Editorial Commentary

Another step forward towards unraveling the biological mechanisms driving breast cancer predisposition: a role for non-coding RNAs

Antoinette Hollestelle

Department of Medical Oncology, Erasmus MC Cancer Institute, Rotterdam, the Netherlands

Correspondence to: Antoinette Hollestelle, PhD. Department of Medical Oncology, Erasmus MC Cancer Institute, Dr. Molewaterplein 40, 3015 GD Rotterdam, the Netherlands. Email: a.hollestelle@erasmusmc.nl.

Comment on: Moradi Marjaneh M, Beesley J, O'Mara TA, *et al.* Non-coding RNAs underlie genetic predisposition to breast cancer. Genome Biol 2020;21:7.

Received: 20 March 2020; Accepted: 06 April 2020; Published: 30 June 2020.

doi: 10.21037/ncri.2020.03.05

View this article at: http://dx.doi.org/10.21037/ncri.2020.03.05

Breast cancer is a complex disease. In addition to lifestyle and environmental factors, genetic factors play a pivotal role in predisposition to breast cancer. Germline mutations in breast cancer predisposition genes *BRCA1*, *BRCA2*, *PALB2*, *CHEK2* and *ATM* confer moderate to high (i.e., >2-fold) risks to develop breast cancer and cluster within breast cancer families. The proteins encoded by these five genes are all involved in the repair of DNA double-strand breaks via homologous recombination repair (1,2).

In addition to these five breast cancer risk genes, more than 170 breast cancer risk variants have been identified that confer a low (i.e., <1.3-fold) risk to develop breast cancer (3-7). Although the breast cancer risk a single variant confers is very low, these variants are common in the population and act multiplicatively. Therefore, when combined in a polygenic risk score (PRS), individuals in the top centile of this PRS have a moderate lifetime overall breast cancer risk of 32.6% (8). This risk is higher than the lifetime risk conferred by, for example, the c.1100delC mutation in the moderate-risk gene *CHEK2* (9).

The 172 low-risk variants were identified through genome-wide association studies (GWASs) in which tens of thousands of breast cancer cases and controls are genotyped for more than half a million SNPs in the human genome. Subsequent imputation of the acquired genotyping data using a human genome reference panel even allows the assessment of more than 10 million SNPs (5). The majority of GWAS-identified variants, however, are not the causal variants conferring the breast cancer risk. Usually, they are tags for the causal variants that lie in close linkage disequilibrium (LD) with the GWAS-identified variants. Therefore, GWASs need to be followed up by fine-mapping of the region and in silico and experimental functional studies to identify the breast cancer risk-causing variants (10,11).

Identifying the causal variants is important for two reasons. First, the causal variant is likely to confer a higher breast cancer risk than the GWAS-identified variant. Therefore, including all casual variants and their corresponding risk in the PRS will lead to more accurate breast cancer risk stratification of individuals in the population. Moreover, the PRS also modifies the breast cancer risk of individuals from breast cancer families either carrying a *BRCA1*, *BRCA2* or *CHEK2* germline mutation or not (12,13). For a 161-SNP PRS this caused, for example, a change in screening recommendation for 11.5% to 19.8% of the familial non-*BRCA1/2* mutation carriers, depending on the breast cancer screening guidelines used (14). Second, identification of the causal variants and the normal cellular functions they disturb will help the research community to understand the biological mechanisms driving breast cancer predisposition. This knowledge could eventually be used to develop novel preventive and therapeutic strategies for breast cancer.

Importantly, the vast majority of GWAS-identified variants are located in non-coding regions, making it not immediately evident which of the variants in close LD is causal and what the function of the variant is. Moreover, it is a common theme among GWAS-identified loci that more

than one causal variant is located in the region, reflected by multiple breast cancer risk signals (11,15). Causal variants are usually involved in regulating target genes, proximal or distal. As a consequence, they are most likely located in regions of open chromatin, binding sites of transcription factors (TFBSs) or histone modification or chromatin interaction sites (10,11). Several regulatory features have been found to be enriched among credible causal variants (CCVs) from breast cancer risk loci as compared with non-CCVs from breast cancer risk loci. These are open chromatin, actively transcribed genes as defined by H3K36me3 or H3K79me2 histone marks, distal gene regulatory regions as defined by H3K4me1 or H3K27ac marks and TFBSs for ER, FOXA1, GATA3, E2F1, TCF7L2, EP300 and 34 other transcription factors (TFs) (5,15,16). The list of TFs is, however, still uncomplete as cistrome (i.e., genome-wide binding site) data from many TFs is currently not available.

Causal variants may also regulate more than one target gene (11). Target gene predictions using the INQUISIT tool have shown that genes regulated by CCVs are enriched for somatic cancer driver genes and TF genes. Moreover, pathway analyses revealed involvement of these genes in developmental processes, the immune system, apoptosis and DNA integrity checkpoints (15). This is a broader spectrum of cellular functions than we observe for the moderate to high-risk breast cancer predisposition genes.

Evidence that breast cancer CCVs may also regulate non-coding (nc) RNAs comes from studying the 11q13 breast cancer risk locus. Betts *et al.* showed that two CCVs at this locus fall within a distal enhancer that does not only regulate *CCND1*, but also two estrogen-regulated long ncRNAs (lncRNAs) named *CUPID1* and *CUPID2* (17). Because novel ncRNAs are continuously being identified and the full catalogue of existing ncRNAs is far from complete, the extent to which deregulation of ncRNAs is underlying predisposition to breast cancer or other complex traits is still unclear.

Therefore, in their recent paper Moradi Marjaneh *et al.* studied the extent of ncRNA deregulation by breast cancer CCVs (18). More specifically, the authors performed RNA CaptureSeq to identify multi-exonic ncRNAs (mencRNAs) that are transcribed from intronic and intragenic regions spanning 750 kb upstream and 750 kb downstream of 139 GWAS-identified breast cancer risk signals. Samples selected for this study included organoids from four normal breast samples, primary human mammary epithelial cells, four breast tumor samples, two normal breast epithelial cell lines and three ER-negative and three ER-positive breast cancer cell lines including MCF-7 cells that were either treated or untreated with estradiol (18). Since the regulation of protein-coding genes, and most likely also mencRNAs, is tissue cell-type specific and tumor-subtype specific and since causal variants may confer risk to either ER-positive or ER-negative breast cancer or both, the authors were wise to include all these different sample types for mencRNA identification.

RNA CaptureSeq of the 139 breast cancer risk signals identified 1,254 (31.2%) known and 2,766 (68.8%) novel mencRNAs which were mostly two exon lncRNAs with a median length of 1,550 base pairs. Interestingly, nearly one third (i.e., 1,189) of the mencRNAs identified in MCF-7 cells were estrogen-regulated. Moreover, all 4,020 mencRNAs clustered the samples not only based on ER status, but also on tumor versus normal status and cell line versus tissue status (18). Although on average the identified mencRNAs had 5.5-fold lower expression than GENCODE lncRNAs and 140-fold lower expression than GENCODE protein-coding genes, roughly 75% (i.e., 3,011/4,020) of the mencRNAs were present in normal breast and breast tumor datasets generated by The Cancer Genome Atlas (TCGA). This allowed the assessment of tissue specificity of these mencRNAs by comparing breast tumors from TCGA with six other tumor types from TCGA. The analysis revealed that tissue specificity for the mencRNAs was even higher than for protein-coding genes (18).

Since the mere identification of novel mencRNAs in regions surrounding GWAS-identified variants does not imply a direct relation, the authors subsequently assessed the frequency of CCVs in mencRNAs as compared to protein-coding genes. Interestingly, CCVs were found to be enriched in the exons, but not the introns of mencRNAs. At half (i.e., 69) of the breast cancer risk signals a total of 119 mencRNAs harbored at least one CCV. In contrast, for protein-coding genes, CCVs were enriched in the introns, but not the exons (18). These results indicate that CCVs may alter mencRNA function more directly instead of influencing expression indirectly via the modulation of TFs or enhancers as for protein-coding genes. An example of a CCV directly impairing a lncRNA can be found at the 2q21.2 GWAS-identified locus associated with celiac disease. Here, the secondary structure of lncRNA *Lnc13* is altered by the CCV and consequently *Lnc13* has lower binding affinity for hnRNPD (19). However, in the study by Betts *et al.* two CCVs were located in a distal enhancer and abolished the ability of this enhancer to interact with

the promoter region of *CUPID1* and *CUPID2* at the 11q13 breast cancer risk locus (17). This suggests that distal regulation of ncRNAs by breast cancer CCVs does occur.

Because of the observed enrichment of CCVs in mencRNA exons, the authors hypothesized that CCVs might modulate the stability of mencRNAs. To verify this, the authors performed an expression quantitative trait loci (eQTL) analysis within the TCGA breast tumor dataset in which mencRNAs were identified whose expression was highly correlated with the genotype of genetic variants. In total, 800 mencRNAs were eQTLs. However, after filtering out the eQTLs which overlapped breast cancer risk signals and further evaluation of co-localization, only seven eQTLs remained. These were *XLOC_022678* at 1q32, *XLOC_142280* at 2q31, *XLOC_169717* at 3p26, *XLOC_195543* at 5q14, *XLOC_209276* at 6p23, *XLOC_093918* at 16q12 and *XLOC_112072* at 18q11 (18). It is important to underscore here that the eQTL analysis was performed in the TCGA breast cancer dataset and not in normal breast tissue. Besides confounding issues with copy number variation in cancer datasets, somatic breast cancer driver mutations and epigenetics have also already deregulated the expression of many genes. As these tumorigenic processes may also affect mencRNAs, eQTLs could have been missed. Unfortunately, large datasets from normal breast tissue with sufficient power for eQTL analyses are lacking at the moment, but are very much needed to allow for the assessment of eQTLs in healthy breast tissue.

Interestingly, out of the seven eQTLs only three (i.e., *XLOC_142280* at 2q31, *XLOC_209276* at 6p23 and *XLOC_112072* at 18q11) had at least one of the eQTL variants located in a mencRNA exon (18). To provide more evidence that these mencRNAs could indeed be the targets for the CCVs at these breast cancer risk loci, the authors subsequently zoomed in on the 2q31 locus where mencRNA *XLOC_142280* is located. Via eQTL analysis the authors had already shown that the risk alleles of the CCVs associate with a reduced expression of *XLOC_142280*. However, they also find that no other annotated protein-coding gene is associated with the CCVs at this breast cancer risk signal, making *XLOC_142280* the only candidate target. Moreover, *XLOC_142280* also appears to be predominantly expressed in ER-positive breast cancers, while the CCVs at this breast cancer risk signal associate with risk for ER-positive breast cancer only (18). More conclusive proof, however, that mencRNA *XLOC_142280* is the actual target at this breast cancer risk signal will require functional studies in which expression of this mencRNA is studied in an isogenic ER-positive breast cell model comparing the risk allele of the exon-localized CCVs rs11675683 and rs2356791 with the reference allele.

For the other four eQTLs (i.e., *XLOC_022678* at 1q32, *XLOC_169717* at 3p26, *XLOC_195543* at 5q14, and *XLOC_093918* at 16q12) none of the eQTL variants was located in a mencRNA exon. Therefore the authors hypothesized that the CCVs might regulate the mencRNAs distally (18). The authors gathered evidence for their hypothesis by zooming in on the 16q12 breast cancer risk locus and making use of the variant Capture Hi-C data that they had generated for the chromatin interactome mapping paper accompanying their original paper (20). By doing so, the authors identified 770 mencRNA promoters that looped to a region containing a CCV. At the 16q12 breast cancer risk locus, one of the two eQTL variants for *XLOC_093918* (i.e., rs11642015) appeared to interact with *XLOC_093918* in normal breast cells, as well as ER-positive and ER-negative breast cancer cell lines. In agreement with this, rs11642015 is located in a putative enhancer as it falls within a region of open chromatin that is marked by H3K27ac and H3K4me1. However, their data also showed that CCV rs11642015 additionally interacts with the bidirectional promoter of *IRX5* and *CRNDE*. Interestingly, there were no eQTLs detected for either of these genes or any other gene, whereas the authors had shown that CCV rs11642015 is an eQTL variant of *XLOC_093918* and that the risk allele increases its expression (18). Similar to the 2q31 breast cancer risk locus, however, more conclusive proof that mencRNA *XLOC_142280* is the actual target at this breast cancer risk signal should come from further functional studies. In these experiments, the expression of *XLOC_093918*, and perhaps also the other genes with which rs11642015 interacts, will need to be measured in an isogenic breast cell model comparing the risk allele of rs11642015 with the reference allele.

Finally, the authors provide evidence that mencRNAs can be a target gene for more than one breast cancer risk signal (18), a concept that has also been shown for protein-coding genes (15,20). The authors identified 222 mencRNAs that had a minimum of two independent CCVs located in either the exon or promoter region, or a region that interacts with the promoter of the mencRNA. As an example, the authors highlight the 18q11 breast cancer risk locus were for signal 3 two CCVs, which are also eQTL variants, are located in an exon of *XLOC_112072*. In addition, the authors show that CCVs at signals 1 and 2 interact with the promoter of *XLOC_112072* in T47D breast cancer cells. In B80T5

normal breast cells, however, there is only an interaction between CCVs at signal 1 and the *XLOC_112072* promoter (18). Since there is no protein-coding eQTL identified for either of the signals, *XLOC_112072* appears to be the candidate target gene for at least two of the three breast cancer risk signals at the 18q11 locus. Again, more conclusive proof that mencRNA *XLOC_112072* is the actual target gene should come from functional studies in which the effect of the reference and risk alleles of the CCVs in the different signals, alone and combined, on the expression of *XLOC_112072* is measured. As we know from the 5q11.2, 6q25.1 and 19p13.1 breast cancer risk loci, it is not uncommon that CCVs at one signal increase the target genes expression, while CCVs at the other signal decrease the target genes expression (21-23).

With their study Moradi Marjaneh *et al.* (18) provide accumulating evidence that deregulation of ncRNAs by causal variants lying in close LD with GWAS-identified variants is an important mechanism underlying predisposition to breast cancer. To date, the focus of post-GWAS studies has been on protein-coding genes as the targets for these causal variants and at several breast cancer risk loci no plausible candidate target genes have been identified. A major reason for this has been the incomplete catalogue of human ncRNAs. The authors have shown with their study, although focusing specifically on mencRNAs, that ncRNAs could potentially fill that void. Together with the interactome data in their accompanying paper and the fine-mapping study of Fachal *et al.* (15,20), valuable datasets have now become available that will drive continuing post-GWAS studies forward. Although we are still years from figuring out the exact mechanism at each of the breast cancer GWAS-identified loci, these datasets will give the research community novel clues for these mechanisms. In their paper, the authors have already highlighted three of these loci for which mencRNAs are candidate targets. Validating these findings in functional studies would add to the study of Betts *et al.* (17) and provide compelling evidence for the role of ncRNAs in breast cancer predisposition.

Acknowledgments

Funding: This work is supported in part by a grant from the Dutch Cancer Society (KWF 10758/2016-2).

Footnote

Provenance and Peer Review: This is an invited article commissioned by the editorial office, *Non-coding RNA Investigation*. The article did not undergo external peer review.

Conflicts of Interest: The author has completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/ncri.2020.03.05). AH reports and that she is a member of the Breast Cancer Association Consortium and consequently collaborates within this context with several authors of the original study.

Ethical Statement: The author is accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. Nat Genet 2008;40:17-22.
2. Hollestelle A, Wasielewski M, Martens JW, Schutte M. Discovering moderate-risk breast cancer susceptibility genes. Curr Opin Genet Dev 2010;20:268-76.
3. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet 2013;45:353-61, 361e1-2.
4. Michailidou K, Beesley J, Lindstrom S, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet 2015;47:373-80.
5. Michailidou K, Lindström S, Dennis J, et al. Association analysis identifies 65 new breast cancer risk loci. Nature 2017;551:92-4.
6. Milne RL, Kuchenbaecker KB, Michailidou K, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet 2017;49:1767-78.
7. Lilyquist J, Ruddy KJ, Vachon CM, et al. Common genetic

variation and breast cancer risk-past, present, and future. Cancer Epidemiol Biomarkers Prev 2018;27:380-94.
8. 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;104:21-34.
9. Schmidt MK, Hogervorst F, van Hien R, et al. Age- and tumor subtype-specific breast cancer risk estimates for CHEK2*1100delC carriers. J Clin Oncol 2016;34:2750-60.
10. Edwards SL, Beesley J, French JD, et al. Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet 2013;93:779-97.
11. Rivandi M, Martens JWM, Hollestelle A. Elucidating the underlying functional mechanisms of breast cancer susceptibility through post-GWAS analyses. Front Genet 2018;9:280.
12. Kuchenbaecker KB, McGuffog L, Barrowdale D, et al. Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. J Natl Cancer Inst 2017. doi: 10.1093/jnci/djw302.
13. Muranen TA, Greco D, Blomqvist C, et al. Genetic modifiers of CHEK2*1100delC-associated breast cancer risk. Genet Med 2017;19:599-603.
14. Lakeman IMM, Hilbers FS, Rodríguez-Girondo 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 2019;56:581-9.
15. Fachal L, Aschard H, Beesley J, et al. Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. Nat Genet 2020;52:56-73.
16. Cowper-Sal lari R, Zhang X, Wright JB, et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nat Genet 2012;44:1191-8.
17. Betts JA, Moradi Marjaneh M, Al-Ejeh F, et al. Long noncoding RNAs CUPID1 and CUPID2 mediate breast cancer risk at 11q13 by modulating the response to DNA damage. Am J Hum Genet 2017;101:255-66.
18. Moradi Marjaneh M, Beesley J, O'Mara TA, et al. Non-coding RNAs underlie genetic predisposition to breast cancer. Genome Biol 2020;21:7.
19. Castellanos-Rubio A, Fernandez-Jimenez N, Kratchmarov R, et al. A long noncoding RNA associated with susceptibility to celiac disease. Science 2016;352:91-5.
20. Beesley J, Sivakumaran H, Moradi Marjaneh M, et al. Chromatin interactome mapping at 139 independent breast cancer risk signals. Genome Biol 2020;21:8.
21. Glubb DM, Maranian MJ, Michailidou K, et al. Fine-scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating MAP3K1. Am J Hum Genet 2015;96:5-20.
22. Dunning AM, Michailidou K, Kuchenbaecker KB, et al. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. Nat Genet 2016;48:374-86.
23. Lawrenson K, Kar S, McCue K, et al. Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast-ovarian cancer susceptibility locus. Nat Commun 2016;7:12675.

doi: 10.21037/ncri.2020.03.05
Cite this article as: Hollestelle A. Another step forward towards unraveling the biological mechanisms driving breast cancer predisposition: a role for non-coding RNAs. Non-coding RNA Investig 2020;4:3.

