SUPPLEMENTARY APPENDIX

Supplementary Methods

Data harmonization and quality control

Raw genomic data of all samples were obtained from the respective data repository as previously described. All FASTQ and BAM files aligned to GRCh37 were realigned to hg19 using Picard tool kits (https://github.com/broadinstitute/picard). GATK (version 3.7) DepthOfCoverage [1] was used to determine the mean target coverage of germline samples. Germline variants in regions with a depth of coverage of fewer than 10 reads and those with a variant allelic fraction (VAF) <30% were excluded from all analyses.

Selection of Mendelian gene sets

In this study, we analyzed disease-causing variants in three gene sets, the germline cancer predisposition genes, the American College of Medical Genetics (ACMG) genes, and the Online Mendelian Inheritance in Men (OMIM) genes (Table S1). The germline cancer predisposition genes were selected based on the level of evidence supporting their Mendelian disease susceptibility. This is composed of the well-curated COSMIC germline cancer census gene set (v86; http://cancer.sanger.ac.uk/census) and the germline cancer gene set listed in Huang et al. 2018 [2] and Rahman 2014 [3]. Cancer genes with preliminary evidence of cancer association or those with no established inheritance pattern were removed. In total, 118 cancer predisposition genes (CPGs) were examined in the original cohort. In addition to cancer predisposition genes, we also examined 59 Mendelian high-penetrance genes associated with severe life-threatening diseases that have been deemed clinically actionable by the American College of Medical Genetics (ACMG). Given the well-established clinical utility, pathogenic variants in the ACMG genes are highly recommended to be disclosed to patients, even if discovered incidentally and regardless of the patient's phenotype [4]. Finally, we also expanded our head-to-head comparison of the examined methods by performing an exome-wide analysis of the clinically relevant genes by evaluating putative loss-of-function (LOF) variants in 5197 Mendelian disease-causing genes in the OMIM database (collectively called the OMIM genes) (https://www.omim.org/). In addition to these 3 gene sets, we also evaluated the number of validated pathogenic variants detected by each method in 12 multi-gene panels clinically used to evaluate cardiovascular disorders, ciliopathies, dermatological disorders, hearing loss, hematological disorders, mitochondrial disorders, neurological disorders, neuromuscular disorders, pulmonary disorders, renal disorders, retinal disorders, and expanded prenatal screening (Table S2).

Functional annotation

Germline variant annotation of all variants was performed using Variant Effect Predictor (VEP) (version 92.0) from Ensembl [5]. Only variants impacting the canonical transcript of the examined genes were included.

Germline variant pathogenicity evaluation

All detected germline variants in cancer-predisposition and ACMG gene sets were classified into 5 categories; benign, likely benign, variants of unknown significance, likely pathogenic and pathogenic using the American College of Medical Genetics and Genomics clinically-oriented guidelines [6]. Only germline variants that had sufficient evidence of pathogenicity, from established epidemiological or functional studies as defined by the widely adopted ACMG

guidelines, to be classified as pathogenic or likely pathogenic variants were included in this study (hereafter collectively referred to as pathogenic germline variants). Variants of unknown significance were excluded from all analyses (Supplementary Methods). For the expanded OMIM gene set (n=5197), we analyzed the performance of each germline variant detection tool by examining putative loss-of-function (pLOF) variants in this expanded gene set. Putative LOF variants were defined as 1) rare variants with minor allele frequency (MAF) <1% in all reference populations in gnomAD that are expected to produce a truncated gene product (i.e. stop codon, frameshift, and canonical splice site variants) and 2) rare missense variants (MAF<1%) that are annotated as pathogenic or likely pathogenic in the Clinical Variation database (ClinVar) (https://www.ncbi.nlm.nih.gov/clinvar/).

Validation of detected germline variants

All unfiltered pathogenic germline variants in the cancer predisposition and ACMG gene sets were validated by examining the Binary Alignment Map (BAM) file using the integrative Genomics Viewer (IGV; v2.3.81) [7]. IGV snapshots of pathogenic variants were generated using the IGV Snapshot Generator (https://github.com/stevekm/IGV-snapshot-automator). IGV snapshots of each called pathogenic variants (in the cancer predisposition gene and ACMG gene sets) were independently manually evaluated by 2 computational biologists with expertise in next-generation sequencing analysis. Variants were marked as “True Positive” or “False Positive” depending on the depth of sequencing, the number of alternative allele reads, the variant allelic fraction (VAF), and the presence of artifacts at or around the examined variant site. Variants that were called “True Positive” by both examiners were considered real variants. Otherwise, the variant was labeled as an artifactual call. In addition to pathogenic variants in the cancer predisposition and ACMG genes, 100 randomly selected pLOF variants in the OMIM gene sets, that were filtered out by GATK-JG in both computational runs, were validated manually using this method.

Performance metrics

We evaluated the ability of the standard variant detection method, GATK-JG, to detect clinically relevant variants in the cancer predisposition genes, the ACMG genes, and the OMIM genes in an original cohort of 239 samples in the presence and absence of additional jointly characterized germline samples. For each gene set, we looked at the absolute number of pathogenic germline variants classified as “high quality” by each computational run. The combined variant callset in the original cohort of 239 samples, generated by running GATK with and without an additional 100 samples, was considered the “reference” against which detection rates of each computational run were calculated.

Evaluation of systematic underdetection of filtered common variants

To evaluate if the probability of germline variants, filtered out by both computational runs, are truly absent in the original cohort (n=239), we used established minor allele frequencies of these variants in the gnomAD dataset [8]. For each variant, we used a two-sided binomial test where the null hypothesis is that the observed frequency in our dataset (zero for filtered our variants) is not significantly different from the smallest MAF of this variant across all continental sub-populations in gnomAD. Bonferroni correction was carried out for the total number of evaluated variants (n=284,515).

Adjusted GATK pipeline parameters for the additive analysis effect

To evaluate if there was an additive effect of detecting high quality germline variants when concurrently analyzing additional samples, we performed computational runs for the following cohorts: the original breast cancer cohort (n = 239), the original cohort plus 50 additional breast cancer samples, and the original cohort plus 100 additional breast cancer samples. Each cohort was analyzed using the Genome Analysis Toolkit (GATK) pipeline detailed in the Supplementary Notes with modifications to the VariantRecalibration (INDELs) step. The modified code is as follows:

```
java -jar -Xmx24G ~/GenomeAnalysisTK.jar \
-T VariantRecalibrator \
-R ~/Homo_sapiens_assembly19.fasta \
-input ~/[original_cohort].gvcf.gz \
-tranche 100.0 \
-tranche 99.9 -tranche 99.9 -tranche 99.8 -tranche 99.7 -tranche 99.6 -
tranche 99.5 \
-tranche 99.4 -tranche 99.3 -tranche 99.2 -tranche 99.1 -tranche 99.0 \
-tranche 98.9 -tranche 98.8 -tranche 98.6 -tranche 98.5 -tranche 98.3 \
-tranche 98.2 -tranche 98.1 -tranche 98.0 -tranche 97.9 -tranche 97.8 \
-tranche 97.5 -tranche 97.0 -tranche 95.0 -tranche 90.0 \
-resource:mills,known=false,training=true,truth=true,prior=12.0
~/Mills_and_1000G_gold_standard.indels.b37.vcf \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
~/dbsnp_138.b37.vcf \
-an FS -an QD -an MQRankSum -an ReadPosRankSum -an InbreedingCoeff \
-mode INDEL \
--maxGaussians 4 \
-recalFile ~/[original_cohort].INDEL.recal \
-tranchesFile ~/[original_cohort].INDEL.tranches \
-rscriptFile ~/[original_cohort].INDEL.R \
-nt 4
```

Supplementary Figures

Figure S1: Mean depth of sequencing coverage for the testicular cancer (A-D) and breast cancer cohorts (E-H).

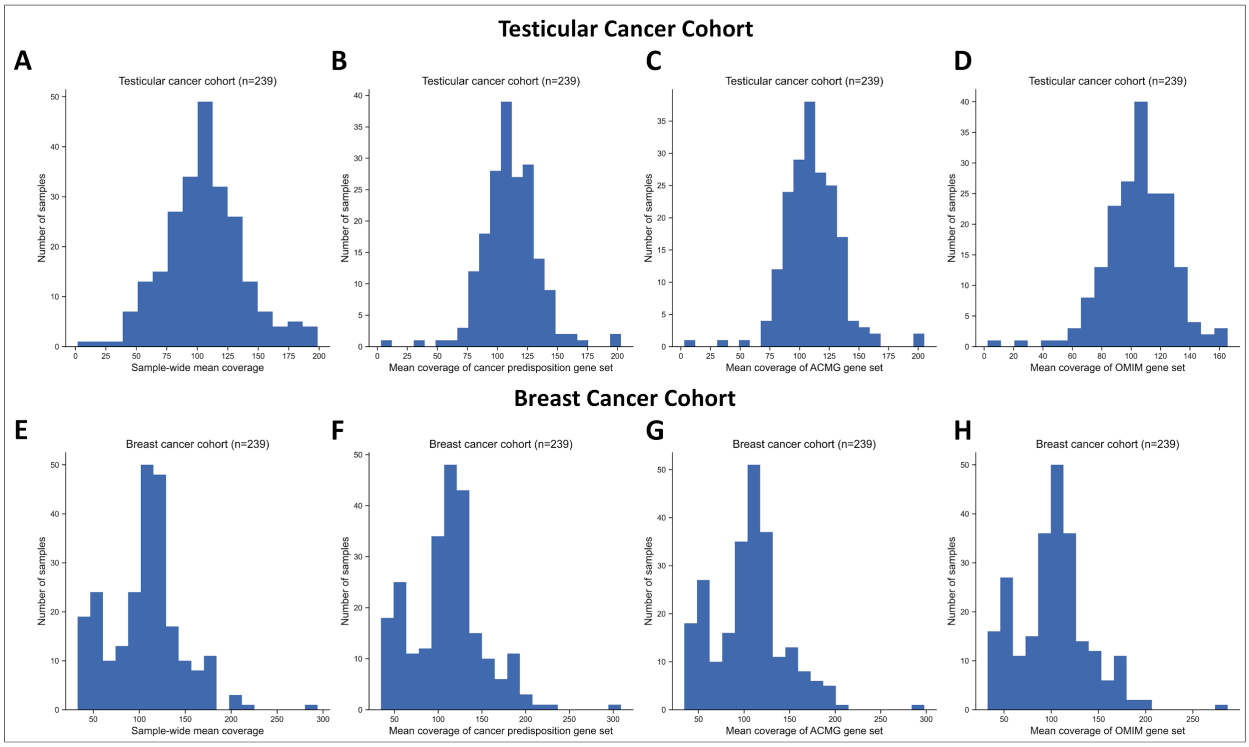


Figure S2: Minor allele frequency of the detected germline variants in the testicular and breast cancer cohorts

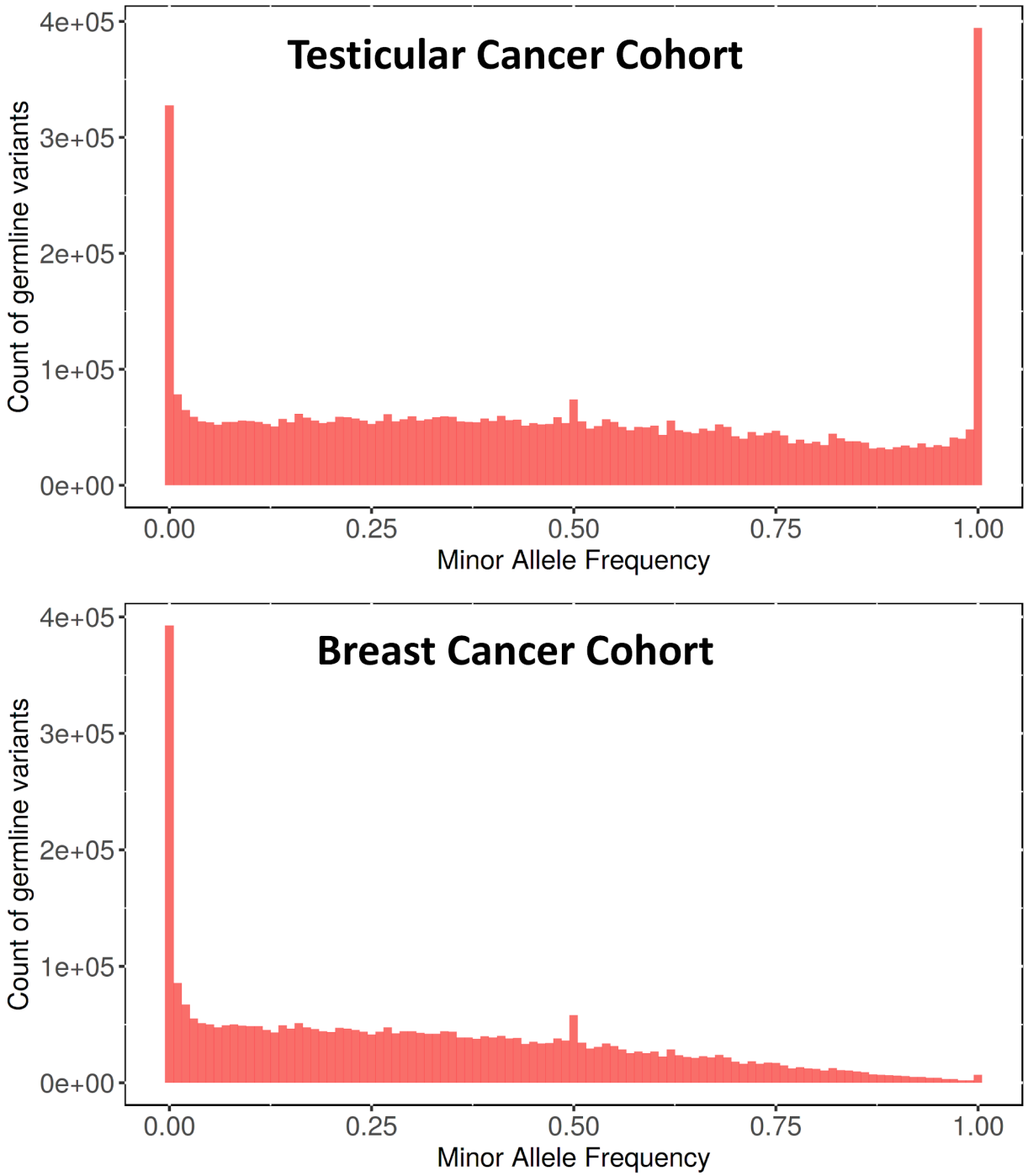


Figure S3: Pathogenic variants that went undetected by one or both computational runs in the clinically actionable cancer-predisposition and ACMG genes. IGV snapshots of the variants shown in Figures 3C-F and Figures 4B and C using GATK HaplotypeCaller "-bamout" option to view the tool's assembled haplotypes and locally aligned reads.

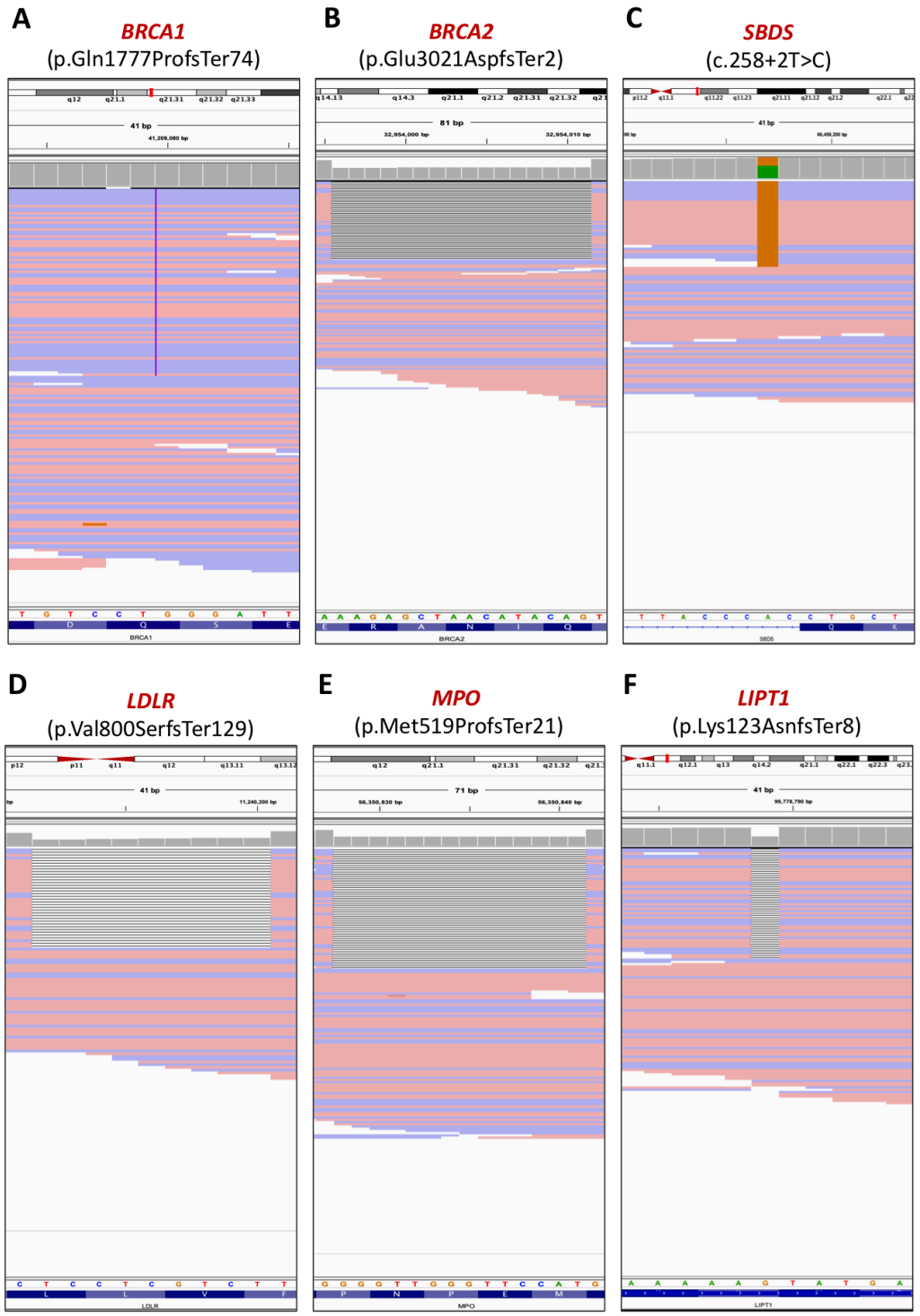
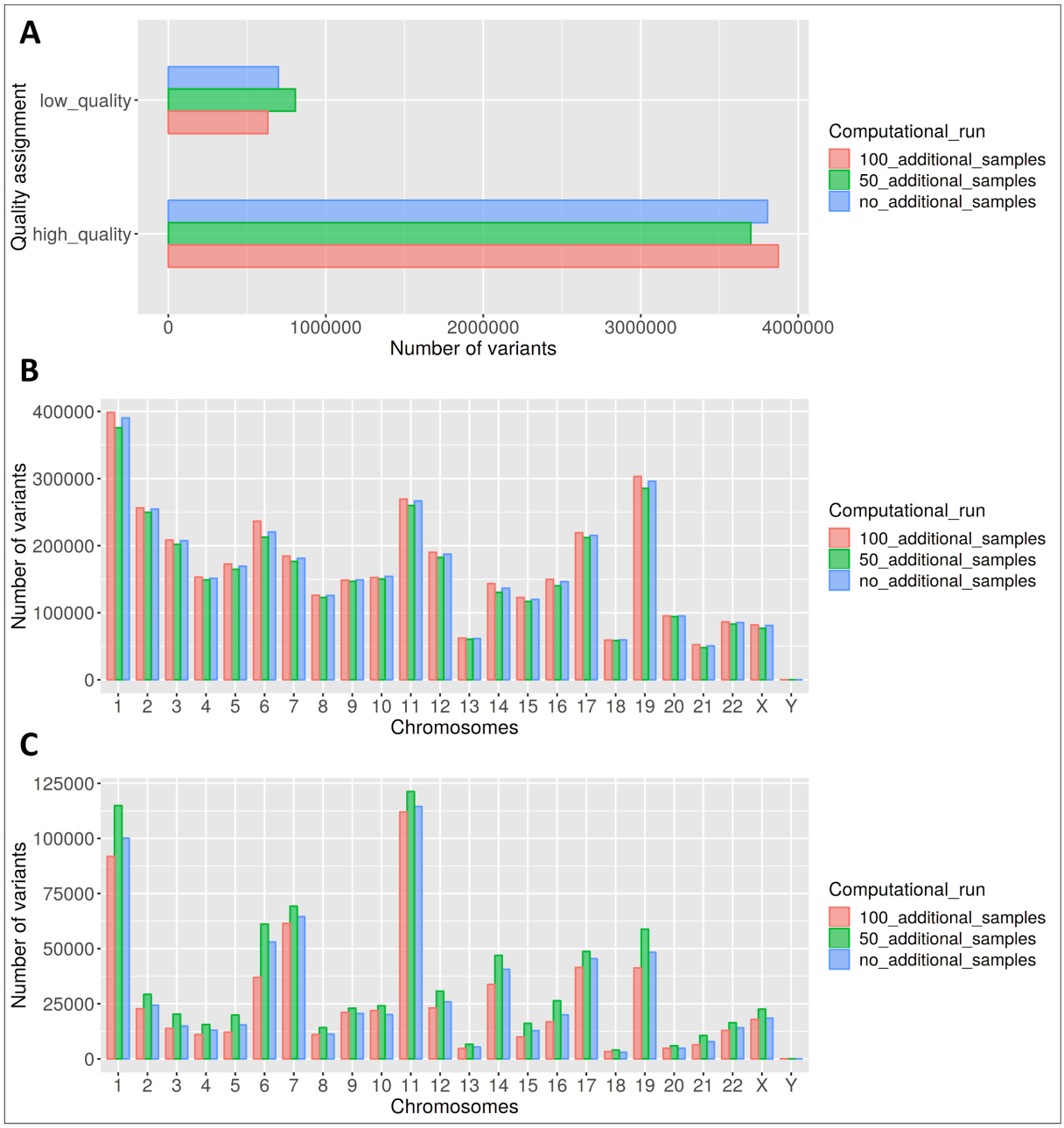


Figure S4: Three-way comparison of germline variant detection in the breast cancer cohort when concurrently analyzed with no additional samples, 50 additional samples, and 100 additional samples. A; Although more germline variants were detected by GATK-JG in the breast cancer cohort (n=239) when 100 additional germline exomes were concurrently characterized, this pattern was not seen when only 50 samples were used for joint genotyping. Indeed, germline variant analysis of the breast cancer cohort along with 50 additional germline exomes resulted in a substantially lower detection rate for germline variants. B & C; This underdetection of germline variants was not confined to particular genes or genomic regions.



Supplementary Notes:

1- Genome Analysis Toolkit (GATK) pipeline

Genome Analysis Toolkit (GATK) HaplotypeCaller (HC) pipeline (version 3.7) was used to call germline variants according to the GATK Best Practices. The following steps and commands were followed:

FIRST COMPUTATIONAL RUN (ORIGINAL COHORT ONLY):

1. HaplotypeCaller (HC): this command is run on each sample individually:

```
java -Xmx12G  -jar ~/GenomeAnalysisTK.jar \
-nct 8 \
-T HaplotypeCaller \
-R  ~/Homo_sapiens_assembly19.fasta \
-I  [single.sample.bam] \
--dbsnp ~/dbsnp_138.hg19.vcf.gz \
--genotyping_mode DISCOVERY \
-variant_index_type LINEAR \
-variant_index_parameter 128000 \
--emitRefConfidence GVCF \
--max_alternate_alleles 6 \
--minPruning 2 \
-stand_call_conf 30.0 \
-A DepthPerSampleHC \
-A StrandBiasBySample \
-A Coverage \
-A StrandBiasBySample \
-o ~/[single.sample].gvcf.gz
```

2. Joint genotyping (GenotypeGVCFs): this step combines all the gVCFs that were generated by the previous step to do cohort-wide genotyping:

```
java -jar -Xmx32G ~/GenomeAnalysisTK.jar \
-R ~/Homo_sapiens_assembly19.fasta -T GenotypeGVCFs \
--variant ~/[list_of_original_cohort_gVCFs].list  \
-L  ~/[capture_region].interval_list \
-o  ~/[original_cohort].gvcf.gz
```

3. VariantRecalibration (SNPs):

```
java -Xmx24G  -jar ~/GenomeAnalysisTK.jar \
-T VariantRecalibrator \
-R ~/Homo_sapiens_assembly19.fasta \
-input ~/[original_cohort].gvcf.gz \
-resource:hapmap,known=false,training=true,truth=true,prior=15.0
~/hapmap_3.3.b37.vcf \
-resource:omni,known=false,training=true,truth=true,prior=12.0
~/1000G_omni2.5.b37.vcf \
```

```
-resource:1000G,known=false,training=true,truth=false,prior=10.0
~/1000G_phase1.snps.high_confidence.b37.vcf \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
~/dbsnp_138.b37.vcf \
-an QD -an MQRankSum -an ReadPosRankSum -an FS -an MQ -an
InbreedingCoeff \
-mode SNP \
-tranche 100.0 \
-tranche 99.9 -tranche 99.9 -tranche 99.8 -tranche 99.7 -tranche
99.6 -tranche 99.5 \
-tranche 99.4 -tranche 99.3 -tranche 99.2 -tranche 99.1 -tranche
99.0 \
-tranche 98.9 -tranche 98.8 -tranche 98.6 -tranche 98.5 -tranche
98.3 \
-tranche 98.2 -tranche 98.1 -tranche 98.0 -tranche 97.9 -tranche
97.8 \
-tranche 97.5 -tranche 97.0 -tranche 95.0 -tranche 90.0 \
-recalFile ~/[original_cohort].SNP.recal \
-tranchesFile ~/[original_cohort].SNP.tranches \
-rscriptFile ~/[original_cohort].SNP.R \
-nt 4
```

4. Apply recalibration (SNP):

```
java -jar -Xmx24G ~/GenomeAnalysisTK.jar \
-T ApplyRecalibration \
-R ~/Homo_sapiens_assembly19.fasta \
-input ~/[original_cohort].gvcf.gz \
--ts_filter_level 99.5 \
-tranchesFile ~/[original_cohort].SNP.tranches \
-recalFile ~/[original_cohort].SNP.recal \
-mode SNP \
-o ~/[original_cohort].snp.recalibrated.vcf.gz
```

5. VariantRecalibration (INDELs):

```
java -jar -Xmx24G ~/GenomeAnalysisTK.jar \
-T VariantRecalibrator \
-R ~/Homo_sapiens_assembly19.fasta \
-input ~/[original_cohort].gvcf.gz \
-tranche 100.0 \
-tranche 99.9 -tranche 99.9 -tranche 99.8 -tranche 99.7 -tranche
99.6 -tranche 99.5 \
-tranche 99.4 -tranche 99.3 -tranche 99.2 -tranche 99.1 -tranche
99.0 \
-tranche 98.9 -tranche 98.8 -tranche 98.6 -tranche 98.5 -tranche
98.3 \
```

```
-tranche 98.2 -tranche 98.1 -tranche 98.0 -tranche 97.9 -tranche
97.8 \
-tranche 97.5 -tranche 97.0 -tranche 95.0 -tranche 90.0 \
-resource:mills,known=false,training=true,truth=true,prior=12.0
~/Mills_and_1000G_gold_standard.indels.b37.vcf \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
~/dbsnp_138.b37.vcf \
-an FS -an QD -an ReadPosRankSum -an InbreedingCoeff \
-mode INDEL \
-recalFile ~/[original_cohort].INDEL.recal \
-tranchesFile ~/[original_cohort].INDEL.tranches \
-rscriptFile ~/[original_cohort].INDEL.R \
-nt 4
```

6. Apply recalibration (INDELS):

```
java -jar -Xmx24G ~/GenomeAnalysisTK.jar \
-T ApplyRecalibration \
-R ~/Homo_sapiens_assembly19.fasta \
-input ~[original_cohort].snp.recalibrated.vcf.gz \
--ts_filter_level 95.0 \
-tranchesFile ~/[cancer_cohort].INDEL.tranches \
-recalFile ~/[cancer_cohort].INDEL.recal \
-mode INDEL \
-o ~/[original_cohort].snp.recalibrated.indel.recalibrated.vcf.gz
```

SECOND COMPUTATIONAL RUN (ORIGINAL COHORT AND THE ADDITIONAL COHORT): differences are highlighted in red.

1. HaplotypeCaller (HC): this command is run on each sample individually:

```
java -Xmx12G  -jar ~/GenomeAnalysisTK.jar \
-nct 8 \
-T HaplotypeCaller \
-R  ~/Homo_sapiens_assembly19.fasta \
-I  [single.sample.bam] \
--dbsnp ~/dbsnp_138.hg19.vcf.gz \
--genotyping_mode DISCOVERY \
-variant_index_type LINEAR \
-variant_index_parameter 128000 \
--emitRefConfidence GVCF \
--max_alternate_alleles 6 \
--minPruning 2 \
-stand_call_conf 30.0 \
-A DepthPerSampleHC \
-A StrandBiasBySample \
-A Coverage \
```

```
-A StrandBiasBySample \
-o ~/[single.sample].gvcf.gz
```

2. Joint genotyping (GenotypeGVCFs): this step combines all the gVCFs that were generated by the previous step to do cohort-wide genotyping:

```
java -jar -Xmx32G ~/GenomeAnalysisTK.jar \
-R ~/Homo_sapiens_assembly19.fasta -T GenotypeGVCFs \
--variant ~/[list_of_cancer_patients_gVCFs].list  \
--variant ~/[original_and_add'l_cohorts].list  \
-L  ~/[capture_region].interval_list \
-o  ~/[original_and_add'l_cohorts].gvcf.gz
```

3. VariantRecalibration (SNPs):

```
java -Xmx24G  -jar ~/GenomeAnalysisTK.jar \
-T VariantRecalibrator \
-R ~/Homo_sapiens_assembly19.fasta \
-input ~/[original_and_add'l_cohorts].gvcf.gz \
-resource:hapmap,known=false,training=true,truth=true,prior=15.0
~/hapmap_3.3.b37.vcf \
-resource:omni,known=false,training=true,truth=true,prior=12.0
~/1000G_omni2.5.b37.vcf \
-resource:1000G,known=false,training=true,truth=false,prior=10.0
~/1000G_phase1.snps.high_confidence.b37.vcf \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
~/dbsnp_138.b37.vcf \
-an QD -an MQRankSum -an ReadPosRankSum -an FS -an MQ -an
InbreedingCoeff \
-mode SNP \
-tranche 100.0 \
-tranche 99.9 -tranche 99.9 -tranche 99.8 -tranche 99.7 -tranche
99.6 -tranche 99.5 \
-tranche 99.4 -tranche 99.3 -tranche 99.2 -tranche 99.1 -tranche
99.0 \
-tranche 98.9 -tranche 98.8 -tranche 98.6 -tranche 98.5 -tranche
98.3 \
-tranche 98.2 -tranche 98.1 -tranche 98.0 -tranche 97.9 -tranche
97.8 \
-tranche 97.5 -tranche 97.0 -tranche 95.0 -tranche 90.0 \
-recalFile ~/[original_and_add'l_cohorts].SNP.recal \
-tranchesFile ~/[original_and_add'l_cohorts].SNP.tranches \
-rscriptFile ~/[original_and_add'l_cohorts].SNP.R \
-nt 4
```

4. Apply recalibration (SNP):

```
java -jar -Xmx24G ~/GenomeAnalysisTK.jar \
```

```
-T ApplyRecalibration \
-R ~/Homo_sapiens_assembly19.fasta \
-input ~/[cancer_cohort].gvcf.gz \
--ts_filter_level 99.5 \
-tranchesFile ~/[original_and_add'l_cohorts].SNP.tranches \
-recalFile ~/[original_and_add'l_cohorts].SNP.recal \
-mode SNP \
-o ~/[original_and_add'l_cohorts].snp.recalibrated.vcf.gz
```

5. VariantRecalibration (INDELs):

```
java -jar -Xmx24G ~/GenomeAnalysisTK.jar \
-T VariantRecalibrator \
-R ~/Homo_sapiens_assembly19.fasta \
-input ~/[original_and_add'l_cohorts].gvcf.gz \
-tranche 100.0 \
-tranche 99.9 -tranche 99.9 -tranche 99.8 -tranche 99.7 -tranche
99.6 -tranche 99.5 \
-tranche 99.4 -tranche 99.3 -tranche 99.2 -tranche 99.1 -tranche
99.0 \
-tranche 98.9 -tranche 98.8 -tranche 98.6 -tranche 98.5 -tranche
98.3 \
-tranche 98.2 -tranche 98.1 -tranche 98.0 -tranche 97.9 -tranche
97.8 \
-tranche 97.5 -tranche 97.0 -tranche 95.0 -tranche 90.0 \
-resource:mills,known=false,training=true,truth=true,prior=12.0
~/Mills_and_1000G_gold_standard.indels.b37.vcf \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
~/dbsnp_138.b37.vcf \
-an FS -an QD -an ReadPosRankSum -an InbreedingCoeff \
-mode INDEL \
-recalFile ~/[original_and_add'l_cohorts].INDEL.recal \
-tranchesFile ~/[original_and_add'l_cohorts].INDEL.tranches \
-rscriptFile ~/[original_and_add'l_cohorts].INDEL.R \
-nt 4
```

6. Apply recalibration (INDELS):

```
java -jar -Xmx24G ~/GenomeAnalysisTK.jar \
-T ApplyRecalibration \
-R ~/Homo_sapiens_assembly19.fasta \
-input ~[original_and_add'l_cohorts].snp.recalibrated.vcf.gz \
--ts_filter_level 95.0 \
-tranchesFile ~/[original_and_add'l_cohorts].INDEL.tranches \
-recalFile ~/[original_and_add'l_cohorts].INDEL.recal \
-mode INDEL \
```

```
    -o
~/[original_and_add'l_cohorts].snp.recalibrated.indel.recalibrated.vcf
.gz
```

7. Filtering out all germline variants discovered in the cancer-free cohort (ESP):

```
java -jar -Xmx24G ~/GenomeAnalysisTK.jar \
-T SelectVariants \
-R ~/Homo_sapiens_assembly19.fasta \
-V
~/[original_and_add'l_cohorts].snp.recalibrated.indel.recalibrate
d.vcf.gz
-o  /[original_cohort].snp.recalibrated.indel.recalibrated.vcf.gz
\
-xl_sf ~/[list_of_add'l_gVCFs].list \
--excludeNonVariants \
--removeUnusedAlternates
```

Supplementary References:

1. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018. p. 201178. doi:10.1101/201178

2. Huang K-L, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, et al. Pathogenic Germline Variants in 10,389 Adult Cancers. Cell. 2018;173: 355–370.e14.

3. Rahman N. Realizing the promise of cancer predisposition genes. Nature. 2014. pp. 302–308. doi:10.1038/nature12981

4. Kalia SS, ; on behalf of the ACMG Secondary Findings Maintenance Working Group, Adelman K, Bale SJ, Chung WK, Eng C, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genetics in Medicine. 2017. pp. 249–255. doi:10.1038/gim.2016.190

5. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016. doi:10.1186/s13059-016-0974-4

6. Richards S, ; on behalf of the ACMG Laboratory Quality Assurance Committee, Aziz N, Bale S, Bick D, Das S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine. 2015. pp. 405–423. doi:10.1038/gim.2015.30

7. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29: 24–26.

8. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536: 285–291.