nature genetics

Supplementary information

https://doi.org/10.1038/s41588-024-02070-5

Functional analysis of cancer-associated germline risk variants

In the format provided by the authors and unedited

SUPPLEMENTARY INFORMATION

Functional analysis of cancer-associated germline risk variants

Laura N. Kellman[1,2], Poornima H. Neela[1], Suhas Srinivasan[1], Zurab Siprashvili[1], Ronald L. Shanderson[1,2], Audrey W. Hong[1], Deepti Rao[1], Douglas F. Porter[1], David L. Reynolds[1], Robin M. Meyers[1,3], Margaret G. Guo[1], Xue Yang[1,2], Yang Zhao[1], Glenn G. Wozniak[1], Laura K. H. Donohue[1,3], Rajani Shenoy[1], Lisa A. Ko[1], Duy T. Nguyen[1], Smarajit Mondal[1], Omar S. Garcia[1], Lara E. Elcavage[1], Ibtihal Elfaki[1,3], Nathan S. Abell[3], Shiying Tao[1], Christopher M. Lopez[1], Stephen B. Montgomery[3,4,5], Paul A. Khavari[1,2,6*]

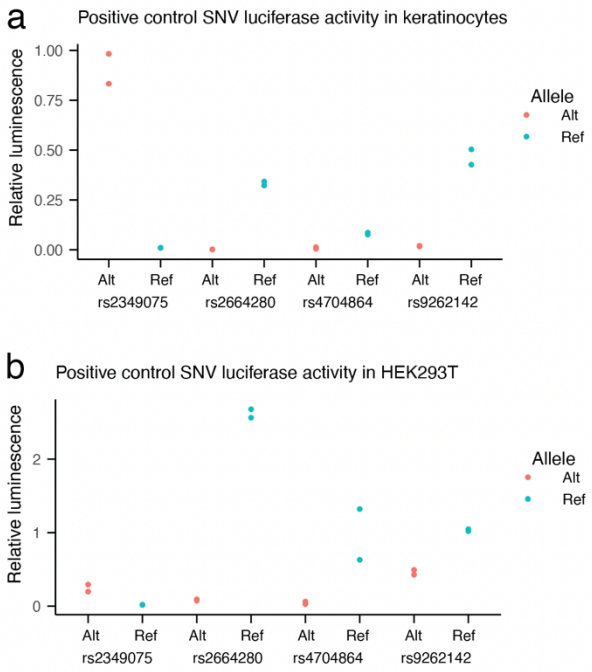
*Correspondence: khavari@stanford.edu

Table of Contents:

SUPPLEMENTARY METHODS

Positive control SNV luciferase assays

4 positive control SNVs – rs2349075, rs9262142, rs4704864 and rs2664280 – were included in the MPRA library because of strong differences in luciferase signal between alleles in HEK293Ts and keratinocytes. rs2349075 luciferase data was previously published in Mondal, et al.[1] The luciferase assays were performed as follows: 21-nucleotide chromosomal sequences containing reference or alternative alleles of both SNPs were triplicated and cloned into pGL4.23(Luc2-minP) vectors. For assays in HEK293T cells, 200 ng pGL4 plasmid with reference or alternative alleles and 2 ng pGL3 plasmid expressing Renilla luciferase were co-transfected into cells in 6-well plates using Lipofectamine 3000 (ThermoFisher) according to the manufacturer's protocol. For primary keratinocytes, 500 ng pGL4 with SNPs and 5 ng pGL3 expressing Renilla luciferase were co-transfected using TransIT-X2 (Mirus Bio) in 6-well plates. Cells were lysed 24 h post-transfection with 1X Passive Lysis Buffer (Promega). Dual-luciferase assays were performed using the Dual-Luciferase Reporter Assay System (Promega) following the manufacturer's instructions. Firefly and Renilla luminescence were measured using a Tecan M200 Infinite Pro microplate reader.



Supplementary Figure 1. Luciferase activity of positive control SNVs in keratinocytes (**a**) and HEK293T cells (**b**).

DHS signal correlation

DHS data from ENCODE relevant to the cell types studied was merged for each cell type using bedtools v2.30.0[2] and signals for merged peaks were averaged. The Pearson correlation between DHS signal and transcriptional activity (alpha, from MPRAnalyze quantitative) for significantly active (FDR<0.05) SNVs in the library that fell in DHS peaks for that cell type (using bedtools intersect) was determined.

Clustering of MPRA transcriptional activity across cell types

For SNVs with significant activity in at least one context (MAD-based p-value from MPRAnalyze's analyzeQuantification function < 0.05), Euclidean distances between z-scores of transcriptional activities were calculated and hierarchical clustering was performed.

Enrichment in TCGA ATAC peaks

ATAC peak calls for each tumor type were obtained from TCGA[3]. Bedtools[2] was used to intersect ATAC peaks for each tumor type with each set of cancer-specific SNVs (e.g., SNVs in the library associated with thyroid cancer, regardless of MPRA activity). The number of daSNVs associated with a given cancer found in a TCGA cancer's ATAC peaks vs. number of non-daSNVs was compared. For example, to calculate enrichment of thyroid daSNVs in colorectal cancer ATAC peaks, we generated a 2x2 table classifying each thyroid associated MPRA SNV by whether it is or is not a daSNV, and whether it is or is not in a colorectal cancer ATAC peak. Fisher's exact test for enrichment was run for each MPRA-cancer/TCGA-cancer comparison, and p-values were FDR-corrected.

daSNV Luciferase assays

401 bp of genomic sequence surrounding each SNV, with both reference and alternate allele versions, were synthesized as GeneBlocks (IDT), then cloned into the pGL4.23 vector (Promega) upstream of a minimal promoter and luciferase sequence using Gibson assembly (NEB). HCT116 (ATCC CCL-247, for colon cancer SNVs) and MCF7 cells (ATCC HTB-22, for breast cancer SNVs) were transfected with each construct using lipofectamine 3000 (Thermo Fisher) with 5 replicates for each allele. 48 hours later, cells were harvested, and fluorescence was measured using the dual luciferase reporter assay system (Promega) and the Tecan Infinite

M1000. Firefly and renilla blanks were subtracted from their respective measurements. Firefly to renilla ratios were calculated, then normalized to the control empty vector. p-values were calculated using a two-tailed t-test.

Allele-specific ATAC analysis

Allele-specific ATAC sites were analyzed using heterozygous samples from Donohue et al. Because most SNVs were not heterozygous in those samples, the number of SNVs that could be assessed was limited. Analysis was performed following the method in Zhang, et al. 2020[4] with the GATK pipelines previously described by van der Auwera and O'Connor[5]. The bwa package[6] was used to generate allele-specific bam files, then Picard[7] was used to build bam indices, remove duplicates and sort the bam file. GATK BaseRecalibrator was used to generate a recalibration table from covariates extracted from known SNVs (dbSNV_138.hg19, 1000G_phase1.SNVs.high_confidence.hg19, 1000G_phase1.indels.hg19, and Mills_and_1000_gold_standard.indels.hg19). All loops and all ATAC peaks were collated for asHiChIP and asATAC analysis, and analysis was restricted to regions of interest. ApplyBQSR was used to apply the base quality score recalibration table to the SNVs. HaplotypeCaller was used to generate allele-specific vcf files for each sample. CombineGVCFs and GentypeVCFs were used to aggregate samples and create the overall SNV vcf. Variant recalibration was called with the following parameters:

```
--resource:hapmap,known=false,training=true,truth=true,prior=15.0
${refSNV_path}/hapmap_3.3.hg19.sites.vcf \
--resource:omni,known=false,training=true,truth=false,prior=12.0
${refSNV_path}/1000G_omni2.5.hg19.sites.vcf \
--resource:1000G,known=false,training=true,truth=false,prior=10.0
${refSNV_path}/1000G_phase1.SNVs.high_confidence.hg19.sites.vcf \
--resource:dbSNV,known=true,training=false,truth=false,prior=2.0
${refSNV_path}/dbSNV_138.hg19.vcf \
-an DP -an QD -an FS -an SOR -an MQ -an ReadPosRankSum \
-tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 \
-mode SNV
```

ApplyVWSR was used to generate the unfiltered SNV vcf file. The file was then filtered for heterozygous SNVswith a depth count >=10, and a minimum reference or alternative allele count >=2. A binomial test was run on each SNV comparing allele counts, then the resulting p-values were Bonferroni corrected. SNVs were considered asATAC if $p_{adj} < 0.01$ and the less abundant allele had half or fewer counts than the more abundant allele (ie, min(Ref/Alt, Alt/Ref) <= 0.5). Enrichment of daSNVs vs. non-daSNVs in the library for tissue-specific or any asATAC SNVs was determined using Fisher's exact test.

Integration of eQTL, GWAS, and colocalization data

For significant BRCA MPRA variants, we extracted expression quantitative trait locus (eQTL) summary statistics for all tissues included in GTEx v8[8]. We further filtered the summary statistics to include only SNV-gene associations with p-value <= 5e-5. To incorporate relevant traits with public association data, we queried SNV-trait associations corresponding to 114 breast cancer GWAS incorporated in GTEx and their colocalizations with eQTLs[9]. We obtained posterior probabilities of colocalization derived from coloc for all genes in all GTEx tissues, and identified loci potentially explained by MPRA variants with the following filters: 1) the colocalized region must contain at least one significant MPRA variant, 2) the region must contain a significant eQTL at p-value < 5e-5, and 3) the colocalization posterior probabilities must have P3 + P4 >= 0.75 and P4 >= 0.1[10]. The final filter is to ensure that substantial overlap between GWAS and eQTL signal exists but allows for cases where individual causal variants are not fully resolved. This procedure returned a list of gene-trait (i.e. eQTL-GWAS) colocalizations which contained eQTLs potentially driven by causal MPRA variants.

Identifying ptGenes implicated in hereditary cancer syndromes or COSMIC

Hereditary cancer syndrome genes were identified from the Online Mendelian Inheritance in Man[11] (OMIM)'s morbid map (a catalog of human genes and genetic disorders) and filtered to remove mildly penetrant cancer susceptibility genes. Genes within the COSMIC cancer gene census (v92)[12] with either Tier 1 or Tier 2 evidence were annotated as COSMIC.

ptGene pathway enrichments

Enriched Reactome pathway terms were found using ReactomePA (v. 1.38.0)[13] for the set of: all ptGenes, ptGenes from different sources, and ptGenes for each cancer.For individual

cancer types, the set of genes expressed (TPM > 0) in RNA-seq data from the corresponding cell type was used as the background[14]. For all ptGenes and the gene source analyses, whole genome background was used. To determine how MPRA influenced the pathways found versus GWAS background, we compared enrichment of the ptGenes for Reactome pathways to enrichment of 2000 simulated gene sets derived from randomly selected SNVs in the library to generate empirical p-values.

Co-essentiality analysis

Co-essential genes were found in Wainberg et al., 2019[15], which used genome-wide CRISPR screens across 485 cancer cell lines to identify co-essentiality using a generalized least-squares regression. In our network, we included non-syntenic co-essential gene pairs with FDR<0.10. For **Fig. S7a**, we seeded the network using all MPRA hit ptGenes (FDR < 0.05). We included any other gene in the network that was annotated as co-essential with at least two of the seed genes. Network nodes represent genes, and edges represent a co-essential relationship at FDR<0.10 with BioGrid[16] or BioPlex[17] interaction support.

Cancer Survival Analysis

Prediction of Clinical Outcomes from Genomic Profiles (PRECOG)[18] was used to correlate expression of eight mitochondrial translation ptGenes with cancer survival across 12 TCGA cancers. Statistical associations between outcomes and gene expression are represented by z-scores.

SUPPLEMENTARY REFERENCES

1. Mondal, S. *et al.* PROBER identifies proteins associated with programmable sequence-specific DNA in living cells. *Nat Methods* **19**, 959–968 (2022).
2. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
3. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, (2018).
4. Zhang, S. *et al.* Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants. *Science (New York, N.Y.)* **369**, 561–565 (2020).
5. Van der Auwera, G., O'Connor, B. & Safari, an O. M. Company. *Genomics in the Cloud*. (O'Reilly Media, Inc, 2020).
6. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
7. Broad Institute. Picard Toolkit. *Broad Institute, GitHub repository* http://broadinstitute.github.io/picard/ (2019).
8. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
9. Barbeira, A. N. *et al.* Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biology 2021 22:1* **22**, 1–24 (2021).
10. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics* **10**, e1004383 (2014).
11. McKusick-Nathans Institute of Genetic Medicine. Online Mendelian Inheritance in Man, OMIM. https://omim.org (2021).

12. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* **47**, D941–D947 (2019).
13. Yu, G. & He, Q. Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems* **12**, 477–479 (2016).
14. Donohue, L. K. H. *et al.* A cis-regulatory lexicon of DNA motif combinations mediating cell-type-specific gene regulation. *Cell genomics* **2**, (2022).
15. Wainberg, M. *et al.* A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. *Nature Genetics 2021 53:5* **53**, 638–649 (2021).
16. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein science : a publication of the Protein Society* **30**, 187–200 (2021).
17. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440 (2015).
18. Gentles, A. J. *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nature Medicine 2015 21:8* **21**, 938–945 (2015).