Targeted resequencing of the coding sequence of 38 genes near breast cancer GWAS loci in a large case-control study

Running title: Genes near breast cancer GWAS loci not associated with risk

Brennan Decker,[1,2,*] Jamie Allen,[1] Craig Luccarini,[3] Karen A. Pooley,[1] Mitul Shah,[3] Manjeet K. Bolla,[1] Qin Wang,[1] Shahana Ahmed,[3] Caroline Baynes,[1] Don M. Conroy,[1] Judith Brown,[1] Robert Luben,[1] Elaine A. Ostrander,[2] Paul D.P. Pharoah,[1,3] Alison M. Dunning,[3] Douglas F. Easton[1,3]

1 Centre for Cancer Genetic Epidemiology, Department of Public and Primary Care, University of Cambridge, Cambridge, UK

2 Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA

3 Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK

* Current affiliation: Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts, USA

Please address correspondence to: Douglas F. Easton (dfe20@medschl.cam.ac.uk)

Disclosure of Potential Conflicts of Interest: No potential conflicts of interest were disclosed.

ABSTRACT

Background: Genes regulated by breast cancer (BC) risk alleles identified through genome-wide association studies (GWAS) may harbour rare coding risk alleles.

Methods: We sequenced the coding regions for 38 genes within 500kb of 38 lead GWAS SNPs in 13,538 breast cancer cases and 5,518 controls.

Results: Truncating variants in these genes were rare, and were not associated with BC risk. Burden testing of rare missense variants highlighted five genes with some suggestion of an association with BC, though none met the multiple testing threshold: *MKL1*, *FTO*, *NEK10*, *MDM4*, and *COX11*. Six common alleles in *COX11, MAP3K1* (two), and *NEK10* (three) were associated at the $P<0.0001$ significance level, but these likely reflect linkage disequilibrium with causal regulatory variants.

Conclusions: There was no evidence that rare coding variants in these genes confer substantial breast cancer risks. However, more modest effect sizes could not be ruled out.

Impact: We tested the hypothesis that rare variants in 38 genes near breast cancer GWAS loci may mediate risk. These variants do not appear to play a major role breast cancer heritability.

Genome-wide association studies (GWAS) have identified ~180 risk loci, (1) all of which are common variants that confer modest disease risks. Fine-mapping and functional analyses suggest that most causal variants modulate risk via regulatory effects, though a few lead SNPs, including in *DCLRE1B* and *EXO1*, are missense substitutions.(1) In some diseases, GWAS association signals have been shown to be mediated, at least in part, by rare, high-risk coding variants in nearby genes.(2) Moreover, even if GWAS signals are due to regulatory variants, rare coding variants in the target genes are biologically plausible candidates for modulating risk. In this study, we tested this hypothesis by sequencing the coding exons and intron-exon boundaries of 38 genes that are potential targets for GWAS-identified causal variants.

The subjects, DNA enrichment, sequencing, and variant calling employed in this study have been described elsewhere.(3) Sequencing primers, coverage statistics, quality metrics, and variants are in **Tables S1-S4**.

A total of 3,839 variants were identified, and most were rare, with 3,564 (92.8%) found in <0.1% of all sequenced subjects (non-coding variants) or ExAC European subjects (coding variants) (**Figure S1**, **Tables S3-S4**). Only 131 truncating variants were identified, and all were uncommon in the population (**Tables S3-S4**). Burden testing showed that truncating variants were not associated with risk for any gene, even at the nominal significance threshold (P<0.05; **Table S5**).

The aggregate of rare missense variants were not associated for any gene at P<0.0001; however, five genes were associated at P<0.05 (**Figure 1, Table 1**). Stratification with SIFT, PolyPhen2, and CADD effect predictions showed that only *NEK10* variants with a CADD score >20 conferred a significantly higher risk than

predicted benign variants (OR=2.73 and 0.86, respectively; P-diff=0.010) (**Table S6**). This signal was partially driven by variants within the highly conserved *NEK10* protein kinase domain, which were more strongly associated than variants outside of the domain (**Table 1;** P-diff=0.033). In contrast, for *MDM4*, the association was stronger for variants outside Pfam-defined domains (**Table 1**).

Six common variants were associated with BC risk at P<0.0001 (**Table S7**), and all were in LD with the reported lead GWAS SNP: a 3'-UTR variant in *COX11* (rs1802212), two synonymous variants in *MAP3K1* (p.Gln1028Gln, rs3822625; and p.Thr522Thr, rs2229882) and one missense and two synonymous variants in *NEK10* (p.Lys513Ser, rs10510592; p.Thr670Thr, rs11129280; and p.Thr687Thr,rs3213930). In each case the strength of the associations were compatible with those seen in Michailidou *et al.* (1), but the associations were much weaker than for the corresponding lead SNP (rs2787486, rs62355902 and rs4973769).

Among variants associated at P<0.05, *DCLRE1B* p.His49Tyr (rs11552449; OR=1.10) was also the lead GWAS SNP and conferred a similar risk to that seen in the initial study (OR=1.07, $P=1.8x10^{-8}$).(4)

Conclusion

Exon sequencing of genes in GWAS regions did not identify clear novel associations. There was no evidence for association with truncating variants in any genes, and while the variants were too rare to establish reliable estimates, these are unlikely to be large contributors to BC risk. There was limited evidence of association for

rare missense variants in five genes, while 1.9 genes would have been expected to be associated by chance. Larger targeted studies will be required to establish whether any of these associations can be confirmed; if so, these may indicate novel associations distinct from the common variant associations identified through GWAS.

Six common variants were associated with BC after correcting for multiple testing, but all were in LD with the lead GWAS SNP, and therefore probably do not represent novel risk loci. Moreover, other non-coding SNPs in these regions were more strongly associated, suggesting that these associations are “passenger” associations reflecting LD with causal regulatory variants.(5)

REFERENCES

1. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature [Internet]. 2017 [cited 2017 Dec 20];551:92–4. Available from: http://www.nature.com/doifinder/10.1038/nature24284
2. van Leeuwen EM, Sabo A, Bis JC, Huffman JE, Manichaikul A, Smith A V, et al. Meta-analysis of 49 549 individuals imputed with the 1000 Genomes Project reveals an exonic damaging variant in ANGPTL4 determining fasting TG levels. J Med Genet [Internet]. 2016 [cited 2016 May 28]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/27036123
3. Decker B, Allen J, Luccarini C, Pooley KA, Shah M, Bolla MK, et al. Rare, protein-truncating variants in ATM, CHEK2 and PALB2, but not XRCC2, are associated with increased breast cancer risks. J Med Genet. 2017;54.
4. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet [Internet]. 2013 [cited 2013 Mar 27];45:353–61. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23535729
5. Milne RL, Burwinkel B, Michailidou K, Arias-Perez J-I, Zamora MP, Menéndez-Rodríguez P, et al. Common non-synonymous SNPs associated with breast cancer susceptibility: findings from the Breast Cancer Association Consortium. Hum Mol Genet [Internet]. 2014 [cited 2016 Jul 9];23:6096–111. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24943594

ACKNOWLEDGEMENTS

BD and EAO were supported by the Intramural Research Program of the National Human Genome Research Institute. SEARCH is funded by a programme grant from Cancer Research UK (C490/A10124) and supported by the UK National Institute for Health Research Biomedical Research Centre at the University of Cambridge. Targeted sequencing in SEARCH was supported by Cancer Research UK grants C1287/A16563 to DFE and C8197/A16565 to AMD.

TABLE

Table 1. *Risk estimates for all rare missense variants in the 38 GWAS genes, as well as subsets of variants that are either localized within or outside of Pfam domain regions.*

	Overall Risk				In Pfam Domain				Not In Pfam Domain				
Gene	Case Carriers	Control Carriers	OR (95%CI)	P	Case Carriers	Control Carriers	OR (95%CI)	P	Case Carriers	Control Carriers	OR (95%CI)	P	P-Diff
C6orf211	60	21	1.17 (0.70-1.92)	0.63	54	17	1.30 (0.75-2.24)	0.42	6	4	0.61 (0.17-2.17)	0.49	0.28
CASP8	57	32	0.72 (0.47-1.12)	0.18	34	18	0.77 (0.43-1.36)	0.45	23	14	0.67 (0.34-1.30)	0.31	0.93
CCDC170	116	50	0.95 (0.68-1.32)	0.81	0	0	NA	1.00	116	50	0.95 (0.68-1.32)	0.81	NA
CCND1	25	14	0.73 (0.38-1.40)	0.44	20	13	0.63 (0.31-1.26)	0.26	5	1	2.04 (0.24-17.5)	0.68	0.55
CDCA7	38	13	1.19 (0.63-2.24)	0.70	3	1	1.22 (0.13-11.8)	1.00	35	12	1.19 (0.62-2.29)	0.72	1.00
CDKN2A	38	22	0.70 (0.42-1.19)	0.24	78	31	1.03 (0.68-1.56)	0.99	3	2	0.61 (0.10-3.66)	0.63	0.96
CDKN2B	14	4	1.42 (0.47-4.34)	0.61	0	0	NA	1.00	14	4	1.43 (0.47-4.34)	0.61	NA
CDYL2	144	46	1.23 (0.92-1.79)	0.17	69	26	1.08 (0.69-1.70)	0.82	75	20	1.53 (0.93-2.51)	0.11	0.40
COX11	**37**	**6**	**2.52 (1.06-5.97)**	**0.045**	9	1	3.67 (0.46-29.0)	0.30	28	5	2.29 (0.88-5.92)	0.086	1.00
DCLRE1B	112	44	1.04 (0.73-1.47)	0.91	36	16	0.92 (0.51-1.65)	0.89	76	28	1.11 (0.72-1.71)	0.73	0.75
DNAJC1	89	47	0.77 (0.54-1.10)	0.18	27	15	0.73 (0.39-1.38)	0.43	62	32	0.79 (0.51-1.21)	0.33	1.00
ESR1	78	34	0.93 (0.62-1.40)	0.82	54	25	0.88 (0.55-1.42)	0.69	24	9	1.09 (0.50-2.34)	0.98	0.82
EXO1	186	85	0.89 (0.69-1.15)	0.42	24	13	0.75 (0.38-1.48)	0.52	162	72	0.92 (0.69-1.21)	0.59	0.73
FGFR2	64	25	1.04 (0.66-1.66)	0.95	33	11	1.22 (0.62-2.42)	0.68	31	14	0.90 (0.48-1.70)	0.88	0.69
FOXQ1	20	9	0.91 (0.41-1.99)	0.97	15	5	1.22 (0.44-3.37)	0.89	5	4	0.51 (0.14-1.90)	0.29	0.40
FTO	**143**	**37**	**1.58 (1.10-2.27)**	**0.016**	**139**	**35**	**1.63 (1.12-2.36)**	**0.012**	4	2	0.82 (0.15-4.45)	1.00	0.78
HNF4G	46	17	1.10 (0.63-1.93)	0.84	27	7	1.57 (0.68-3.61)	0.37	19	10	0.77 (0.36-1.67)	0.65	0.34
IGFBP5	43	15	1.17 (0.65-2.11)	0.71	22	6	1.50 (0.61-3.69)	0.50	21	9	0.95 (0.44-2.08)	1.00	0.66
MAP3K1	154	50	1.26 (0.91-1.73)	0.18	14	7	0.81 (0.33-2.02)	0.84	140	43	1.33 (0.94-1.88)	0.12	0.47
MDM4	**54**	**11**	**2.00 (1.05-3.84)**	**0.045**	5	5	0.41 (0.12-1.41)	0.17	**49**	**6**	**3.34 (1.43-7.79)**	**5.0E-03**	**9.0E-03**
MKL1	**310**	**90**	**1.41 (1.12-1.79)**	**4.8E-03**	9	2	1.83 (0.40-8.49)	0.74	**301**	**88**	**1.40 (1.10-1.78)**	**6.4E-03**	1.00
MRPS30	79	39	0.82 (0.56-1.21)	0.38	79	39	0.82 (0.56-1.21)	0.38	0	1	NA	0.29	0.73
MYC	65	26	1.02 (0.65-1.61)	1.00	65	26	1.02 (0.65-1.61)	1.00	0	0	NA	1.00	NA
NEK10	**118**	**30**	**1.61 (1.08-2.40)**	**0.025**	**32**	**2**	**6.53 (1.57-27.3)**	**1.9E-03**	86	28	1.25 (0.82-1.92)	0.35	**0.033**
NRBF2	32	11	1.19 (0.60-2.35)	0.75	28	10	1.14 (0.55-2.35)	0.86	4	1	1.63 (0.18-14.6)	1.00	1.00
NTN4	104	42	1.01 (0.70-1.45)	1.00	68	26	1.07 (0.68-1.68)	0.87	36	16	0.92 (0.51-1.65)	0.89	0.84
PDE4D	81	40	0.82 (0.56-1.21)	0.37	18	4	1.84 (0.62-5.43)	0.35	63	36	0.71 (0.47-1.07)	0.13	0.13
PTH1R	78	26	1.22 (0.78-1.91)	0.43	40	15	1.09 (0.60-1.97)	0.90	38	11	1.41 (0.72-2.76)	0.40	0.73
PTHLH	13	2	2.65 (0.60-11.9)	0.26	6	0	5.30 (NA)	0.19	7	2	1.43 (0.30-6.87)	1.00	0.49
RAD51B	76	29	1.07 (0.70-1.64)	0.85	44	18	1.00 (0.58-1.73)	1.00	32	11	1.19 (0.60-2.35)	0.75	0.87
RNF115	51	18	1.16 (0.67-1.98)	0.69	8	3	1.09 (0.29-4.10)	1.00	43	15	1.17 (0.65-2.11)	0.71	1.00
TBX3	88	29	1.24 (0.81-1.89)	0.37	46	18	1.04 (0.60-1.80)	0.99	59	27	0.92 (0.58-1.47)	0.83	0.80
TERT	81	42	0.78 (0.54-1.14)	0.24	**22**	**2**	**4.49 (1.06-19.1)**	**0.024**	59	27	0.89 (0.56-1.41)	0.70	**0.034**
TET2	255	117	0.89 (0.71-1.11)	0.31	66	37	0.73 (0.48-1.09)	0.15	189	80	0.96 (0.74-1.25)	0.83	0.31
TGFBR2	45	18	1.02 (0.59-1.76)	1.00	28	14	0.81 (0.43-1.55)	0.65	17	4	1.73 (0.58-5.15)	0.47	0.37
TOX3	158	52	1.24 (0.91-1.70)	0.20	3	1	1.22 (0.13-11.8)	1.00	155	51	1.24 (0.90-1.72)	0.21	1.00
ZNF365	84	22	1.56 (0.97-2.50)	0.078	0	0	NA	1.00	84	22	1.56 (0.97-2.50)	0.078	NA

FIGURE

Figure 1. *Variant position and frequency for rare missense variants in* ***(A)*** MKL1; ***(B)*** FTO; ***(C)*** NEK10; ***(D)*** MDM4; *and* ***(E)*** COX11*. Variants are color-coded by CADD prediction. ZnF=Zinc Finger Domain; SAP=Putative DNA/RNA binding domain.*

Figure 1.

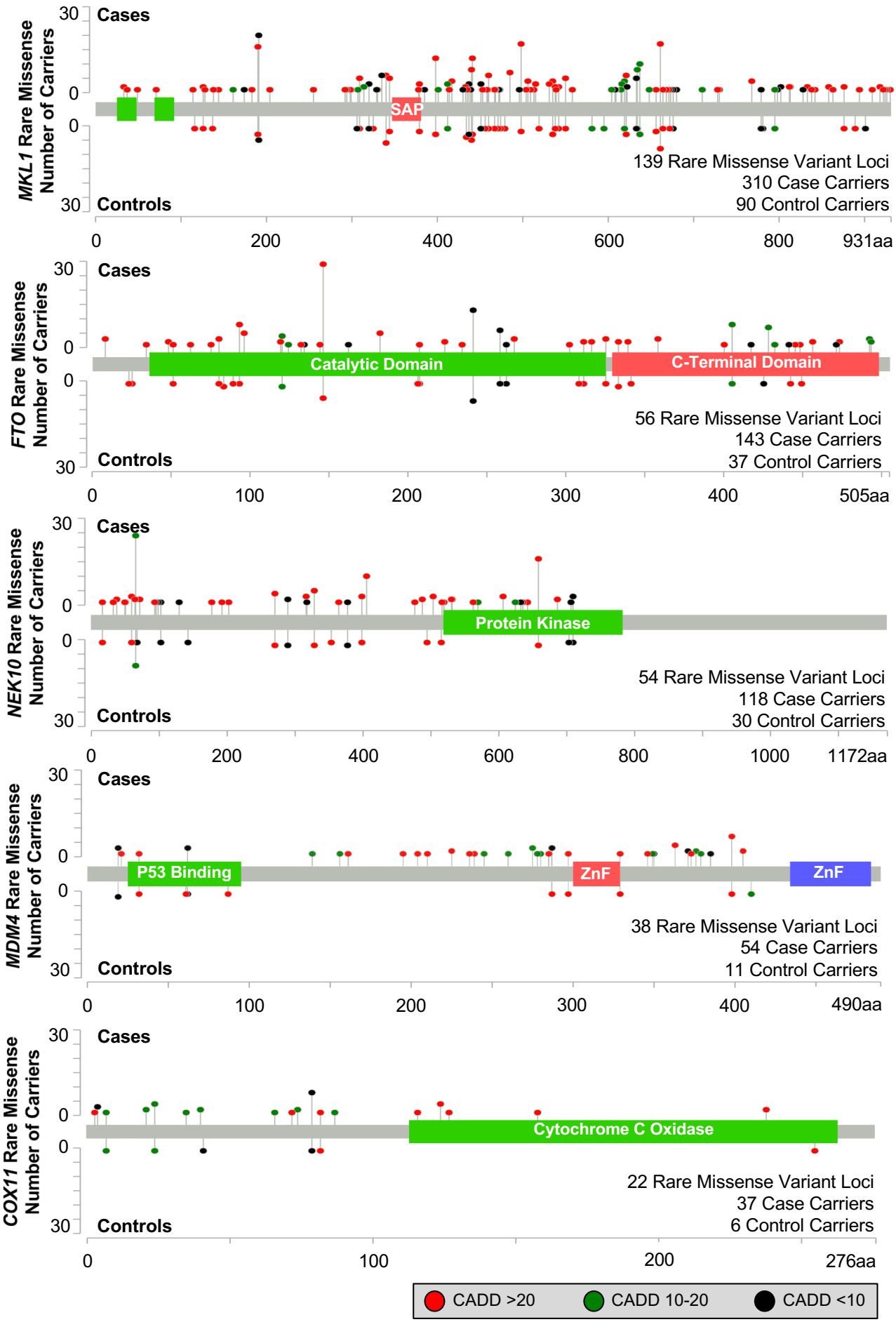
MKL1 Rare Missense Number of Carriers
Cases
Controls
30
0
0
30
SAP
139 Rare Missense Variant Loci
310 Case Carriers
90 Control Carriers
0
200
400
600
800
931aa
FTO Rare Missense Number of Carriers
Catalytic Domain
C-Terminal Domain
56 Rare Missense Variant Loci
143 Case Carriers
37 Control Carriers
100
300
505aa
NEK10 Rare Missense Number of Carriers
Protein Kinase
54 Rare Missense Variant Loci
118 Case Carriers
30 Control Carriers
1000
1172aa
MDM4 Rare Missense Number of Carriers
P53 Binding
ZnF
ZnF
38 Rare Missense Variant Loci
54 Case Carriers
11 Control Carriers
490aa
COX11 Rare Missense Number of Carriers
Cytochrome C Oxidase
22 Rare Missense Variant Loci
37 Case Carriers
6 Control Carriers
276aa
CADD >20
CADD 10-20
CADD <10

Figure S1.

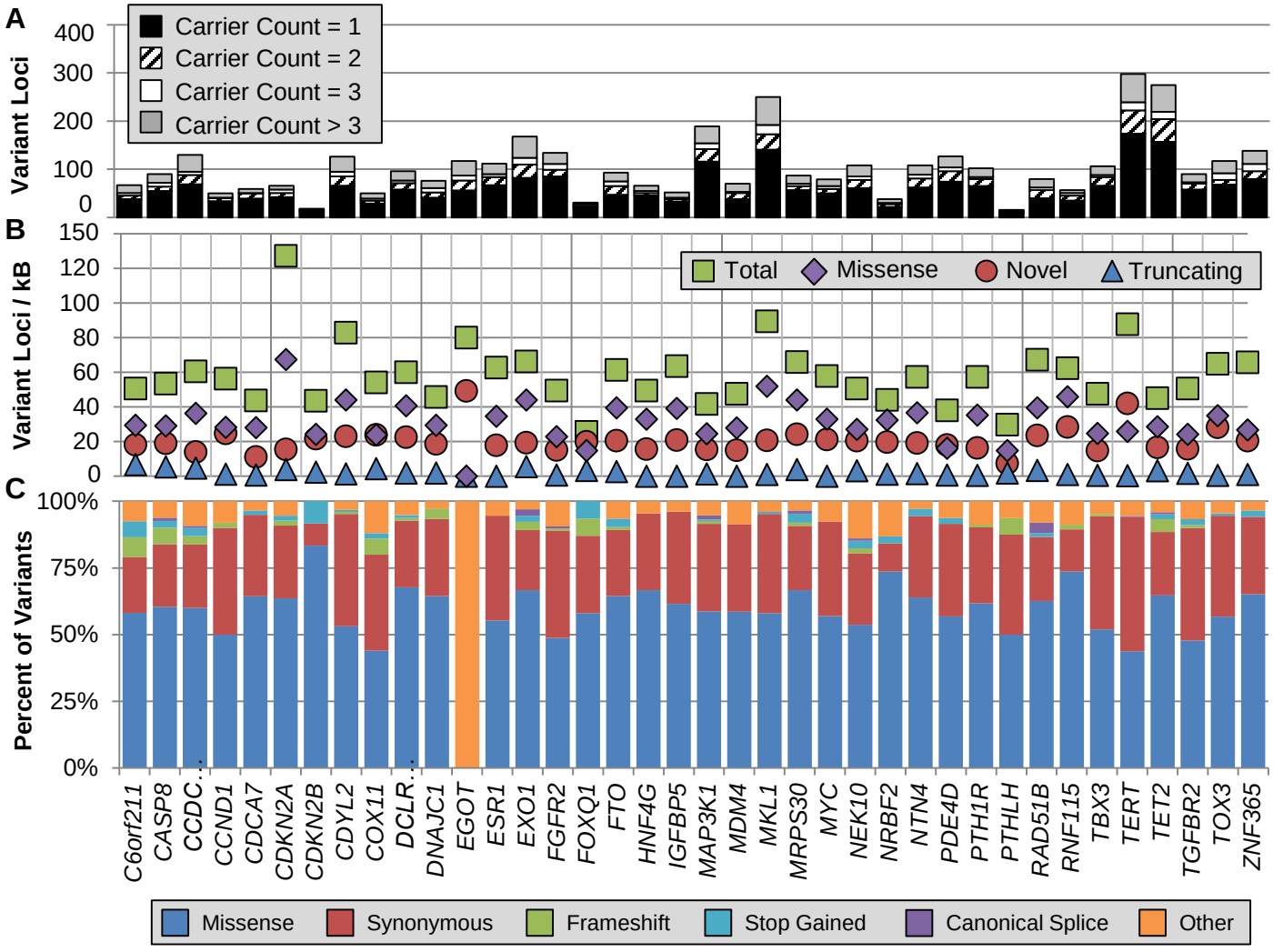

Figure S1. *Spectrum of variation in the 38 BC GWAS genes from Experiment 2.* ***(A)*** *The total variant burden varied significantly from gene to gene. More than half of all variants were singletons, and few variants were found in more than three subjects.* ***(B)*** *The rates of total, novel, and truncating variants were more similar after correcting for gene length, but some genes had higher or lower variant rates per kb.* ***(C)*** *The contribution of each variant type was similar from gene to gene, though some genes had a higher contribution from truncating variants.* EGOT *is a long, non-coding RNA and therefore variants in that gene all fall into the "Other" category, which is otherwise largely comprised of UTR and non-canonical splice variants. While they were very rare, start lost, stop lost, and stop retained variants were also included in this category.*