplified under our PCR conditions. The transcription initiation site of the p53I4 transcript is located in intron 4 and involves a different promoter than the one involved in transcription of FSp53 and p53I2 transcripts. In addition, Δ133p53 cannot be produced by internal initiation of translation from FSp53 transcript [Marcel and Hainaut, 2009]. Thus, since our marker SNP is located in exon 4, we were not able to test our samples set for DAE of the p53I4 transcript variant, using our experimental conditions.

In the present study, we did not identify statistically significant DAE that would have led to investigation of the transcriptional regulatory regions of the studied genes. However, the experimental approach for DAE screening that we have described elsewhere and applied here on the *TP53* and *ATM* genes can be used to assess other breast cancer susceptibility genes, such as *RAD50*, *BRIP1*, or *RAD51*. Genes for which DAE is observed can be further screened for sequence variants in putative transcriptional regulatory regions identified by computational or comparative genomics methods. Various web-based resources exist for investigation of gene regulation. These include resources for promoter and transcription factor binding site predictions, transcription factor binding profile databases, and alignment of non-coding genome sequences or orthology resources [Wasserman and Sandelin, 2004, Maston et al., 2006].

Identifying functional elements in the human genome, including those that regulate gene expression, is a major challenge that presents great interest especially since numerous diseases have been associated with mutations in both transcriptional regulatory elements and various components of the transcriptional machinery. Although computational approaches have been developed to identify transcriptional regulatory elements on a genome-wide scale, it is likely that bioinformatics methods will not replace the need for experimental verification of regulatory elements. A predicted transcription factor binding site is not necessarily a genuine binding site, and binding does not demonstrate a functional role for that site or necessarily indicate which gene is regulated by the site [Maston et al., 2006].

Conclusion

Allele-specific expression assays allow detection of the existence of regulatory variations without directly identifying or requiring prior knowledge of specific *cis*-regulatory SNPs. DAE testing of genes may yield a considerable reduction of the amount of work in gene expression studies, by focusing discovery effort on a subset of genes that are most likely to harbor coding or regulatory variants that may alter gene expression. HRM analysis is an appropriate approach for this type of study, as it is accurate, rapid and inexpensive. The assays are easy to set up for a large number of genes, once a large samples set of LCLs has been gathered.

Figures and Tables

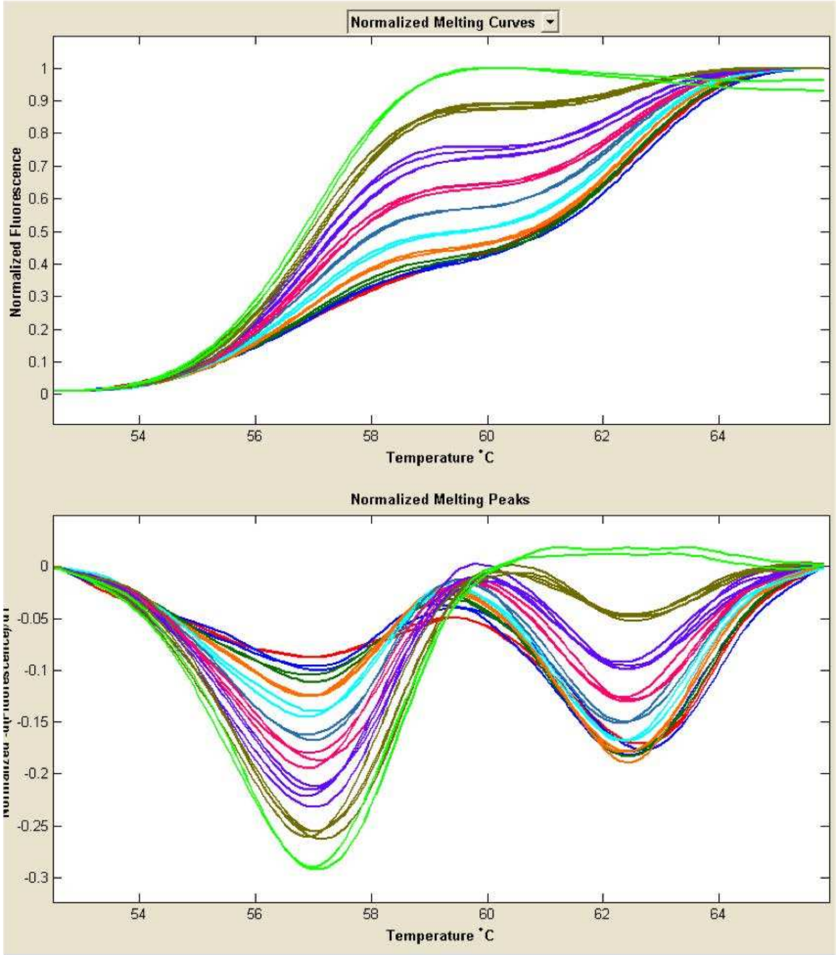


Figure III.1: Fluorescein probe melting curves generated on the LightScanner® instrument from mixing series of opposite homozygous genomic DNAs , using the marker SNP D1853N in *ATM*. Artificial bi-allelic templates with decreasing major allele:minor allele proportions were produced (9:1, 8:2, 7:3, 6:4, 5:5, 4:6, 3:7, 2:8, 1:9). Each bi-allelic mixture was assessed in 4 replicate measurements.

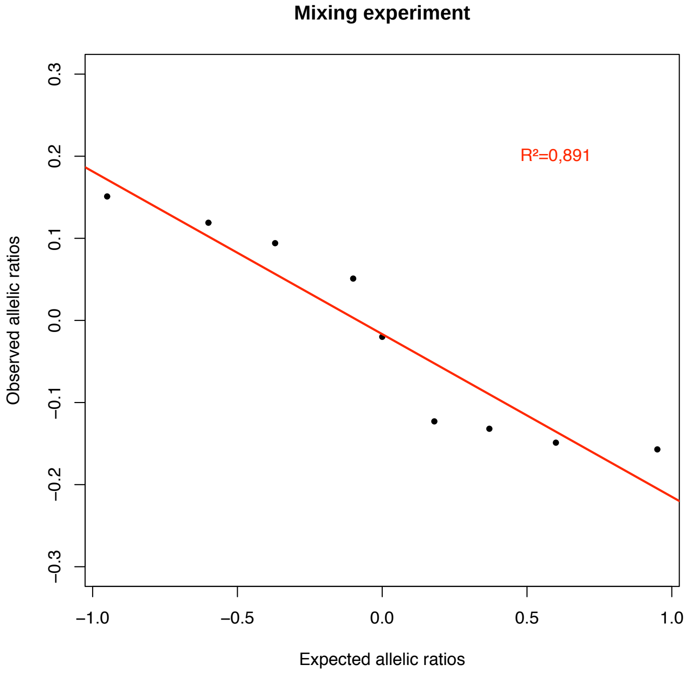


Figure III.2: Standard curve from the mixing experiment to assess efficiency of HRM for detection of differential allelic expression in the *ATM* gene. Levels of DAE were measured and the determination coefficient R^2 between the expected and the observed allelic ratios was 0.891. Each estimate corresponds to the mean value of 4 replicate measurements.

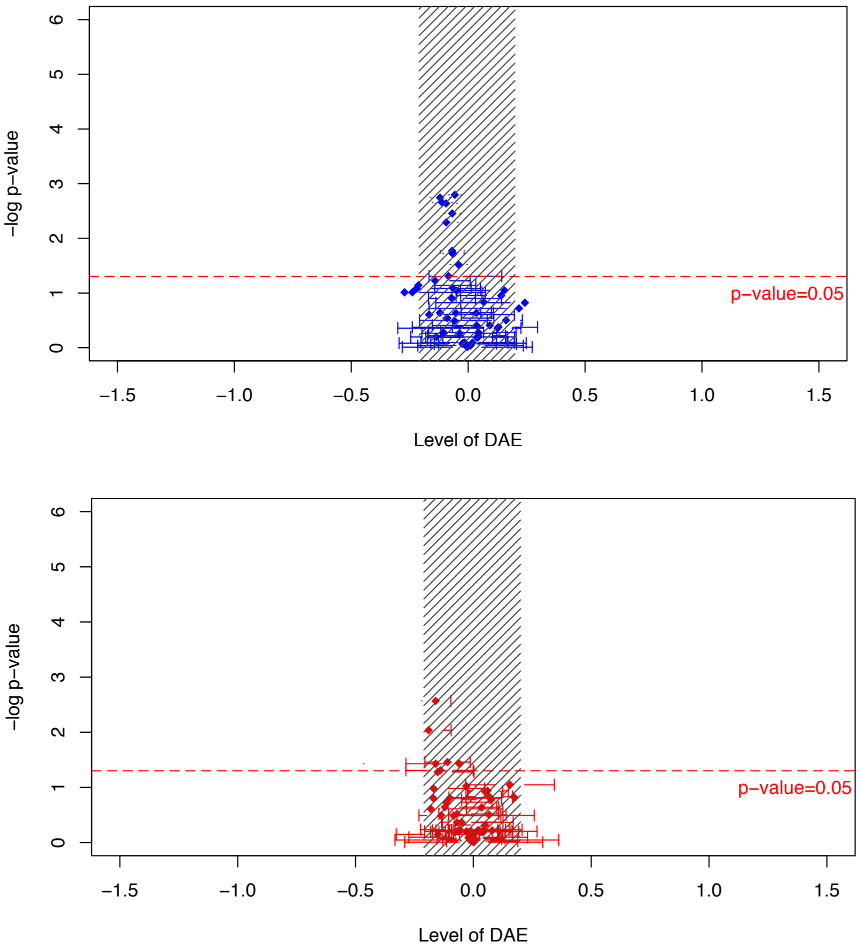


Figure III.3: R plots showing the DAE assay results for the 52 heterozygous individuals enrolled in the study of *TP53*. Results from the assessment of the FSp53 and p53I2 transcripts are in blue and red respectively. The level of DAE is calculated by dividing the allelic ratio in cDNA by the corresponding ratio in genomic DNA (log cDNA-log gDNA). Statistical significance for DAE is evaluated using Student's *t*-test. Evidence for DAE is reached when i) the point estimate of the level of DAE (plotted on the horizontal axis) is greater than 20%, ii) the Student's *t*-test p-value (plotted on the vertical axis) is ≤ 0.05, and iii) the 95% confidence interval of the point estimate (based on 4 replicate assays) does not include 0 (given the high number of samples, the plot does not show the CIs). Samples above the horizontal line and outside the hatched area reached the statistical threshold for DAE. In our experiment, we did not find any samples satisfying the 3 criteria.

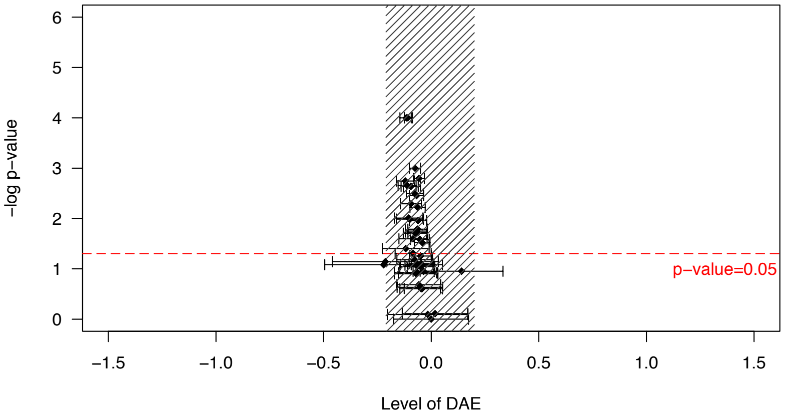


Figure III.4: R plot showing the DAE assay results for the 50 heterozygous individuals enrolled in the study of *ATM*. Statistical criteria were the same as above. We did not find any samples satisfying the 3 criteria for statistically significant DAE.

Primer or probe		Oligonucleotide sequence
PCR_1	Forward primer	CCTATGGAAACTACTTCCTG
FSp53 variant	Reverse primer	AGGGGAGTACGTGCAAGT
	Forward primer	ATGGGACTGACTTTCTGCTC
p53I2 variant	Reverse primer	AGGGGAGTACGTGCAAGT
	Forward primer	GTCAGATCCTAGCGTCG
genomic DNA	Reverse primer	AGAATGCAAGAAGCCCA
PCR_2	Forward primer	AGATGAAGCTCCCAGAA
	Reverse primer	CTGGTAGGTTTTCTGGGAAG
Probe		GGCTGCTCCCCGCGTGGC

Table III.1: Primers and fluorescein probe used for DAE assessment of the *TP53* gene.

Primer or probe		Oligonucleotide sequence
PCR_1	Forward primer	CCAATGTGTGAAGTGAAAACT
cDNA	Reverse primer	TTTGCGAGAAGTGTCGAA
	Forward primer	CTTTTGTCAGACTGTACTTCCATA
genomic DNA	Reverse primer	GGTGAAAAATCCCTGAACA
PCR_2	Forward primer	CTTTTGTCAGACTGTACTTCCATA
	Reverse primer	GGTGAAAAATCCCTGAACA
Probe		CCATGATTCATTTGTATCTTGGAG

Table III.2: Primers and fluorescein probe used for DAE assessment of the *ATM* gene.

Discussion and Conclusion

Discussion

Technical hurdles to Aim 1: gathering informative LCLs

To test a candidate gene for DAE, the following criteria for the choice of the polymorphic markers should be observed. The SNP has to 1) be located in the transcribed region of the gene of interest and 2) have a sufficiently high minor allele frequency to gather a reasonable number of informative samples. This latter point will therefore depend on the number of samples available for assessment. Whenever possible, the position of the SNP should be far enough from exon boundaries to allow use of the same set of primers for PCR_2 amplification from both cDNA and genomic DNA. However, our personal experience has shown that it is possible to use different primer sets for cDNA and genomic DNA (results not shown).

In a DAE study, the main setup issue is the ability to select a subset of LCLs that will support informative assessment. For the study of the *ATM* and *TP53* genes, we were able to gather at least 50 informative LCLs for DAE analysis. In the case of *CHEK2*, it was not possible to find a samples set that was as large. Thus, for this gene, we sought to reach our target number of heterozygous individuals by using two different marker SNPs. In fact, selecting 2 or 3 marker SNPs (that are not in strong disequilibrium with each other) provides a reasonable strategy because it increases the number of informative samples and, if some of the samples are double heterozygotes, allows cross-checking and error reduction.

Addressing Aim 2: Implementing an appropriate assay

In DAE analysis by HRM, the peak heights ratio obtained from the melting curve of a given sample reflects the relative abundance of each allele's transcript. The reproducibility and precision of the assay are supported by the small standard deviations associated with the DAE calculations. The accuracy of the method was illustrated by the consistency of the allelic expression estimates across multiple replicate assays within the same individual sample. In all three assessed genes, genomic DNA ratios varied within a very narrow range, showing the excellent reproducibility and precision of the assay on DNA derived from LCL. The intra-sample variation in replicate analysis was higher for mRNA ratios than for DNA ratios, possibly owing to RNA stability. At low copy numbers of mRNA, the stochastic distribution of the RNA templates may be a major source of variation and hence affect the accuracy of DAE analysis, by for instance, generating discordant replicate results [Pastinen et al., 2004].

Assay optimization from the LightCycler 2.0 to the HR-1™, then to the LightScanner® instrument allowed us to increase i) the resolution of the assay initially used by Ware et al., and ii) its throughput. In parallel to technical up-scaling of the method, development of an analysis tool was essential for efficient determination of which specific samples showed evidence of DAE. Once HRM data were acquired, the normalization of the curves, peak height measurements, ratio calculations and statistical analysis were performed automatically within less than 15 minutes for a set of 96 samples when using the LightScanner® instrument. The script developed using R dramatically reduced the length of the analysis, as compared for instance to our initial study using the HRM approach [Jordheim et al., 2008].

Access to DAE assessment technology can be cost prohibitive for many laboratories. HRM analysis is a simple approach that can sensitively visualize expressed allelic variants and the transcript abundance dynamics in high-throughput, using a small amount of RNA. The protocol is relatively inexpensive since it only requires standard PCR reagents and a small amount of fluorescent probe. In our study, enhancing throughput was not detrimental to the accuracy of the assay. The results obtained with the LightScanner® instrument showed that this methodology can successfully be applied to larger-scale studies. Blood tissue can be used in such high-throughput contexts. DAE in breast tissue would also be interesting to assess but these are difficult to collect.

The main challenge of Aim 3: The determination of the statistical criteria for presence of DAE

As DAE assays are an emerging field, no routine or standardized statistical analytical approaches have been developed to address the specific attributes of this assay.

- One approach that could be envisaged is to estimate the natural variation of the peak height ratios in a situation where both alleles are theoretically equally abundant to determine a cut-off for calling DAE. To do so, one would calculate the average of the peak heights ratios from all genomic DNAs enrolled in a given DAE study and call outlier any sample falling outside the 95% confidence interval [Pastinen and Hudson, 2004].
- We sought advices from statisticians who suggested the following test where the null hypothesis is that both alleles are equally expressed in a series of LCLs, where:

  1. N is the number of heterozygous LCLs used in the statistical test
  2. n is the number of replicates per cDNA sample
  3. X is the level of DAE for a cDNA sample calculated through n replicates ($i.e$ the average of n ratios observed in cDNA, normalized to the average of all N genomic DNAs)
  4. $\overline{X}$ is the global average level of DAE observed through n replicates of N heterozygous cDNA samples.

A global standard deviation $\overline{D}$ is calculated as the average of the standard deviations from each normalized measurement taken into account in $\overline{X}$ calculation. The standard deviation of each series of n replicates is calculated as: $SDev_n = \sqrt{\frac{1}{n-1}(n * \overline{D})}$.

For each individual LCL, the point estimate of X must to be located within the 95% confidence interval of $\overline{X}$. Any point estimate lying outside this range leads to the rejection of the null hypothesis so this sample needs to be subtracted from $\overline{X}$ and the 95% CI calculations. The remaining samples are to be re-assessed. This procedure is to be repeated until the null hypothesis is verified in the population of remaining samples. Any rejected sample would be candidate examples of DAE.

- In the approach we chose, the statistical significance for DAE was calculated using Student's *t*-test to compare cDNA measurements of a given LCL to genomic DNA measurements from the same LCL. In this case, genomic DNA is again used to normalize cDNA measurements. Criteria for statistically significant DAE were: i) the point estimate of the difference between cDNA and genomic DNA ratios should be greater than 20%; ii) the Student's *t*-test p-value should be ≤ 0.05, and iii) the 95% confidence interval of the point estimate should not include 0. Application of another statistical test may have identified more samples with DAE. However, results from the assessment of *CHEK2* showed that the +/-20% limit allows to observe DAE due to a biological mechanism such as NMD. For borderline samples, we suppose that their imbalanced allelic expression would be too weak to justify further investigation at this stage. Thus, although it might not be the most appropriate strategy, our approach to identify the possible outliers was very conservative.

Aim 4: Elucidating the mechanism of DAE

DAE that is observed in all the heterozygous samples and in one direction suggests that i) the marker SNP itself is causal or that ii) the causal variant is in complete LD with the marker, *i.e.* on a shared haplotype. If DAE is observed in both directions, several mechanisms or causal variants (showing incomplete LD with the marker SNP) might be involved [Wang and Sadée, 2006, Wilkins et al., 2007].

Genes for which DAE is evidenced are to be further screened for sequence variants in putative transcriptional regulatory regions identified by comparative genomics methods. Various web-based resources exist for investigation of gene regulation. These include resources for promoter and transcription factor binding sites predictions, transcription factor binding profile databases, alignment of non-coding genome sequences or orthology resources [Wasserman and Sandelin, 2004]. Additional sequence variants identified at this stage would be incorporated into the haplotypes defined following mutation screening analysis of the genes of interest.

Heritability of *cis*-acting effects can also be studied by investigating pedigrees of DAE carriers [Yan et al., 2002*b*]. If no evidence of Mendelian inheritance can be found, allele-specific epigenetic mechanisms might be involved in the observed DAE phenotype. One can then search for differential methylation of regulatory sequences, post-translational histone modification or replication asynchrony.

Unfortunately, in the present study, we did not identify statistically significant DAE that pointed towards useful investigations of the transcription regulatory regions of the studied genes. However, the assay we have implemented for screening for DAE can be applied to test other breast cancer susceptibility genes from the case-control mutation screening project, such as *RAD50*, *BRIP1*, or *RAD51*, using the large set of available LCLs, with "ready to use" genomic DNA, cDNA and puro-cDNA.

Conclusion

Although a number of cancer susceptibility genes have been successfully identified, study design and analytic approaches issues remain a challenge for future disease-related gene/mechanism discovery.

The strength of our approach is really to combine mutation screening and haplotype information with screening for DAE and expression data, in order to identify specific alleles/haplotypes that are differentially regulated. Analysis of the relative allelic ratios of marker SNPs circumvents the issue of confounding *trans*-acting factors. Any significant differences in these ratios support the existence of DAE and hence, *cis*-acting genetic variants determining gene expression.

Allele-specific expression assays allow detection of the existence of regulatory variations without directly identifying or requiring prior knowledge of specific *cis*-regulatory SNPs. DAE testing of genes examined in large-scale case-control mutation screening takes advantage of the fact that the mutation screening will identify reasonable numbers of subjects who are heterozygous for useful marker alleles, even if those alleles are not particularly common. In principle, many genes could then be studied for DAE with good sensitivity, allowing the search for regulatory variants that may alter gene expression to focus on the subset of genes that actually display DAE.

If one aims at establishing DAE as a high-throughput analysis of dysfunctional variants in large cohorts, LCLs and blood are accessible tissues that are suitable resources for DAE analysis [Maia et al., 2009]. Although one advantage from the use of LCLs is to permit performing NMD analyses, there are also limitations to the use of such material. Absence of DAE in LCLs does not exclude the possibility that there is relevant differential expression in tissues of higher relevance to the disease in question (*i.e* breast epithelial lineage), but variation is not recapitulated in LCLs. On the other hand, it is also controversial whether culture conditions could artefactually create DAE. It has been nevertheless reported that DAE is little influenced by variation in culture environment of LCLs and LCLs harvested from different passages yielded similar results [Serre et al., 2008].

The primary goal of DAE study is to identify sequence variants that are likely to alter gene expression and gene product function, and thereby influence susceptibility to breast cancer. However, to demonstrate that some of these variants actually show disease association, large-scale epidemiological studies would be required and might ultimately lead towards the identification of causal genetic factors responsible for susceptibility to disease. Identification and elucidation of rare intermediate-risk genetic variants associated with susceptibility to cancer could contribute to a better understanding of the etiology of the disease.

One important challenge today is to find the locations of the genes or to identify additional mechanisms involved in predisposition to cancer, with little or no knowledge at all of how many genes are involved, how they interact with each other or with environmental factors, and what, if any, the genotype-phenotype relationship is. Variation in expression adds another level of complexity. A few genes have already been shown to exhibit DAE patterns apparently predisposing to cancer. Determining the genetic causes of cancers has immense public health benefits including prevention, early detection and improved treatment. Current genetic tools offer much promise to this research but the complexities of common cancers remain challenging.

Annexes

Annex I – Article V

Determining the effectiveness of high-resolution melting analysis for SNP genotyping and mutation scanning at the TP53 locus

Garritano S, Gemignani F, Voegele C, **Nguyen-Dumont T**, Le Calvez-Kelm F, De Silva D, Lesueur F, Landi S, Tavtigian SV.

BMC Genetics, 2009; 10:5

BMC Genetics

Methodology article

Determining the effectiveness of High Resolution Melting analysis for SNP genotyping and mutation scanning at the *TP53* locus

Sonia Garritano[1,2], Federica Gemignani[1], Catherine Voegele[2], Tú Nguyen-Dumont[2], Florence Le Calvez-Kelm[2], Deepika De Silva[3], Fabienne Lesueur[2], Stefano Landi[1] and Sean V Tavtigian*[2]

Address: [1]Department of Biology – Genetics via Derna 1, 56126 University of Pisa, Pisa, Italy, [2]Genetic Susceptibility group, International Agency for Research on Cancer, Lyon, France and [3]Idaho Technology Inc, Salt Lake City, Utah, USA

Email: Sonia Garritano - garritanos@iarc.fr; Federica Gemignani - fgemignani@biologia.unipi.it; Catherine Voegele - voegele@iarc.fr; Tú Nguyen-Dumont - nguyent@students.iarc.fr; Florence Le Calvez-Kelm - lecalvez@iarc.fr; Deepika De Silva - Deepika@idahotech.com; Fabienne Lesueur - lesueurf@iarc.fr; Stefano Landi - slandi@biologia.unipi.it; Sean V Tavtigian* - tavtigian@iarc.fr

* Corresponding author

Published: 17 February 2009

BMC Genetics 2009, **10**:5 doi:10.1186/1471-2156-10-5

Received: 8 September 2008
Accepted: 17 February 2009

This article is available from: http://www.biomedcentral.com/1471-2156/10/5

Abstract

Background: Together single nucleotide substitutions and small insertion/deletion variants are the most common form of sequence variation in the human gene pool.

High-resolution SNP profile and/or haplotype analyses enable the identification of modest-risk susceptibility genes to common diseases, genes that may modulate responses to pharmaceutical agents, and SNPs that can affect either their expression or function. In addition, sensitive techniques for germline or somatic mutation detection are important tools for characterizing sequence variations in genes responsible for tumor predisposition. Cost-effective methods are highly desirable. Many of the recently developed high-throughput technologies are geared toward industrial scale genetic studies and arguably do not provide useful solutions for small laboratory investigator-initiated projects. Recently, the use of new fluorescent dyes allowed the high-resolution analysis of DNA melting curves (HRM).

Results: Here, we compared the capacity of HRM, applicable to both genotyping and mutation scanning, to detect genetic variations in the tumor suppressor gene *TP53* with that of mutation screening by full resequencing. We also assessed the performance of a variety of available HRM-based genotyping assays by genotyping 30 *TP53* SNPs. We describe a series of solutions to handle the difficulties that may arise in large-scale application of HRM to mutation screening and genotyping at the *TP53* locus. In particular, we developed specific HRM assays that render possible genotyping of 2 or more, sometimes closely spaced, polymorphisms within the same amplicon. We also show that simultaneous genotyping of 2 SNPs from 2 different amplicons using a multiplex PCR reaction is feasible; the data can be analyzed in a single HRM run, potentially improving the efficiency of HRM genotyping workflows.

Conclusion: The HRM technique showed high sensitivity and specificity (1.0, and 0.8, respectively, for amplicons of <400 bp) for mutation screening and provided useful genotyping assays as assessed by comparing the results with those obtained with Sanger sequencing. Thus, HRM is particularly suitable for either performing mutation scanning of a large number of samples, even in the situation where the amplicon(s) of interest harbor a common variant that may disturb the analysis, or in a context where gathering common SNP genotypes is of interest.

Background

Together Single Nucleotide Polymorphisms (SNPs), rare single nucleotide substitutions, and small insertion/deletion mutations constitute the most common forms of sequence variation in the human genome. For example, Nickerson *et al.* [1] have estimated that the density of common SNPs (with a frequency greater than 1%) is about 1 per 300 bp in the overall human gene pool. Furthermore, deep resequencing studies have demonstrated that the number of rare single nucleotide substitutions and small insertion/deletion variants vastly outnumber common SNPs [2,3].

During the last decade, SNPs have essentially replaced microsatellites for linkage and/or association studies [4,5] and genome-wide association studies with phase 2 and phase 3 confirmations have now provided overwhelming evidence of association on common SNPs with a number of diseases [6,7]. SNPs are also becoming of interest in pharmacogenetics, because some of them are associated with significant differences in biological response to pharmaceutical agents [8,9].

Heavy interest in SNPs has led to the development of different genotyping methods: some of them are targeted to the analysis of one or few SNPs [10,11], and others are designed to scan the whole genome [12,13]. Modern genotyping equipment has driven the per genotype cost for very large-scale SNP genotyping studies quite low. In addition, clonal sequencing technologies may drive the cost of moderate sensitivity resequencing studies very low [14,15]. However, these technologies are actually geared to what are essentially industrial scale genetic studies and arguably to not provide useful solutions for small laboratory investigator-initiated projects.

Interest in fast and reliable methods of mutation screening is increasing as well. Such methods are desirable for case-control mutation screening studies and high-throughput somatic (tumor) mutation screening studies [16,17], aiding the identification of new genes involved in carcinogenesis. They are also desirable for detecting genes responsible for drug-resistance in micro-organisms [18], and for detecting genes that modify growth, resistance to parasites, or yield in plants [19].

Many techniques have been developed to discover genomic variation, including those based on HPLC (High Performance Liquid Chromatography), electrophoretic conformational changes, and enzymatic or chemical cleavage reactions [20]. The goal of these screening techniques is to reduce the use of DNA sequencing and control costs while maintaining sensitivity and specificity. The HRM technique has been used to mutation scan the coding sequences of several clinically important genes [21-26]. For instance, 3 studies have reported mutation screening of *TP53* exonic regions [21,22,27]. In this manuscript, we describe lessons learned from a larger scale application of HRM to mutation screening and genotyping at the entire *TP53* locus. First, we assayed (in terms of sensitivity and specificity) the HRM technique, by comparing the results with the classic Sanger sequencing method, used here as the gold standard reference. Second, we propose solutions for genotyping challenges (discrimination of the 3 genotypes, simultaneous genotyping of 2 or more SNPs) that are sometimes encountered when using a classical HRM approach.

Methods

Origin of DNA samples

Mutation screening of the entire *TP53* locus was performed on 47 DNA samples including lymphocyte DNA from 25 Li-Fraumeni patients, DNA from lymphoblastoid cell lines derived from 15 familial breast cancer patients, and DNA from 7 hemizygous (at the *TP53* locus) breast tumor cell lines (Garritano et al, in preparation).

Genotyping of 30 SNPs located within the *TP53* locus was performed on 270 DNA samples from the Coriell Repository, corresponding to 90 Caucasians, 90 East Asians, and 90 Africans.

This mutation screening and genotyping project received approval from the IARC Institutional Review Board and from the Brazilian center from which we received the Li-Fraumeni patient samples. It was conducted according to the Declaration of Helsinki Principles.

Mutation screening/SNP discovery using HRM

PCR was performed in 8 μl reactions containing 20 ng of template DNA, 1.5 mM $MgCl_2$, 265 μM dNTP, 400 nM forward and reverse primers, 0.8X LCGreen® Plus (Idaho Technology, Salt Lake City, Utah, USA), 0.04 U/μl of Platinum®Taq Polymerase, and 1× PCR buffer supplied by the manufacturer (Invitrogen, Paisley, Scotland).

The HRM process consists in performing the PCR in the presence of the DNA binding dye LC Green®, monitoring the progressive change in fluorescence caused by release of the dye from a DNA duplex as it is denatured by increasing the temperature, collecting a high resolution melting curve, and identifying the samples with melting curve aberrations indicative of the presence of a sequence variant. Fluorescence intensity as a function of temperature, monitored by the LightScanner® instrument (Idaho Technology, Salt Lake City, Utah, USA), can reveal very small changes in the melting curve shape, when analyzed with the LightScanner® software using the "Scanning" mode (Idaho Technology, Salt Lake City, Utah, USA).

Genotyping using HRM

We designed pairs of primers flanking each SNP [See Additional file 1] to amplify DNA fragments shorter than 400 bp. In some instances, HRM can directly discriminate all 3 genotypes (common homozygotes, heterozygotes and rare homozygotes) of a polymorphism. However, for the majority of *TP53* amplicons, genotyping using spike-in control DNA was performed to allow distinction of rare homozygotes from common homozygotes. In brief, genomic DNAs were mixed with an equal amount of DNA from a known major allele homozygous subject to allow formation of heteroduplexes. This strategy converts the minor allele homozygotes into heterozygotes, rendering them distinguishable from the major allele homozygous samples. The scoring of genotypes obtained with spike-in experiments was managed via automated procedures. For instance, we have developed a Laboratory Information Management Systems (LIMS) where results generated from a standard HRM genotyping plate and a corresponding spike-in genotyping plate are automatically converted into a final genotype call [28]. The program is also capable of rejecting samples that show unacceptable calls.

For amplicons containing two or more SNPs, sensitivity of mutation scanning may be decreased by producing complex melting curve data, and a different genotyping strategy had to be applied. This second strategy relies on an unlabelled probe-based genotyping analysis followed by mutation scanning, where the probe is designed to target the SNP(s) of interest. Both probe-amplicon duplex and whole amplicon duplex melting regions can be observed from the same melting run, in two distinct temperature windows, allowing genotyping and mutation scanning analyses to be performed simultaneously. Stratification of the samples according to their genotypes at the common variant positions prior to mutation scanning analysis reduces the noise and enhances the sensitivity for the detection of rare or unknown variants.

In practice, unlabeled 3' blocked probes targeting each common SNP were designed. PCR were performed in presence of a DNA dye (Here LC Green®) and oligonucleotides serving as probes were blocked at the 3' end to prevent extension during amplification. All genotyping assays were performed as a nested PCR, to ensure a good amplification of the region of interest. The primary PCR used standard conditions, whereas the secondary PCR included the unlabelled probe (500 nM) and was asymmetric so that more copies of the strand to which the probe anneals were produced. The ratio between the nested PCR primers was 1:5 (100 nM:500 nM) with an excess of the primer for the strand that is complementary to the probe. This favours probe-target annealing and reduces competition with the complementary strand [29]. Thus, this protocol produced sufficient double-stranded product for amplicon melting and enough single stranded product for probe annealing [30]. The analysis proceeds in two steps. The first step consists in analyzing the melting curve in the region corresponding to the probe T_m. This step stratifies the samples into three groups based on the genotypes of the common SNP. The second analysis step consists in performing mutation scanning of the genotype-defined subgroups in the region corresponding to the amplicon T_m, to identify the samples that are heterozygous for any rare sequence variants. For the amplicon containing SNPs rs9894946 (common) and rs17883532 (rare) the probe was designed to perfectly complement the rs9894946 T allele (GGAGCTCAGTAC**T**GCCTGCCC, the variable nucleotide is indicated in bold). For the amplicon containing SNPs rs858528, rs1641548, and rs1641549, two probes were designed. The first probe was designed to perfectly complement the rs858528 G allele (GCAGAGC**G**AGACTCAAAA). The second probe was designed to complement the rs1641548 G allele and rs1641549 A allele (TTAACC**G**GGC**A**TGATGGCAG, the variable nucleotides corresponding to SNPs rs1641548 and rs1641549 are indicated in bold). Probes were designed to have different T_m (63°C and 54°C, respectively), in order not to interfere with each other in the melting data analysis.

Results

Mutation Scanning

During the course of a project to mutation screen the entire *TP53* locus by direct resequencing from a set of 47 samples, we took delivery of a High Resolution Melt instrument. To assess the sensitivity and specificity of HRM for mutation scanning, we undertook mutation screening of the last 21 *TP53* amplicons and of 1 amplicon corresponding to the proximal promoter region of the gene (from a total of 67 amplicons) by both full-sequence resequencing and HRM in a single pass experiment (Table 1).

Nine of the amplicons were <400 bp in length, with an average length of 286 bp. For these, the sensitivity and specificity of HRM for sequence variant detection were 1.0 (38 true positive/(38 true positive + 0 false negative)), and 0.83, (295 true negative/(60 false positive + 295 true negative)), respectively.

Thirteen of the amplicons were >400 bp in length, with an average length of 544 bp. For these, the sensitivity and specificity of HRM for sequence variant detection were 0.81 (105 true positive/(105 true positive + 23 false negative)), and 0.84, (339 true negative/(69 false positive + 339 true negative)), respectively. Of note, the variant rs17551157, insertion of a cytosine following a 7 cytosine mononucleotide run in the TP53 promoter region, was undetectable in an amplicon of 653 bp.

Table 1: Oligonucleotide primer sequences used for comparison of HRM and sequencing sensitivity and specificity.

Amplicon	Forward sequence 5'>3'	Reverse sequence 5'>3'	Location	size
3	CGGGACGTGAAAGGTTAGAA	TTTTGGGGTGGAAAATTCTG	promoter	653
39	TGGCCATCTACAAGCAGTCA	ACACGCAAATTTCCTTCCAC	exon5-intron5	211
40	CATGAGCGCTGCTCAGATAG	CAGTTGCAAACCAGACCTCA	exon6	234
41	GTGGAAGGAAATTTGCGTGT	TTGCACATCTCATGGGGTTA	intron6	212
43	TGGCTCTGACTGTACCACCA	TCTACTCCCAACCACCCTTG	intron 7	371
44	CTGGAAGACTCCAGGTCAGG	AGCTGTTCCGTCCCAGTAGA	intron7	383
46	GCGCACAGAGGAAGAGAATC	TGAAAGCTGGTCTGGTCCTT	intron9	452
47	GCAGTGATGCCTCAAAGACA	GCAGGCTAGGCTAAGCTATGA	intron9	280
48	TGACTTTGCCTGATACAGATGC	TAGCTACTGGGGAGGCAGAG	intron9	596
49	GGCCTGCCTAGCCTACTTTT	GTAGCAGGCGCTTGTAGTCC	intron9	578
50B	GACTACAAGCGCCTGCTACC	TTTCATGCAACCATGCTGTT	intron9	614
51	CCCTACAGTTGGGCAAAGTC	CGACTGTGCCTCGTTTCTTT	intron9	491
52A	CCTGGGCGATAGAGTGAGAC	GGCTGGACTCAAACTCTTGG	intron9	134
52B	GTCGCATGCACATGTAGTCC	CTTGAGTTCCAAGGCCTCAT	intron9	635
53	ACTTCTCCCCCTCCTCTGTT	CCTGGGTTTGGATGTTCTGT	exon10-intron10	348
55	TATACTCAGCCCTGCCATGC	GGACTTCAGGTGGCTGTAGG	intron10	603
57	TTTGGGTCTTTGAACCCTTG	GTGGTTTCAAGGCCAGATGT	exon11 (3'UTR)	400
58	GGCCCACTTCACCGTACTAA	AAGCGAGACCCAGTCTCAAA	exon11 (3'UTR)	485
59	AAGGAAATCTCACCCCATCC	AAATGCAGATGTGCTTGCAG	exon11 (3'UTR)	456
60	TTGAGACTGGGTCTCGCTTT	CAGTCTCCAGCCTTTGTTCC		566
61	AAAACTTTGCTGCCACCTGT	ATCCTGCCACTTTCTGATGG		415
62	GCCTCTCACCAAGGATTACG	CCTGGACAGTAGCACCCACT		535

The joint dropout rate from PCR, sequencing, and or HRM was 6.8%. Neither PCR-sequencing nor PCR-HRM had a single pass dropout rate exceeding 5%, thus staying above our general research mutation screening success rate target of 95%.

Genotyping

We performed genotyping of 30 SNPs located within the *TP53* locus on 270 DNA samples from the Coriell Repository.

In some cases, it was possible to distinguish directly the three different genotypes of a SNP using standard HRM analysis of the amplicon of interest. Figure 1 displays the melting curve analysis of the SNP rs9903378, a T>G substitution. In this experiment, it was possible to discriminate the three groups corresponding to each genotype directly (common homozygotes TT, heterozygotes TG, and rare homozygotes GG) (Figure 1). This SNP resides in a T-rich sequence that has a low melting temperature. The GG samples have a melting curve different from the common homozygotes TT; evidently, the G interrupts the long poli-Ts and markedly increases the melting point of the amplicon.

However, we found direct detection of minor allele homozygotes to be the exception rather than the rule, at least when using the mutation scanning approach for amplicons in the 200 bp to 400 bp length range. In this context, spike-in experiments provided an approach to detection of minor allele homozygotes. As an example, melting analyses for the SNP rs17881035 are displayed in Figure 2 (panels A, B). Applying standard melting curve analysis (Figure 2A), we observed that the heterozygous AG samples have a distinct melting curve profile compared to the common homozygote AA samples. However, the minor allele homozygote GG samples were not distinguished from the common homozygous AA samples because the difference between their T_m was insufficient. In a second experiment (Figure 2B), each sample was mixed with an equal quantity of DNA from an AA homozygote (a pre-PCR spike-in experiment). This strategy in effect converts the minor allele GG homozygotes into GT heterozygotes, rendering them distinguishable from AA samples. In some instances, HRM can directly discriminate all of the genotypes of an amplicon that contains two SNPs. As an example, a 130 bp amplicon carrying SNPs rs17880560 and rs1614984 is displayed on Figure 3.

Nevertheless, we have encountered examples where HRM cannot discriminate heterozygous samples for SNP1 from heterozygous samples for SNP2. In Figure 4 (panels A, B, C), we present an example and solution for an amplicon that contains two SNPs (rs9894946 and rs17883532) not directly distinguishable from each other. Heterozygous samples for either the first or the second SNP show almost indistinguishable melting curves (Figure 4A). We then

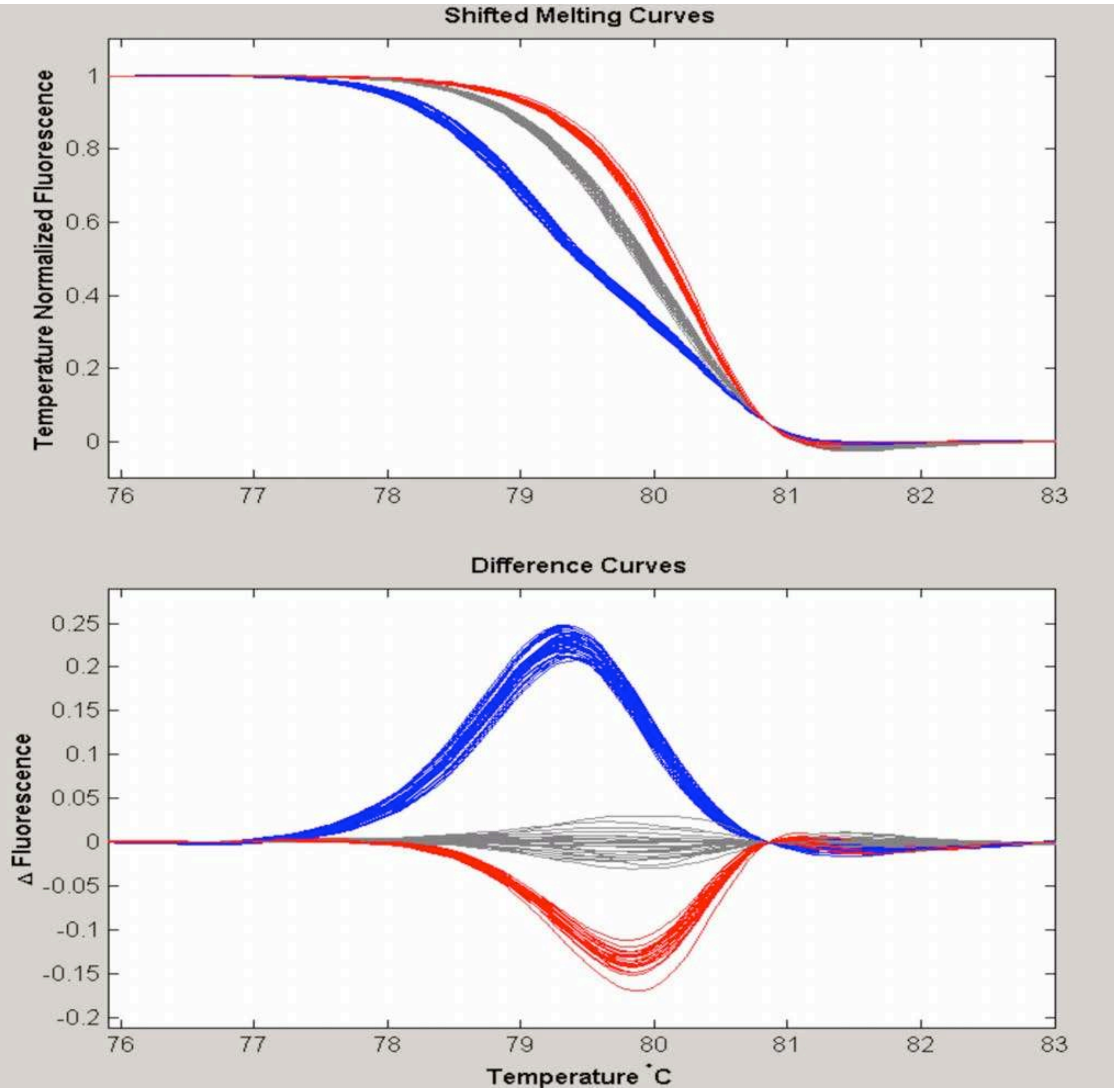


Figure 1
Genotyping of SNP rs9903378. The three groups are well distinguished: TT in grey, GG in red and TG in blue.

designed an unlabeled 3' blocked probe that hybridizes to the region of sequence specific for the more common SNP rs9894946 [31]. Results are displayed in Figure 4 panels B, C. The first analysis step stratifies the samples into three groups based on the genotypes of rs9894946 (Figure 4B). In a second analysis step, mutation scanning of the genotype-defined subgroups in the region corresponding to the amplicon T_m is performed, in order to identify the samples that are heterozygous for any other rare sequence variants. In this example, we found three heterozygous subjects for the SNP rs17883532 in 270 samples (Figure 4C). All these heterozygous subjects were homozygous for the major allele of SNP rs9894946. This solution is acceptable only if the second SNP is rare because the mutation scanning applied after the stratification of samples according the melting profile of the common SNP targeted by the unlabeled probe may not distinguish rare from common homozygous samples.

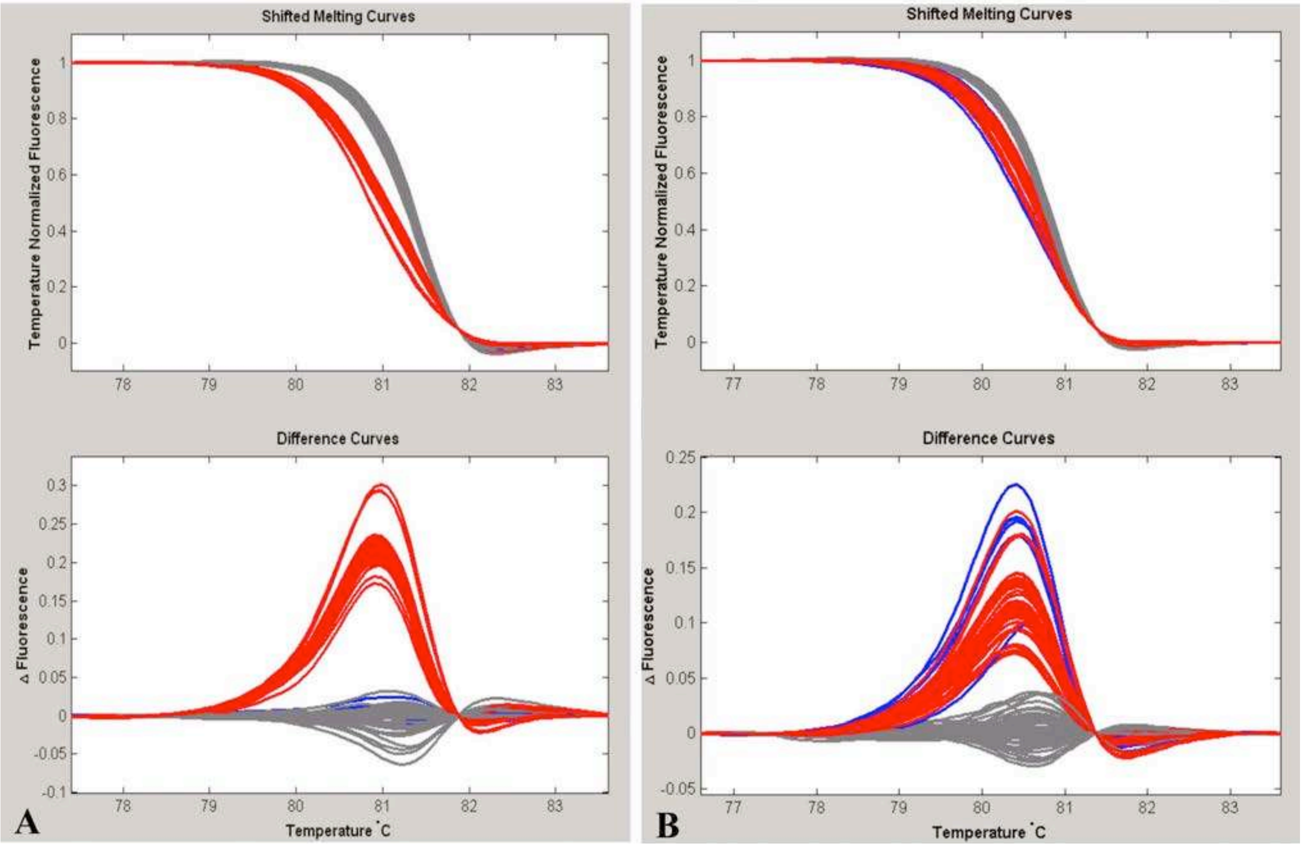


Figure 2

Genotyping using spike-in control DNA to distinguish common homozygotes from rare homozygotes. A. The melting curves of AG heterozygotes (in red) are distinguished from homozygous AA (in grey). Homozygous GG samples (in blue) are not distinguished from the common AA homozygous samples.**B** After spike-in of a homozygous AA sample, the GG samples are converted into AG heterozygotes (in blue), and they are now distinguished from AA samples.

During the course of our TP53 study, we faced another particular situation, where two or more common SNPs lie in the same amplicon. In such a case, it may be necessary to use more than one unlabelled probe. For instance, one of our *TP53* amplicons contains two SNPs: rs1641548 and rs1641549. These are only 4 bp apart, and both are A-to-G variants. Samples that are heterozygous for either the first or the second SNP have essentially indistinguishable melting curves (Figure 5A). Moreover, this amplicon also contains the SNP rs858528. In this case, two probes were designed. The first probe was designed to perfectly complement the rs858528 G allele. The second probe was designed to complement the rs1641548 G allele and rs1641549 A allele. Thus, homozygotes for the rs858528/rs1641548/rs1641549 haplotype G-G-A match both probes exactly and therefore have the highest T_m across the compound-melting interval. Figure 5B and 5C show various genotypes combinations of the three SNPs found in our sample series.

Finally, we evaluated whether the different genotypes from 2 independent PCR products could be discriminated from one single melting curve analysis. Since the amplicon containing SNP rs9903378 (which can be directly genotyped, see Figure 1) and the amplicon containing SNP rs9894946 (for which a specific probe had to be designed, see Figure 4) showed different T_m (range 75–80°C and 90–95°C, respectively), they were selected to conduct the experiment. Both SNP containing fragments were amplified in a single PCR, and HRM analysis was conducted on the multiplex PCR product. In this last experiment, conditions of the multiplex PCR slightly differed from conditions of the simplex PCR, in order to achieve simultaneous amplification of both amplicons in a single reaction. In particular, the primer concentration for the amplicon containing the rs9903378 was decreased from 400 nM to 300 nM because at higher concentration only this amplicon was amplified (data not shown). Melting curves of the multiplex PCR products showed different patterns depending on the genotype combinations for the

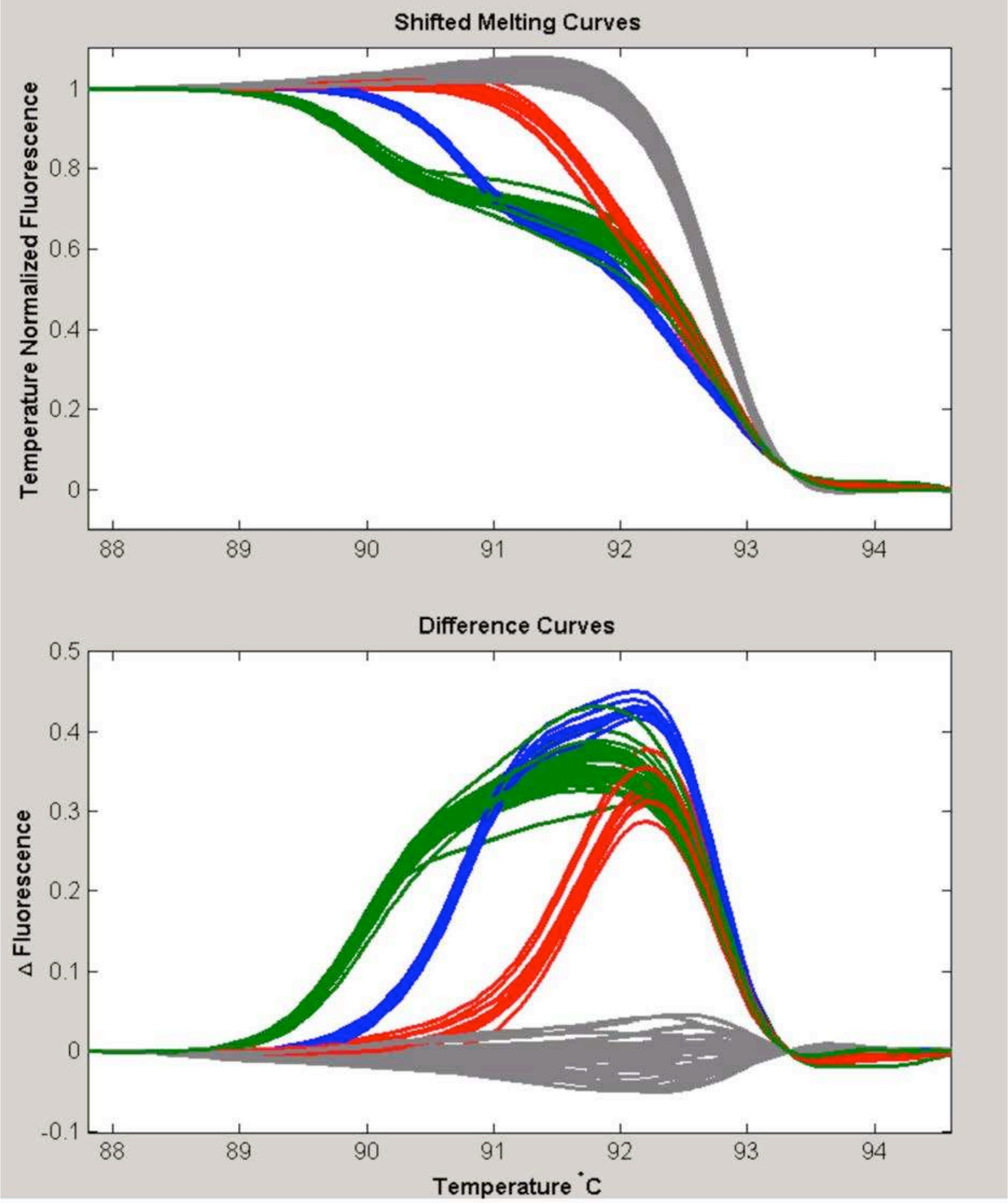


Figure 3

Simultaneous genotyping of two or more SNPs within the same amplicon using the classic HRM approach. There are two SNPs in this amplicon: rs17880560 and rs1614984. In grey, samples homozygous for both SNPs; in red, samples homozygous for rs17880560 (delCACGGC/delCACGGC) and heterozygous for rs1614984 (C/T). In blue, samples homozygous for rs1614984 (C/C) and heterozygous for rs17880560 (insCACGGC/delCACGGC). In green, samples heterozygous for both SNPs.

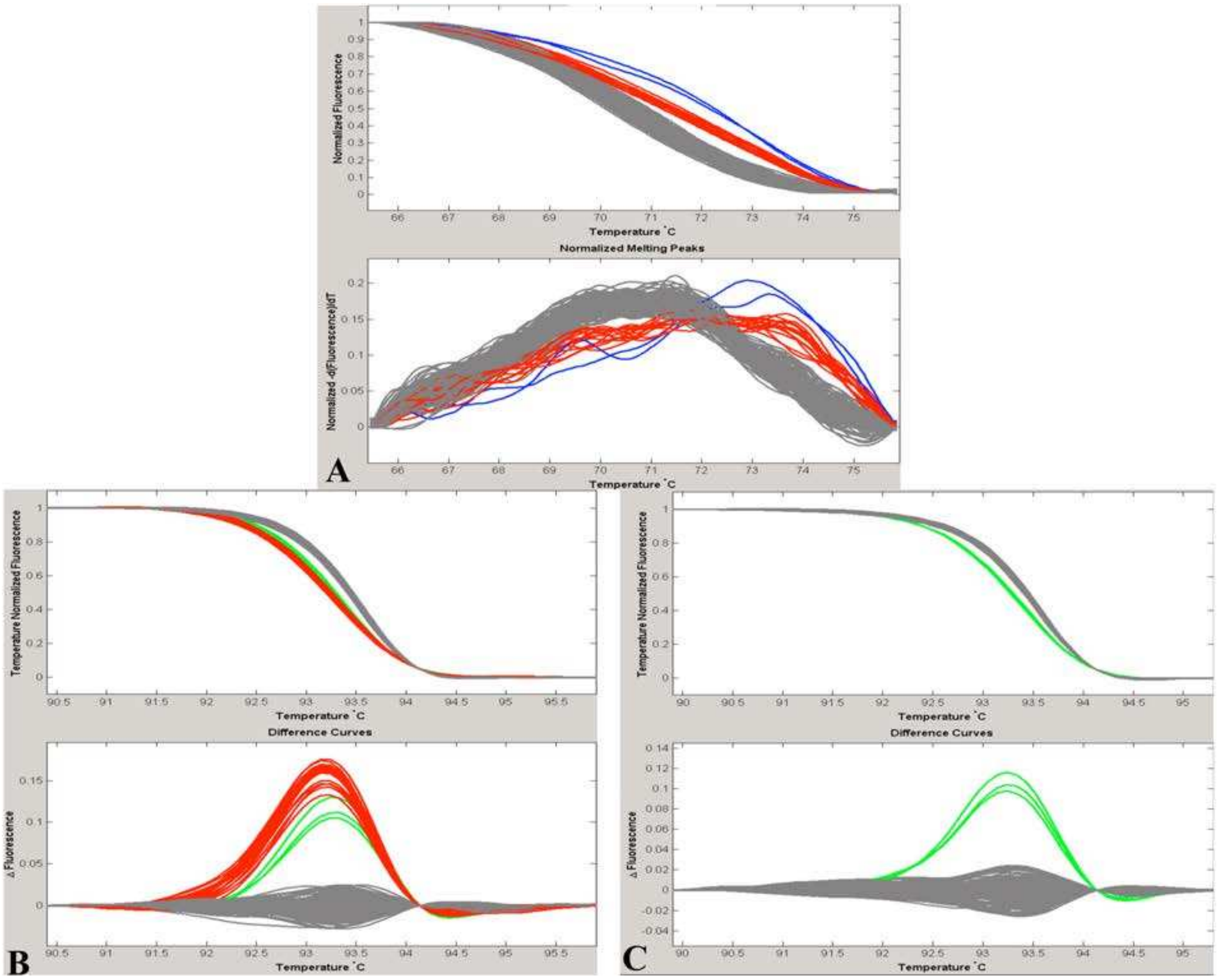


Figure 4
Simultaneous genotyping of common SNP rs9894946 and rare SNP rs17883532. A. In mutation scanning mode, heterozygous samples for either the first (in red) or the second SNP (in green) have virtually indistinguishable melting curves. **B**. In genotyping mode using an unlabeled probe for rs9894946, the 3 genotypes are distinguisable (CC in gray, CT in red, TT in blue). **C**. Mutation scanning of homozygous rs9894946 CC subset reveals heterozygous rs17883532 CT (In green).

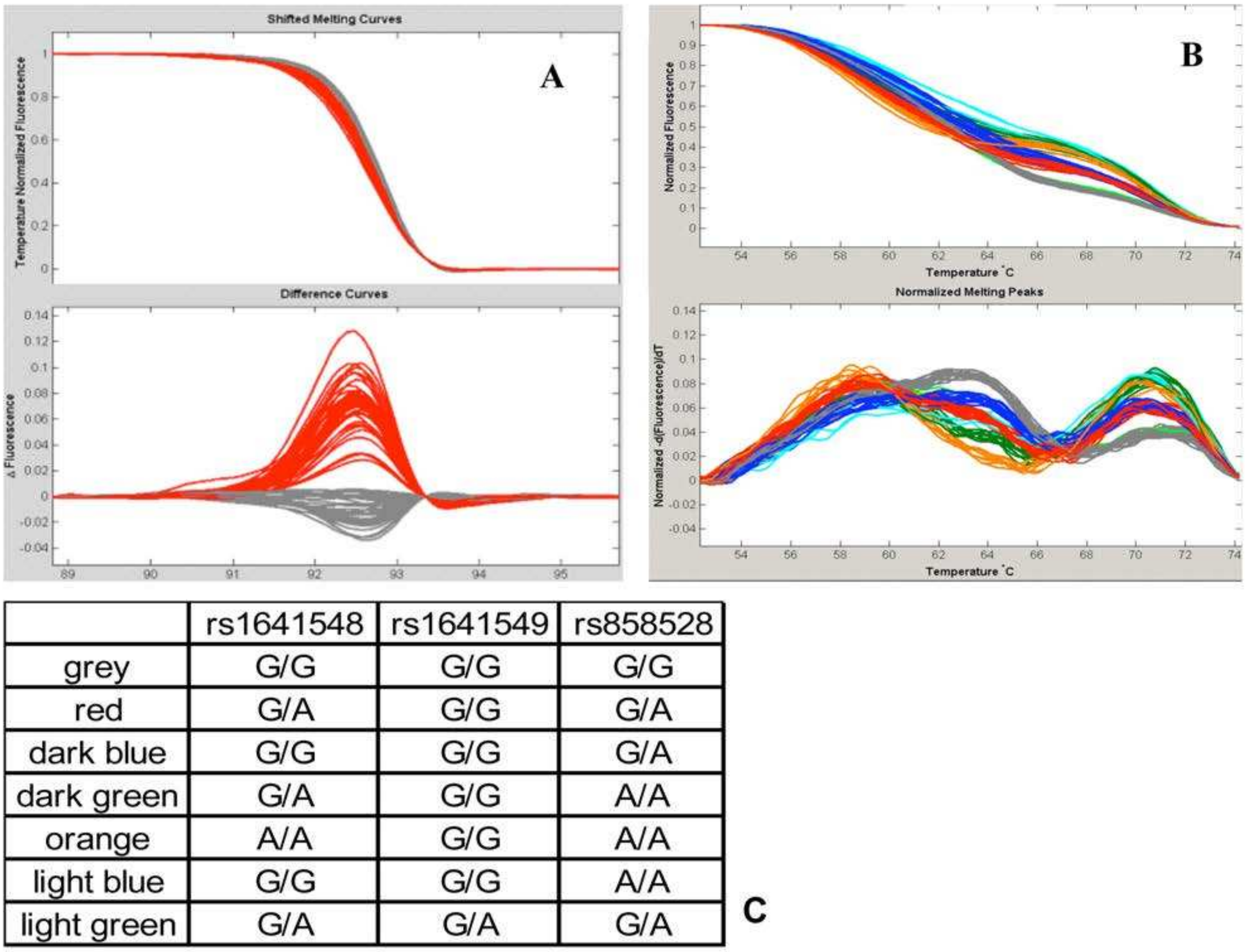
2 SNPs (Figure 6). Using the "genotyping" mode of LightScanner® software, it was possible to distinguish them from each other in a single melting curve analysis (Figure 6A). However, especially when analyzing a large number of samples (>80), resolution can be improved by performing the HRM analysis of the multiplex PCR products in two steps, that is by analyzing the 2 melting regions corresponding to the 2 DNA fragments separately (Figure 6B and 6C). Thus, our results demonstrate that HRM genotyping of multiplex PCR products is feasible and cost and time effective.

Discussion

In this manuscript, we assessed the sensitivity and specificity of HRM for mutation screening by comparing it head to head with the direct resequencing of 21 *TP53* locus amplicons on 47 DNA samples. A second application of the HRM analysis was the genotyping of 30 known SNPs within this gene on 270 DNA samples.

In mutation scanning mode, the sensitivity and specificity of HRM were 1.0 and 0.80, respectively, for amplicons of <400 bp, and 0.81 and 0.84, respectively, for amplicons of >400 bp.

Recent studies have validated HRM for screening of number genes of clinical significance [21-26]. These studies also reported a sensitivity of HRM close to 100%, except in the situation were amplicons have a high GC content [26]. We also encountered a similar situation with one *TP53* GC-rich amplicon (see below). However, in the previous studies, the HRM technique was evaluated only



	rs1641548	rs1641549	rs858528
grey	G/G	G/G	G/G
red	G/A	G/G	G/A
dark blue	G/G	G/G	G/A
dark green	G/A	G/G	A/A
orange	A/A	G/G	A/A
light blue	G/G	G/G	A/A
light green	G/A	G/A	G/A

Figure 5

Simultaneous genotyping of rs858528, rs1641548, and rs1641549 using two unlabeled probes. **A** Samples that are heterozygous for rs858528, rs1641548, and rs1641549 have essentially indistinguishable melting curves (in red). **B** Genotyping using two unlabeled probes. Probe 1 targets the G allele of rs858528 and probe 2 targets the G allele of rs1641548 and the A allele of rs1641549. **C** Each distinct melting profile from panel B corresponds to a combination of genotypes of the three SNPs found in our population.

on partial or full coding sequence(s) of the genes of interest. For instance, more than 80% of *TP53* mutation studies focus on exons 5–8 (residues 126–306) because most mutations are localized in the DNA binding domain of the protein (residues 100–300) [21,22]. However, in one study where the HRM analysis was extended to the entire coding exon of *TP53*, 41% of the alterations fall outside exons 5–8 [22]. Thus partial scanning of *TP53* sequence may lead to a bias in the mutation analysis. Following this idea, we aimed to evaluate a set of HRM assays that sample the entire *TP53* locus, including one or more amplicons from the proximal promoter, coding exons, introns, and 3' UTR. Our study thereby provides a broader view of the strengths and limitations of HRM-based techniques. However, the *TP53* amplicons used in the present study were not designed specifically for HRM but rather for mutation screening of the entire *TP53* locus by direct resequencing in the context of a Li-Fraumeni syndrome-related study (Garritano *et al*, in preparation). Consequently, some of the amplicons were longer than the optimum for HRM mutation scanning. Nonetheless, we have shown that the sensitivity of HRM for mutation screening is very high, especially for amplicons <400 bp.

For genotyping applications, especially for intronic SNPs, primers were sometimes designed quite far from the SNP of interest to avoid unspecific amplification. Despite the length of the amplicons used, we obtained full concordances between HRM genotyping calls and results of direct sequencing. Moreover, using an amplicon of >400 bp, we

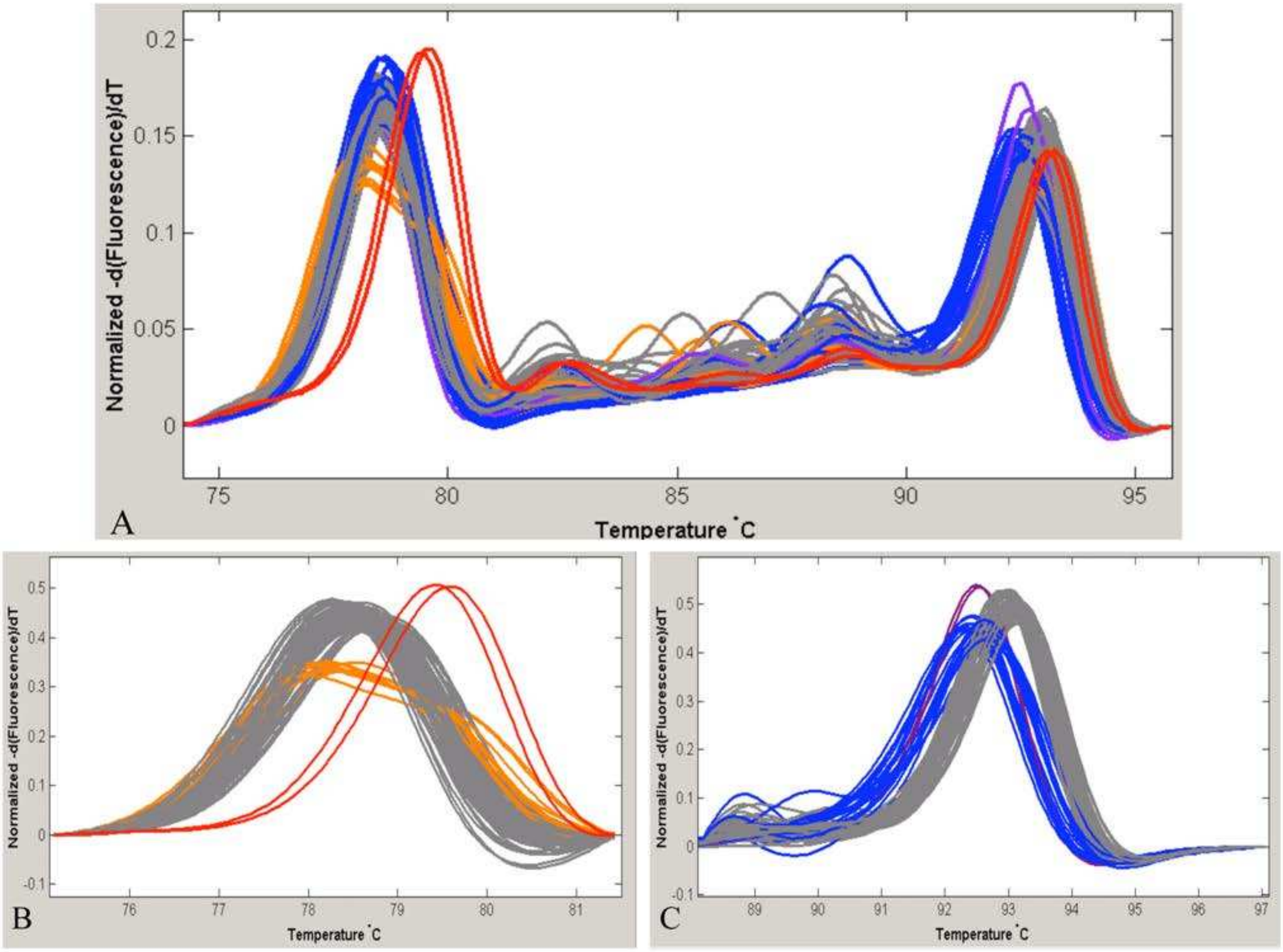


Figure 6

An amplicon containing rs9903378 and an amplicon containing rs9894946 were amplified in a multiplex PCR. A The melting curves of the different genotype combinations of the two SNPs show different profiles. In gray TT-CC, in blue TT-CT, in red GG-CC, in orange TG-CC and in purple TT-TT, respectively for rs9903378 and rs9894946. **B** The analysis was performed in the region of melting temperature of rs9903378 (75–80°C). in gray TT, in orange TG in red GG. **C** The analysis was performed in the region of melting temperature of rs9894946 (90–95°C). in gray CC, in blue CT in purple TT.

succeeded to simultaneously genotype two SNPs located approximately 200 bp apart from each other (SNP rs858528 and rs1641549), thus reducing the number of PCR reactions and improving time and cost effectiveness.

In mutation scanning mode, HRM tended to call as "variant" some DNA samples that actually were wild type. To minimize the frequency of false positive "variant" calls, it is recommended to standardize DNA preparation, storage methods, and storage conditions. Because the sensitivity and specificity of HRM are exquisitely dependent on the melting temperature of each individual sample, variation in salt or buffer concentration carried into PCR reactions along with substrate DNA can generate heterogeneous melting profiles. If needed, to reduce sample-to-sample heterogeneity, it can also be useful to perform a nested PCR and the HRM assay on the secondary PCR.

In the course of our work, we have observed one potential weakness in HRM: the technique may have limited sensitivity for single nucleotide insertion-deletion variants located immediately adjacent to mononucleotide runs of sufficient length that they stutter during PCR. In this work, SNP rs17551157, an insertion of a cytosine adjacent to a 7 cytosine repeat within TP53 proximal promoter, was undetectable in a 653 bp HRM amplicon. A similar situation was also encountered during a large-scale case/control mutation scanning of *ATM* performed in our laboratory, where insertion of an adenosine adjacent to an intronic run of 10 thymines (rs3218681) was also undetectable by HRM mutation scanning. In both cases, the sequencing chromatograms revealed PCR stuttering at the mononucleotide run (data not shown). Such mononucleotide runs are relatively uncommon within the ORF of protein coding genes. Nonetheless, we suggest that at the outset of an HRM based mutation-screening project, investigators check the ORF of the gene of interest for mononucleotide runs that could create such a problem. If any are present, apply a different mutation screening technique to the relevant amplicon.

From our extensive mutation screening of the *TP53* locus, we found that HRM provides sensitive assays both for detection of new sequence variants and genotyping of known polymorphisms. Table 2 summarizes the various HRM approaches for the different genetic contexts that we have considered, according to the number of SNPs present in each amplicon and to their frequencies in the studied population. In our experience, selecting an appropriate HRM analysis strategy depends both on study size and the number of known common polymorphisms in a given amplicon. For relatively small mutation screening studies, it may be reasonable to sequence all samples that appear to contain a sequence variant. In this case, amplicons known (or found) to contain a common variant can be PCR amplified in duplicate, once as a standard analysis and once as a spike-in analysis. The former will detect the presence of heterozygous variants and the latter will detect the presence of minor allele homozygotes. All samples with HRM curves that differ from the major allele homozygotes curves would then be queued for sequencing.

In large-scale mutation screening studies, there may be cost benefit to enabling HRM determination of common SNP genotypes prior to mutation scanning, so that only samples showing a variant HRM curve not attributable to the presence of a common SNP are queued for sequencing. If an amplicon contains a common variant, this variant can mask the presence of a rare variant that might have the same melting profile. Without a discrimination step, one has to either 1) sequence all the heterozygous samples, even though most will be due to a common SNP or 2) accept failure to detect rare variants that have the same melting profile as the common SNP. Inclusion of a discrimination step, which can be achieved with little added cost and with no extra PCR reactions, allows assigning the common SNP by genotyping and simultaneously queuing the rare variant heterozygotes for identification by sequencing.

Conclusion

HRM is a simple and cost effective post-PCR technique that can be used for high-throughput mutation scanning and genotyping in a small laboratory environment. It is inexpensive, flexible, and only mildly constrained by primer design. HRM reactions are closed-tube, which reduces risk of contamination. In addition, HRM assays are non-destructive so that the actual sample used in mutation scanning can serve as a sequencing template.

Abbreviations

dHPLC: denaturated High-Performance Liquid Chromatography; HRM: High-Resolution Melting curve analysis; SNP: Single Nucleotide Polymorphisms.

Authors' contributions

SG carried out the experiments and drafted the manuscript; FG participated in its coordination and helped to draft the manuscript; CV developed the Laboratory Information Management Systems (LIMS); TND participated in design of the unlabeled 3' blocked probe assays for each common SNP; FLK participated in the development of the laboratory workflow; DD participated in the experiment design; FL participated in its coordination and helped to draft the manuscript; SL participated in its coordination and helped to draft the manuscript; SVT conceived the study, participated in its design and

Table 2: Summary of different approaches utilized for each genetic situations.

Genetic situation	HRM solutions
No common SNP in the amplicon	Direct HRM analysis
Only one common SNP in the amplicon	Sometimes it is possible to distinguish directly the three groups (Figure 1) Otherwise, spike-in or unlabelled probe is needed (Figure 2)
Two SNPs in the amplicon, one is common the other is rare	Spike-in when possible (Figure 3); otherwise, use unlabeled probe for the common SNP, and perform mutation scanning for the rare one (Figure 4)
Two or more common SNPs in the amplicon Two SNPs in two different amplicons	Use an unlabeled probe for each SNP (Figure 5) When T_m of 2 amplicons is different, it is possible to perform multiplex PCR (Figure 6)

coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Primers sequences for TP53 *SNPs analyzed in the present study. The data provided correspond to the oligonucleotide sequences used to analyze* TP53 *SNPs.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2156-10-5-S1.doc]

Acknowledgements

SG is a recipient of a fellowship from the University of Pisa and a Special Training Award from the International Agency for Research on Cancer. TN is a recipient of a fellowship from the Fondation de France. We would like to thank Dr Pierre Hainaut for providing breast cancer cell-lines and Dr Maria Isabel Waddington Achatz for supplying us with Li Fraumeni samples. We also thank Ms Nathalie Forey and Ms Sandrine McKay-Chopin for their technical assistance.

References

1. Nickerson DA, Rieder MJ, Crawford DC, Carlson CS, Livingston RJ: **An overview of the environmental genome project.** *Essays on the Future of Environmental Health Research: A Tribute to Dr Keneth Olden* 2005:42-53.
2. Guthery SL, Salisbury BA, Pungliya MS, Stephens JC, Bamshad M: **The structure of common genetic variation in United States populations.** *Am J Hum Genet* 2007, **81:**1221-1231.
3. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI: **Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms.** *Am J Hum Genet* 2008, **82:**100-112.
4. Lin HF, Juo SH, Cheng R: **Comparison of the power between microsatellite and single-nucleotide polymorphism markers for linkage and linkage disequilibrium mapping of an electrophysiological phenotype.** *BMC Genet* 2005, **6(Suppl 1):**S7.
5. Papachristou C, Lin S: **Microsatellites versus Single-Nucleotide Polymorphisms in confidence interval estimation of disease loci.** *Genet Epidemiol* 2006, **30:**3-17.
6. Gray IC, Campbell DA, Spurr NK: **Single nucleotide polymorphisms as tools in human genetics.** *Hum Mol Genet* 2000, **9:**2403-2408.
7. Dong LM, Potter JD, White E, Ulrich CM, Cardon LR, Peters U: **Genetic susceptibility to cancer: the role of polymorphisms in candidate genes.** *JAMA* 2008, **299:**2423-2436.
8. Lamba JK, Crews K, Pounds S, Schuetz EG, Gresham J, Gandhi V, Plunkett W, Rubnitz J, Ribeiro R: **Pharmacogenetics of deoxycytidine kinase: identification and characterization of novel genetic variants.** *J Pharmacol Exp Ther* 2007, **323:**935-945.
9. Oldenburg J, Bevans CG, Fregin A, Geisen C, Muller-Reible C, Watzka M: **Current pharmacogenetic developments in oral anticoagulation therapy: the influence of variant VKORC1 and CYP2C9 alleles.** *Thromb Haemost* 2007, **98:**570-578.
10. Zhu X, Yan D, Cooper RS, Luke A, Ikeda MA, Chang YP, Weder A, Chakravarti A: **Linkage disequilibrium and haplotype diversity in the genes of the renin-angiotensin system: findings from the family blood pressure program.** *Genome Res* 2003, **13:**173-181.
11. Yan D, Ouyang XM, Zhu X, Du LL, Chen ZY, Liu XZ: **Refinement of the DFNA41 locus and candidate genes analysis.** *J Hum Genet* 2005, **50:**516-522.
12. Meaburn E, Butcher LM, Schalkwyk LC, Plomin R: **Genotyping pooled DNA using 100 K SNP microarrays: a step towards genomewide association scans.** *Nucleic Acids Res* 2006, **34:**e27.
13. Rauch A, Rüschendorf F, Huang J, Trautmann U, Becker C, Thiel C, Jones KW, Reis A, Nürnberg P: **Molecular karyotyping using an SNP array for genomewide genotyping.** *J Med Genet* 2004, **41:**916-922.
14. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA: **Direct selection of human genomic loci by microarray hybridization.** *Nat Methods* 2007, **4:**903-905.
15. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME: **Microarray-based genomic selection for high-throughput resequencing.** *Nat Methods* 2007, **4:**907-909.
16. Mariadason JM, Augenlicht LH, Arango D: **Microarray analysis in the clinical management of cancer.** *Hematol Oncol Clin North Am* 2003, **17:**377-387.
17. Kashiwagi H, Uchida K: **Genome-wide profiling of gene amplification and deletion in cancer.** *Hum Cell* 2000, **13:**135-141.
18. Wade MM, Volokhov D, Peredelchuk M, Chizhikov V, Zhang Y: **Accurate mapping of mutations of pyrazinamide-resistant Mycobacterium tuberculosis strains with a scanning-frame oligonucleotide microarray.** *Diagn Microbiol Infect Dis* 2004, **49:**89-97.
19. Skot L, Humphreys J, Humphreys MO, Thorogood D, Gallagher J, Sanderson R, Armstead IP, Thomas ID: **Association of candidate genes with flowering time and water-soluble carbohydrate content in Lolium perenne (L.).** *Genetics* 2007, **177:**535-547.
20. Tavtigian SV, Le Calvez-Kelm F: **Molecular Diagnostics: Methods and Limitations.** In *Hereditary Breast Cancer* Informa healthcare.Isaacs and Rebbeck.; 2008:179-205.
21. Krypuy M, Ahmed AA, Etemadmoghadam D, Hyland SJ, Australian Ovarian Cancer Study Group, DeFazio A, Fox SB, Brenton JD, Bowtell DD, Dobrovic A: **High resolution melting for mutation scanning of TP53 exons 5–8.** *BMC Cancer* 2007, **7:**168.
22. Bastien R, Lewis TB, Hawkes JE, Quackenbush JF, Robbins TC, Palazzo J, Perou CM, Bernard PS: **High-throughput amplicon scanning of the TP53 gene in breast cancer using high-resolution fluorescent melting curve analyses and automatic mutation calling.** *Hum Mutat* 2008, **29:**757-764.
23. Millat G, Chanavat V, Rodriguez-Lafrasse C, Rousson R: **Rapid, sensitive and inexpensive detection of SCN5A genetic variations by high resolution melting analysis.** *Clin Biochem* 2008.
24. Takano EA, Mitchell G, Fox SB, Dobrovic A: **Rapid detection of carriers with BRCA1 and BRCA2 mutations using high resolution melting analysis.** *BMC Cancer* 2008, **8:**59.
25. Audrezet MP, Dabricot A, Le MC, Ferec C: **Validation of high-resolution DNA melting analysis for mutation scanning of the cystic fibrosis transmembrane conductance regulator (CFTR) gene.** *J Mol Diagn* 2008, **10:**424-434.
26. Laurie AD, George PM: **Evaluation of high-resolution melting analysis for screening the LDL receptor gene.** *Clin Biochem* 2008.
27. Sarvary E, Nagy P, Benjamin A, Szoke M, Remport A, Jansen J, Nemes B, Kobori L, Fehervari I, Sulyok B, Perner F, Varga M, Fazakas J, Lakatos M, Szabo M, Toth A, Járay J: **Mutation scanning of the p53 tumor suppressor gene in renal and liver transplant patients in Hungary.** *Transplant Proc* 2005, **37:**969-972.
28. Voegele C, Tavtigian SV, de Silva D, Cuber S, Thomas A, Le Calvez-Kelm F: **A Laboratory Information Management System (LIMS) for a high throughput genetic platform aimed at candidate gene mutation screening.** *Bioinformatics* 2007, **23:**2504-2506.
29. Zhou L, Myers AN, Vandersteen JG, Wang L, Wittwer CT: **Closed-tube genotyping with unlabeled oligonucleotide probes and a saturating DNA dye.** *Clin Chem* 2004, **50:**1328-1335.
30. Zhou L, Wang L, Palais R, Pryor R, Wittwer CT: **High-resolution DNA melting analysis for simultaneous mutation scanning and genotyping in solution.** *Clin Chem* 2005, **51:**1770-1777.
31. Montgomery J, Wittwer CT, Palais R, Zhou L: **Simultaneous mutation scanning and genotyping by high-resolution DNA melting analysis.** *Nat Protoc* 2007, **2:**59-66.

Annex II – Article VI

Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer susceptibility: results from a breast cancer family registry case-control mutation-screening study

Le Calvez-Kelm F, Lesueur F, Damiola F, Vallée M, Voegele C, Babikyan D, Durand G, Forey N, McKay-Chopin S, Robinot N, **Nguyen-Dumont T**, Thomas A, Byrnes GB, Breast Cancer Family Registry, Hopper JL, Southey MC, Andrulis IL, John EM, Tavtigian SV.

Breast Cancer Res. 2011 Jan 18;13(1):R6

Breast Cancer
RESEARCH

RESEARCH ARTICLE Open Access

Rare, evolutionarily unlikely missense substitutions in *CHEK2* contribute to breast cancer susceptibility: results from a breast cancer family registry case-control mutation-screening study

Florence Le Calvez-Kelm[1†], Fabienne Lesueur[1†], Francesca Damiola[1], Maxime Vallée[1], Catherine Voegele[1], Davit Babikyan[2], Geoffroy Durand[1], Nathalie Forey[1], Sandrine McKay-Chopin[1], Nivonirina Robinot[1], Tù Nguyen-Dumont[1], Alun Thomas[3], Graham B Byrnes[1], Breast Cancer Family Registry[4,5,6], John L Hopper[4], Melissa C Southey[7], Irene L Andrulis[5], Esther M John[6,8], Sean V Tavtigian[9*]

Abstract

Introduction: Both protein-truncating variants and some missense substitutions in *CHEK2* confer increased risk of breast cancer. However, no large-scale study has used full open reading frame mutation screening to assess the contribution of rare missense substitutions in *CHEK2* to breast cancer risk. This absence has been due in part to a lack of validated statistical methods for summarizing risk attributable to large numbers of individually rare missense substitutions.

Methods: Previously, we adapted an *in silico* assessment of missense substitutions used for analysis of unclassified missense substitutions in *BRCA1* and *BRCA2* to the problem of assessing candidate genes using rare missense substitution data observed in case-control mutation-screening studies. The method involves stratifying rare missense substitutions observed in cases and/or controls into a series of grades ordered *a priori* from least to most likely to be evolutionarily deleterious, followed by a logistic regression test for trends to compare the frequency distributions of the graded missense substitutions in cases versus controls. Here we used this approach to analyze *CHEK2* mutation-screening data from a population-based series of 1,303 female breast cancer patients and 1,109 unaffected female controls.

Results: We found evidence of risk associated with rare, evolutionarily unlikely *CHEK2* missense substitutions. Additional findings were that (1) the risk estimate for the most severe grade of *CHEK2* missense substitutions (denoted C65) is approximately equivalent to that of *CHEK2* protein-truncating variants; (2) the population attributable fraction and the familial relative risk explained by the pool of rare missense substitutions were similar to those explained by the pool of protein-truncating variants; and (3) *post hoc* power calculations implied that scaling up case-control mutation screening to examine entire biochemical pathways would require roughly 2,000 cases and controls to achieve acceptable statistical power.

Conclusions: This study shows that *CHEK2* harbors many rare sequence variants that confer increased risk of breast cancer and that a substantial proportion of these are missense substitutions. The study validates our analytic approach to rare missense substitutions and provides a method to combine data from protein-truncating variants and rare missense substitutions into a one degree of freedom per gene test.

* Correspondence: sean.tavtigian@hci.utah.edu
† Contributed equally
[9]Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah School of Medicine, 2000 Circle of Hope, Salt Lake City, UT 84112, USA
Full list of author information is available at the end of the article

Introduction

Familial clustering of breast cancer is well recognized, having been described over 140 years ago [1]; the familial relative risk of breast cancer is on average about twofold and is higher among relatives of patients with early-onset cases [2,3]. Three classes of breast cancer susceptibility sequence variants with different levels of risk and prevalence in the population are now well established [4,5]: rare high-risk variants, such as protein-truncating mutations in *BRCA1*, *BRCA2*, *PTEN* and *TP53* (Mendelian Inheritance in Man numbers (MIMs) 113705, 600185, 601728 and 191170, respectively); rare intermediate-risk variants, such as protein-truncating mutations in *ATM* [6,7], *BRIP1* [8], *CHEK2* [9] and *PALB2* [10,11] (MIMs 208900, 605882, 604373 and 610355 respectively); and, more recently, common modest penetrance variants such as the risk single-nucleotide polymorphisms (SNPs) detected by genome-wide association studies (GWASs) in *FGFR2*, *TOX3* (*TNRC9*), *MAP3K1* and *LSP1* [12-14] (MIMs 176943, 611416, 600982 and 153432, respectively). High-risk variants in the known major breast cancer susceptibility genes *BRCA1*, *BRCA2*, *TP53* and *PTEN* account for approximately 20% to 25% of the familial risk of breast cancer, and adding the known intermediate-risk genes increases the proportion by perhaps 1% for each gene [15]. Moreover, the panoply of known modest-risk SNPs account for about 8% of the familial relative risk [16]. Thus known genetic effects account for about one-third of the familial relative risk of breast cancer, leaving two-thirds unaccounted for, a phenomenon referred to as the "problem of missing heritability." Some of this so-called missing "heritability" is of course due to the familial component of environmental risk factors; the measured surrogates for these factors probably explain about 5% of the familial relative risk, but if measured more specifically and more precisely, they may explain considerably more familial aggregation [17].

The gene *CHEK2* encodes a serine/threonine kinase, CHK2, that functions in the signaling pathways activated by DNA damage, particularly DNA double-stranded breaks [18]. Inheritance of a *CHEK2* protein-truncating mutation such as the relatively well investigated Northern European founder mutation *c.1100delC* confers a two- to threefold increased risk of breast cancer, an increased risk of a number of other cancer types and perhaps a decreased risk of some smoking-related cancers [9,19-21]. Some missense substitutions in *CHEK2* also alter cancer risk, as exemplified by the Ashkenazi *CHEK2* missense substitution p.S428F and the Slavic substitution p.I157T [22-26]. Most large-scale genetic studies of *CHEK2* conducted to date have focused on genotyping known variants, such as founder mutations. Consequently, there has been little opportunity to assess the role of the potentially more numerous, rarer variants of this gene.

During the 1990s, linkage analysis proved to be an effective genome-wide approach for finding high-risk susceptibility genes for breast and colon cancer. Over the past few years, GWASs have proved to be an effective genome-wide approach to finding common, not necessarily causal, SNPs associated with modest risk. Case-control mutation screening, or its quantitative trait homolog of comparative mutation screening of individuals from the opposite ends of a trait spectrum, is emerging as a useful strategy for identifying and characterizing intermediate-risk susceptibility genes [6-8,10,27-29]. While case-control mutation screening has been, to date, too technically demanding to examine a whole biochemical pathway, let alone the entire exome, one can imagine combining exon hybridization capture and massively parallel sequencing to accomplish such a study design. Beyond the laboratory challenge imposed by the implied scale of resequencing, a second challenge is to conduct a statistically powerful analysis of the large number of rare sequence variants that would be revealed if such a study design were applied to a common disease such as breast or colon cancer. Previously, we used data from mutation screening of *ATM* in breast cancer patients and controls to demonstrate the ability to detect evidence of pathogenicity from both truncating and splice junction variants (T+SJV) and rare missense substitutions (rMS) [7]. Here we apply the same analytic strategy to *CHEK2* and then extrapolate the results to determine the requirements for much larger-scale studies.

Materials and methods

Ethics statement

The *CHEK2* mutation-screening studies and analyses described here were approved by the institutional review board (IRB) of the International Agency for Research on Cancer, the University of Utah IRB and the local IRBs of the Breast Cancer Family Registry (Breast CFR) centers from which we received samples. All participants gave written, informed consent.

Subjects

Patients were selected from among women gathered by population-based sampling by the Breast CFRs at three centers (Cancer Care Ontario, the Cancer Prevention Institute of California (formerly the Northern California Cancer Center) and the University of Melbourne) [30]. Patients were recruited between 1995 and 2005.

Selection criteria for cases (N = 1,313) were diagnosis at or before age 45 years and self-reported race or ethnicity plus grandparents' country of origin consistent with

Caucasian, East Asian, Hispanic/Latino or African American racial or ethnic heritage.

The controls (N = 1,123) were frequency matched to cases within each center on racial or ethnic group, with age at selection not more than ± 10 years difference the age range at diagnosis of the patients gathered from the same center. Because of the shortage of available controls in some ethnic and age groups, the frequency matching was not one-to-one in all subgroups.

Mutation screening

Mutation screening started from whole-genome amplified (WGA) DNA for coding exons 1-9 and from genomic DNA for exons 10-14. A nested polymerase chain reaction (PCR) strategy was used, followed by high-resolution melting (HRM) curve analysis [31,32] and then dye terminator resequencing of samples that contained a melt curve aberration indicative of the presence of a sequence variant. For *CHEK2* amplicons harboring SNPs with a frequency ≥1% in either the Single Nucleotide Polymorphism Database (dbSNP) [33] or initial amplicon testing, we applied a simultaneous mutation scanning and genotyping approach using HRM curve analysis to improve the sensitivity and efficiency of the mutation screening [34]. The laboratory process used was the same as that described in detail for our recent case-control mutation screening for *ATM* [7], except that primary PCR assays for *CHEK2* exons 10-14 (which are involved in a subtelomeric repeat) relied on a long-range PCR assay as described by Sodha *et al.* [35].

All exonic sequence variants, plus splice junction consensus sequence variants that reduced splice junction sequence similarity to the standard consensus sequences AG^GTRRGT (donor) or Y_{16}NYAG^ (acceptor) (where ^ indicates the position of the splice junction), were reamplified from genomic DNA for confirmation of the presence of the variant. Because of the presence of pseudogenes that partially matched the sequence of the *CHEK2* long-range PCR exons (exons 10-14), sequence variants identified within these exons were subsequently tested using allele-specific PCR assays for the primary PCR to confirm that the sequence variants initially identified were true *CHEK2* variants. To ensure amplification of the *CHEK2* DNA sequence and not amplification of the potentially interfering *CHEK2* pseudogenes, the positions of the specific primers were chosen so that the 3' extremity bases perfectly matched the *CHEK2* wild-type sequence, while they mismatched the corresponding position of the pseudogenes.

All samples that failed at the primary PCR, secondary PCR or sequencing reaction stage were reamplified from WGA DNAs or genomic DNA. Samples that still did not provide satisfactory mutation-screening results for at least 80% of the *CHEK2* coding sequence were excluded from further analyses (n = 24). Process and data management of the mutation screening were carried out as described by Voegele *et al.* [36]. Primer and probe sequences are available from FLCK upon request.

Alignments and scoring of missense substitutions

Previously, we used the T-Coffee (Tree-based Consistency Objective Function for alignment Evaluation) software suite of alignment tools [37,38] to prepare a CHK2 protein multiple sequence alignment in which the most diverged sequence was from sea urchin (*Strongylocentrotus purpuratus*) to analyze a small number of *CHEK2* missense substitutions and in-frame deletions [39]. We updated this alignment by replacing the partial pufferfish (*Tetraodon nigroviridis*) sequence with a full-length zebrafish (*Danio rerio*) sequence and including predicted CHK2 sequences from elephant (*Loxodonta africana*), platypus (*Ornithorhynchus anatinus*), tunicate (*Ciona intestinalis*) and fruit fly (*Drosophila melanogaster*). The alignment was characterized by (1) determining percentage sequence identity between each pair of sequences in the alignment, (2) using the Protpars routine of Phylogeny Inference Package version 3.2 software (PHYLIP; free software developed by Felsenstein [40]) to make a maximum parsimony estimate of the number of substitutions that occurred along each clade of the underlying phylogeny and (3) recording the "median sequence conservation score" reported by the missense substitution analysis program Sorting Intolerant from Tolerant (SIFT) [41,42]. The sequence alignment, or updated versions thereof, is available at the Align-GVGD website [43]. Missense substitutions observed during our mutation screening of *CHEK2* were scored using the Align-GVGD [43-45] and SIFT [41,42] software programs with our curated alignments and with Polymorphism Phenotyping version 2 software, or PolyPhen-2, using its precompiled alignment [46,47].

Statistical analysis and power calculations

To assess risk associations using the case-control frequency distribution of T+SJVs and rMSs, we constructed a single table with one entry per participant; zero or one rare sequence variant per participant; and annotations for type of sequence variants, study center, case-control status, race or ethnicity, and age. For the two participants who carried more than one rare variant of interest (one participant carried p.I448S (C15) plus p. E394D (C35), and one participant carried p.E239K (C15) plus p.R346H (C25)), only the variant belonging to the more likely evolutionarily deleterious grade (that is, higher C-number as scored by Align-GVGD) was considered.

Most analyses were performed using multivariable unconditional logistic regression using Stata version 11 software (StataCorp, College Station, TX, USA). Differences in the case-control ratio between ethnic groups and age categories were accounted for by including categorical variables for each age category and ethnic group. Adjustment was also made for study center. We explored the possibility of interactions between ethnic group and study center, checking both improvement of model fit by the likelihood ratio statistic and comparing the estimates of the parameter of interest (log odds ratio (OR) per Align-GVGD grade) in different models. Adjustment for ethnic group should also capture confounding of genetic and social factors with interaction terms, allowing that this confounding effect may be different for the broadly labeled ethnic groups in different centers. Because the Breast CFR matched cases and controls for age in 5-year categories, and because the maximum age of Breast CFR patients included in this study was 45 years, all participants ages 41 years and older (at diagnosis for patients and at ascertainment for controls) were combined into a single age category.

Logistic regression trend tests were formatted such that participants who did not carry any T+SJV or any rMS, as well as carriers of the seven grades of rMSs (C0, C15, C25, C35, C45, C55 and C65) defined by Align-GVGD [45], were assigned the default row labels 0, 1, 2, 3, 4, 5, 6 and 7, respectively. These row labels were then used as a continuous variable in the logistic regression analyses. Regression coefficients and trend test *P* values (P_{trend}) were estimated from the resulting lognormal ORs using the logit function of Stata software. Carriers of T+SJVs were analyzed against the same noncarrier group defined above. Two strategies were used to combine evidence of association with T+SJV and rMS variants: (1) carriers of T+SJVs were combined with carriers of C65 rMSs in category 7, and (2) T+SJV carriers were assigned row label 8. We used the Fisher's exact test to obtain the lower bound of the 95% confidence interval (95% CI) for associations with categories that contained one or more patients but zero controls.

Post hoc power calculations were performed by specifying a hypothetical OR and population prevalence for each class of variant, together with the cumulative probability of breast cancer prior to age 70 years. The ORs and control carrier frequencies that we specified for the individual grades of sequence variants, relative to the noncarriers, were based on data from the population-based Breast CFR sample series. For the grades for which there were a reasonable number of observations, that is, C0, C15, C25, C65 and T+SJV, we used the adjusted ORs and observed carrier frequencies. Because of the very low numbers of observations in grades C35-C55, ORs for these categories were estimated from the logistic regression OR coefficient and population carrier frequencies defined to obtain the specified OR, given the number of observations in patients. On the basis of these OR and frequency estimates, we calculated expected values and variances of the test statistics for the types of test considered: Pearson's χ^2 test for the two-category tests and the Wald statistic from a logistic regression for the trend test. We then calculated the probability of these statistics exceeding a series of desired *P* value thresholds using a normal approximation.

Attributable fractions were estimated according to the method described by Greenland [48], and familial relative risks were estimated according to the methods described by Goldgar [49]. Both calculations used the same frequency and risk association estimates as those used for the *post hoc* power calculations.

Results

Number of subjects included in the analysis

Of the 2,436 Breast CFR participants, 24 (10 patients and 14 controls) were excluded because their PCR failure rate for *CHEK2* mutation-screening amplicons was greater than 20% (Table 1). The distributions of the remaining cases and controls by age, race or ethnicity, and study center are detailed in Table 2.

Analysis of protein-truncating variants

Full open reading frame mutation screening of *CHEK2* revealed three distinct nonsense substitutions and four distinct small insertion deletion variants that should result in a truncated protein. One of these, *c.1100delC*, a well-known Northern European founder mutation that has been shown beyond any reasonable doubt to confer a moderately increased risk of breast cancer [50], was observed in 11 patients compared with three controls. The other six protein-truncating variants were observed once each, always in a patient (Supplementary Table S1 in Additional file 1). The overall OR associated with T+SJVs was 6.18 (*P* = 0.005) (Table 3). However, as *1100delC* genotyping has already been reported for most of the Breast CFR participants included in this study [50,51], we note that the combination of the other six

Table 1 Participants excluded because of poor mutation-screening performance by study center[a]

Center	Patients, *n* (%)	Controls, *n* (%)
Breast CFR Australia	5 (0.8%)	11 (2.1%)
Breast CFR Canada	1 (0.3%)	2 (0.4%)
Breast CFR Northern California	4 (1.0%)	1 (0.7%)
Total	10 (0.8%)	14 (1.2%)

[a] All 10 excluded patients were <42 years old, and all 14 excluded controls were <45 years old; percentage data are the percentages of the total number of patient or control DNA provided by the indicated Breast CFR center; Breast CFR, Breast Cancer Family Registry.

Table 2 Distribution of patients and controls by age, race or ethnicity, and study center[a]

Distributions	Patients, *n* (%)	Controls, *n* (%)
Age range, yr		
≤30	106 (8.1%)	66 (6.0%)
31-35	322 (24.7%)	171 (15.4%)
36-40	434 (33.3%)	231 (20.8%)
41-45	441 (33.8%)	199 (17.9%)
46-50	0 (0.0%)	230 (20.7%)
51-55	0 (0.0%)	212 (19.1%)
Total	1,303 (100.0%)	1,109 (100.0%)
Race or ethnicity		
Caucasian	843 (64.7%)	956 (86.2%)
East Asian	204 (15.7%)	70 (6.3%)
Latina	158 (12.1%)	47 (4.2%)
Recent African ancestry	98 (7.5%)	36 (3.2%)
Total	1,303 (100.0%)	1,109 (100.0%)
Study center		
Breast CFR Australia	588 (45.1%)	513 (46.3%)
Breast CFR Canada	302 (23.2%)	461 (41.6%)
Breast CFR Northern California	413 (31.7%)	135 (12.2%)
Total	1,303 (100.0%)	1,109 (100.0%)

[a] Patients and controls excluded because of poor mutation-screening performance are not included; percentage data are the percentages of the total number of patient or control DNA in the category indicated that met the mutation-screening quality control criterion; Breast CFR, Breast Cancer Family Registry.

protein-truncating variants was marginally significant by itself (*P* = 0.033), but since none of this set of controls were found to carry such a variant, we could not estimate the OR.

Analysis of rare missense substitutions

In the course of this mutation screening, we observed 34 distinct *CHEK2* missense substitutions (Supplementary Table S1 in Additional file 1). The majority (24 of 34) of these were observed once each. The most common one, p.I448S, was observed 10 times, and none had an overall frequency greater than 1% in this sample series. Overall, 42 of the patients carried one rMS, 2 of the patients carried two rMSs, and 17 controls carried one rMS. Thus, there was a significant excess of rMS carriers among the patients (OR = 2.20, *P* = 0.010).

To analyze the rMSs in more detail, we prepared and characterized a protein multiple sequence alignment containing CHK2 sequences from seven mammals, three additional vertebrates, two additional deuterostomates and one protostomate. Ordering the nonmammalian sequences by decreasing identity to human CHK2 and sequentially assessing overall sequence diversity, the alignment exceeded a maximum parsimony estimate of an average of three substitutions per position upon inclusion of the sea urchin (*Strongylocentrotus purpuratus*) sequence (Supplementary Table S2 in Additional file 1). Three substitutions per position was suggested as a criterion of sequence diversity for analysis of missense substitutions, and we have adopted it as our criterion for use with Align-GVGD in case-control mutation-screening applications [7,52,53].

Using this alignment, we scored the 34 missense substitutions with Align-GVGD [43-45] and SIFT [41,42] (Supplementary Table S1 in Additional file 1). Rather than generating a binary classification, Align-GVGD categorizes missense substitutions into seven grades ordered from evolutionarily most likely (C0) to least likely (C65) [45]. Align-GVGD scored 14 of the rMSs as C0, with 12 patients versus 9 controls carrying a C0 rMS as their highest-grade *CHEK2* variant. The OR for this grade of rMS was near 1.0 (OR, 1.39; 95% CI, 0.55 to 3.56) (Table 3). In contrast, five different rMSs scored as C65, with nine patients versus one control carrying a C65 rMS (again, as their highest-grade *CHEK2* variant). The OR for C65 rMSs was 8.75 (*P* = 0.044) (Table 3). Exploiting the intrinsic ordering of the Align-GVGD grades, we performed a logistic regression test for log-linear OR trends across noncarriers and carriers of the seven grades of rMSs. This test yielded a lognormal OR increase of 0.33/grade (P_{trend} = 0.0055) (Table 4). Thus the statistical evidence in favor of pathogenicity from the trend test was stronger than that generated by either the binary test over all the missense substitutions or the test for any individual grade of missense substitution. These results include adjustments for age category, study center and ethnic group. Neither the removal of the study center nor the inclusion of interactions between center and ethnic group changed the first two digits of these estimates. The interaction terms did not significantly improve the model fit (*P* = 0.18) and were omitted. While removing the study center did not significantly reduce the goodness of fit (*P* = 0.12), this adjustment was retained on the grounds of prior plausibility.

We emphasize that our preplanned rMS analysis was based on rMS grading using Align-GVGD with a *CHEK2* protein multiple sequence alignment having an average of at least three substitutions per position and in which the farthest diverged sequence was from the (deuterostomate) sea urchin (*Strongylocentrotus purpuratus*). Our analysis thus conformed to the conditions under which Align-GVGD was calibrated and was used to grade missense substitutions in *ATM* [7,45]. In addition to the pre-planned Align-GVGD analysis, we carried out corresponding analyses on the basis of rMS grading with SIFT [41,42] and PolyPhen-2 [46,47]. With SIFT, we set up three rMS grades: (1) the program's standard likely neutral grade of SIFT score >0.05, (2) a

Table 3 Analyses of rare variants with missense substitutions stratified by Align-GVGD grade[a]

Class	Patients, *n*	Controls, *n*	Crude OR (95% CI)	Adjusted OR (95% CI)
Noncarriers	1,242	1,089		
T+SJV	17	3	4.97 (1.45 to 17.0)	6.18 (1.76 to 21.8)
Any rMS	44	17	2.27 (1.29 to 4.00)	2.20 (1.20 to 4.01)
rMS stratified by Align-GVGD grade[b]				
C0	12	9	1.17 (0.49 to 2.79)	1.39 (0.55 to 3.56)
C15	14	5	2.46 (0.88 to 6.84)	1.82 (0.62 to 5.34)
C25	7	2	3.07 (0.64 to 14.8)	2.47 (0.45 to 13.49)
C35	1	0	-	
C45	0	0	-	
C55	1	0	-	
C65	9	1	7.89 (1.00 to 62.4)	8.75 (1.06 to 72.2)
rMS stratified by SIFT grade[c]				
S > 0.05	21	8	2.30 (1.02 to 5.22)	1.99 (0.83 to 4.77)
0.05 ≥ S > 0.00	12	5	2.10 (0.74 to 5.99)	1.91 (0.63 to 5.86)
S = 0.00	11	4	2.41 (0.77 to 7.59)	3.03 (0.91 to 10.0)
rMS stratified by PolyPhen-2 grade				
Benign	16	7	2.00 (0.82 to 4.89)	1.69 (0.64 to 4.41)
Possibly D[d]	10	6	1.46 (0.53 to 4.03)	1.65 (0.55 to 4.89)
Probably D[e]	18	4	3.95 (1.33 to 11.7)	3.87 (1.25 to 12.0)

[a] Odds ratios are adjusted for race or ethnicity (Caucasian, East Asian, African American or Latina), study center, and age as categorical variables; OR, odds ratio; 95% CI, 95% confidence interval; T+SJV, protein-truncating variants plus splice junction variant; rMS, rare missense substitution; S, SIFT score; [b]Using the *CHEK2* sequence alignment through *S. purpuratus* (sea urchin); [c]Using the *CHEK2* sequence alignment through *D. melanogaster* (fruit fly); [d]PolyPhen-2 grade "Possibly Damaging"; [e]PolyPhen-2 grade "Probably Damaging."

likely deleterious grade of 0.05 ≥ SIFT score ≥ 0.01, and (3) a more likely deleterious grade of SIFT score 0.00. Using a *CHEK2* alignment in which the farthest diverged sequence was from the (protostomate) fruit fly (*Drosophila melanogaster*), which reached SIFT's median sequence conservation score threshold of 3.25, the OR for the SIFT score 0.00 grade was 3.03 and the logistic regression trend test gave $P_{\text{trend}} = 0.012$ (Table 3). Using the slightly less informative alignment in which the most diverged sequence was from the sea urchin, the logistic regression trend test gave $P_{\text{trend}} = 0.014$ (data not shown). PolyPhen-2 uses a combination of its own precompiled protein multiple sequence alignments and crystal structure information to score missense substitutions. Using PolyPhen-2, we also set up three rMS grades: (1) the program's standard "Benign" grade, (2) its standard "Possibly Damaging" grade, and (3) its standard "Probably Damaging" grade. The OR for the Probably Damaging grade was 3.87, and the logistic regression trend test gave $P_{\text{trend}} = 0.0070$. The rMS grades obtained with SIFT and PolyPhen-2 are also included in Supplementary Table S1 in Additional file 1.

Table 4 Results from logistic regression tests for loglinear odds ratio trends[a]

	Loglinear OR regression coefficient (95% CI) and *P* value	
Grouping of rMS and/or T+SJV	Crude	Adjusted[b]
rMS only (that is, excluding T+SJV)	0.35 (0.12 to 0.58)	0.33 (0.09 to 0.55)
(note that C65 is grade 7)	*P* = 0.0029	*P* = 0.0055
C65 rMS and T+SJV	0.28 (0.14 to 0.43)	0.29 (0.14 to 0.43)
pooled in grade 7	*P* = 0.00013	*P* = 0.000088
C65 rMS in grade 7 and	0.26 (0.12 to 0.39)	0.26 (0.13 to 0.40)
T+SJV in grade 8	*P* = 0.00017	*P* = 0.00011

[a] OR, odds ratio; 95% CI, 95% confidence interval; rMS, rare missense substitution; T+SJV, protein-truncating variants plus splice junction variant; [b]Adjusted for race or ethnicity (Caucasian, East Asian, African American or Latina), study center and age as categorical variables.

One question that arises from this approach to missense substitution analysis is whether the rMSs that drive the difference between patients and controls are truly evolutionarily unlikely, which is shorthand for "subject to purifying selection such that they are disproportionately unlikely ever to become fixed as major alleles." To address this question, we waited until after our primary protein multiple sequence alignment had been created and the rare human missense substitutions had been scored, then we assembled an additional mammalian *CHEK2* gene model (from Guinea pig, *Cavia porcellus*). Insertion of the *C. porcellus* CHK2 sequence into our alignment and comparison with the other placental mammalian CHK2 sequences revealed 34 *C. porcellus*-specific amino acid substitutions (that is, apparently wild-type *C. porcellus* CHK2 amino acid residues that differ from the residues present at that position in the other placental mammalian CHK2 sequences). We then scored these residues with Align-GVGD as if they were amino acid substitutions in the human *CHEK2* sequence. All 34 scored C0, the most evolutionarily likely grade and the grade that contributes least to the difference that we observe between breast cancer patients and controls. Simulating and scoring all possible single-nucleotide substitutions to the canonical human *CHEK2* cDNA sequence, we found that 57.2% of possible missense substitutions are C0. Taking into account differing probabilities of these substitutions due to their underlying sequence contexts as estimated by dinucleotide substitution rate constants [54], 58.6% of a random draw of missense substitutions would be C0. Therefore, ignoring the effects of purifying selection, the probability that 34 of 34 *C. porcellus*-specific substitutions would be C0 is $\sim 0.586^{34} = 1.3 \times 10^{-8}$. Thus selection acts against the rMSs of grade >C0. As these grades have sequentially increasing leverage (toward C65) on the test for trends, evolutionarily unlikely rMSs indeed drive the observed difference between patients and controls.

Combined evidence

Looking forward to candidate gene studies, it could be useful to combine evidence from both T+SJVs and rMSs. The loglinear OR trend test provides a simple mechanism by which to achieve this end: observations of T+SJVs can either be combined with observations of the highest grade of missense substitutions (C65s) or we can add an eighth (even higher) carrier grade for the T+SJVs. For this data set, combining T+SJVs and C65 rMSs in grade 7 appeared to be slightly more effective: lognormal OR increased by 0.29/grade ($P_{\text{trend}} = 8.8 \times 10^{-5}$) as opposed to 0.26/grade ($P_{\text{trend}} = 1.1 \times 10^{-4}$) with the alternative approach. The important point is that the data were less compatible with chance when combined than when they were considered as either T+SJVs or rMSs alone.

Extrapolation to pathway and whole-exome case-control mutation-screening projects

Massively parallel sequencing has evolved to the point where it is being used to identify susceptibility genes for rare diseases, and one can imagine study designs where it could be used to identify or characterize intermediate-risk susceptibility genes for common diseases. Using rare variant carrier frequencies of 0.0045, 0.0018, 0.00021*, 0.00011*, 0.00090 and 0.0027 for the rMS grades C15, C25, C35*, C55*, C65 and T+SJV, respectively, as well as ORs of 1.82, 2.47, 3.74*, 7.24*, 8.75 and 6.18 for the same series of grades, we estimated the number of participants required for a reasonably powered many-gene case-control mutation-screening study. (Note that these frequency and OR values were taken or calculated directly from Tables 3 and 4 unless marked with an asterisk; marked values were estimated from the lognormal OR regression coefficient given in Table 4 and the number of observations in patients.) Setting a Bonferroni-adjusted *P* value threshold of 0.0005 for a study of the ~100 genes in the DNA double-stranded break repair and allied cell cycle checkpoint pathways, we estimate that ~2,000 cases and a similar number of controls would be required for 80% power in a combined analysis of T+SJVs and rMSs (Table 5). An analysis based on T+SJVs alone would require 3,400 each of patients and controls, and an analysis based on rMSs alone would require 4,700 each of patients and controls. Setting a *P* value threshold of 2.5×10^{-6}, which might be considered appropriate for a whole-exome study, 3,350 each of patients and controls would be required for 80% power.

Discussion

That protein-truncating variants in *CHEK2* confer a moderately increased risk of breast cancer is well established. The OR that we observed for T+SJVs is numerically somewhat higher than that reported in the 2004 CHEK2 Breast Cancer Case-Control Consortium study of *c.1100delC* [50], but not significantly, as our 95% CIs do include the point estimate from that study. Moreover, as previous studies have observed higher ORs for *c.1100delC* in familial versus sporadic cases and in

Table 5 Number of patients and frequency-matched controls required for various scales of future intermediate-risk gene case-control mutation-screening studies[a]

Study scale	Single genes	Whole pathways[b]	Whole exome[c]
Type I error	0.05	0.0005	2.5×10^{-6}
Power	0.80	0.80	0.80
rMS alone, *n*	1,975	4,700	7,725
T+SJV alone, *n*	1,425	3,400	5,600
rMS plus T+SJV, *n*	850	2,025	3,350

[a] rMS, rare missense substitution; T+SJV, protein-truncating variants plus splice junction variant; [b] Calculated for 100 genes, approximately the gene count of DNA double-stranded break repair and associated cell cycle checkpoints; [c]Calculated for 20,000 genes.

early-onset versus later-onset cases [9,50], we should expect that this study's focus on early-onset breast cancer cases with oversampling of familial cases would result in relatively high OR estimates.

Previous studies have shown that some *CHEK2* missense substitutions are pathogenic, but the scale of their contribution to breast cancer susceptibility relative to that of T+SJVs is not known. Although we hesitate to extrapolate our current data to true population-attributable risks (within the age groups that we sampled) or familial relative risks, the data do provide a basis on which to compare the relative contributions of these two classes of variants. Working from the control carrier frequencies and the OR point estimates (adjusted for race or ethnicity, study center, and age) observed from the population-based Breast CFR sample series, we calculate attributable fractions of 0.014 for T+SJVs as compared with 0.015 for the sum of C15-C65 rMSs. In addition, we calculate a familial relative risk among first-degree relatives of 1.036 for T+SJVs as compared with 1.033 for a product across the C15-C65 rMSs. Thus, as a first approximation, the attributable fractions and familial relative risks of truncating variants and rare missense substitutions are virtually identical. It is important to remember that these attributable fraction and familial relative risk point estimates are inflated compared with those that would be obtained from a population-based study that included patients diagnosed in their 70s or older. In addition, as more than 25% of the T+SJVs observed in this study were nonsense and frame shift mutations other than *c.1100delC*, these data also speak to the importance of full open reading frame mutation screening to observe the majority of genetically relevant sequence variants in this cancer susceptibility gene.

Several of the missense substitutions observed in this study have been subjected to functional assays in one or more published works. For the 14 missense substitutions that Align-GVGD scored C0 and which we would consequently predict to be neutral or nearly so, assay results have been reported for 4 (p.P85L, p.R137Q, p.R180H and p.T323P). Using a *Saccharomyces cerevisiae* Rad53 complementation assay, Shaag *et al.* [22] found that *p. P85L* is equivalent to wild-type *CHEK2.* While Bell *et al.* [55] found this allele to have modestly reduced activity in an *in vitro* kinase function assay, both Bell *et al.* and Shaag *et al.* concluded that the allele is effectively neutral. Sodha *et al.* [39] assayed the *p.R137Q* allele and found that it encodes a protein with normal stability and normal response to DNA damage. Bell *et al.* [55] also assayed the *p.R137Q* allele and found that it has normal kinase activity. In addition, Sodha *et al.* [39] assayed the *p.R180H* allele and found that it encodes a protein with slightly reduced stability but normal response to DNA damage. Thus existing functional assay results for these three variants are consistent with their being either neutral or at most weakly pathogenic. Wu *et al.* [56] found the fourth C0 substitution, p. T323P, to have moderately reduced autophosphorylation and Cdc25C kinase activity. Classification of this substation as C0 is probably a true Align-GVGD error, because the crystal structure of the protein reveals that T323 is located in an α-helix, which would not typically be permissive of substitution to proline. The algorithmic problem is that the atomic composition and polarity of proline (the amino acid side chain characteristics considered by the original Grantham difference [57] and Align-GVGD are atomic composition, polarity and volume) are intermediate between those of threonine and isoleucine, which are the two amino acids observed at position 323 in our alignment. The consequence is that proline is only slightly outside the range of variation represented by these two wild-type residues and is consequently predicted to be neutral or nearly so. Although unpublished, misclassification of substitutions to proline that map within an α-helix is a problem that we have observed before and is an obvious issue to bear in mind when considering missense substitution analyses made using Align-GVGD. p.I157T is perhaps the most interesting of the substitutions observed in our study that have been subjected to functional assays. Align-GVGD scores the variant as C15, indicative of modest evidence in favor of pathogenicity. Initially, Lee *et al.* [58] found that kinase activity of the *p.I157T* allele was comparable to the wild type. More recent studies have reported that the allele is at least partially defective in dimerization and autophosphorylation, binding and phosphorylating Cdc25, and binding *BRCA1* [59-62]. In populations in which *p.I157T* and *c.1100delC* are both present at appreciable frequencies and have been subject to independent risk estimates, *p.I157T* does appear to confer increased risk of breast cancer, but the OR or penetrance associated with the missense substitution appears to be more modest than that associated with the frame shift *c.1100delC* [63]. At the other end of the spectrum, of the five C65 substitutions that we observed, only one, p.R117G, has been subjected to functional assays. Summing across several studies, the protein encoded by this allele is phosphorylated by ATM in response to DNA damage, shows slightly to markedly reduced autophosphorylation, probably fails to oligomerize and has severely compromised kinase activity toward Cdc25C [39,56,62]. Therefore, the *p.R117G* allele encodes a functionally defective protein and is in all likelihood pathogenic. Thus, for the missense substitutions that were observed in our mutation-screening study and subjected to functional assays, there is a qualitative trend toward agreement between the Align-

GVGD classification and the functional assay result, consistent with the trend in ORs that we observed across the Align-GVGD-defined ordered series of missense substitution grades. However, since concordant results between *in silico* assessments and functional assays are not yet considered sufficient for formal clinical classification of missense substitutions observed in *BRCA1* and *BRCA2* [64-66], it does not appear that the state-of-the-art of *CHK2* functional assays has reached the point at which concordant results from an *in silico* assessment and a functional assay would be sufficient for clinically relevant classification of a *CHEK2* missense substitution.

The genetic results described in this work, combined with the above functional assay summary, have implications for potential clinical genetic susceptibility tests that might include *CHEK2* and other genes with similar mutation profiles. In the 2003 American Society of Clinical Oncology Policy Statement Update on Genetic Testing for Cancer Susceptibility, the second and third "indications for genetic testing for cancer susceptibility" were that "2) the genetic test can be adequately interpreted, and 3) the test results will aid in diagnosis or influence the medical or surgical management of the patient or family members at hereditary risk of cancer" (pp. 2398) [67]. With regard to the third criterion, some investigators have argued that in the context of a high-risk family, the difference in risk between carriers and noncarriers of clearly pathogenic *CHEK2* sequence variants is sufficient to justify a difference in cancer surveillance strategies [68-70]. However, our results in addition to similar work regarding *ATM* [7,71] point toward an issue under the second criterion. If roughly one-half of the genetically relevant risk that the test can pick up actually resides in rare missense substitutions that will be considered unclassified variants at their initial detection, it may not currently be possible to adequately interpret the test results. Therefore, while it is now technically feasible to design a massively parallel sequencing-based test that can accurately and relatively inexpensively identify mutations in a panel of breast cancer susceptibility genes that includes *ATM* and *CHEK2* [72], it may be inappropriate to introduce such a test into widespread use before a clinically validated method of assessing unclassified missense substitutions in these genes has been developed.

The rare missense substitution analysis model combining Align-GVGD with the logistic regression test for trends grew out of the *in silico* analysis of missense substitutions that has now become a standard component in the integrated evaluation of unclassified variants in *BRCA1* and *BRCA2* [65,73]. We proposed the model on the basis of clinical *BRCA1* and *BRCA2* mutation-screening data and then demonstrated its effectiveness by an analysis of *ATM* case-control mutation-screening data [7,45]. Thus the *CHEK2* analysis presented here stands as a methodological confirmation of our approach to the inclusion of rare missense substitution data in case-control mutation-screening studies. The logistic regression test for trends that we used also provides a simple approach to combining evidence from rare missense substitutions with evidence from protein-truncating sequence variants to build a more complete and statistically powerful approach to assessing case-control mutation-screening data than would be afforded by either method alone. From a technological perspective, we can envision combining exon capture and massively parallel sequencing to extend case-control mutation screening to entire biochemical pathways and beyond. On the basis of our *post hoc* power calculations, at least 2,000 patients and 2,000 controls would be required for a whole pathway (such as DNA double-stranded break repair and allied cell cycle checkpoints) study, and 3,300 patients and 3,300 controls would be required to undertake a whole-exome study. On the one hand, these numbers could be an underestimate because *CHEK2* might be among the most important (in terms of familial relative risk) of the intermediate-risk class of breast cancer susceptibility genes. On the other hand, it could turn out that a test based on observations of evolutionarily unlikely sequence variants has an intrinsically lower false-positive rate than anonymous marker GWASs and consequently would not require a full Bonferroni multiple testing correction to reasonably constrain the rate of false-positive results.

Conclusions

This case-control mutation-screening study of *CHEK2* shows that the gene harbors many different rare pathogenic sequence variants, a substantial proportion of which are missense substitutions. From a clinical perspective, the risk of breast cancer conferred by some pathogenic sequence variants in *CHEK2* may be great enough to be of use in a clinical cancer genetics setting, and we note that the technical capability of offering a multigene breast cancer susceptibility testing panel at relatively low per gene laboratory cost is in place. Yet, our results with both *CHEK2* and *ATM* suggest that such a test would create a severe burden of unclassified missense substitutions and that a large fraction of the genetically relevant risk would reside in those unclassified missense substitutions. Paradoxically, on the basis of the research perspective of susceptibility gene identification and characterization, this study validates our approach to the analysis of rare missense substitutions observed during case-control mutation screening and provides a method to combine data from protein-truncating variants and rare missense substitutions into a one degree of freedom per gene test.

Additional material

Additional file 1: Supplementary Tables S1 and S2. Supplementary Table S1: Missense, nonsense, frame shift, and splice junction variants. Supplementary Table S2: CHEK2 protein multiple sequence alignment characterization.

Abbreviations

Breast CFR: Breast Cancer Family Registry; GWAS: genome-wide association study; IRB: institutional review board; OR: odds ratio; PCR: polymerase chain reaction; rMS: rare missense substitution; SNP: single-nucleotide polymorphism; T+SJV: truncating and splice junction variant; WGA: whole genome amplification.

Acknowledgements

We thank David Goldgar for help with familial relative risk calculations and note that a Microsoft Excel-based familial relative risk calculator is available from him at david.goldgar@hsc.utah.edu. We also thank James McKay for help with mutation screening amplicon design and Mia Hashibe and Deborah Neklason for helpful comments on the manuscript. This work was supported by the National Cancer Institute, National Institutes of Health (NIH) grant R01 CA121245, and Canadian Institute for Health Research (CIHR) grant CRN-87521-IC089832. The Breast CFR was funded by the NIH under RFA-CA-06-503 and through cooperative agreements with Breast CFR members, including Cancer Care Ontario (U01 CA69467), the Northern California Cancer Center (U01 CA69417) and the University of Melbourne (U01 CA69638). The content of this article does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast CFR, nor does the mention of trade names, commercial products or organizations imply endorsement by the U.S. government or the Breast CFR.

Author details

[1]International Agency for Research on Cancer, 150 Cours Albert Thomas, Lyon CEDEX 08, F-69372, France. [2]Laboratory of Cancer Genetics, Center of Medical Genetics and Primary Health Care, 4 Tigran Mets Avenue, Yerevan 375010, Armenia. [3]Department of Internal Medicine, University of Utah School of Medicine, 391 Chipeta Way, Suite D, Salt Lake City, UT 84108, USA. [4]Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, The University of Melbourne, 723 Swanston Street, Melbourne, Victoria 3010, Australia. [5]Cancer Care Ontario, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Department of Molecular Genetics, University of Toronto, 60 Murray Street, Toronto, ON M5T 3L9, Canada. [6]Cancer Prevention Institute of California, 2201 Walnut Avenue, Suite 300, Fremont, CA 94538, USA. [7]Department of Pathology, The University of Melbourne, Medical Building 181, Melbourne, Victoria 3010, Australia. [8]Stanford University School of Medicine, 300 Pasteur Drive, Stanford, CA 94305, USA. [9]Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah School of Medicine, 2000 Circle of Hope, Salt Lake City, UT 84112, USA.

Authors' contributions

FLCK contributed to study design, led the laboratory team and helped to draft the manuscript. FL contributed to study design, led the data analysis and helped to draft the manuscript. FD contributed to the mutation screening and data analysis and helped to refine the laboratory platform. MV contributed to the sequence alignment and data analysis. CV was responsible for data management throughput for the project and helped to refine the laboratory platform. DB contributed to the sequence alignment and method for analysis of rare missense substitutions. GD contributed to the mutation screening and data analysis and helped to refine the laboratory platform. NF contributed to the mutation screening and data analysis and helped to refine the laboratory platform. SMC contributed to the mutation screening and data analysis and helped to refine the laboratory platform. NR contributed to the mutation screening and data analysis and helped to refine the laboratory platform. TND contributed to the sequence alignment and data analysis. AT contributed to statistical analyses and helped to draft the manuscript. GBB contributed to statistical analyses and helped to draft the manuscript. JLH was responsible for patients gathered through the University of Melbourne and helped to draft the manuscript. MCS contributed to study design and contributed to the management of samples obtained through the University of Melbourne. ILA was responsible for patients gathered through Cancer Care Ontario. EMJ was responsible for patients gathered through the Northern California Cancer Center (now the Cancer Prevention Institute of California). SVT was responsible for overall study design, contributed to data analysis and helped to draft the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 2 September 2010 Revised: 24 November 2010
Accepted: 18 January 2011 Published: 18 January 2011

References

1. Broca PP: **Traite des Tumeurs.** Paris: Asselin; 1866.
2. Goldgar DE, Easton DF, Cannon-Albright L, Skolnick MH: **Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands.** *J Natl Cancer Inst* 1994, **86**:1600-1608.
3. Amundadottir LT, Thorvaldsson S, Gudbjartsson DF, Sulem P, Kristjansson K, Arnason S, Gulcher JR, Bjornsson J, Kong A, Thorsteinsdottir U, Stefansson K: **Cancer as a complex phenotype: pattern of cancer distribution within and beyond the nuclear family.** *PLoS Med* 2004, **1**:e65.
4. Stratton MR, Rahman N: **The emerging landscape of breast cancer susceptibility.** *Nat Genet* 2008, **40**:17-22.
5. **Genetic susceptibility.** In *World Cancer Report 2008.* Edited by: Boyle P, Levin B. Lyon, France: International Agency for Research on Cancer (IARC); 2008:182-185.
6. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, North B, Jayatilake H, Barfoot R, Spanova K, McGuffog L, Evans DG, Eccles D, Easton DF, Stratton MR, Rahman N: **ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles.** *Nat Genet* 2006, **38**:873-875.
7. Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F, Byrnes GB, Chuang SC, Forey N, Feuchtinger C, Gioia L, Hall J, Hashibe M, Herte B, McKay-Chopin S, Thomas A, Vallee MP, Voegele C, Webb PM, Whiteman DC, Sangrajrang S, Hopper JL, Southey MC, Andrulis IL, John EM, Chenevix-Trench G: **Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer.** *Am J Hum Genet* 2009, **85**:427-446.
8. Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, Chagtai T, Jayatilake H, Ahmed M, Spanova K, North B, McGuffog L, Evans DG, Eccles D, Breast Cancer Susceptibility Collaboration (UK), Easton DF, Stratton MR, Rahman N: **Truncating mutations in the Fanconi anemia J gene *BRIP1* are low-penetrance breast cancer susceptibility alleles.** *Nat Genet* 2006, **38**:1239-1241.
9. Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M, Elstrodt F, van Duijn C, Bartels C, Meijers C, Schutte M, McGuffog L, Thompson D, Easton D, Sodha N, Seal S, Barfoot R, Mangion J, Chang-Claude J, Eccles D, Eeles R, Evans DG, Houlston R, Murday V, Narod S, Peretz T, CHEK2-Breast Cancer Consortium, *et al*: **Low-penetrance susceptibility to breast cancer due to *CHEK2**1100delC in noncarriers of *BRCA1* or *BRCA2* mutations.** *Nat Genet* 2002, **31**:55-59.
10. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T, Jayatilake H, McGuffog L, Hanks S, Evans DG, Eccles D, Easton DF, Stratton MR: ***PALB2*, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene.** *Nat Genet* 2006, **39**:165-167.
11. Erkko H, Dowty JG, Nikkila J, Syrjakoski K, Mannermaa A, Pylkas K, Southey MC, Holli K, Kallioniemi A, Jukkola-Vuorinen A, Kataja V, Kosma VM, Xia B, Livingston DM, Winqvist R, Hopper JL: **Penetrance analysis of the *PALB2* c.1592delT founder mutation.** *Clin Cancer Res* 2008, **14**:4667-4671.
12. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, *et al*: **Genome-wide association study identifies novel breast cancer susceptibility loci.** *Nature* 2007, **447**:1087-1093.

13. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JFJ, Hoover RN, Thomas G, Chanock SJ: **A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer.** *Nat Genet* 2007, **39**:870-874.
14. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, Masson G, Jakobsdottir M, Thorlacius S, Helgason A, Aben KK, Strobbe LJ, Albers-Akkers MT, Swinkels DW, Henderson BE, Kolonel LN, Le Marchand L, Millastre E, Andres R, Godino J, Garcia-Prats MD, Polo E, Tres A, Mouy M, Saemundsdottir J, Backman VM, Gudmundsson L, Kristjansson K, Bergthorsson JT, Kostic J, *et al*: **Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer.** *Nat Genet* 2007, **39**:865-869.
15. Thompson D, Easton D: **The genetic epidemiology of breast cancer genes.** *J Mammary Gland Biol Neoplasia* 2004, **9**:221-236.
16. Mavaddat N, Pharoah PD, Blows F, Driver KE, Provenzano E, Thompson D, Macinnis RJ, Shah M, Search SO, Easton DF, Antoniou AC: **Familial relative risks for breast cancer by pathological subtype: a population-based cohort study.** *Breast Cancer Res* 2010, **12**:R10.
17. Hopper JL, Carlin JB: **Familial aggregation of a disease consequent upon correlation between relatives in a risk factor measured on a continuous scale.** *Am J Epidemiol* 1992, **136**:1138-1147.
18. Antoni L, Sodha N, Collins I, Garrett MD: **CHK2 kinase: cancer susceptibility and cancer therapy: two sides of the same coin?** *Nat Rev Cancer* 2007, **7**:925-936.
19. Bell DW, Varley JM, Szydlo TE, Kang DH, Wahrer DC, Shannon KE, Lubratovich M, Verselis SJ, Isselbacher KJ, Fraumeni JF, Birch JM, Li FP, Garber JE, Haber DA: **Heterozygous germ line *hCHK2* mutations in Li-Fraumeni syndrome.** *Science* 1999, **286**:2528-2531.
20. Cybulski C, Gorski B, Huzarski T, Masojc B, Mierzejewski M, Debniak T, Teodorczyk U, Byrski T, Gronwald J, Matyjasik J, Zlowocka E, Lenner M, Grabowska E, Nej K, Castaneda J, Medrek K, Szymanska A, Szymanska J, Kurzawski G, Suchy J, Oszurek O, Witek A, Narod SA, Lubinski J: ***CHEK2* is a multiorgan cancer susceptibility gene.** *Am J Hum Genet* 2004, **75**:1131-1135.
21. Cybulski C, Masojc B, Oszutowska D, Jaworowska E, Grodzki T, Waloszczyk P, Serwatowski P, Pankowski J, Huzarski T, Byrski T, Gorski B, Jakubowska A, Debniak T, Wokolorczyk D, Gronwald J, Tarnowska C, Serrano-Fernandez P, Lubinski J, Narod SA: **Constitutional *CHEK2* mutations are associated with a decreased risk of lung and laryngeal cancers.** *Carcinogenesis* 2008, **29**:762-765.
22. Shaag A, Walsh T, Renbaum P, Kirchhoff T, Nafa K, Shiovitz S, Mandell JB, Welcsh P, Lee MK, Ellis N, Offit K, Levy-Lahad E, King MC: **Functional and genomic approaches reveal an ancient *CHEK2* allele associated with breast cancer in the Ashkenazi Jewish population.** *Hum Mol Genet* 2005, **14**:555-563.
23. Laitman Y, Kaufman B, Lahad EL, Papa MZ, Friedman E: **Germline *CHEK2* mutations in Jewish Ashkenazi women at high risk for breast cancer.** *Isr Med Assoc J* 2007, **9**:791-796.
24. Cybulski C, Górski B, Huzarski T, Byrski T, Gronwald J, Debniak T, Wokolorczyk D, Jakubowska A, Kowalska E, Oszurek O, Narod SA, Lubinski J: ***CHEK2*-positive breast cancers in young Polish women.** *Clin Cancer Res* 2006, **12**:4832-4835.
25. Cybulski C, Wokolorczyk D, Kladny J, Kurzawski G, Suchy J, Grabowska E, Gronwald J, Huzarski T, Byrski T, Gorski B, D Ecedil Bniak T, Narod SA, Lubinski J: **Germline *CHEK2* mutations and colorectal cancer risk: different effects of a missense and truncating mutations?** *Eur J Hum Genet* 2007, **15**:237-241.
26. Brennan P, McKay J, Moore L, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Chow WH, Rothman N, Chabrier A, Gaborieau V, Odefrey F, Southey M, Hashibe M, Hall J, Boffetta P, Peto J, Peto R, Hung RJ: **Uncommon *CHEK2* mis-sense variant and reduced risk of tobacco-related cancers: case-control study.** *Hum Mol Genet* 2007, **16**:1794-1801.
27. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH: **Multiple rare alleles contribute to low plasma levels of HDL cholesterol.** *Science* 2004, **305**:869-872.
28. Kanetsky PA, Rebbeck TR, Hummer AJ, Panossian S, Armstrong BK, Kricker A, Marrett LD, Millikan RC, Gruber SB, Culver HA, Zanetti R, Gallagher RP, Dwyer T, Busam K, From L, Mujumdar U, Wilcox H, Begg CB, Berwick M: **Population-based study of natural variation in the melanocortin-1 receptor gene and melanoma.** *Cancer Res* 2006, **66**:9330-9337.
29. Fernandez L, Milne R, Bravo J, Lopez J, Avilés J, Longo M, Benítez J, Lázaro P, Ribas G: ***MC1R*: three novel variants identified in a malignant melanoma association study in the Spanish population.** *Carcinogenesis* 2007, **28**:1659-1664.
30. John EM, Hopper JL, Beck JC, Knight JA, Neuhausen SL, Senie RT, Ziogas A, Andrulis IL, Anton-Culver H, Boyd N, Buys SS, Daly MB, O'Malley FP, Santella RM, Southey MC, Venne VL, Venter DJ, West DW, Whittemore AS, Seminara D: **The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer.** *Breast Cancer Res* 2004, **6**:R375-R389.
31. Reed GH, Wittwer CT: **Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis.** *Clin Chem* 2004, **50**:1748-1754.
32. Takano EA, Mitchell G, Fox SB, Dobrovic A: **Rapid detection of carriers with *BRCA1* and *BRCA2* mutations using high resolution melting analysis.** *BMC Cancer* 2008, **8**:59.
33. **Single Nucleotide Polymorphism Database (dbSNP).** [http://www.ncbi.nlm.nih.gov/projects/SNP/].
34. Nguyen-Dumont T, Calvez-Kelm FL, Forey N, McKay-Chopin S, Garritano S, Gioia-Patricola L, De Silva D, Weigel R, Sangrajrang S, Lesueur F, Tavtigian SV: **Description and validation of high-throughput simultaneous genotyping and mutation scanning by high-resolution melting curve analysis.** *Hum Mutat* 2009, **30**:884-890.
35. Sodha N, Houlston RS, Williams R, Yuille MA, Mangion J, Eeles RA: **A robust method for detecting *CHK2/RAD53* mutations in genomic DNA.** *Hum Mutat* 2002, **19**:173-177.
36. Voegele C, Tavtigian SV, de Silva D, Cuber S, Thomas A, Le Calvez-Kelm F: **A Laboratory Information Management System (LIMS) for a high throughput genetic platform aimed at candidate gene mutation screening.** *Bioinformatics* 2007, **23**:2504-2506.
37. **T-Coffee Multiple Sequence Alignment Tools.** [http://www.tcoffee.org/Projects_home_page/t_coffee_home_page.html].
38. Wallace IM, O'Sullivan O, Higgins DG, Notredame C: **M-Coffee: combining multiple sequence alignment methods with T-Coffee.** *Nucleic Acids Res* 2006, **34**:1692-1699.
39. Sodha N, Mantoni TS, Tavtigian SV, Eeles R, Garrett MD: **Rare germ line *CHEK2* variants identified in breast cancer families encode proteins that show impaired activation.** *Cancer Res* 2006, **66**:8966-8970.
40. Felsenstein J: **PHYLIP: Phylogeny Inference Package (version 3.2).** *Cladistics* 1989, **5**:164-166.
41. Ng PC, Henikoff S: **Accounting for human polymorphisms predicted to affect protein function.** *Genome Res* 2002, **12**:436-446.
42. **SIFT.** [http://sift.jcvi.org/].
43. **Align-GVGD.** [http://agvgd.iarc.fr/].
44. Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A: **Comprehensive statistical study of 452 *BRCA1* missense substitutions with classification of eight recurrent substitutions as neutral.** *J Med Genet* 2006, **43**:295-305.
45. Tavtigian SV, Byrnes GB, Goldgar DE, Thomas A: **Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications.** *Hum Mutat* 2008, **29**:1342-1354.
46. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7**:248-249.
47. **PolyPhen-2: prediction of functional effects of human nsSNPs.** [http://genetics.bwh.harvard.edu/pph2/].
48. Greenland S: **Applications of stratified analysis methods.** In *Modern Epidemiology.* 2 edition. Edited by: Rothman KJ, Greenland S. Philadelphia: Lippincott-Raven; 1998:281-300.
49. Goldgar DE: **Population aspects of cancer genetics.** *Biochimie* 2002, **84**:19-25.
50. CHEK2 Breast Cancer Case-Control Consortium: **CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies.** *Am J Hum Genet* 2004, **74**:1175-1182.
51. Bernstein JL, Teraoka SN, John EM, Andrulis IL, Knight JA, Lapinski R, Olson ER, Wolitzer AL, Seminara D, Whittemore AS, Concannon P: **The**

*CHEK2*1100delC* allelic variant and risk of breast cancer: screening results from the Breast Cancer Family Registry. *Cancer Epidemiol Biomarkers Prev* 2006, **15**:348-352.

52. Greenblatt MS, Beaudet JG, Gump JR, Godin KS, Trombley L, Koh J, Bond JP: Detailed computational study of p53 and p16: using evolutionary sequence analysis and disease-associated mutations to predict the functional consequences of allelic variants. *Oncogene* 2003, **22**:1150-1163.
53. Cooper GM, Brudno M, Green ED, Batzoglou S, Sidow A: Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res* 2003, **13**:813-820.
54. Lunter G, Hein J: A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics* 2004, **20(Suppl 1)**:I216-I223.
55. Bell DW, Kim SH, Godwin AK, Schiripo TA, Harris PL, Haserlat SM, Wahrer DC, Haiman CA, Daly MB, Niendorf KB, Smith MR, Sgroi DC, Garber JE, Olopade OI, Le Marchand L, Henderson BE, Altshuler D, Haber DA, Freedman ML: Genetic and functional analysis of *CHEK2* (*CHK2*) variants in multiethnic cohorts. *Int J Cancer* 2007, **121**:2661-2667.
56. Wu X, Dong X, Liu W, Chen J: Characterization of *CHEK2* mutations in prostate cancer. *Hum Mutat* 2006, **27**:742-747.
57. Grantham R: Amino acid difference formula to help explain protein evolution. *Science* 1974, **185**:862-864.
58. Lee SB, Kim SH, Bell DW, Wahrer DC, Schiripo TA, Jorczak MM, Sgroi DC, Garber JE, Li FP, Nichols KE, Varley JM, Godwin AK, Shannon KM, Harlow E, Haber DA: Destabilization of *CHK2* by a missense mutation associated with Li-Fraumeni Syndrome. *Cancer Res* 2001, **61**:8062-8067.
59. Cai Z, Chehab NH, Pavletich NP: Structure and activation mechanism of the *CHK2* DNA damage checkpoint kinase. *Mol Cell* 2009, **35**:818-829.
60. Falck J, Mailand N, Syljuåsen RG, Bartek J, Lukas J: The ATM-Chk2-Cdc25A checkpoint pathway guards against radioresistant DNA synthesis. *Nature* 2001, **410**:842-847.
61. Li J, Williams BL, Haire LF, Goldberg M, Wilker E, Durocher D, Yaffe MB, Jackson SP, Smerdon SJ: Structural and functional versatility of the FHA domain in DNA-damage signaling by the tumor suppressor kinase Chk2. *Mol Cell* 2002, **9**:1045-1054.
62. Chrisanthar R, Knappskog S, Lokkevik E, Anker G, Ostenstad B, Lundgren S, Berge EO, Risberg T, Mjaaland I, Maehle L, Engebretsen LF, Lillehaug JR, Lonning PE: *CHEK2* mutations affecting kinase activity together with mutations in *TP53* indicate a functional pathway associated with resistance to epirubicin in primary breast cancer. *PLoS One* 2008, **3**:e3062.
63. Nevanlinna H, Bartek J: The *CHEK2* gene and inherited breast cancer susceptibility. *Oncogene* 2006, **25**:5912-5919.
64. Couch FJ, Rasmussen LJ, Hofstra R, Monteiro AN, Greenblatt MS, de Wind N: Assessment of functional effects of unclassified genetic variants. *Hum Mutat* 2008, **29**:1314-1326.
65. Goldgar DE, Easton DF, Byrnes GB, Spurdle AB, Iversen ES, Greenblatt MS: Genetic evidence and integration of various data sources for classifying uncertain variants into a single model. *Hum Mutat* 2008, **29**:1265-1272.
66. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, Tavtigian SV: Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat* 2008, **29**:1282-1291.
67. American Society of Clinical Oncology: American Society of Clinical Oncology policy statement update: genetic testing for cancer susceptibility. *J Clin Oncol* 2003, **21**:2397-2406.
68. Johnson N, Fletcher O, Naceur-Lombardelli C, dos Santos Silva I, Ashworth A, Peto J: Interaction between *CHEK2*1100delC* and other low-penetrance breast-cancer susceptibility genes: a familial study. *Lancet* 2005, **366**:1554-1557.
69. Byrnes GB, Southey MC, Hopper JL: Are the so-called low penetrance breast cancer genes, *ATM*, *BRIP1*, *PALB2* and *CHEK2*, high risk for women with strong family histories? *Breast Cancer Res* 2008, **10**:208.
70. Narod SA: Testing for *CHEK2* in the cancer genetics clinic: ready for prime time? *Clin Genet* 2010, **78**:1-7.
71. Bernstein JL, Haile RW, Stovall M, Boice JDJ, Shore RE, Langholz B, Thomas DC, Bernstein L, Lynch CF, Olsen JH, Malone KE, Mellemkjaer L, Borresen-Dale AL, Rosenstein BS, Teraoka SN, Diep AT, Smith SA, Capanu M, Reiner AS, Liang X, Gatti RA, Concannon P, WECARE Study Collaborative Group: Radiation exposure, the *ATM* gene, and contralateral breast cancer in the Women's Environmental Cancer and Radiation Epidemiology Study. *J Natl Cancer Inst* 2010, **102**:475-483.
72. Walsh T, Lee MK, Casadei S, Thornton AM, Stray SM, Pennil C, Nord AS, Mandell JB, Swisher EM, King MC: Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proc Natl Acad Sci USA* 2010, **107**:12629-12633.
73. Spurdle AB, Lakhani SR, Healey S, Parry S, Da Silva LM, Brinkworth R, Hopper JL, Brown MA, Babikyan D, Chenevix-Trench G, Tavtigian SV, Goldgar DE: Clinical classification of *BRCA1* and *BRCA2* DNA sequence variants: the value of cytokeratin profiles and evolutionary analysis: a report from the kConFab Investigators. *J Clin Oncol* 2008, **26**:1657-1663.

doi:10.1186/bcr2810
Cite this article as: Le Calvez-Kelm *et al.*: Rare, evolutionarily unlikely missense substitutions in *CHEK2* contribute to breast cancer susceptibility: results from a breast cancer family registry case-control mutation-screening study. *Breast Cancer Research* 2011 **13**:R6.

Bibliography

[Abzhanov et al., 2004] Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ (2004): Bmp4 and morphological variation of beaks in darwin's finches. Science 305 (5689):1462–5.

[Ahlbom et al., 1997] Ahlbom A, Lichtenstein P, Malmström H, Feychting M, Hemminki K, Pedersen NL (1997): Cancer in twins: genetic and nongenetic familial risk factors. J Natl Cancer Inst 89 (4):287–93.

[Ahmed and Rahman, 2006] Ahmed M, Rahman N (2006): Atm and breast cancer susceptibility. Oncogene 25 (43):5906–11.

[Antoni et al., 2007] Antoni L, Sodha N, Collins I, Garrett MD (2007): Chk2 kinase: cancer susceptibility and cancer therapy - two sides of the same coin? Nat Rev Cancer 7 (12):925–36.

[Antoniou et al., 2002] Antoniou AC, Pharoah PDP, McMullan G, Day NE, Stratton MR, Peto J, Ponder BJ, Easton DF (2002): A comprehensive model for familial breast cancer incorporating brca1, brca2 and other genes. Br J Cancer 86 (1):76–83.

[Antoniou et al., 2003] Antoniou A, Pharoah PDP, Narod S, Risch HA, Eyfjord JE, Hopper JL, Loman N, Olsson H, Johannsson O, Borg A, Pasini B, Radice P, Manoukian S, Eccles DM, Tang N, Olah E, Anton-Culver H, Warner E, Lubinski J, Gronwald J, Gorski B, Tulinius H, Thorlacius S, Eerola H, Nevanlinna H, Syrjäkoski K, Kallioniemi OP, Thompson D, Evans C, Peto J, Lalloo F, Evans DG, Easton DF (2003): Average risks of breast and ovarian cancer associated with brca1 or brca2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet 72 (5):1117–30.

[Antoniou and Easton, 2006] Antoniou AC, Easton DF (2006): Models of genetic susceptibility to breast cancer. Oncogene 25 (43):5898–905.

[Auranen et al., 2005] Auranen A, Song H, Waterfall C, Dicioccio RA, Kuschel B, Kjaer SK, Hogdall E, Hogdall C, Stratton J, Whittemore AS, Easton DF, Ponder BAJ, Novik KL, Dunning AM, Gayther S, Pharoah PDP (2005): Polymorphisms in dna repair genes and epithelial ovarian cancer risk. Int J Cancer 117 (4):611–8.

[Azzato et al., 2010] Azzato EM, Lee AJX, Teschendorff A, Ponder BAJ, Pharoah P, Caldas C, Maia AT (2010): Common germ-line polymorphism of c1qa and breast cancer survival. Br J Cancer 102 (8):1294–9.

[Bellini et al., 2010] Bellini I, Pitto L, Marini MG, Porcu L, Moi P, Garritano S, Boldrini L, Rainaldi G, Fontanini G, Chiarugi M, Barale R, Gemignani F, Landi S (2010): Deltan133p53 expression levels in relation to haplotypes of the tp53 internal promoter region. Hum Mutat 31 (4):456–65.

[Berman et al., 1996] Berman DB, Costalas J, Schultz DC, Grana G, Daly M, Godwin AK (1996): A common mutation in brca2 that predisposes to a variety of cancers is found in both jewish ashkenazi and non-jewish individuals. Cancer Res 56 (15):3409–14.

[Boyle and Levin, 2008] Boyle P, Levin B (2008):. World cancer report. WHO Press.

[Bray et al., 2003] Bray NJ, Buckland PR, Owen MJ, O'Donovan MC (2003): Cis-acting variation in the expression of a high proportion of genes in human brain. Hum Genet 113 (2):149–153.

[Brem et al., 2002] Brem RB, Yvert G, Clinton R, Kruglyak L (2002): Genetic dissection of transcriptional regulation in budding yeast. Science 296 (5568):752–5.

[Buckland, 2004] Buckland PR (2004): Allele-specific gene expression differences in humans. Hum Mol Genet 13 Spec No 2:R255–60.

[Chen et al., 2008] Chen X, Weaver J, Bove BA, Vanderveer LA, Weil SC, Miron A, Daly MB, Godwin AK (2008): Allelic imbalance in brca1 and brca2 gene expression is associated with an increased breast cancer risk. Hum Mol Genet 17 (9):1336–1348.

[Cheung et al., 2003] Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS (2003): Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet 33 (3):422–425.

[Chou et al., 2005] Chou LS, Lyon E, Wittwer CT (2005): A comparison of high-resolution melting analysis with denaturing high-performance liquid chromatography for mutation scanning: cystic fibrosis transmembrane conductance regulator gene as a model. Am J Clin Pathol 124 (3):330–338.

[Chun and Gatti, 2004] Chun HH, Gatti RA (2004): Ataxia-telangiectasia, an evolving phenotype. DNA Repair (Amst) 3 (8-9):1187–96.

[Clark et al., 2006] Clark RM, Wagler TN, Quijada P, Doebley J (2006): A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat Genet 38 (5):594–7.

[Collins et al., 1998] Collins FS, Brooks LD, Chakravarti A (1998): A dna polymorphism discovery resource for research on human genetic variation. Genome Res 8 (12):1229–31.

[Conti and Izaurralde, 2005] Conti E, Izaurralde E (2005): Nonsense-mediated mrna decay: molecular insights and mechanistic variations across species. Curr Opin Cell Biol 17 (3):316–25.

[Cookson et al., 2009] Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009): Mapping complex disease traits with global gene expression. Nat Rev Genet 10 (3):184–94.

[Cowles et al., 2002] Cowles CR, Hirschhorn JN, Altshuler D, Lander ES (2002): Detection of regulatory variation in mouse genes. Nat Genet 32 (3):432–437.

[Crockett and Wittwer, 2001] Crockett AO, Wittwer CT (2001): Fluorescein-labeled oligonucleotides for real-time pcr: using the inherent quenching of deoxyguanosine nucleotides. Anal Biochem 290 (1):89–97.

[Cybulski et al., 2007] Cybulski C, Wokołorczyk D, Huzarski T, Byrski T, Gronwald J, Górski B, Debniak T, Masojć B, Jakubowska A, van de Wetering T, Narod SA, Lubiński J (2007): A deletion in chek2 of 5,395 bp predisposes to breast cancer in poland. Breast Cancer Res Treat 102 (1):119–22.

[Dobrowolski et al., 2009] Dobrowolski SF, Hendrickx ATM, van den Bosch BJC, Smeets HJM, Gray J, Miller T, Sears M (2009): Identifying sequence variants in the human mitochondrial genome using high-resolution melt (hrm) profiling. Hum Mutat 30 (6):891–8.

[Easton et al., 2007] Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R, SEARCH collaborators, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, Johnson N, Seal S, Stratton MR, Rahman N, Chenevix-Trench G, Bojesen SE, Nordestgaard BG, Axelsson CK, Garcia-Closas M, Brinton L, Chanock S, Lissowska J, Peplonska B, Nevanlinna H, Fagerholm R, Eerola H, Kang D, Yoo KY, Noh DY, Ahn SH, Hunter DJ, Hankinson SE, Cox DG, Hall P, Wedren S, Liu J, Low YL, Bogdanova N, Schürmann P, Dörk T, Tollenaar RAEM, Jacobi CE, Devilee P, Klijn JGM, Sigurdson AJ, Doody MM, Alexander BH, Zhang J, Cox A, Brock IW, MacPherson G, Reed MWR, Couch FJ, Goode EL, Olson JE, Meijers-Heijboer H, van den Ouweland A, Uitterlinden A, Rivadeneira F, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Hopper JL, McCredie M, Southey M, Giles GG, Schroen C, Justenhoven C, Brauch H, Hamann U, Ko YD, Spurdle AB, Beesley J, Chen X, kConFab, AOCS Management Group, Mannermaa A, Kosma VM, Kataja V, Hartikainen J, Day

NE, Cox DR, Ponder BAJ (2007): Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447 (7148):1087–93.

[Erali et al., 2008] Erali M, Voelkerding KV, Wittwer CT (2008): High resolution melting applications for clinical laboratory medicine. Exp Mol Pathol 85 (1):50–8.

[Fackenthal et al., 2001] Fackenthal JD, Marsh DJ, Richardson AL, Cummings SA, Eng C, Robinson BG, Olopade OI (2001): Male breast cancer in cowden syndrome patients with germline pten mutations. J Med Genet 38 (3):159–64.

[Feinberg and Tycko, 2004] Feinberg AP, Tycko B (2004): The history of cancer epigenetics. Nat Rev Cancer 4 (2):143–53.

[Fransen et al., 2010] Fransen K, Visschedijk MC, van Sommeren S, Fu JY, Franke L, Festen EAM, Stokkers PCF, van Bodegraven AA, Crusius JBA, Hommes DW, Zanen P, de Jong DJ, Wijmenga C, van Diemen CC, Weersma RK (2010): Analysis of snps with an effect on gene expression identifies ube2l3 and bcl3 as potential new risk genes for crohn's disease. Hum Mol Genet 19 (17):3482–8.

[Gabriel et al., 2002] Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002): The structure of haplotype blocks in the human genome. Science 296 (5576):2225–9.

[Gameau et al., 2005] Gameau LJ, Brown LD, Moore MA, E DJ, DemYan WB (2005): Optimization of lighttyper genotyping assays. Biochemica 3.

[Garritano et al., 2009] Garritano S, Gemignani F, Voegele C, Nguyen-Dumont T, Le Calvez-Kelm F, De Silva D, Lesueur F, Landi S, Tavtigian SV (2009): Determining the effectiveness of high resolution melting analysis for snp genotyping and mutation scanning at the tp53 locus. BMC Genet 10:5.

[Gemignani et al., 2004] Gemignani F, Moreno V, Landi S, Moullan N, Chabrier A, Gutiérrez-Enríquez S, Hall J, Guino E, Peinado MA, Capella G, Canzian

F (2004): A tp53 polymorphism is associated with increased risk of colorectal cancer and with reduced levels of tp53 mrna. Oncogene 23 (10):1954–6.

[Gilad et al., 2006] Gilad Y, Oshlack A, Rifkin SA (2006): Natural selection on gene expression. Trends Genet 22 (8):456–61.

[Gilad et al., 2008] Gilad Y, Rifkin SA, Pritchard JK (2008): Revealing the architecture of gene regulation: the promise of eqtl studies. Trends Genet 24 (8):408–15.

[Gompel et al., 2005] Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB (2005): Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in drosophila. Nature 433 (7025):481–7.

[Graham et al., 2005] Graham R, Liew M, Meadows C, Lyon E, Wittwer CT (2005): Distinguishing different dna heterozygotes by high-resolution melting. Clin Chem 51 (7):1295–1298.

[Hammock and Young, 2005] Hammock EAD, Young LJ (2005): Microsatellite instability generates diversity in brain and sociobehavioral traits. Science 308 (5728):1630–4.

[Hanahan and Weinberg, 2000] Hanahan D, Weinberg RA (2000): The hallmarks of cancer. Cell 100 (1):57–70.

[He and Hannon, 2004] He L, Hannon GJ (2004): Micrornas: small rnas with a big role in gene regulation. Nat Rev Genet 5 (7):522–31.

[Hemminki and Dong, 2000] Hemminki K, Dong C (2000): Lifestyle and cancer: protection from a cancer-free spouse. Int J Cancer 87 (2):308–9.

[Herrmann et al., 2006] Herrmann MG, Durtschi JD, Bromley LK, Wittwer CT, Voelkerding KV (2006): Amplicon dna melting analysis for mutation scanning and genotyping: cross-platform comparison of instruments and dyes. Clin Chem 52 (3):494–503.

[Herrmann et al., 2007] Herrmann MG, Durtschi JD, Bromley LK, Wittwer CT, Voelkerding KV (2007): Instrument comparison for heterozygote scanning of single and double heterozygotes: a correction and extension of herrmann et al., clin chem 2006;52:494-503. Clin Chem 53 (1):150–2.

[Honrado et al., 2006] Honrado E, Osorio A, Palacios J, Benitez J (2006): Pathology and gene expression of hereditary breast tumors associated with brca1, brca2 and chek2 gene mutations. Oncogene 25 (43):5837–5845.

[Houlston and Peto, 2004] Houlston RS, Peto J (2004): The search for low-penetrance cancer susceptibility alleles. Oncogene 23 (38):6471–6.

[Isaacs and Rebbeck, 2008] Isaacs C, Rebbeck TR (2008):. Hereditary breast cancer. Informa Healthcare.

[Jaenisch and Bird, 2003] Jaenisch R, Bird A (2003): Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 33 Suppl:245–54.

[Jansen and Nap, 2001] Jansen RC, Nap JP (2001): Genetical genomics: the added value from segregation. Trends Genet 17 (7):388–91.

[Johnson and Porter, 2000] Johnson NA, Porter AH (2000): Rapid speciation via parallel, directional selection on regulatory genetic pathways. J Theor Biol 205 (4):527–42.

[Jordheim and Dumontet, 2007] Jordheim LP, Dumontet C (2007): Review of recent studies on resistance to cytotoxic deoxynucleoside analogues. Biochim Biophys Acta 1776 (2):138–59.

[Jordheim et al., 2008] Jordheim LP, Nguyen-Dumont T, Thomas X, Dumontet C, Tavtigian SV (2008): Differential allelic expression in leukoblast from patients with acute myeloid leukemia suggests genetic regulation of cda, dck, nt5c2, nt5c3, and tp53. Drug Metab Dispos 36 (12):2419–2423.

[King and Wilson, 1975] King MC, Wilson AC (1975): Evolution at two levels in humans and chimpanzees. Science 188 (4184):107–16.

[Knight et al., 2003] Knight JC, Keating BJ, Rockett KA, Kwiatkowski DP (2003): In vivo characterization of regulatory polymorphisms by allele-specific quantification of rna polymerase loading. Nat Genet 33 (4):469–75.

[Kurreeman et al., 2004] Kurreeman FAS, Schonkeren JJM, Heijmans BT, Toes REM, Huizinga TWJ (2004): Transcription of the il10 gene reveals allele-specific regulation at the mrna level. Hum Mol Genet 13 (16):1755–1762.

[Levine, 2002] Levine M (2002): How insects lose their limbs. Nature 415 (6874):848–9.

[Lewin, 2004] Lewin B (2004):. Genes VIII. Pearson Practice Hall.

[Li et al., 2006] Li Y, Grupe A, Rowland C, Nowotny P, Kauwe JSK, Smemo S, Hinrichs A, Tacey K, Toombs TA, Kwok S, Catanese J, White TJ, Maxwell TJ, Hollingworth P, Abraham R, Rubinsztein DC, Brayne C, Wavrant-De Vrièze F, Hardy J, O'Donovan M, Lovestone S, Morris JC, Thal LJ, Owen M, Williams J, Goate A (2006): Dapk1 variants are associated with alzheimer's disease and allele-specific expression. Hum Mol Genet 15 (17):2560–8.

[Liaw et al., 1997] Liaw D, Marsh DJ, Li J, Dahia PL, Wang SI, Zheng Z, Bose S, Call KM, Tsou HC, Peacocke M, Eng C, Parsons R (1997): Germline mutations of the pten gene in cowden disease, an inherited breast and thyroid cancer syndrome. Nat Genet 16 (1):64–7.

[Lichtenstein et al., 2000] Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K (2000): Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from sweden, denmark, and finland. N Engl J Med 343 (2):78–85.

[Liew et al., 2004] Liew M, Pryor R, Palais R, Meadows C, Erali M, Lyon E, Wittwer C (2004): Genotyping of single-nucleotide polymorphisms by high-resolution melting of small amplicons. Clin Chem 50 (7):1156–1164.

[Liew et al., 2006] Liew M, Nelson L, Margraf R, Mitchell S, Erali M, Mao R, Lyon E, Wittwer C (2006): Genotyping of human platelet antigens 1 to 6 and 15 by high-resolution amplicon melting and conventional hybridization probes. J Mol Diagn 8 (1):97–104.

[Liew et al., 2007] Liew M, Seipp M, Durtschi J, Margraf RL, Dames S, Erali M, Voelkerding K, Wittwer C (2007): Closed-tube snp genotyping without labeled probes/a comparison between unlabeled probe and amplicon melting. Am J Clin Pathol 127 (3):341–8.

[Lim et al., 2005] Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM (2005): Microarray analysis shows that some micrornas downregulate large numbers of target mrnas. Nature 433 (7027):769–73.

[Lipsky et al., 2001] Lipsky RH, Mazzanti CM, Rudolph JG, Xu K, Vyas G, Bozak D, Radel MQ, Goldman D (2001): Dna melting analysis for detection of single nucleotide polymorphisms. Clin Chem 47 (4):635–44.

[Lo et al., 2003] Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, Lee MP (2003): Allelic variation in gene expression is common in the human genome. Genome Res 13 (8):1855–1862.

[Maia et al., 2009] Maia AT, Spiteri I, Lee AJX, O'Reilly M, Jones L, Caldas C, Ponder BAJ (2009): Extent of differential allelic expression of candidate breast cancer genes is similar in blood and breast. Breast Cancer Res 11 (6):R88.

[Marcel and Hainaut, 2009] Marcel V, Hainaut P (2009): p53 isoforms - a conspiracy to kidnap p53 tumor suppressor activity? Cell Mol Life Sci 66 (3):391–406.

[Margraf et al., 2006] Margraf RL, Mao R, Highsmith WE, Holtegaard LM, Wittwer CT (2006): Mutation scanning of the ret protooncogene using high-resolution melting analysis. Clin Chem 52 (1):138–141.

[Maston et al., 2006] Maston GA, Evans SK, Green MR (2006): Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet 7:29–59.

[Meijers-Heijboer et al., 2002] Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M, Elstrodt F, van Duijn C, Bartels C, Meijers C, Schutte M, McGuffog L, Thompson D, Easton D, Sodha N, Seal S, Barfoot R, Mangion J, Chang-Claude J, Eccles D, Eeles R, Evans DG, Houlston R, Murday V, Narod S, Peretz T, Peto J, Phelan C, Zhang HX, Szabo C, Devilee P, Goldgar D, Futreal PA, Nathanson KL, Weber B, Rahman N, Stratton MR (2002): Low-penetrance susceptibility to breast cancer due to chek2(*)1100delc in noncarriers of brca1 or brca2 mutations. Nat Genet 31 (1):55–59.

[Meyer et al., 2008] Meyer KB, Maia AT, O'Reilly M, Teschendorff AE, Chin SF, Caldas C, Ponder BAJ (2008): Allele-specific up-regulation of fgfr2 increases susceptibility to breast cancer. PLoS Biol 6 (5):e108.

[Miki et al., 1994] Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W (1994): A strong candidate for the breast and ovarian cancer susceptibility gene brca1. Science 266 (5182):66–71.

[Milani et al., 2007] Milani L, Gupta M, Andersen M, Dhar S, Fryknäs M, Isaksson A, Larsson R, Syvänen AC (2007): Allelic imbalance in gene expression as a guide to cis-acting regulatory single nucleotide polymorphisms in cancer cells. Nucleic Acids Res 35 (5):e34.

[Moffatt et al., 2007] Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E, Heinzmann A, Simma B, Frischer T, Willis-Owen SAG, Wong KCC, Illig T, Vogelberg C, Weiland SK, von Mutius E, Abecasis GR, Farrall M, Gut IG, Lathrop GM, Cookson WOC (2007): Genetic variants regulating ormdl3 expression contribute to the risk of childhood asthma. Nature 448 (7152):470–3.

[Monks et al., 2004] Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs A, Schadt EE (2004): Genetic inheritance of gene expression in human cell lines. Am J Hum Genet 75 (6):1094–105.

[Montgomery et al., 2007] Montgomery J, Wittwer CT, Palais R, Zhou L (2007): Simultaneous mutation scanning and genotyping by high-resolution dna melting analysis. Nat Protoc 2 (1):59–66.

[Morgan et al., 1998] Morgan WF, Corcoran J, Hartmann A, Kaplan MI, Limoli CL, Ponnaiya B (1998): Dna double-strand breaks, chromosomal rearrangements, and genomic instability. Mutat Res 404 (1-2):125–8.

[Morley et al., 2004] Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004): Genetic analysis of genome-wide variation in human gene expression. Nature 430 (7001):743–7.

[Nagy et al., 2004] Nagy R, Sweet K, Eng C (2004): Highly penetrant hereditary cancer syndromes. Oncogene 23 (38):6445–70.

[Narod, 2006] Narod SA (2006): Modifiers of risk of hereditary breast cancer. Oncogene 25 (43):5832–6.

[Nevanlinna and Bartek, 2006] Nevanlinna H, Bartek J (2006): The chek2 gene and inherited breast cancer susceptibility. Oncogene 25 (43):5912–5919.

[Nguyen-Dumont et al., 2009] Nguyen-Dumont T, Calvez-Kelm FL, Forey N, McKay-Chopin S, Garritano S, Gioia-Patricola L, De Silva D, Weigel R, Sangrajrang S, Lesueur F, Tavtigian SV, Breast Cancer Family Registries (BCFR), Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer (kConFab) (2009): Description and validation of high-throughput simultaneous genotyping and mutation scanning by high-resolution melting curve analysis. Hum Mutat 30 (6):884–90.

[Oldenburg et al., 2003] Oldenburg RA, Kroeze-Jansema K, Kraan J, Morreau H, Klijn JGM, Hoogerbrugge N, Ligtenberg MJL, van Asperen CJ, Vasen HFA, Meijers C, Meijers-Heijboer H, de Bock TH, Cornelisse CJ, Devilee P

(2003): The chek2*1100delc variant acts as a breast cancer risk modifier in non-brca1/brca2 multiple-case families. Cancer Res 63 (23):8153–8157.

[Oldenburg et al., 2007] Oldenburg RA, Meijers-Heijboer H, Cornelisse CJ, Devilee P (2007): Genetic susceptibility for breast cancer: how many more genes to be found? Crit Rev Oncol Hematol 63 (2):125–49.

[Oleksiak et al., 2002] Oleksiak MF, Churchill GA, Crawford DL (2002): Variation in gene expression within and among natural populations. Nat Genet 32 (2):261–6.

[Olivier et al., 2003] Olivier M, Goldgar DE, Sodha N, Ohgaki H, Kleihues P, Hainaut P, Eeles RA (2003): Li-fraumeni and related syndromes: correlation between tumor type, family structure, and tp53 genotype. Cancer Res 63 (20):6643–50.

[Palais et al., 2005] Palais RA, Liew MA, Wittwer CT (2005): Quantitative heteroduplex analysis for single nucleotide polymorphism genotyping. Anal Biochem 346 (1):167–175.

[Pant et al., 2006] Pant PVK, Tao H, Beilharz EJ, Ballinger DG, Cox DR, Frazer KA (2006): Analysis of allelic differential expression in human white blood cells. Genome Res 16 (3):331–9.

[Pastinen and Hudson, 2004] Pastinen T, Hudson TJ (2004): Cis-acting regulatory variation in the human genome. Science 306 (5696):647–50.

[Pastinen et al., 2004] Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, Lepage P, Lavergne K, Villeneuve A, Gaudin T, Brandstrom H, Beck A, Verner A, Kingsley J, Harmsen E, Labuda D, Morgan K, Vohl MC, Naumova AK, Sinnett D, Hudson TJ (2004): A survey of genetic and epigenetic variation affecting human gene expression. Physiol Genomics 16 (2):184–193.

[Pennacchio et al., 2006] Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel

A, Rubin EM (2006): In vivo enhancer analysis of human conserved non-coding sequences. Nature 444 (7118):499–502.

[Perera and Bapat, 2007] Perera S, Bapat B (2007):. Genetic instability in cancer.

[Peto et al., 1999] Peto J, Collins N, Barfoot R, Seal S, Warren W, Rahman N, Easton DF, Evans C, Deacon J, Stratton MR (1999): Prevalence of brca1 and brca2 gene mutations in patients with early-onset breast cancer. J Natl Cancer Inst 91 (11):943–9.

[Pharoah and Caldas, 1999] Pharoah PD, Caldas C (1999): Molecular genetics and the assessment of human cancers. Expert Rev Mol Med 1999:1–19.

[Rahman et al., 2007] Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T, Jayatilake H, McGuffog L, Hanks S, Evans DG, Eccles D, Breast Cancer Susceptibility Collaboration (UK), Easton DF, Stratton MR (2007): Palb2, which encodes a brca2-interacting protein, is a breast cancer susceptibility gene. Nat Genet 39 (2):165–7.

[Reed and Wittwer, 2004] Reed GH, Wittwer CT (2004): Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. Clin Chem 50 (10):1748–1754.

[Renwick et al., 2006] Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, North B, Jayatilake H, Barfoot R, Spanova K, McGuffog L, Evans DG, Eccles D, Breast Cancer Susceptibility Collaboration (UK), Easton DF, Stratton MR, Rahman N (2006): Atm mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet 38 (8):873–5.

[Risch, 2001] Risch N (2001): The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. Cancer Epidemiol Biomarkers Prev 10 (7):733–41.

[Rockman and Kruglyak, 2006] Rockman MV, Kruglyak L (2006): Genetics of global gene expression. Nat Rev Genet 7 (11):862–72.

[Ronald et al., 2005] Ronald J, Akey JM, Whittle J, Smith EN, Yvert G, Kruglyak L (2005): Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. Genome Res 15 (2):284–91.

[Rouleau et al., 2009] Rouleau E, Lefol C, Bourdon V, Coulet F, Noguchi T, Soubrier F, Bièche I, Olschwang S, Sobol H, Lidereau R (2009): Quantitative pcr high-resolution melting (qpcr-hrm) curve analysis, a new approach to simultaneously screen point mutations and large rearrangements: application to mlh1 germline mutations in lynch syndrome. Hum Mutat 30 (6):867–75.

[Savitsky et al., 1995] Savitsky K, Bar-Shira A, Gilad S, Rotman G, Ziv Y, Vanagaite L, Tagle DA, Smith S, Uziel T, Sfez S, Ashkenazi M, Pecker I, Frydman M, Harnik R, Patanjali SR, Simmons A, Clines GA, Sartiel A, Gatti RA, Chessa L, Sanal O, Lavin MF, Jaspers NG, Taylor AM, Arlett CF, Miki T, Weissman SM, Lovett M, Collins FS, Shiloh Y (1995): A single ataxia telangiectasia gene with a product similar to pi-3 kinase. Science 268 (5218):1749–53.

[Schadt et al., 2003] Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH (2003): Genetics of gene expression surveyed in maize, mouse and man. Nature 422 (6929):297–302.

[Schroeder et al., 2006] Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T (2006): The rin: an rna integrity number for assigning integrity values to rna measurements. BMC Mol Biol 7:3.

[Seal et al., 2006] Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, Chagtai T, Jayatilake H, Ahmed M, Spanova K, North B, McGuffog L, Evans DG, Eccles D, Breast Cancer Susceptibility Collaboration (UK), Easton DF, Stratton MR, Rahman N (2006): Truncating mutations in the fanconi anemia j gene brip1 are low-penetrance breast cancer susceptibility alleles. Nat Genet 38 (11):1239–41.

[Secko, 2005] Secko D (2005): How an imprint can lead to cancer. CMAJ 172 (10):1286.

[Serre et al., 2008] Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, Bibikova M, Chudin E, Barker DL, Dickinson T, Fan JB, Hudson TJ (2008): Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. PLoS Genet 4 (2):e1000006.

[Shaag et al., 2005] Shaag A, Walsh T, Renbaum P, Kirchhoff T, Nafa K, Shiovitz S, Mandell JB, Welcsh P, Lee MK, Ellis N, Offit K, Levy-Lahad E, King MC (2005): Functional and genomic approaches reveal an ancient chek2 allele associated with breast cancer in the ashkenazi jewish population. Hum Mol Genet 14 (4):555–63.

[Simpson, 1997] Simpson AJ (1997): The natural somatic mutation frequency and human carcinogenesis. Adv Cancer Res 71:209–40.

[Smith et al., 2006] Smith P, McGuffog L, Easton DF, Mann GJ, Pupo GM, Newman B, Chenevix-Trench G, kConFab Investigators, Szabo C, Southey M, Renard H, Odefrey F, Lynch H, Stoppa-Lyonnet D, Couch F, Hopper JL, Giles GG, McCredie MRE, Buys S, Andrulis I, Senie R, BCFS, BRCAX Collaborators Group, Goldgar DE, Oldenburg R, Kroeze-Jansema K, Kraan J, Meijers-Heijboer H, Klijn JGM, van Asperen C, van Leeuwen I, Vasen HFA, Cornelisse CJ, Devilee P, Baskcomb L, Seal S, Barfoot R, Mangion J, Hall A, Edkins S, Rapley E, Wooster R, Chang-Claude J, Eccles D, Evans DG, Futreal PA, Nathanson KL, Weber BL, Breast Cancer Susceptibility Collaboration (UK), Rahman N, Stratton MR (2006): A genome wide linkage search for breast cancer susceptibility genes. Genes Chromosomes Cancer 45 (7):646–55.

[Spielman et al., 2007] Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG (2007): Common genetic variants account for differences in gene expression among ethnic groups. Nat Genet 39 (2):226–31.

[Stamatoyannopoulos, 2004] Stamatoyannopoulos JA (2004): The genomics of gene expression. Genomics 84 (3):449–57.

[Stewart and Kleihues, 2003] Stewart BW, Kleihues P (2003):. World cancer report. IARCPress.

[van der Stoep et al., 2009] van der Stoep N, van Paridon CDM, Janssens T, Krenkova P, Stambergova A, Macek M, Matthijs G, Bakker E (2009): Diagnostic guidelines for high-resolution melting curve (hrm) analysis: an interlaboratory validation of brca1 mutation scanning using the 96-well lightscanner. Hum Mutat 30 (6):899–909.

[Storey et al., 2007] Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM (2007): Gene-expression variation within and among human populations. Am J Hum Genet 80 (3):502–9.

[Stranger et al., 2007] Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavaré S, Deloukas P, Dermitzakis ET (2007): Population genomics of human gene expression. Nat Genet 39 (10):1217–24.

[Stratton and Rahman, 2008] Stratton MR, Rahman N (2008): The emerging landscape of breast cancer susceptibility. Nat Genet 40 (1):17–22.

[Struewing, 2004] Struewing JP (2004): Genomic approaches to identifying breast cancer susceptibility factors. Breast Dis 19:3–9.

[Szymańska and Hainaut, 2003] Szymańska K, Hainaut P (2003): Tp53 and mutations in human cancer. Acta Biochim Pol 50 (1):231–8.

[Takano et al., 2008] Takano EA, Mitchell G, Fox SB, Dobrovic A (2008): Rapid detection of carriers with brca1 and brca2 mutations using high resolution melting analysis. BMC Cancer 8:59.

[Tavtigian et al., 2009] Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F, Byrnes GB, Chuang SC, Forey N, Feuchtinger C, Gioia L, Hall J, Hashibe M, Herte B, McKay-Chopin S, Thomas A, Vallée MP, Voegele C, Webb PM, Whiteman DC, Australian Cancer Study, Breast Cancer Family Registries (BCFR), Kathleen Cuningham Foundation Consortium for Research into Familial Aspects of Breast Cancer (kConFab), Sangrajrang S, Hopper JL, Southey MC, Andrulis IL, John EM, Chenevix-Trench G (2009): Rare, evolutionarily unlikely missense substitutions in atm confer increased risk of breast cancer. Am J Hum Genet 85 (4):427–46.

[Taylor and Byrd, 2005] Taylor AMR, Byrd PJ (2005): Molecular pathology of ataxia telangiectasia. J Clin Pathol 58 (10):1009–15.

[Thompson and Easton, 2004] Thompson D, Easton D (2004): The genetic epidemiology of breast cancer genes. J Mammary Gland Biol Neoplasia 9 (3):221–36.

[Thompson et al., 2005*a*] Thompson D, Antoniou AC, Jenkins M, Marsh A, Chen X, Wayne T, Tesoriero A, Milne R, Spurdle A, Thorstenson Y, Southey M, Giles GG, Khanna KK, Sambrook J, Oefner P, Goldgar D, Hopper JL, Easton D, Chenevix-Trench G, KConFab Investigators (2005*a*): Two atm variants and breast cancer risk. Hum Mutat 25 (6):594–5.

[Thompson et al., 2005*b*] Thompson D, Duedal S, Kirner J, McGuffog L, Last J, Reiman A, Byrd P, Taylor M, Easton DF (2005*b*): Cancer risks and mortality in heterozygous atm mutation carriers. J Natl Cancer Inst 97 (11):813–22.

[Vasemägi and Primmer, 2005] Vasemägi A, Primmer CR (2005): Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. Mol Ecol 14 (12):3623–42.

[Visel et al., 2009] Visel A, Rubin EM, Pennacchio LA (2009): Genomic views of distant-acting enhancers. Nature 461 (7261):199–205.

[Vossen et al., 2009] Vossen RHAM, Aten E, Roos A, den Dunnen JT (2009): High-resolution melting analysis (hrma): more than just sequence variant screening. Hum Mutat 30 (6):860–6.

[Walsh and King, 2007] Walsh T, King MC (2007): Ten genes for inherited breast cancer. Cancer Cell 11 (2):103–5.

[Wang and Sadée, 2006] Wang D, Sadée W (2006): Searching for polymorphisms that affect gene expression and mrna processing: example abcb1 (mdr1). AAPS J 8 (3):E515–20.

[Ware et al., 2006] Ware MD, DeSilva D, Sinilnikova OM, Stoppa-Lyonnet D, Tavtigian SV, Mazoyer S (2006): Does nonsense-mediated mrna decay explain the ovarian cancer cluster region of the brca2 gene? Oncogene 25 (2):323–328.

[Wasserman and Sandelin, 2004] Wasserman WW, Sandelin A (2004): Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5 (4):276–87.

[Webb, 2002] Webb T (2002): Snps: can genetic variants control cancer susceptibility? J Natl Cancer Inst 94 (7):476–8.

[Whitney et al., 2003] Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, Brown PO (2003): Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci U S A 100 (4):1896–901.

[Wilkins et al., 2007] Wilkins JM, Southam L, Price AJ, Mustafa Z, Carr A, Loughlin J (2007): Extreme context specificity in differential allelic expression. Hum Mol Genet 16 (5):537–46.

[Williams et al., 2007] Williams RBH, Chan EKF, Cowley MJ, Little PFR (2007): The influence of genetic variation on gene expression. Genome Res 17 (12):1707–16.

[Wittwer et al., 1997] Wittwer CT, Herrmann MG, Moss AA, Rasmussen RP (1997): Continuous fluorescence monitoring of rapid cycle dna amplification. Biotechniques 22 (1):130–1, 134–8.

[Wittwer et al., 2003] Wittwer CT, Reed GH, Gundry CN, Vandersteen JG, Pryor RJ (2003): High-resolution genotyping by amplicon melting analysis using lcgreen. Clin Chem 49 (6 Pt 1):853–860.

[Wittwer, 2009] Wittwer CT (2009): High-resolution dna melting analysis: advancements and limitations. Hum Mutat 30 (6):857–9.

[Wojdacz and Dobrovic, 2007] Wojdacz TK, Dobrovic A (2007): Methylation-sensitive high resolution melting (ms-hrm): a new approach for sensitive and high-throughput assessment of methylation. Nucleic Acids Res 35 (6):e41.

[Wooster et al., 1995] Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C, Micklem G (1995): Identification of the breast cancer susceptibility gene brca2. Nature 378 (6559):789–92.

[Worm et al., 2001] Worm J, Aggerholm A, Guldberg P (2001): In-tube dna methylation profiling by fluorescence melting curve analysis. Clin Chem 47 (7):1183–1189.

[Wray et al., 2003] Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA (2003): The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20 (9):1377–419.

[Yan et al., 2002*a*] Yan H, Dobbie Z, Gruber SB, Markowitz S, Romans K, Giardiello FM, Kinzler KW, Vogelstein B (2002*a*): Small changes in expression affect predisposition to tumorigenesis. Nat Genet 30 (1):25–26.

[Yan et al., 2002*b*] Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW (2002*b*): Allelic variation in human gene expression. Science 297 (5584):1143.

[Yvert et al., 2003] Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L (2003): Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. Nat Genet 35 (1):57–64.

[Zhou et al., 2004] Zhou L, Myers AN, Vandersteen JG, Wang L, Wittwer CT (2004): Closed-tube genotyping with unlabeled oligonucleotide probes and a saturating dna dye. Clin Chem 50 (8):1328–35.

[CHEK2 Breast Cancer Case-Control Consortium, 2004] CHEK2 Breast Cancer Case-Control Consortium (2004): Chek2*1100delc and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. Am J Hum Genet 74 (6):1175–82.