SUPPLEMENTAL METHODS

Whole-exome sequencing

DNA samples were collected from whole blood and exome capture was performed using a modified version of the IDT xGen probe library (Integrated DNA Technologies, Boulder, CO, USA). Samples were sequenced on the Illumina NovaSeq 6000 (Illumina Inc., San Diego, CA, USA) platform to a mean depth of 55x using a 75 base-pair paired-end protocol. Sequencing reads were aligned to the hg38 human reference genome using BWA-MEM (1). Samples with low sequence coverage (<80% of target bases with ≥20x coverage), discordance with microarray genotyping data or reported sex, high rates of heterozygosity/contamination (FREEMIX score >5%), or duplication were removed. Polymerase chain reaction duplicate reads were filtered using Genome Analysis Toolkit (GATK) MarkDuplicates v2.21.2 (2).

SNP array

Genotyping was performed by Affymetrix using the UK BiLEVE Axiom (~50,000 participants) and UK Biobank Axiom (~450,000 participants) arrays. The genotype data contains a total of 805,426 unique variants from both arrays. Quality control and imputation (with IMPUTE4 using the Haplotype Reference Consortium and UK10K data as imputation reference panels) were performed as previously described (3).

CH variant calling

After base quality score recalibration [GATK BQSRPipelineSpark, v4.2.0.0 (2)], single-nucleotide variants (SNVs) and insertions/deletions (indels) were called using Mutect2 v4.2.1.0 (4) and VarDictJava v1.6.0 (5). Variants were retained that were supported by both variant callers. We only retained variants with an alternative allele frequency (VAF) ≥2% and with ≥2 supporting reads including at least one from both the forward and reverse strands.

CH post-processing filtering criteria

We used the sequencing data from 50 individuals under the age of 41 from the UKBB without known CH hotspot mutations as an internal panel of normals (PON). Variants that were found with a VAF of >2% in two or more samples from the PON were removed. For every variant that passed consensus variant calling, we used the PON samples to empirically estimate the sequencing error at that position through the following approach. First, we interrogated each PON sample for the presence of an alternative allele supporting the putative CH mutation. Second, we summed the total number of alternative counts across the PON, then tested for an enrichment in the alternative alleles supporting the CH mutation in the sample compared to the PON using Fisher’s exact test. We adjusted for multiple hypothesis testing using a Bonferroni correction based on the size of the WES capture region, including splice site acceptor and donor sites (since CH variants could have been called at every position on the panel of 39.5 Mbps) yielding a p-value of $1.3x10^{-9}$. Variants where the VAF was significantly greater in the sample compared to the PON were retained.

We removed variants as possible artifacts if they met the following criteria:

1) Variants with <2 reads supporting the call or variants not supported by both forward and reverse reads

2) Any variant recurrent in >1% of samples that had never been previously reported in large-scale CH sequencing studies

3) Variants with evidence from only the forward or reverse strand

To filter possible germline polymorphisms, we removed the following variants:

1) Any variant reported in the gnomAD database [exome v2.1.1 and genome v3.1.2 (6)] that has a population allele frequency > $5×10^{-4}$ or maximum sub-population allele frequency > $5×10^{-3}$

2) All variants with a VAF ≥35% unless it was a clear somatic hotspot in hematologic malignancies with no or minimal evidence of the variant being seen as a germline mutation (e.g., SRSF2 P95H)

3) Given the low depth of the UKBB data and coverage variation, we removed recurrent variants having VAF ≥25% where the median VAF for that variant was >35%; we also removed variants with a VAF of >25% that were not previously reported in large CH studies [Bolton et al. (7) and Bick et al. (8)] or COSMIC v95 (9).

We exempted loss of function variants in *DNMT3A, TET2, ASXL1 and PPM1D* from the second and third germline filters given that loss of function variants in these genes have not been previously reported as germline.

CH variant annotation

Variants were annotated with Ensembl Variant Effect Predictor v109 (10) using the canonical transcript. Putative driver (PD) mutations were annotated according to prior evidence for functional relevance in hematologic cancer as previously described (7). We included 939 genes with known relevance to cancer (Supplemental Table 11). In brief variants were classified as PD if they met the following criteria:

1) Truncating variants (nonsense, essential splice site or frameshift indel) in tumor suppressor genes (Supplemental Table 11)

2) Truncating mutations within exon 6 of PPM1D

3) Missense mutations at amino acid positions 95 (SRSF2); 622-626, 662, 663-666, 700-704, 740-742 (SF3B1); 132 (IDH1); 140, 172 (IDH2); or in exon 6 (PPM1D)

4) Any variant reported as somatic in the COSMIC at least 10 times or in ‘hematopoietic and lymphoid’ tumors at least 5 times

5) Variants previously reported as a CH-PD mutation by Bick et al. (8) or Bolton et al (7).

6) Any variant noted as oncogenic or likely oncogenic in OncoKB v3.10 (11).

7) Missense mutations within 3 amino acid residues or within 9 nucleotides of hotspot variants from Bolton, et al. (7), OncoKB (11), Bick, et al. (8), or COSMIC (9) and were predicted as deleterious by SIFT v6.2.1 (12) and PolyPhen v2.2.3 (13).

8) Splice variants within 2 nucleotides of an intron in tumor suppressor genes

9) Variants reported as pathogenic in ClinVar (April 30, 2023 release) (14).

Mosaic chromosomal alteration calling

We used MoChA v10.2.0 (15) to identify acquired (i.e. mosaic) chromosomal alterations (mCAs) in the UKBB genotyping array data. The analysis was performed using non-imputed data with default parameters, and limited to autosomal chromosomes. Genotypes were phased using SHAPEIT (v4.2) prior to calling chromosomal alterations with MoChA. We followed the recommended post-processing steps outlined on the MoChA GitHub repository (https://github.com/freeseek/mocha) to filter likely germline copy number events and artifacts.

REFERENCES

1. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. arXiv; 2013 [cited 2023 Apr 30]. Available from: http://arxiv.org/abs/1303.3997

2. Van Der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. CP in Bioinformatics [Internet]. 2013 [cited 2023 Apr 30];43. Available from: https://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi1110s43

3. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562:203–9.

4. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31:213–9.

5. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44:e108–e108.

6. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes [Internet]. Genetics; 2022 Mar. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.03.20.485034

7. Bolton KL, Ptashkin RN, Gao T, Braunstein L, Devlin SM, Kelly D, et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. Nat Genet. 2020;52:1219–26.

8. Bick AG, Weinstock JS, Nandakumar SK, Fulco CP, Bao EL, Zekavat SM, et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. Nature. 2020;586:763–8.

9. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18:696–705.

10. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.

11. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol. 2017;2017:PO.17.00011.

12. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31:3812–4.

13. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.

14. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research. 2018;46:D1062–7.

15. Loh P-R, Genovese G, Handsaker RE, Finucane HK, Reshef YA, Palamara PF, et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature. 2018;559:350–5.