The genetic etiology of familial breast cancer: Assessing the role of rare genetic variation using next generation sequencing

Hilbers, F.S.M.

Chapter 7

Summary and Discussion

Introduction

A family history of breast cancer is one of the most important risk factors for the disease.[1] Over the last decades many genetic loci associated with breast cancer risk have been discovered. In spite of this, only approximately half of the familial relative risk (FRR) for breast cancer can be explained by the currently known genetic risk factors.[2–4] In families where no genetic explanation has been found for the clustering of breast cancer, uncertainty remains about who is at increased risk and to which extend this risk is increased. This hampers decisions on screening strategies and preventive measures. Next generation sequencing (NGS) offers new possibilities to explore the genetic etiology of unexplained familial clustering of breast cancer. It allows for the detection of genetic variants regardless of their frequency in the general population and without the need of a prior hypothesis about which genes or genomic regions might be involved in breast cancer susceptibility. The aim of this thesis was to get a better understanding of the genetic etiology of non-*BRCA1/2* familial breast cancer with the help of NGS.

The selection of a homogeneous phenotype: a failed strategy?

The average exome of an individual from European descent has approximately 12,000 non-synonymous genetic variants.[5] However, at the time the studies described in this thesis were conducted, the cost of NGS only allowed us to sequence the exomes of a relatively small number of familial cases. Therefore, further association analysis using different techniques was necessary to follow up on potentially interesting genetic variants. However, also the number of familial cases available in downstream case-control analyses was limited. Thus, it was crucial to carefully select and strongly reduce the candidate variants for follow-up analysis. Selection based on the predicted effect of a genetic variant on protein function is often not able to sufficiently reduce the number of potentially interesting variants, while selection based on the pathways in which a gene is involved, strongly depend on our assumptions on which pathways play a role in breast cancer. Similar to classical association studies, an exome sequencing effort would ideally find multiple families with a mutation in the same gene while these mutations are absent or extremely rare in the general population. However, as the genetic etiology of breast cancer is already known to be very heterogeneous, a large number of familial cases would need to be sequenced in order to find two of them with a mutation in the same gene. Interestingly, mutations in *BRCA1* are strongly associated with a number of tumor characteristics. Notably, tumors of *BRCA1* mutation carriers are strongly enriched for the "triple negative" (lacking the receptors ER, PR and HER2) immunohistochemistry phenotype and a basal-like expression profile.[6] Based on this, we hypothesized that by selecting non-*BRCA1/2* breast cancer patients or families that share a certain phenotype, we would also select for a more homogenous genetic etiology and increase our chances of finding multiple cases with mutations in the same gene.

In **Chapter 2** of this thesis we selected six non-*BRCA1/2* families in which the majority of tumors show a specific, previously identified array comparative genome hybridization (CGH) profile.[7] Subsequent linkage analysis in these families showed a peak with a LOD score of 2.49 on chromosome 4, which suggested that the clustering of breast cancer in these families might be caused by mutations in a gene in this linkage region. Therefore, whole-exome sequencing was performed on two individuals per family. However, no genes with a likely pathogenic variant in more than one family were found. Not on chromosome 4, nor

elsewhere in the genome. Similarly, **Chapter 5** describes an exome sequencing study in which we focused on families with a possible recessive mode of inheritance. For this study we selected 19 non-*BRCA1/2* breast cancer families in which at least three siblings were affected, while no first-degree relatives in the previous or following generation had breast cancer. The germline DNA from one of the siblings was subjected to exome sequencing, while all affected siblings were genotyped using a SNP arrays in order to assess haplotype sharing. This allowed us to focus on the exome sequencing variants in the regions where all affected siblings shared two haplotypes. However, also this exome sequencing study did not yield any potential novel susceptibility genes. It is possible that in these two studies we have missed high-risk susceptibility alleles that were in fact present in the sequenced individuals. We might have discarded a variant as it seemed unlikely to affect protein function or we might not have detected it at all as it resides outside the protein-coding regions. Moreover, the results of these studies do not exclude that there are additional high-risk breast cancer susceptibility genes that are strongly associated with a specific phenotype. We might simply have selected the wrong tumor and family characteristics. As conventionally a LOD score greater than 3.0 is considered evidence for linkage, our LOD score of 2.49 in **Chapter 2** might have been spurious. And, while our selection of families with at least three affected siblings in **Chapter 5** had important advantages for the variant filtering, other selection criteria, such as very early onset cases, might have been more likely to enrich for recessive susceptibility alleles. Moreover, as the families we selected had a large number of breast cancer cases in one generation, it is not impossible that these families are in fact explained by a dominant allele.

Biologically, there are several ways in which an inherited genetic variant can be associated with a tumor phenotype. Cancer susceptibility genes are typically thought to be tumor suppressors, which require the loss of both copies for malignant transformation. However, on a cellular level the loss of one copy of a gene can already have effects on downstream signalling, gene expression and cellular functions. This is called haploinsufficiency. For example, lymphoblastic cell lines derived from carrier of a heterozygous deleterious *PALB2* mutation show aberrant DNA replication and a shift to error-prone DNA repair mechanisms.[8,9] This might result in characteristics that most tumors associated with a specific susceptibility gene have in common, for example, altered expression of genes controlled by the pathway in which the gene with the inherited mutation functions or altered phosphorylation of proteins in this same pathway. However, due to the large number of genetic and epigenetic changes a tumor cell acquires during tumorigenesis, characteristics associated with an inherited mutation might be partly masked and become difficult to detect. Association between inherited mutations and tumor phenotype can also occur indirectly due to synergy with other genetic, epigenetic or microenvironmental changes, which are subsequently selected for because of increased fitness. One of the best-studied examples is so-called loss of heterozygosity (LOH). As mentioned above, while a cell is thought to be relatively unaffected by the loss of one copy of a tumor suppressor gene, loss of the second copy will have a much more dramatic effect and contribute significantly to tumorigenesis. Somatic loss of one copy of a susceptibility gene therefore will be selected for in the context of an inherited pathogenic mutation, but not in the absence of such a mutation. This phenomenon is frequently observed in high-risk breast cancer susceptibility genes *BRCA1*, *BRCA2*[10] and *PALB2*.[11] LOH at these loci is very high in tumors of gene carriers, but much lower in sporadic cases. In genes associated with a more moderate increase in risk of breast cancer this association is less clear: while loss of the wild-type allele is frequently observed in the breast tumors of *ATM* mutation carriers,[12,13] tumors of *CHEK2* mutation carriers show no strong enrichment for the loss of the wild-type *CHEK2*

allele.[14–16] More complex relationships between germline mutations and tumor characteristics have also been described. For example, tumors of *BRCA1* and *BRCA2* mutation carriers have been found to be associated with two specific mutation signatures and two rearrangement signatures, which are thought to be the "genomic scar" shaped by the absence of *BRCA1* or *BRCA2* function and the resulting DNA repair deficiencies.[17] Interestingly, this also shows that mutations in two different genes can result in a similar phenotype. Lastly, it has been hypothesized that part of the phenotypical heterogeneity of breast cancer stems from the existence of two different of cell types of origin, myoepithelial and luminal cells.[18] In a full-grown tumor, epigenetic features and gene expression patterns, might still be traced back to this cell of origin. Therefore, if a genetic risk factor more strongly predisposes to cancer in one of these two cell types, these epigenetic and gene expression features would also be associated with this genetic risk factor.

Although the two studies described in **Chapters 2** and **5** of this thesis have been unsuccessful in discovering novel breast cancer risk alleles, it might be too early to completely dismiss the strategy of selecting a more homogeneous group of familial breast cancer patients when aiming to find new breast cancer risk alleles. Several of the known breast cancer susceptibility genes are associated with a specific phenotype (see **Chapter 1** of this thesis), although, apart from a few of the cancer syndromes, these phenotypes were typically discovered after the association of breast cancer with a specific gene or genomic region had been detected. The CGH profile as used in **Chapter 2** of this thesis provides a relatively low-resolution picture of the tumor genome. Over the past years, technical advances and decreasing sequencing costs have provided new opportunities to assess tumor characteristic and therefore to potentially select tumors with a more homogeneous etiology. Most importantly, it has become possible to apply massive parallel sequencing on DNA and RNA from formalin fixed paraffin embedded (FFPE) material. In addition, copy number aberrations can now be characterized more precisely using SNPs array-based techniques. The molecular tumor characteristics described in the introduction of this thesis, e.g. based on mutations, copy number variations, "intrinsic" gene expression-based subtypes and mutational signatures, could potentially be used to select for a more homogeneous population of familial breast cancer cases. Currently, little or no data exists on whether these molecular features cluster within families. Exploring this would be a first step to decide if it is worthwhile further pursuing the "homogeneous phenotype strategy".

The challenge of establishing the risk associated with extremely rare variants

As discussed above, arguably the most difficult and laborious step in the analysis of exome sequencing data is the variant filtering. If no genes are identified in which multiple families carry a variant that is likely to affect protein function, other strategies are needed to select those variants most likely to be associated with breast cancer risk for downstream validation. **Chapter 6** of this thesis reviews several approaches. As a first step, genetic variants are often filtered based on minor allele frequency in one of the many available reference data sets of healthy individuals. This is rationalized by the prevalence of breast cancer in the general population, which must be consistent with the presumed risk associated with the variant, i.e. a high-risk variant cannot be too common, otherwise breast cancer would be more prevalent than it is. Although somewhat arbitrary, a cut-off of 0.1% allele frequency is often used.

However, it is important to keep in mind that there are several examples of founder mutations, such as *CHEK2**1100delC, that are associated with a moderately or strongly increased breast cancer risk and have an allele frequency larger than 0.1%. Moreover, due to recent explosive human population growth, most variants are rare regardless of association with disease.[19] For example, within the ExAC dataset containing the exome data of more than 60,000 individuals, approximately 99% of detected high quality variants had an allele frequency of less than 1%.[5] Hence, selecting for variants occurring <0.1% in reference data sets will not dismiss a large proportion of candidates. A next obvious step is to focus on protein truncating variants as they are almost certain to affect protein function. However, these variants make up only a small minority of the detected variants. Further filtering can be done using in silico prediction algorithms. These tools use information such as evolutionary conservation, known functional domains and three-dimensional structure, to estimate the likelihood of a missense variant to affect protein function. Unfortunately, the sensitivity and specificity of these tools is known to be far from optimal.[20,21] However, the limited number of alternative filtering options besides in silico prediction algorithms make that these algorithms are frequently used. Lastly, variants are often filtered based on the function of the gene in which the variant is found, where genes with roles cancer-related processes such as cell proliferation and DNA repair are prioritized. This filter depends heavily on our knowledge of the pathways involved in carcinogenesis. In addition, one could argue that if we are only considering variants in genes related to carcinogenesis, a sequencing of a cancer specific gene panel might be a better approach than exome sequencing.

Follow-up association analysis can be done on the variant level, i.e. by genotyping a set of cases and controls for that specific variant only. However, as explained above many variants will be very rare and even large case control sets will often lack the power to detect a significant association. An alternative approach to a variant-level analysis is a so-called burden analysis, where in every case and control the whole coding region of the respective gene is sequenced. In this case, the association analysis is not done based on individual variants, but rather compares the total number of likely damaging variants between cases and controls. This requires a decision on which variants to consider as (likely) damaging. Some of the same filters as discussed above, such as allele frequency and in silico prediction can be used. However, as the number of variants to be assessed is now considerably smaller, additional options are available. These include co-segregation analysis in families with a variant, assessment of loss of heterozygosity in the tumors of carriers, and functional assays. It is worth investing time and effort in the selection of variants, as both the inclusion of benign variants and the exclusion of variants that truly affect the protein of interest reduce the power of the association analysis.

Chapter 3 and **4** of this thesis give a good example of the difficulties associated with establishing the risk associated with very rare variants, in this case in the gene *XRCC2*. The possible association between variants in *XRCC2* and familial breast cancer was first reported by a research group at the university of Melbourne, Australia. They had found one protein-truncating variant in the exome of a familial breast cancer patient and, based on the function of XRCC2 in DNA repair, had decided to further explore the possibility of this being a breast cancer susceptibility allele. Subsequently, they requested access to the exome sequencing data from several other groups, among which the data from our study reported in **Chapter 2** of this thesis. This pooled analysis of exomes and a subsequent case-control study provided a suggestion that variants in *XRCC2* might indeed be associated with breast cancer risk and the results were subsequently published.[22] **Chapter 3** of this thesis reports the data of a large

international case-control study aiming to validate the results of this initial publication. This study applied a burden analysis strategy, classifying variants based on in silico prediction algorithms (Polyphen-2[23], SIFT[24] and AlignGVGD[25]). Regardless of the prediction algorithm used, this study did not find an association with breast cancer. However, as prediction algorithms are known to be imperfect, we decided to further explore this result using functional assays to assess the effect of the genetic variants on the XRCC2 protein and the DNA repair pathways in which it functions. Chapter 4 of this thesis reports the results of this effort. It showed that, based on a RAD51 foci formation assay and two reporter constructs, the SCR reporter and the DR-GFP reporter, only a limited number of variants actually affected protein function. When only taking into account these variants, again no association with breast cancer risk was found, although an association could not be ruled out for those variants which strongly affect XRCC2 function, due to the low number of variants in this group.

This example underlines the difficulties with variant selection and the issues related to very rare genetic variants. A big challenge for the future will be to conduct exome sequencing studies with sufficient sample size to at least allow for gene-level association analyses. Recently, an exome sequencing study in over 20,000 cases with type 2 diabetes showed that it was able to identify four susceptibility genes at exome-wide significance based on a rare variant (<0.5% minor allele frequency) burden analysis.[26] Of note, the effective power of this study was increased by selecting an ethnically diverse population, thereby sampling a broader range of haplotypes. In the near future, similar efforts for breast cancer susceptibility will likely be pursued in the context of existing international collaborations, such as the Breast Cancer Association Consortium (BCAC). However, in order to solve variant-level associations for very rare variants, case-control studies with the currently available sample sizes will likely not suffice. A possible alternative method of classifying rare variants of uncertain significance (VUS) is by way of co-segregation analysis. Co-segregation analysis assesses the association between a genetic variant and a disease by quantifying, based on a pedigree and the breast cancer cases occurring within it, the extent to which a genetic variant co-occurs with disease more often than expected. However, co-segregation analysis requires extensive DNA sampling within families carrying a VUS. It will therefore be crucial to invest in the collection of such DNA samples.

The updated landscape of breast cancer susceptibility

Before embarking on further efforts to discover novel breast cancer risk alleles, it is important to reflect on what we have learned about the landscape of genetic susceptibility over the last few years. To tailor our efforts, we need to understand how likely it is that further high-risk alleles explain a considerable proportion of the currently unexplained familial clustering of breast cancer.

Historically, research into the genetic etiology of breast cancer mainly focused on the discovery of high-risk genes. Naturally, families with very strong clustering of breast cancer are a logical starting point for the discovery of genetic risk factors. Moreover, for a long time, research on the genetic etiology of breast cancer was limited by the fact that the sequence of most the human genome was unknown, and no technologies existed that allowed for the analysis of large genomic regions. Therefore, the region of interest first needed to be narrowed using linkage analysis with low-density microsatellite genotyping, after which the regions significantly more often shared by the affected individuals in a set of families could be further explored. As linkage analysis is only able to find regions in which alleles with a

relatively high penetrance are located, this limited the scope of early research into the genetic etiology of breast cancer. When in 2001 the first draft of the human genome was released, it was accompanied by a manuscript describing the construction of a map of 1.4 million SNPs in the human genome, providing for the first time sufficient density to study human haplotype structure and allowing for subsequent genome-wide studies assessing the association between common genetic variation and disease in the general population. From 2007 onward, several genome-wide association studies (GWAS) together have reported over 300

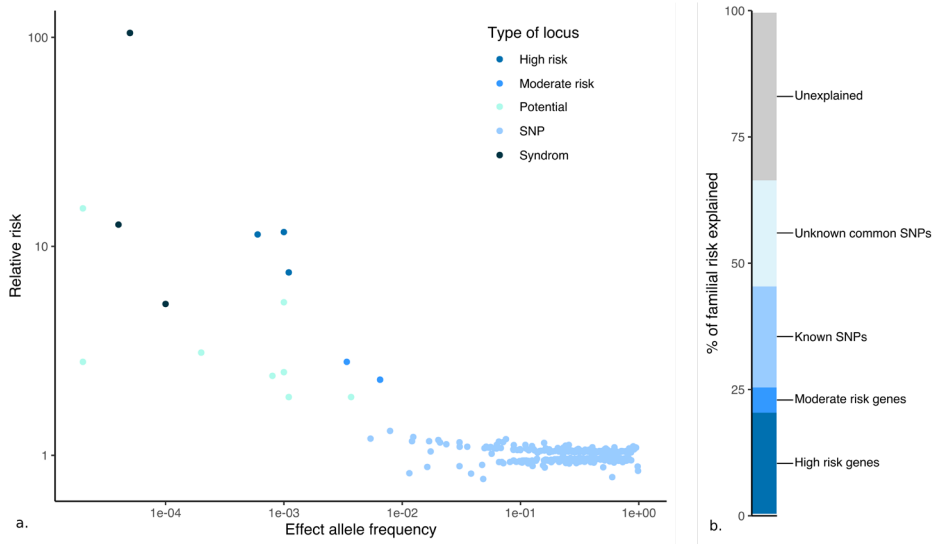

Figure 1. The current landscape of breast cancer susceptibility alleles
a. Relative risk and effect allele frequency for the currently known breast cancer susceptibility alleles. For underlying data please see supplementary table 1. b. percentage of familial risk explained by the currently known breast cancer susceptibility alleles. For references and underlying data see supplementary table 1.

The invention of NGS brought new hope for the discovery of novel high-risk susceptibility genes as it allows for cost-effective genome-wide detection of genetic variants in individual familial breast cancer cases. However, as outlined above, we have not been able to find any novel high-risk breast cancer alleles in the two exome sequencing studies described in this thesis, **Chapter 2** and **Chapter 5**. In **Chapter 5** of this thesis, besides exploring the potential role of recessive high-risk alleles, we genotyped the families in this study for all the known and suspected moderate and high-risk genes in addition to genotyping the 160 SNPs currently known to be associated with a small increase in breast cancer risk. This study found that the average normalized PRS of the familial cases was significantly higher than that in both general population cases and controls. Indicating that the low risk variants do contribute to familial clustering of breast cancer, although it is difficult to estimate to which extend due to the atypical breast cancer families represented in this study. Moreover, in several families we detected a moderate risk variant in *ATM* or *CHEK2*. In another study (not included in this thesis), we have reported that in a set of 101 unselected non-*BRCA1/2* breast cancer families, familial breast cancer cases have on average a higher PRS. Moreover, taking into account the PRS can change risk management recommendations in 10-20% of the women in these families depending on the guideline used.[29] Interestingly, also work from others has shown limited value of NGS for the discovery of novel high-risk genes. Although

over 30 exome sequencing studies have been published to date, the number of potential high-risk susceptibility genes identified is limited. Several genes, such as *KAT6B*,[30] *RINT1*,[31] *APOBEC3B*,[32] *XRCC2*[22] and *RCC1*,[33] have been suggested as novel susceptibility genes but external validation has either not been performed or resulted in conflicting results. Only two promising novel susceptibility genes coming out of exome sequencing studies have now been validated independently by several other studies: *FANCM*[34] and *RECQL*.[35,36] Remarkably, many exome sequencing studies report pathogenic variants in known moderate risk genes such as *ATM,*[37–42] *CHEK2*,[39,41,43–45] and *PALB2*,[37,39,41,45,46] suggesting that indeed a substantial proportion of familial cases might be explained by a combination of low and moderate risk susceptibility alleles. In fact, already shortly after the discovery of *BRCA1* and *BRCA2* as breast cancer susceptibility genes, several segregation analyses have suggested a that a polygenic model would best explain the remaining familial clustering of breast cancer.[47,48] This now seems to be confirmed by the results of the latest GWAS analysis in which it is estimated that that all, currently known and yet to be discovered, common low risk alleles explain together approximately 41% of the familial risk of breast cancer.[2] Figure 1 provides an overview of our current understanding of the genetic landscape of breast cancer. While moderate and high-risk alleles are thought to explain 5% and 20% respectively,[49] 41% is thought to be explained by low risk, common variants of which ~20% by the currently known low risk susceptibility loci.[2] The remaining 35% would then be explained by currently unknown factors such as, rare variants, interactions between risk factors, inherited epigenetic factors and environmental risk factors that are shared between family members.

Now that it has become clear that many genetic factors contribute to familial clustering of breast cancer and that individual factors are often associated with just a small increase in risk, there is a clear need to combine these factors into risk prediction models that are able to provide insights in an individual's risk of breast cancer. Several studies have aimed to combine the effects of low risk loci into a polygenic risk score (PRS). The latest of these, combines the effect of 313 SNPs.[28] Predating the GWAS era, there are also many models aiming to predict the risk of breast cancer based on non-genetic factors and high-risk mutations. Arguably the most extensive model to date is the BOADICEA model, which has very recently been updated to include the effect of the 313 currently known low risk loci. This model now uses truncating mutations in *BRCA1*, *BRCA2*, *ATM*, *CHEK2* and *PALB2*; 313 low risk loci; age at menarche; age at menopause; parity; age at first live birth; oral contraceptive (OC) use; hormone replacement therapy (HRT) use; height; BMI; alcohol intake; family history and a residual polygenic component to predict lifetime breast cancer risk.[50] In the UK population, this model would predict approximately 15% of women to have moderate lifetime risk of breast cancer (≥17% and <30% according to the NICE guidelines) and approximately 1% to have a high risk (>30%). This model has not yet been prospectively tested. Moreover, it does not take into account interactions between risk factors beyond the log-additive model, nor variant-specific risks in moderate and high risk genes, genetic variants in genes associated with cancer syndromes i.e. *TP53*, *CDH1* and *PTEN*, genetic variants in likely novel breast cancer susceptibility genes such as *FANCM* and *RECQL*, subtype specific effects, time varying variables for BMI, alcohol, OC and HRT use and the exact timing of pregnancies. A large prospective validation effort could be combined with an attempt to include these factors and improve the model, either through classic association analysis or deep learning. An advantage of the latter method is that it allows for continuous learning, making use of all available data, although the lack of formal statistics to express uncertainty is a disadvantage. The most optimal approach would probably be to prospectively calculate, for example for a large cohort of women in the context of population

screening, risk of breast cancer based on the current BOADICEA model in order to validate it and to simultaneously try to optimize the model using deep learning on prospectively collected data from these same women. The use of a model like BOADICEA would also make it possible to focus future research on (familial) cases that are unexplained by the currently known risk factors, e.g. familial or early onset cases with a very low predicted risk

Conclusions

To summarize, the work reported in this thesis has not been able to identify any novel high-risk breast cancer susceptibility alleles. Although there are likely still several extremely rare risk alleles to be discovered and the presence of high-risk alleles outside of protein-coding regions cannot be excluded, it seems presently unlikely that these will explain a substantial proportion of familial breast cancer. Both our work and that of others has suggested that most non-*BRCA1/2* familial breast cancer cases are likely explained by a combination of low-, and moderate-risk susceptibility alleles.

As expected, the largest challenge associated with the use of exome sequencing in the context of familial breast cancer has been the large number of genetic variants detected in a relatively small set of familial cases, which, with the sample sizes used to date, prohibits any formal association testing in the variant selection process. Therefore, until we are able to conduct exome sequencing studies with at least sufficient power to allow for exome-wide gene-level association analyses, the discovery of novel risk alleles in an assumption-free manner is still not a reality. The selection of a more homogeneous phenotype in hopes of selecting for a more homogeneous genetic etiology, has not resulted in any potential risk alleles being detected in more than one family. Although, with more advance techniques becoming available for the phenotyping of tumors, there might still be value in this approach for future attempts to discover novel high-risk alleles. Our experience attempting to validate rare genetic variants in *XRCC2* as breast cancer susceptibility alleles has served as a reminder of the limited value of in silico prediction algorithms, which can lead to misleading results of burden analyses and incorrect conclusions about a gene's role in breast cancer susceptibility. Although functional assays can give important insights, in many cases the time needed for the set-up and conduct of these assays, makes that they will likely only be used for strong candidate genes.

Taken together, these findings lead to the conclusion that if we want to be able to provide better risk prediction to the familial breast cancer cases who remain unexplained by susceptibility alleles currently tested in clinical practice, and to any woman for that matter, future efforts should focus on developing models that combine all currently known susceptibility alleles and take into account other risk factors. After initial validation of such a model, deep-learning techniques could be employed to continuously improve them based on real-world data. This prospective might mean that the dichotomy of sporadic and familial breast cancer with regard to genetic susceptibility disappears, which would require a change in perspective for both breast cancer susceptibility research and genetic counseling.

References:

1. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. Lancet Lond. Engl. 358, 1389–1399 (2001).
2. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. Nature (2017) doi:10.1038/nature24284.
3. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group. Br. J. Cancer 83, 1301–1308 (2000).
4. Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat. Genet. 45, 353–361, 361e1-2 (2013).
5. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).
6. Mavaddat, N. et al. Pathology of breast and ovarian cancers among BRCA1 and BRCA2 mutation carriers: results from the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA). Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol. 21, 134–147 (2012).
7. Didraga, M. A. et al. A non-BRCA1/2 hereditary breast cancer sub-group defined by aCGH profiling of genetically related patients. Breast Cancer Res. Treat. 130, 425–436 (2011).
8. Nikkilä, J. et al. Heterozygous mutations in PALB2 cause DNA replication and damage response defects. Nat. Commun. 4, 2578 (2013).
9. Obermeier, K. et al. Heterozygous PALB2 c.1592delT mutation channels DNA double-strand break repair into error-prone pathways in breast cancer patients. Oncogene 35, 3796–3806 (2016).
10. Maxwell, K. N. et al. BRCA locus-specific loss of heterozygosity in germline BRCA1 and BRCA2 carriers. Nat. Commun. 8, 319 (2017).
11. Lee, J. E. A. et al. Molecular analysis of PALB2-associated breast cancers. J. Pathol. 245, 53–60 (2018).
12. Weigelt, B. et al. The Landscape of Somatic Genetic Alterations in Breast Cancers From ATM Germline Mutation Carriers. J. Natl. Cancer Inst. 110, 1030–1034 (2018).
13. Renault, A.-L. et al. Morphology and genomic hallmarks of breast tumours developed by ATM deleterious variant carriers. Breast Cancer Res. BCR 20, 28 (2018).
14. Muranen, T. A. et al. Breast tumors from CHEK2 1100delC-mutation carriers: genomic landscape and clinical implications. Breast Cancer Res. BCR 13, R90 (2011).
15. Oldenburg, R. A. et al. The CHEK2*1100delC variant acts as a breast cancer risk modifier in non-BRCA1/BRCA2 multiple-case families. Cancer Res. 63, 8153–8157 (2003).
16. Massink, M. P. G., Kooi, I. E., Martens, J. W. M., Waisfisz, Q. & Meijers-Heijboer, H. Genomic profiling of CHEK2*1100delC-mutated breast carcinomas. BMC Cancer 15, 877 (2015).
17. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47–54 (2016).
18. Anderson, W. F., Rosenberg, P. S., Prat, A., Perou, C. M. & Sherman, M. E. How Many Etiological Subtypes of Breast Cancer: Two, Three, Four, Or More? JNCI J. Natl. Cancer Inst. 106, (2014).
19. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336, 740–743 (2012).
20. Miosge, L. A. et al. Comparison of predicted and actual consequences of missense mutations. Proc. Natl. Acad. Sci. U. S. A. 112, E5189-5198 (2015).
21. Ghosh, R., Oak, N. & Plon, S. E. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol. 18, 225 (2017).
22. Park, D. J. et al. Rare mutations in XRCC2 increase the risk of breast cancer. Am. J. Hum. Genet. 90, 734–739 (2012).
23. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. Chapter 7, Unit7.20 (2013).

24. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. Genome Res. 11, 863–874 (2001).
25. Tavtigian, S. V. et al. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J. Med. Genet. 43, 295–305 (2006).
26. Flannick, J. et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. Nature 570, 71–76 (2019).
27. Easton, D. F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447, 1087–1093 (2007).
28. Mavaddat, N. et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am. J. Hum. Genet. 104, 21–34 (2019).
29. Lakeman, I. M. M. et al. Addition of a 161-SNP polygenic risk score to family history-based risk prediction: impact on clinical management in non-BRCA1/2 breast cancer families. J. Med. Genet. 56, 581–589 (2019).
30. Lynch, H. et al. Can unknown predisposition in familial breast cancer be family-specific? Breast J. 19, 520–528 (2013).
31. Park, D. J. et al. Rare mutations in RINT1 predispose carriers to breast and Lynch syndrome-spectrum cancers. Cancer Discov. 4, 804–815 (2014).
32. Radmanesh, H. et al. Assessment of an APOBEC3B truncating mutation, c.783delG, in patients with breast cancer. Breast Cancer Res. Treat. 162, 31–37 (2017).
33. Riahi, A. et al. Exome sequencing and case-control analyses identify RCC1 as a candidate breast cancer susceptibility gene. Int. J. Cancer 142, 2512–2517 (2018).
34. Kiiski, J. I. et al. Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. Proc. Natl. Acad. Sci. U. S. A. 111, 15172–15177 (2014).
35. Cybulski, C. et al. Germline RECQL mutations are associated with breast cancer susceptibility. Nat. Genet. 47, 643–646 (2015).
36. Sun, J. et al. Mutations in RECQL Gene Are Associated with Predisposition to Breast Cancer. PLoS Genet. 11, e1005228 (2015).
37. Cybulski, C. et al. Mutations predisposing to breast cancer in 12 candidate genes in breast cancer patients from Poland. Clin. Genet. 88, 366–370 (2015).
38. Määttä, K. et al. Whole-exome sequencing of Finnish hereditary breast cancer families. Eur. J. Hum. Genet. EJHG 25, 85–93 (2016).
39. Maxwell, K. N. et al. Evaluation of ACMG-Guideline-Based Variant Classification of Cancer Susceptibility and Non-Cancer-Associated Genes in Families Affected by Breast Cancer. Am. J. Hum. Genet. 98, 801–817 (2016).
40. Tavera-Tapia, A. et al. Almost 2% of Spanish breast cancer families are associated to germline pathogenic mutations in the ATM gene. Breast Cancer Res. Treat. 161, 597–604 (2017).
41. Lu, H.-M. et al. Association of Breast and Ovarian Cancers With Predisposition Genes Identified by Large-Scale Sequencing. JAMA Oncol. 5, 51–57 (2019).
42. Guo, X. et al. Discovery of a Pathogenic Variant rs139379666 (p. P2974L) in ATM for Breast Cancer Risk in Chinese Populations. Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol. 28, 1308–1315 (2019).
43. Snape, K. et al. Predisposition gene identification in common cancers by exome sequencing: insights from familial breast cancer. Breast Cancer Res. Treat. 134, 429–433 (2012).
44. Gracia-Aznarez, F. J. et al. Whole exome sequencing suggests much of non-BRCA1/BRCA2 familial breast cancer is due to moderate and low penetrance susceptibility alleles. PloS One 8, e55681 (2013).
45. Shahi, R. B. et al. Identification of candidate cancer predisposing variants by performing whole-exome sequencing on index patients from BRCA1 and BRCA2-negative breast cancer families. BMC Cancer 19, 313 (2019).
46. Silvestri, V. et al. Whole-exome sequencing and targeted gene sequencing provide insights into the role of PALB2 as a male breast cancer susceptibility gene. Cancer 123, 210–218

(2017).
47. Antoniou, A. C. et al. Evidence for further breast cancer susceptibility genes in addition to BRCA1 and BRCA2 in a population-based study. Genet. Epidemiol. 21, 1–18 (2001).
48. Antoniou, A. C. et al. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. Br. J. Cancer 86, 76–83 (2002).
49. Antoniou, A. C. & Easton, D. F. Models of genetic susceptibility to breast cancer. Oncogene 25, 5898–5905 (2006).
50. Lee, A. et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genet. Med. Off. J. Am. Coll. Med. Genet. 21, 1708–1718 (2019).

Supplementary data

locus	EAF	RR	group
BRCA1	6.00E-04[1]	11.4[2]	High risk
BRCA2	0.001[1]	11.7[2]	High risk
PALB2	0.0011[3]	7.5[3]	High risk
ATM	0.0034[3]	2.8[3]	Moderate risk
CHEK2	0.0065[3]	2.3[3]	Moderate risk
TP53	5.00E-05[4]	4.5[5]	Syndrome
CDH1	1.00E-04[3]	5.3[3]	Syndrome
PTEN	4.00E-05[3]	12.7[3]	Syndrome
BARD1	0.001[6]	5.4[6]	Potential
FANCC	8.00E-04[7]	2.4[8]	Potential
FANCM	0.0037[9]	1.9[10]	Potential
MEN1	2.00E-05[11]	2.8[12]	Potential
MSH6	0.001[13]	1.9[3]	Potential
RECQL	0.001[13]	2.5[14]	Potential
STK11	2.00E-05[15]	15.2[16]	Potential
N=313	various[17]	various[17]	SNP

Supplementary table 1.
References for the effect allele frequencies (EAF) and relative risks (RR) for the breast cancer susceptibility alleles from Figure 1.

Supplementary references

1. Antoniou AC, Cunningham AP, Peto J, et al. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. Br J Cancer. 2008;98(8):1457-1466. doi:10.1038/sj.bjc.6604305
2. Easton DF, Pharoah PDP, Antoniou AC, et al. Gene-panel sequencing and the prediction of breast-cancer risk. N Engl J Med. 2015;372(23):2243-2257. doi:10.1056/NEJMsr1501341
3. Couch FJ, Shimelis H, Hu C, et al. Associations Between Cancer Predisposition Testing Panel Genes and Breast Cancer. JAMA Oncol. 2017;3(9):1190-1196. doi:10.1001/jamaoncol.2017.0424
4. Gonzalez KD, Noltner KA, Buzin CH, et al. Beyond Li Fraumeni Syndrome: clinical characteristics of families with p53 germline mutations. J Clin Oncol. 2009;27(8):1250-1256. doi:10.1200/JCO.2008.16.6959
5. LaDuca H, Polley EC, Yussuf A, et al. A clinical guide to hereditary cancer panel testing: evaluation of gene-specific cancer associations and sensitivity of genetic testing criteria in a cohort of 165,000 high-risk patients. Genet Med. August 2019. doi:10.1038/s41436-019-0633-8
6. Weber-Lassalle N, Borde J, Weber-Lassalle K, et al. Germline loss-of-function variants in the BARD1 gene are associated with early-onset familial breast cancer but not ovarian cancer. Breast Cancer Res. 2019;21(1):55. doi:10.1186/s13058-019-1137-9
7. FANCC | gnomAD. https://gnomad.broadinstitute.org/gene/ENSG00000158169. Accessed December 8, 2019.
8. Berwick M, Satagopan JM, Ben-Porat L, et al. Genetic heterogeneity among Fanconi anemia

SNPs associated with breast cancer.[27,28] Together these low risk alleles explain approximately 20% of the familial relative risk of breast cancer.[28]

heterozygotes and risk of cancer. Cancer Res. 2007;67(19):9591-9596. doi:10.1158/0008-5472.CAN-07-1501

9. FANCM | gnomAD. https://gnomad.broadinstitute.org/gene/ENSG00000187790. Accessed December 8, 2019.
10. Kiiski JI, Pelttari LM, Khan S, et al. Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. Proc Natl Acad Sci USA. 2014;111(42):15172-15177. doi:10.1073/pnas.1407909111
11. Kamilaris CDC, Stratakis CA. Multiple Endocrine Neoplasia Type 1 (MEN1): An Update and the Significance of Early Genetic and Clinical Diagnosis. Front Endocrinol (Lausanne). 2019;10:339. doi:10.3389/fendo.2019.00339
12. Dreijerink KMA, Goudet P, Burgess JR, Valk GD. Breast-Cancer Predisposition in Multiple Endocrine Neoplasia Type 1. N Engl J Med. 2014;371(6):583-584. doi:10.1056/NEJMc1406028
13. RECQL | gnomAD. https://gnomad.broadinstitute.org/gene/ENSG00000004700. Accessed December 8, 2019.
14. Bogdanova N, Pfeifer K, Schurmann P, et al. Analysis of a RECQL splicing mutation, c.1667_1667+3delAGTA, in breast cancer patients and controls from Central Europe. Fam Cancer. 2017;16(2):181-186. doi:10.1007/s10689-016-9944-y
15. Beggs AD, Latchford AR, Vasen HFA, et al. Peutz-Jeghers syndrome: a systematic review and recommendations for management. Gut. 2010;59(7):975-986. doi:10.1136/gut.2009.198499
16. Giardiello FM, Brensinger JD, Tersmette AC, et al. Very high risk of cancer in familial Peutz-Jeghers syndrome. Gastroenterology. 2000;119(6):1447-1453. doi:10.1053/gast.2000.20228
17. Mavaddat N, Michailidou K, Dennis J, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet. 2019;104(1):21-34. doi:10.1016/j.ajhg.2018.11.002