

Supplementary Information

Alignment of short reads from the *PRSSI-PRSS2* locus to the human reference genome

To determine the copy-number of 10.6-kb tandem duplication in the human reference genome GRCh38, we aligned the sequence of the *PRSSI-PRSS2* locus from the primary scaffold of chromosome 7 to that in the alternate (ALT) contig (chr7_KI270803v1_alt), where RefSeq annotates another location of the *PRSSI-PRSS2* locus. The alignment result shows that the primary scaffold of chromosome 7 is a 3-gene haplotype while the ALT contig is a 5-gene haplotype (Figure S1A). Each gene unit is approximately 10.6-kb and with an identity of ~90%, consistent with the previous reported haplotype structures [1]. We also investigated the haplotype structure in the GRCh37 build. Although *PRSS2* is included in the fixed patched contig, this gene is missing in the primary scaffold of GRCh37 (Figure S1B). Based on these observations, we chose GRCh38 as reference in the subsequent analysis.

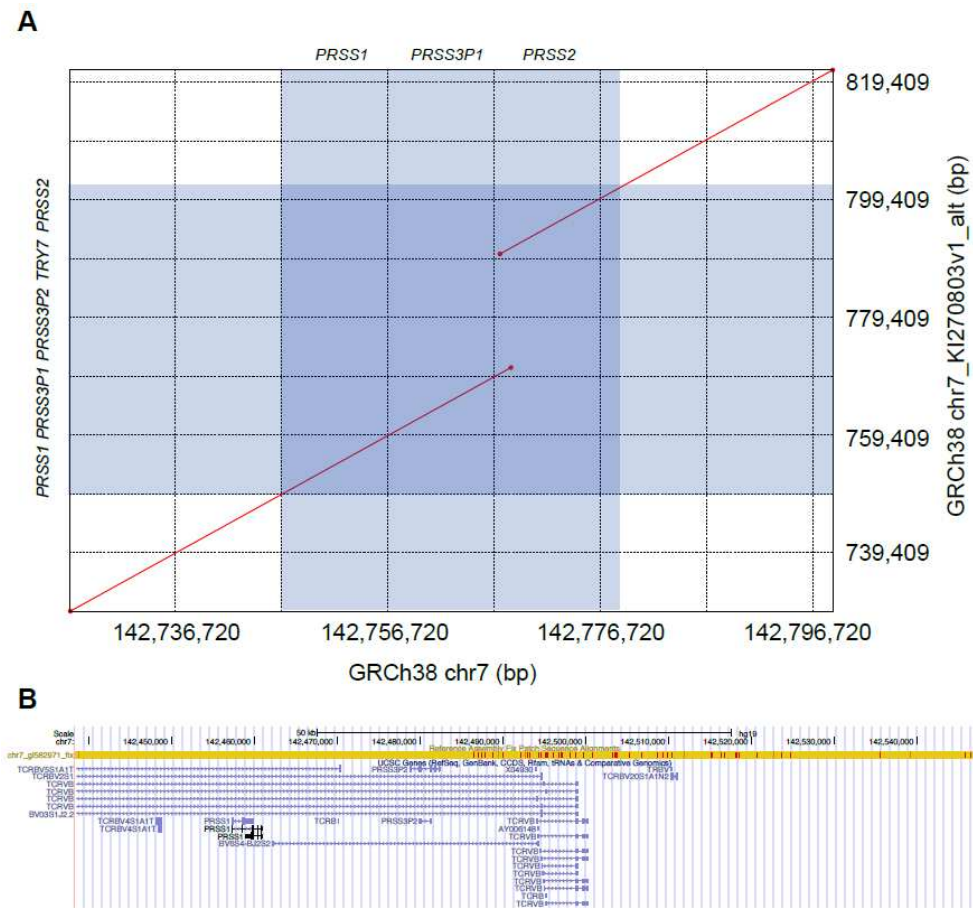


Figure S1. *PRSSI-PRSS2* sequence on the human reference genome.

(A) Alignment between the primary and the alternate (ALT) contig in the GRCh38 build. (B) Missing *PRSS2* in the primary scaffold in the GRCh37 build.

Next, taking advantage of the high-quality genomes that were *de novo* assembled by long-read

and other technologies, we conducted simulation analysis to assess the short-read mapping to the GRCh38 reference. By sequence alignment, we identified the haplotype structure of two high-quality European genome assembly CHM13 [2] and NA12878 (GCA_002077035.3), which were typical 3-gene and 5-gene haplotype samples, respectively (Figure S2). Note that CHM13 is a complete hydatidiform mole, which is essentially a haploid genome; while NA12878 is a diploid individual, and the de novo assembly of GCA_002077035.3 is a collapsed diploid genome, which to some extent is equivalent to a haploid genome. To evaluate the mapping performance, we employed EAGLE (<https://github.com/sequencing/EAGLE>) to simulate the paired-end sequencing data with read-length of 150-bp and depth of 30× from the CHM13 and NA12878 assembly. These simulated reads were aligned to GRCh38 primary contigs with BWA [3]. The 3-gene sample (CHM13) showed an overall good-alignment with few mismatches, but 5-gene sample (NA12878) produced substantial mismatches on the primary contig (Table S1), which was resulted from the mis-aligned reads of the additional two copies (*PRSS3P2* and *TRY7*) in 5-gene haplotype (Figure 1A and Figure S3). Then, we aligned the simulated reads to GRCh38 with ALT contigs and further processed with post-alt script. Both 3-gene and 5-gene samples were well-aligned to ALT contig, and the alignment between the 3-gene haplotype and the ALT contigs showed 21-kb deletion at the *PRSS3P2* and *TRY7* regions (Figure 1A and Figure S3). Furthermore, we observed a small portion of the reads mapped to *PRSS3* on chromosome 9, due to the homology between *PRSS3* and *PRSS2* [4]. In light of this, we used ALT contig instead of the primary scaffold as the reference sequence for short-read mapping.

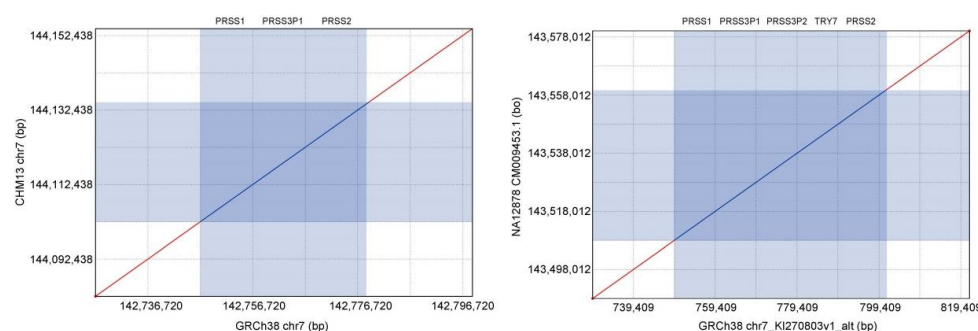


Figure S2. Sequence alignment between the assembly of CHM13, NA12878 and GRCh38.

Left panel, CHM13 aligned to GRCh38 primary chromosome 7. Right panel, NA12878 aligned to GRCh38 alternate chromosome 7.

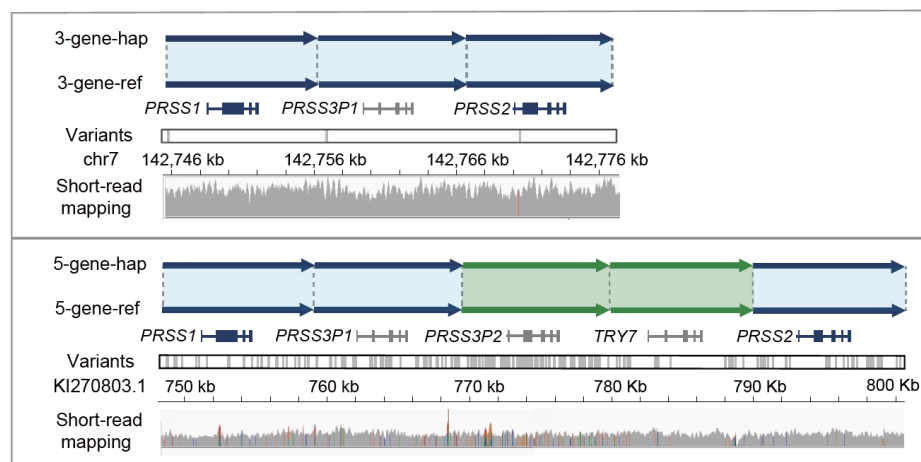


Figure S3. Sequence alignment between the assembly of CHM13, NA12878 and GRCh38.

Short-read mapping of a 3-gene trypsinogen haplotype (CHM13) to a 3-gene trypsinogen reference (GRCh38 primary chromosome 7; upper panel) and a 5-gene trypsinogen haplotype (NA12878) to a 5-gene trypsinogen reference (GRCh38 alternate contig; lower panel). See [Figure 1A](#) for other two cases. The color bars in the ‘Variants’ track denote the mismatches to the reference. The color peak in the ‘Short-read mapping’ track denotes the alignment of the nucleotide: gray, matched; other colors, mismatched.

NGS.PRSS1-2caller toolkit

Based on the above observations, we developed a pipeline to call small variants as well as the copy number at the *PRSS1-PRSS2* locus. Briefly, we extracted the aligned short reads on primary and alternate chromosome 7 together with those on *PRSS3* regions (chromosome 9) with SAMtools [5], and re-aligned all these extracted reads to the ALT contig with bwa-mem. After removing the PCR duplicated reads by Picard (<http://broadinstitute.github.io/picard/>), the copy number for each of the 10.6-kb tandem gene duplicates was estimated based on the read-depth information. Then we applied a Bayesian genetic variant detector freebayes [6] to call small variants given the copy number of each gene unit. Notably, the NGS.PRSS1-2caller can handle whole genome sequencing aligned data (bam files) regardless of whether the GRCh38 ALT contig was included in the initial mapping. The final output vcf file (variant call file) provides both the genotype information of small variants and the copy number state as well as their predicted biological consequences. The effect of the variants were annotated by SnpEff [7]. Furthermore, we included an option for phasing the population data with Beagle [8].

Evaluation of NGS.PRSS1-2caller

To evaluate the performance of NGS.PRSS1-2caller, we started with the simulation data. The variants derived from sequence alignment between the test sample (CHM13 and NA12878) and the reference ALT contig were used as the true call-set. Here the true call-set was built by using MUMmer [9]. The combinations of the two haploid genomes were used as test diploid samples, i.e., CHM13/CHM13 for a homologous 3-gene trypsinogen sample, CHM13/NA12878 for a heterozygous 3-gene/5-gene trypsinogen sample and NA12878/NA12878 for a homologous

5-gene trypsinogen sample. The simulated short-reads were generated with EAGLE as described above. The precision and the recall rate were calculated by comparing the simulation results and the true call-set. Additionally, we compared NGS.PRSS1-2caller with variant toolkit GATK [10] based on *PRSS1-PRSS2* locus of the ALT contig. Since the current version of GATK does not support variant calling on ALT contigs, we followed the instructions of GATK and modified the bam file to flag all the paired-end reads as single-end reads so that GATK could call the variants (<https://gatk.broadinstitute.org/hc/en-us/articles/360037498992--How-to-Map-reads-to-a-reference-with-alternate-contigs-like-GRCH38>). However, no copy number estimate was provided by GATK, and GATK assumes a typical diploid mode to detect small variants. Therefore, the results of GATK generated more false-positives and false-negatives, especially for heterozygous haplotype structure (Figure S4).

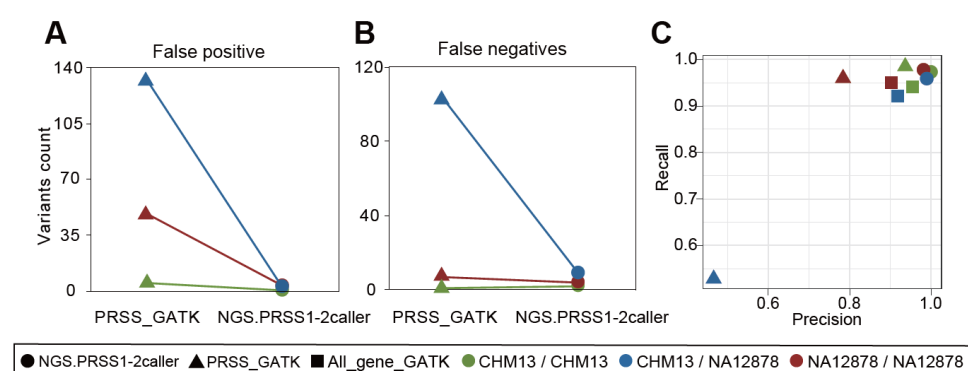


Figure S4. Evaluation of NGS.PRSS1-2caller with simulation data.

(A) False-positive and (B) false-negative variant count detected by GATK and NGS.PRSS1-2caller with GRCh38 ALT contig as reference. (C) Comparison of precision and recall rate of variant discovery between NGS.PRSS1-2caller and GATK. Circle, NGS.PRSS1-2caller; Triangle, GATK at *PRSS1-PRSS2* locus (PRSS_GATK); Square, GATK at genome-wide genic regions (All_gene_GATK); Green, CHM13/CHM13; blue, CHM13/NA12878; red, NA12878/NA12878. The poor performance of CHM13/NA12878 with GATK (at *PRSS1-PRSS2* locus) was due to incorrect genotypes on *PRSS3P2* and *TRY7* where the heterozygous variants were falsely genotyped as homozygous.

Next, we used the real data from a subset panel of haplotype-resolved *de novo* assemblies [11] to evaluate the NGS.PRSS1-2caller performance. These genomes were assembled with PacBio long-read and strand sequencing, and their short read sequencing data were also available in the 1000 Genomes Project [12]. A total of ten samples (HG00513, HG00864, NA18534, NA18939, HG01596, NA20847, NA19238, NA19239, NA12878, HG00171) covered all three haplotype structures were used in our evaluation. Since the long-read assembly might still contain sequencing errors, we applied Pilon [13] to polish these assemblies with the short-reads from the same individual in a diploid mode. The true variant set for evaluation was constructed by alignment between the polished assembly and the ALT contig. The results in F1-score were shown in Figure 1B.

Finally, to test whether NGS.PRSS1-2caller could accurately capture the reported pathogenic variants documented in the pancreatitis database (www.pancreasgenetics.org), we replaced the alleles in the sequence of CHM13 and NA12878 with the pathogenic alleles, and simulated the

short reads of three haplotype structures (3-gene/3-gene, 3-gene/5-gene and 5-gene/5-gene). A total of 20 pathogenic variants were included. For each haplotype structure, we performed 20 simulations with the number of pathogenic variants from 1 to 20. Once the number (N) of the pathogenic variants was set, we randomly replaced N alleles in the haplotype with the pathogenic alleles to simulate the short-reads and processed the calling with NGS.PRSS1-2caller. The simulation results showed that 95% (19/20) of the known pathogenic alleles were perfectly recalled by NGS.PRSS1-2caller, the recall rate of the remaining one was 0.84 which might be due to the low depth coverage of the variant simulated (Table S2).

Haplotype structure frequency and the linkage with pancreatitis protective alleles

We detected genetic variants from the high-coverage sequencing data from the 1000 Genomes Project [12] and calculated the frequency for different haplotype combinations (Figure S5). The pancreatitis protective alleles documented in pancreatitis database (www.pancreasgenetics.org) were used in this analysis. LDBlockShow was used to calculate and visualize the linkage disequilibrium (LD) between these variants and the deletion genotype [14]. The LD heatmap was displayed in Figure 2 and Figure S6.

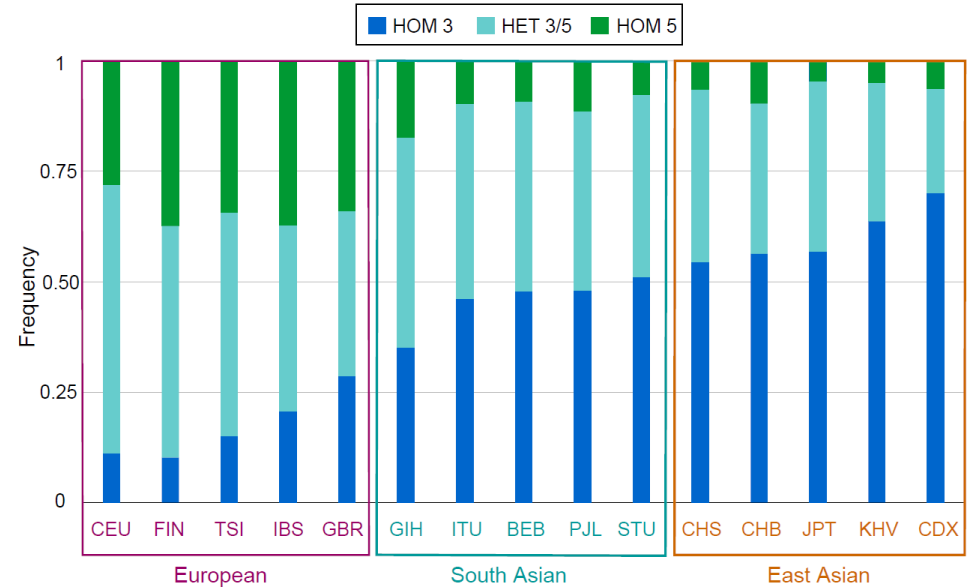


Figure S5. Haplotype structure frequency of European and Asian populations.

HOM 5, homozygous 5-gene structure; HOM 3, homozygous 3-gene structure; HET 3/5 heterozygous 3-gene/5-gene structure. CEU, Utah Residents with European ancestry; FIN, Finnish in Finland; TSI, Toscani in Italy; IBS, Iberian population in Spain; GBR, British in England and Scotland; GIH, Gujarati Indian from Houston; ITU, Indian Telugu from the UK; BEB, Bengali from Bangladesh; PJJ, Punjabi from Lahore; STU, Sri Lankan Tamil from the UK; CHS, Southern Han Chinese; CHB, Han Chinese in Beijing; JPT, Japanese in Tokyo; KHV, Kinh in Ho Chi Minh City, Vietnam; CDX, Chinese Dai in Xishuangbanna.

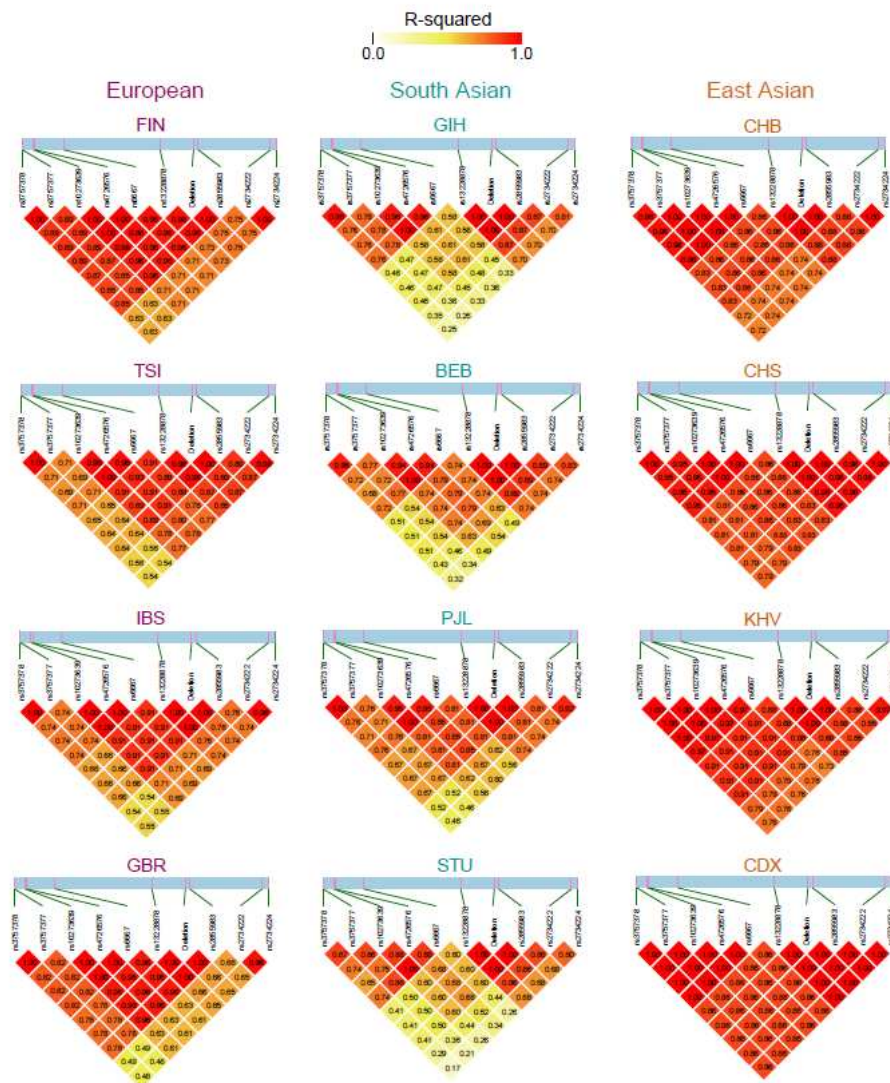


Table S1. False positive variants annotated with moderate and high impact when aligning 5-gene haplotype to GRCh38 primary chromosome 7.

SNV	missense	frameshift	splice_donor	splice_acceptor	pathogenic
chr7:142749524-C-G	√				
chr7:142750558-C-G	√				
chr7:142750561-C-T	√				√
chr7:142750563-C-T	√				√
chr7:142750660-G-T	√				
chr7:142750672-T-A	√				
chr7:142750675-A-G	√				
chr7:142750676-C-G	√				
chr7:142750680-C-T					
chr7:142750699-G-C	√				
chr7:142750715-G-A			√		
chr7:142751076-AAG-A		√			
chr7:142751081-A-AG		√			
chr7:142751082-A-G	√				
chr7:142751775-C-T	√				
chr7:142752476-G-C	√				
chr7:142752490-G-A	√				
chr7:142752505-G-T					
chr7:142752506-G-T	√				
chr7:142752522-C-G	√				
chr7:142752913-G-A	√				
chr7:142760456-C-T			√		
chr7:142772094-T-C	√				
chr7:142772142-A-C	√				
chr7:142772154-G-T	√				
chr7:142772166-T-A	√				
chr7:142772170-C-G	√				
chr7:142772174-C-T					
chr7:142772193-G-C	√				
chr7:142772207-T-C	√				
chr7:142772209-G-A			√		
chr7:142773267-C-T	√				
chr7:142773330-A-T	√				
chr7:142773337-C-T	√				
chr7:142773369-C-T	√				
chr7:142773408-T-A	√				
chr7:142773417-G-A	√				

chr7:142773426-T-G	√		
chr7:142773430-G-A	√		
chr7:142773438-G-A	√		
chr7:142773456-G-A	√		
chr7:142773476-C-T			√
chr7:142773481-C-G	√		
chr7:142773976-G-A	√		
chr7:142774011-A-G	√		
chr7:142774423-GT-G		√	
chr7:142774435-A-G	√		
chr7:142774436-C-A	√		

The effect of the variants were annotated by SnpEff [7].

Table S2. Evaluation of detecting pathogenic alleles on *PRSS1* using PRSS-call.

Region	Nucleotide change	SNV ID	GRCh38 Position	Clinical Significance	Recall rate
exon 2	c.47C>T	rs202003805	142750561	Pathogenic	1
exon 2	c.49C>A		142750563	Pathogenic	1
exon 2	c.56A>C		142750570	Pathogenic	1
exon 2	c.62A>C		142750576	Pathogenic	1
exon 2	c.65A>G		142750579	Pathogenic	1
exon 2	c.68A>G	rs111033567	142750582	Pathogenic	1
exon 2	c.86A>C	rs111033566	142750600	Pathogenic	1
exon 2	c.116T>C		142750630	Pathogenic	1
exon 3	c.235G>A	rs111033564	142751808	Likely pathogenic	0.84
exon 3	c.276G>T		142751849	Pathogenic	1
exon 3	c.298G>C		142751871	Pathogenic	1
exon 3	c.311T>C		142751884	Pathogenic	1
exon 3	c.346C>T		142751919	Pathogenic	1
exon 3	c.364C>T	rs111033568	142751937	Pathogenic	1
exon 3	c.365G>A	rs111033565	142751938	Pathogenic	1
exon 3	c.371C>T		142751944	Pathogenic	1
exon 3	c.415T>A		142751988	Pathogenic	1
exon 3	c.416G>T		142751989	Pathogenic	1
exon 4	c.508A>G		142752484	Likely pathogenic	1
exon 5	c.623G>C	rs189270875	142752899	Pathogenic	1

Variants in this table are derived from 'pathogenic' and 'likely pathogenic' variants documented in the pancreatitis genetic database (www.pancreasgenetics.org).

References:

- 1 Rowen L, Koop BF, Hood L. The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Science* 1996;**272**:1755-62.
- 2 Nurk S, Koren S, Rhie A, Rautiainen M, Bizkadze AV, Mikheenko A, *et al.* The complete sequence of a human genome. *bioRxiv* 2021.
- 3 Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;**26**:589-95.
- 4 Rowen L, Williams E, Glusman G, Linardopoulou E, Friedman C, Ahearn ME, *et al.* Interchromosomal segmental duplications explain the unusual structure of PRSS3, the gene for an inhibitor-resistant trypsinogen. *Mol Biol Evol* 2005;**22**:1712-20.
- 5 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078-9.
- 6 Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012:arXiv:1207.3907.
- 7 Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 2012;**6**:80-92.
- 8 Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics* 2007;**81**:1084-97.
- 9 Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, *et al.* Versatile and open software for comparing large genomes. *Genome Biol* 2004;**5**:R12.
- 10 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. *Curr Protoc Bioinformatics* 2013;**11**:11.0.1-0.33.
- 11 Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 2021;**372**.
- 12 Byrka-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, *et al.* High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* 2021.
- 13 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;**9**:e112963.
- 14 Dong SS, He WM, Ji JJ, Zhang C, Guo Y, Yang TL. LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *Brief Bioinform* 2021;**22**.