Published in final edited form as:
Breast Cancer Res Treat. 2012 July ; 134(1): 429–433. doi:10.1007/s10549-012-2057-x.

Predisposition gene identification in common cancers by exome sequencing: insights from familial breast cancer

Katie Snape[1], **Elise Ruark**[1], **Patrick Tarpey**[2], **Anthony Renwick**[1], **Clare Turnbull**[1], **Sheila Seal**[1], **Anne Murray**[1], **Sandra Hanks**[1], **Jenny Douglas**[1], **Michael R. Stratton**[2], and **Nazneen Rahman**[1]

[1]Division of Genetics and Epidemiology, Institute of Cancer Research and Royal Marsden Hospital Foundation Trust, Sutton, Surrey UK

[2]The Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Abstract

The genetic component of breast cancer predisposition remains largely unexplained. Candidate-gene case-control resequencing has identified predisposition genes characterised by rare, protein truncating mutations that confer moderate risks of disease. In theory, exome sequencing should yield additional genes of this class. Here, we explore the feasibility and design considerations of this approach.

We performed exome sequencing in 50 individuals with familial breast cancer, applying frequency and protein function filters to identify variants most likely to be pathogenic. We identified 867,378 variants that passed the call quality filters of which 1,296 variants passed the frequency and protein truncation filters. The median number of validated, rare, protein truncating variants (PTVs) was 10 in individuals with, and without, mutations in known genes. The functional candidacy of mutated genes was similar in both groups. Without prior knowledge, the known genes would not have been recognisable as breast cancer predisposition genes. Everyone carries multiple rare mutations that are plausibly related to disease. Exome sequencing in common conditions will therefore require intelligent sample and variant prioritisation strategies in large case-control studies to deliver robust genetic evidence of disease association.

Keywords

breast cancer predisposition; exome sequencing; common disease genetics; missing heritability

INTRODUCTION

Exome sequencing has proved highly successful in the identification of genes that cause rare Mendelian diseases. In such conditions the underlying genetic model is usually known and the mutational spectrum is distinctive and readily distinguishable from the pattern in unaffected individuals (reviewed in Ku et al. [1]).

Corresponding author: Professor Nazneen Rahman MD PhD, Professor of Human Genetics, The Institute of Cancer Research; 15 Cotswold Road; Sutton SM2 5NG; United Kingdom. nazneen.rahman@icr.ac.uk, Tel: +44 (0)20 8722 4145 Fax: +44 (0)20 8722 4026.

Competing interests: None

Ethics Approval: NHS National Research Ethics Service (MREC/01/2/18).

The identification of rare genetic variants that contribute to common disorders has proved more challenging. The underlying genetic architecture is typically complex and often poorly understood, the penetrance may be modest and/or incomplete, and our ability to robustly predict the impact of genetic variation on gene function and disease causation is still limited. Nevertheless, a component of the missing heritability of many common disorders is likely to reside in rare gene variants of moderate/low penetrance that are potentially tractable by exome sequencing.

Breast cancer is one of the few common conditions for which such variants have already been identified. Using candidate gene case-control resequencing, DNA repair genes such as *CHEK2, PALB2, BRIP1* and *ATM* have been shown to be breast cancer predisposition genes [2-5]. These genes are characterised by multiple, very rare inactivating (primarily truncating) mutations associated with moderate risks of disease (RR 2-4). Together with high-penetrance genes and low-penetrance variants, these moderate-penetrance genes are estimated to account for only ~35% of the familial risk of breast cancer [6]. Thus a substantial proportion of the genetic contribution to breast cancer remains unexplained.

Given that a small number of candidate gene studies have already yielded rare, moderate-penetrance predisposition genes in breast cancer, it is highly likely that other genes of this class exist. Such genes are not identifiable by linkage analyses (the risks are not high enough) nor genome-wide association studies (the mutations are not common enough), but should be detectable by suitably powered exome sequencing studies. Exome sequencing offers the potential to apply an agnostic rather than a candidate gene approach to their discovery and is therefore a highly attractive strategy. However, interrogating the vast datasets generated to provide robust evidence of association of a given gene with breast cancer is daunting. To explore the feasibility of using exome sequencing in the identification of breast cancer susceptibility genes, we sequenced the exomes of 50 individuals with familial breast cancer. We applied frequency and protein truncation filters to prioritise variants most likely to act as moderate-penetrance breast cancer susceptibility alleles, based on existing paradigms [6],[7]. We identified mutations in known breast cancer predisposition genes in four individuals, demonstrating the utility of this approach for mutation detection in already established predisposition genes. We then compared the mutational profiles in these cases with eight individuals without mutations in known genes to investigate the utility in discovering novel disease predisposition genes.

PATIENTS AND METHODS

Full details of the samples and methods are given in the online supplemental material. Briefly, we undertook exome sequencing in 50 individuals recruited to the Familial Breast Cancer Study (FBCS). A summary of the characteristics of these families is given in Table 1, and fuller details are given in online Supplementary Table 1. All individuals had breast cancer and were negative for *BRCA1* and *BRCA2* mutations (by Sanger sequencing and/or heteroduplex analysis and MLPA). We used a commercially available 38 Mb exome array in 30 individuals and a 47.9 Mb custom GENCODE exome array in 20 individuals [8]. Sequencing was performed on an Illumina Genome Analyzer IIx platform. We undertook read mapping and variant analysis using NextGENe software (version 2.10) and applied call quality, frequency and protein truncation filters to prioritise variants for further consideration. We selected 12 cases for detailed analyses; four with mutations in known breast cancer predisposition genes and eight without. We performed validation analyses of all the prioritised variants in the 12 samples by Sanger sequencing. We undertook gene list enrichment analysis using the ToppGene Suite [9].

RESULTS

Exome Sequencing

Overall, a median of 53.5 million reads were generated per sample and typically, 99% of reads mapped to the reference genome. A median of 83% (Range 41%-88%) of bases within the target region had coverage of ≥ 15 per sample (online Supplementary Table 2). There was considerable inter-sample variation because two different exome arrays were used and the sequencing was performed over several months. Overall, we identified 1,592,412 variants in the 50 exomes under NextGENe default settings. 353,948 variants remained after we excluded all variants with read coverage <15 reads, base substitutions with a mutant:wild-type read % of <30%, intronic variants (except those at splice junctions) and synonymous variants. To further prioritise variants most likely to predispose to disease we applied a filter to detect sequence variants that result in protein truncation, as previously described [7]. This identifies all variants predicted to result in premature protein truncation: frameshifting insertions and deletions, nonsense mutations and mutations at consensus splicing residues. The script also removes variants in genes with 5 or more different truncating variants (as these are likely to be pseudogenes or to tolerate haploinsufficiency without causing disease). The filter identified 15,784 truncating variants. To prioritise variants for follow-up we next applied a frequency filter to identify variants present in 1 of the 50 familial breast cancer cases, consistent with the mutation prevalence of known breast cancer predisposition genes [6]. After this filter 1,296 variants remained.

Variant validation in 12 exomes

Within the 1,296 variants we identified four mutations in known predisposition genes which we confirmed by Sanger sequencing; three were in the moderate-penetrance genes *CHEK2* (n=2) and *ATM* (n=1). The fourth was a splicing mutation in *BRCA2* that had evaded detection by heteroduplex analysis, which is recognised to have reduced sensitivity for base substitutions (Table 2).

We undertook Sanger sequencing evaluation of all 316 variants passing all filters in 12 samples (four with mutations in known genes, eight without) in 292 amplicons. Sequencing was successful for 241 amplicons. 51 amplicons failed the automated design and sequencing process. 127 variants (68 base substitutions, 59 indels) were confirmed, although for three variants Sanger sequencing revealed the deletions to be inframe. These were removed from the final analysis as they do not result in premature protein truncation. No variant was detected in the remaining 114 amplicons, i.e. these were false positive calls (23 base substitutions, 91 indels). This relatively high false positive rate reflects our deliberate lower call quality filter settings for insertion and deletion variants; such variants have a strong prior likelihood of being associated with disease, but are challenging to call in short read data. There was no difference between the number of truncating variants seen in the samples with known gene mutations (median = 10, range 5-13) and those without (median = 10, range 7-15, p = 0.55). Only two genes contained two truncating variants; *CHEK2* and *USP45*, with the remaining 122 truncating variants occurring in distinct genes (online Supplementary Table 3).

Gene list enrichment analysis of validated truncating variants

We undertook gene enrichment analysis of all 122 genes in which we identified truncating variants and of the subset of 85 genes with truncating mutations in the 8 cases without mutations in known genes, using the ToppGene Suite ToppFun software [9]. No gene ontology term was identified as significant under a Bonferroni correction at a *P* value cut-off of 0.05 in either analysis.

DISCUSSION

Exome sequencing is revolutionising our ability to identify rare genetic variants that predispose to disease. However, the interrogation and interpretation of the resulting data outside the context of rare, Mendelian syndromes is very challenging. Here we have undertaken exome sequencing in familial breast cancer, one of the few diseases for which there is compelling evidence of rare moderate/low penetrance predisposition genes. We used a number of strategies to empower the analyses. Firstly, we used cases enriched for genetic susceptibility factors, specifically individuals with bilateral breast cancer and/or a family history of breast cancer. This significantly improves power for gene discovery as previously demonstrated [2-5, 10]. An alternative approach that is often considered in disease gene identification studies is to prioritise variants shared by distantly related affected individuals for further evaluation. This strategy is most powerful in the identification of highly penetrant mutations in rare conditions. In common conditions, such as breast cancer, the phenocopy rate is often high and the penetrance of predisposing mutations often intermediate/low, both of which act to reduce the utility of this strategy.

Secondly, we used a data filtering strategy that allows prioritisation of rare, protein truncating mutations; this class of mutation has strong prior evidence of disease association, particularly in breast cancer [2-5]. Moreover, simulation-based analysis of NGS data filtering in complex disorders supports prioritisation of variants predicted to result in premature protein truncation as a useful strategy for disease gene identification [11]. Even after this stringent filtering, 1,296 PTVs were identified in the 50 cases. This included four mutations in known breast cancer predisposition genes further demonstrating the utility of exome sequencing for the identification of disease-associated mutations.

To explore the feasibility of identifying novel breast cancer predisposition genes we first performed validation experiments in 12 of the 50 cases to establish which PTVs were real. In total, we confirmed 124 PTVs in the 12 samples. The median number of PTVs was similar in the cases with and without mutations in known genes, indicating that simply identifying a rare PTV is not sufficient to prove causality, as has been implied by some papers [12]; additional evidence is required. This is further supported by the observation that cases with mutations in known genes also carried other PTVs in genes plausibly related to disease. For example, the individual with a *BRCA2* mutation (Case 1) also carries PTVs in the regulator of apoptosis, *CASP5*, and the transcriptional regulators *SMARCD2* and *SSX9,* all of which are plausibly related to oncogenesis (Table 2). Similarly, Case 3, carries a *CHEK2* mutation and PTVs in five other genes implicated in a variety of diseases, including the DNA repair gene *WRN,* which causes Werner syndrome in biallelic mutation carriers [13]. It is possible that some of these additional mutations are also contributing to breast cancer, indeed it is anticipated that individuals will have multiple genetic variants that confer susceptibility to disease, particularly carriers of moderate-penetrance mutations. However, we identified PTVs in 122 different genes in just 12 cases indicating that, firstly, most of the mutations must be unrelated to the cancer and secondly, the burden of proof required to demonstrate a disease association, even for rare truncating mutations, is very substantial. This is further supported by studies demonstrating rare PTVs in healthy individuals [14]. Comparison of case data with control data acquired by similar methods and matched for metrics such as coverage, will be required to reliably distinguish genes which tolerate haploinsufficiency from disease predisposition genes.

Consideration of gene function has proved a useful prioritisation strategy in gene identification studies. For breast cancer, mutational analyses of DNA repair genes, particularly those that interact with the high-penetrance breast cancer susceptibility genes *BRCA1* and *BRCA2*, was fundamental to the identification of breast cancer predisposition genes such as *PALB2* and *BRIP1* [2, 4]. However, our *in silico* analyses did not reveal enrichment of any group of functionally related genes amongst genes with PTVs in the 12 familial breast cancer cases in which we performed comprehensive validation.

Only two genes contained two different truncating variants, one of which was *CHEK2,* a bone-fide breast cancer predisposition gene. This suggests that in a larger experiment, genes with multiple, different truncating mutations may serve as a useful filter to identify genuine predisposition genes. This pattern was crucial to the identification of other genes of this class in studies of 1000-3000 samples [2-5]. The number of samples required in an exome sequencing study is not known, and will be influenced by multiple factors including the prevalence and penetrance of mutations in the relevant gene, the type of samples analysed (genetically enriched vs unselected) and correction for multiple testing. However, it is very likely that exomic analysis of many hundreds / thousands of samples will be required. Follow up studies, analogous to the staged approach of some GWAS, may be helpful in replicating exome findings and providing definitive proof that a gene predisposes to disease. Replication sequencing studies of single genes or small sets of genes in thousands of samples is becoming feasible and could be targeted, for example, at genes in the exome study with the distinctive pattern of multiple, different, rare truncating variants in cases compared to controls.

In summary, our experiment provides further evidence that exome analyses can identify pathogenic mutations in known disease-associated genes. The potential for this technology to be utilised in gene discovery in common, complex conditions is high. However, it will require carefully designed large-scale experiments, maximally powered through judicious sample selection and analytical prioritisation approaches, coupled with replication analyses, to provide robust evidence of disease-association.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank all the patients and families that participated in the research. We thank Bernadette Ebbs for assistance in DNA extraction and running the ABI sequencers. We thank Margaret Warren-Perry, Anna Zachariou, Jessie Bull, Darshna Dudakia, Jessie Bull and Rachel Linger for coordination of recruitment and sample collection, and Ann Strydom for assistance with the manuscript. We are very grateful to all the clinicians and counsellors in the Breast Cancer Susceptibility Collaboration UK (BCSC) that have contributed to the recruitment and collection of the FBCS samples. The full list is the Supplementary Appendix.

Funding Statement

This work was supported by the US Military Acquisition (ACQ) Activity, Era of Hope Award (W81XWH-05-1-0204), Cancer Research UK (C8620/A8372 and C8620/A8857) and the Institute of Cancer Research (UK). We acknowledge National Health Service funding to the Royal Marsden / Institute of Cancer Research National Institute of Health Research Specialist Biomedical Research Centre for Cancer. K.S. is supported by the Michael and Betty Kadoorie Cancer Genetics Research Programme.

Abbreviations

RR	Relative risk
FBCS	Familial Breast Cancer Study
PTV	Protein Truncating Variant

REFERENCES

1. Ku CS, Naidoo N, Pawitan Y. Revisiting Mendelian disorders through exome sequencing. Hum Genet. 2011; 129:351–370. [PubMed: 21331778]
2. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T, et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat Genet. 2007; 39:165–167. [PubMed: 17200668]
3. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, North B, Jayatilake H, Barfoot R, Spanova K, et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet. 2006; 38:873–875. [PubMed: 16832357]
4. Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, Chagtai T, Jayatilake H, Ahmed M, Spanova K, et al. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nat Genet. 2006; 38:1239–1241. [PubMed: 17033622]
5. 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, 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. [PubMed: 11967536]
6. Turnbull C, Rahman N. Genetic predisposition to breast cancer: past, present, and future. Annu Rev Genomics Hum Genet. 2008; 9:321–345. [PubMed: 18544032]
7. Snape K, Hanks S, Ruark E, Barros-Nunez P, Elliott A, Murray A, Lane AH, Shannon N, Callier P, Chitayat D, et al. Mutations in CEP57 cause mosaic variegated aneuploidy syndrome. Nature Genetics. 2011; 43:527–529. [PubMed: 21552266]
8. Coffey AJ, Kokocinski F, Calafato MS, Scott CE, Palta P, Drury E, Joyce CJ, Leproust EM, Harrow J, Hunt S, et al. The GENCODE exome: sequencing the complete human exome. Eur J Hum Genet. 2011; 19:827–831. [PubMed: 21364695]
9. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009; 37:W305–311. [PubMed: 19465376]
10. Antoniou AC, Easton DF. Polygenic inheritance of breast cancer: Implications for design of association studies. Genet Epidemiol. 2003; 25:190–202. [PubMed: 14557987]
11. Feng BJ, Tavtigian SV, Southey MC, Goldgar DE. Design considerations for massively parallel sequencing studies of complex human disease. PLoS One. 2011; 6:e23221. [PubMed: 21850262]
12. Walsh T, Casadei S, Lee MK, Pennil CC, Nord AS, Thornton AM, Roeb W, Agnew KJ, Stray SM, Wickramanayake A, et al. Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. Proc Natl Acad Sci U S A. 2011
13. Yu CE, Oshima J, Fu YH, Wijsman EM, Hisama F, Alisch R, Matthews S, Nakura J, Miki T, Ouais S, et al. Positional cloning of the Werner's syndrome gene. Science. 1996; 272:258–262. [PubMed: 8602509]
14. MacArthur DG, Tyler-Smith C. Loss-of-function variants in the genomes of healthy humans. Hum Mol Genet. 2010; 19:R125–130. [PubMed: 20805107]

Table 1

Summary of probands in familial breast cancer exome study

Characteristics of breast cancer cases	
Total number of cases	**50**
Bilateral cases	42
Unilateral cases	8
Median age of diagnosis	
First breast cancer	53
Second breast cancer	60
Median Family History Score[*] (FHS)	**3**

[*] see Supplementary material – an individual with bilateral breast cancer and two first degree relatives with breast cancer (or equivalent) has a FHS = 3

Table 2

Confirmed heterozygous truncating variants in familial breast cancer probands with mutations in known breast cancer predisposition genes.

ID	Gene	Truncating mutation	Disease Association
1	**BRCA2**	**c.7977-1G>C**	Breast +ovarian cancer (monoallelic), FA-D1 (biallelic)
	BRIX1	c.793-2_793-1insA	
	CASP5	c.1135+1C>T	
	CXCL6	c.239_240insT	
	FILIP1	c.303delG	
	HEATR7B	c.2214+5A>G	
	IGSF22	c.479-2T>A	
	MLL4	c.3059_3060dupG	
	PTCHD3	c.923_924dupG	
	SLAMF6	c.321G>C, p.Y107X	
	SMARCD2	c.574G>A, p.R136X	
	SSX9	c.110delC	
	TNFAIP6	c.90G>A, p.W30X	
2	**CHEK2**	**c.1100delC**	Breast cancer (monoallelic)
	C2orf63	c.1384+2A>T	
	CFHR5	c.486_487insA	Membranoproliferative Glomerulonephritis, Type II
	PPEF2	c.1960G>A, p.R654X	
	SERPINI2	c.628_629delAC	
3	**CHEK2**	**c.658T>A, p.K220X**	Breast cancer (monoallelic)
	ABCC11	c.2813C>G, p.S938X	
	DNMT3A	c.1025_1026insC	AML
	EPS8L1	c.1514_1515dupT	
	FTMT	c.436A>T, p.K146X	
	LOC64702	c.303_304delAT	
	MCAT	c.729+1G>T	
	NOD2	c.3019_3020dupC	Crohn disease (monoallelic)
	PRMT7	c.1056-1G>T	
	PRSS7	c.2042_2043dupT	Enterokinase deficiency (biallelic)
	VPS13B	c.6732+1G>A	Cohen syndrome (biallelic)
	WRN	c.1230_1231insA	Werner syndrome (bilallelic)
	ZNF451	c.488G>G/A, p.W163X	
	ZNF582	c.136+1G>T	
4	**ATM**	**c.4396C>T, p.R1466X**	Breast cancer (monoallelic), ataxia telengiectasia (biallelic)
	FETUB	c.127_128insCA	
	KIAA1919	c.614delT	
	SLC26A10	c.1483C>T, p.R495X	

ID	Gene	Truncating mutation	Disease Association
	TAOK1	c.2544+5A>G	
	ZIM2	c.1513C>T, p.R505X	