HGG Advances

ARTICLE

High-quality read-based phasing of cystic fibrosis cohort informs genetic understanding of disease modification

Scott Mastromatteo,[1,2] Angela Chen,[1] Jiafen Gong,[1] Fan Lin,[1] Bhooma Thiruvahindrapuram,[1,2] Wilson W.L. Sung,[1,2] Joe Whitney,[1,2] Zhuozhi Wang,[1,2] Rohan V. Patel,[1,2] Katherine Keenan,[1] Anat Halevy,[1] Naim Panjwani,[1] Julie Avolio,[1] Cheng Wang,[1] Guillaume Côté-Maurais,[3] Stéphanie Bégin,[3] Damien Adam,[3,4] Emmanuelle Brochiero,[3,4] Candice Bjornson,[5] Mark Chilvers,[6] April Price,[7] Michael Parkins,[8] Richard van Wylick,[9] Dimas Mateos-Corral,[10] Daniel Hughes,[10] Mary Jane Smith,[11] Nancy Morrison,[12] Elizabeth Tullis,[13] Anne L. Stephenson,[13] Pearce Wilcox,[14] Bradley S. Quon,[14] Winnie M. Leung,[15] Melinda Solomon,[16] Lei Sun,[17,18] Felix Ratjen,[1,19] and Lisa J. Strug[1,2,17,18,20,*]

Summary

Phasing of heterozygous alleles is critical for interpretation of *cis*-effects of disease-relevant variation. We sequenced 477 individuals with cystic fibrosis (CF) using linked-read sequencing, which display an average phase block N50 of 4.39 Mb. We use these samples to construct a graph representation of *CFTR* haplotypes, demonstrating its utility for understanding complex CF alleles. These are visualized in a Web app, CFTbaRcodes, that enables interactive exploration of *CFTR* haplotypes present in this cohort. We perform fine-mapping and phasing of the chr7q35 trypsinogen locus associated with CF meconium ileus, an intestinal obstruction at birth associated with more severe CF outcomes and pancreatic disease. A 20-kb deletion polymorphism and a *PRSS2* missense variant p.Thr8Ile (rs62473563) are shown to independently contribute to meconium ileus risk ($p = 0.0028$, $p = 0.011$, respectively) and are *PRSS2* pancreas eQTLs ($p = 9.5 \times 10^{-7}$ and $p = 1.4 \times 10^{-4}$, respectively), suggesting the mechanism by which these polymorphisms contribute to CF. The phase information from linked reads provides a putative causal explanation for variation at a CF-relevant locus, which also has implications for the genetic basis of non-CF pancreatitis, to which this locus has been reported to contribute.

Introduction

Current genetic epidemiological studies often fail to capture the complete diploid nature of the human genome,[1] largely because of a reliance on genotyping arrays and short-read whole-genome sequencing. These technologies can identify heterozygous alleles but provide little to no information regarding the *cis* or *trans* phase relationships of their heterozygous allele pairs. Accurate haplotype information can be essential in informing phenotype-genotype relationships such as complex alleles and compound heterozygotes for recessive diseases as well as resolving allelic heterogeneity of SNPs identified though a genome-wide association study (GWAS).

One of the most well-known examples of compound heterozygosity comes from cystic fibrosis (CF).[2] CF is caused by mutations in the *CF transmembrane conductance regulator* (*CFTR*).[3] Over 2,100 variants have been identified in *CFTR*,[4] and more than 400 of these have been shown to be disease causing, while others are reported to have varying clinical consequence and are CF causing only when in *cis* with another deleterious variant.[5] Meanwhile, individuals with identical CF-causing alleles display variable disease severity and response to *CFTR*-targeting therapies.[6] CF co-morbidities and variation in disease severity are complex genetic traits,[7] presumed to be due to the impact of other genes beyond *CFTR* collectively referred to as modifier genes. For example, GWAS of CF meconium ileus, an intestinal obstruction seen at birth in 13%–21% of individuals with CF,[8] has identified associated loci[9,10] that correspond to altered comorbidity risk.

Meconium ileus occurs almost exclusively in individuals with severe *CFTR* mutations and correlates with disease in other CF affected organs, in particular pancreatic disease[8,11] and CF-related diabetes.[12,13] Additionally, the pancreas is one of the earliest-affected organs in CF,[14]

[1]Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON, Canada; [2]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON, Canada; [3]Centre de recherche du centre hospitalier de l'Université de Montréal (CRCHUM), Montréal, QC, Canada; [4]Department of Medicine, Faculty of Medicine, Université de Montréal, Montréal, QC, Canada; [5]Alberta Children's Hospital, Calgary, AB, Canada; [6]British Columbia Children's Hospital, Vancouver, BC, Canada; [7]The Children's Hospital, London Health Science Centre, London, ON, Canada; [8]University of Calgary, Department of Medicine, Calgary, AB, Canada; [9]Kingston Health Sciences Centre, Kingston, ON, Canada; [10]IWK Health Centre, Halifax, NS, Canada; [11]Memorial University of Newfoundland, Faculty of Medicine, St. John's, NL, Canada; [12]Queen Elizabeth II Health Sciences Centre, Halifax, NS, Canada; [13]St. Michael's Hospital, Toronto, ON, Canada; [14]St. Paul's Hospital, Vancouver, BC, Canada; [15]University of Alberta Hospital, Edmonton, AB, Canada; [16]Division of Respiratory Medicine, The Hospital for Sick Children, Toronto, ON, Canada; [17]Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada; [18]Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada; [19]Department of Paediatrics, University of Toronto, Toronto, ON, Canada; [20]Department of Computer Science, University of Toronto, Toronto, ON, Canada
*Correspondence: lisa.strug@utoronto.ca
https://doi.org/10.1016/j.xhgg.2022.100156.

and it has been observed that individuals with meconium ileus have lower immunoreactive trypsinogen levels at birth compared with individuals with CF without meconium ileus.[15] CF meconium ileus GWAS has identified three genome-wide significant loci, with a fourth suggestive intergenic locus within the T cell receptor beta region (chr7q35),[9] which was replicated in independent samples.[10] Despite failing to reach genome-wide significance, the chr7q35 locus is tantalizing because it contains pancreatic trypsinogen and also affects non-CF pancreatitis risk. An amino acid substitution in *PRSS1* (p.R122H) is the most common cause of hereditary pancreatitis in Europeans.[16] This small change alters a trypsin cleavage site that is important for regulation of trypsin activity through autoinactivation of trypsinogen.[17] Similarly, chronic pancreatitis has been shown to be associated with a common T>C variant (rs10273639) near *PRSS1*,[18] thought to be associated with altered risk by tagging a promoter SNP (rs4726576) that increases *PRSS1* expression.[19]

Early sequencing work in the chr7q35 region identified five trypsinogen paralogs (originally annotated as T4 to T8) with approximately 90%–91% nucleotide similarity.[20] *PRSS1* (T4) and anionic trypsinogen *PRSS2* (T8) are major forms of trypsin expressed exclusively in the pancreas. The other three genes are pseudogenes: *PRSS3P1* (T5), *PRSS3P2* (T6), and *TRY7* (T7). Of the three pseudogenes, there is only evidence for *PRSS3P2* transcription but a protein product has not been observed.[21] The GRCh38 reference genome only includes three of these genes (T4, T5, T8), which is an accurate representation of a common deletion polymorphism that removes T6 and T7. This approximately 20-kb deletion appears to have arisen via non-allelic homologous recombination,[22] and is a common variation found in approximately 41% of individuals with European ancestry.[23] The GRCh38 alternative contig, `KI270803.1`,[24] represents the non-deleted haplotype and contains genes T4–T8. This is further complicated by reference assembly GRCh37 being erroneously structured (T4, T5, T6) and excluding *PRSS2*; a correction was later released (`chr7_gl582971_fix`) that included all five genes.

Given the structural complications mentioned above, past GWAS work provides limited insight into the complex variation at this locus because genotype arrays generally contain common SNPs in easily accessible regions of the genome by design. The meconium ileus-associated SNPs evaluated were not found in high linkage disequilibrium (LD) with protein-coding variations, suggesting their impact could be through gene regulation.[9,10] However, much remains to be learned about the variation in *cis* with these associated SNP risk alleles or whether combinations of multiple *cis*-acting variants contribute to meconium ileus risk; for this, genotype data at the associated loci must be phased.

In a typical epidemiological study, data external to the target individual are used to reconstruct maternal and paternal haplotypes. Pedigree-based phasing offers a high degree of accuracy[25] but requires a family-based experimental design and cannot resolve phase for variants that are heterozygous for all members. Population-based phasing is a cost-effective alternative that exploits shared ancestry information and LD patterns to statistically infer haplotypes. However, the statistical nature of population-based phasing makes it vulnerable to frequent switch errors: accidental transitions from maternal to paternal haplotypes between neighboring heterozygous sites.[1] Phasing rare variants can also be problematic, requiring inference when few or no copies of that rare variant are present within the reference population.

In contrast, sequenced-based phasing approaches determine phase relationships for a target individual without reliance on an external dataset. Long-read sequencing technologies such as Pacific Biosciences (PacBio) single molecule, real-time (SMRT) sequencing and Oxford Nanopore generate longer reads capable of phasing longer distances, but, until recently, these technologies were associated with being error prone and costly relative to short-read sequencing. Other alternatives utilize a standard short-read sequencing pipeline with an additional experimental step that introduces long-range information into the read data. For example, the 10x Genomics (10XG) linked-read technology[26] and Universal Sequencing Technologies TELL-Seq[27] tag reads are derived from a single DNA molecule with a shared nucleotide barcode.

We sequenced 477 individuals with CF from the Canadian CF Gene Modifier Study Consortium (CGMS) cohort[10] using 10XG linked-read technology at approximately 30× coverage. We summarize the phasing quality achieved and use the phase information to unravel the complex genomic architecture. First, we examine *CFTR* as an example of how phase enables improved understanding of complex alleles and compound heterozygosity in recessive disease and then consider the insights for allelic heterogeneity at the chr7q35 modifier locus for the CF comorbidity meconium ileus.

Materials and methods

High-molecular-weight DNA extraction methods

CGMS was approved by the Research Ethics Board of the Hospital for Sick Children (# 0020020214 from 2012-2019 and #1000065760 from 2019-present). Blood samples were extracted from patients with CF across Canada (Table S1) and sent for processing to The Hospital for Sick Children in Toronto, Canada. Written informed consent was obtained from all participants, or parents/guardians/substitute decision makers. High-molecular-weight (HMW) DNA was extracted from fresh or frozen blood aliquots using the MagAttract HMW DNA Kit (Qiagen, catalog no. [Cat#] 67563) as per supplier recommendations. Quantitation was determined by Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, Cat# P11469), as recommended by the supplier. Quality of DNA was then further assessed by electrophoretic migration in 0.4% agarose gel and run at 50 V for 18 h at 4°C in Tris-acetate buffer at pH 8.0 with comparison with Quick-Load 1-kb Extend DNA ladder (NEB, Cat# N3239S). Unless otherwise stated, only

samples indicating that bulk DNA was larger than 50 kb (>80% by visual inspection of agarose gel) were submitted for sequencing.

We also investigated three other DNA extraction methods, including two Autopure methods (Maxi, 7–10 mL of blood; Midi, 3–4 mL of blood) and Puregene (0.3–1 mL of blood, manual extraction) (Qiagen, Cat# 1057048, 949006, 949008, 949016, 949018, and 949010). These samples were prepared as recommended by the kit supplier, but typically failed the HMW quality control assessment by the 0.4% agarose gel.

Library preparation and 10XG sequencing

Approximately 1 μg of genomic DNA was submitted to The Centre for Applied Genomics (TCAG) at The Hospital for Sick Children for genomic library preparation and whole-genome sequencing. DNA samples were quantified using Qubit High Sensitivity Assay and sample purity was checked using Nanodrop OD260/280 ratio. DNA was run on the Genomic Tape on Tapestation (Agilent, Cat# 5067–5365 and 5067–5366) to check DNA fragment size. Ten nanograms of DNA was used as input material for library preparation using the 10XG Library Kit (PN-120258 andPN-120257) following the manufacturer's recommended protocol. In brief, DNA was denatured and mixed with gel beads to form emulsion droplets using the Chromium Controller (PN-110203), emulsion droplets were tagged with barcodes and amplified by PCR, and emulsions were broken and DNA captured and cleaned using magnetic beads. DNA was checked on the Bioanalyzer DNA High Sensitivity chip to ensure fragment size, and the DNA proceeds to library preparation. DNA was end-repaired, A-tailed, ligated with Illumina-compatible adapters, and PCR amplified with indexed Chromium i7 primers (PN-120262). Libraries are validated on a Bioanalyzer DNA High Sensitivity chip to check for size and absence of primer dimers and quantified by qPCR using Kapa Library Quantification Illumina/ABI Prism Kit protocol (KAPA Biosystems). Validated libraries were paired-end sequenced on an Illumina HiSeq X platform following Illumina's recommended protocol to generate paired-end reads of 150 bases in length.

Variant calling and phasing metrics for 10XG samples

Long Ranger 2.2.2 and GRCh38 reference version 2.1.0 were used to process 10XG reads. Base calling was performed using the `mkfastq` command. VCF files were generated using the `wgs` command to call and phase variants; GATK 4.0.0.0 was used internally by Long Ranger to call variants. Alignment and phasing statistics were also generated by Long Ranger as output to the `wgs` command. The `stats` command from WhatsHap v0.18 was applied to the Long Ranger VCF files to produce additional phasing statistics. When both Long Ranger and WhatsHap reported the same metric, we took the values reported by Long Ranger. For causal CF variants, chart review and manual inspection of the Long Ranger alignment file with the Integrative Genomics Viewer (IGV) was performed to investigate disagreements between clinical records and called variants. Linked-read data for reference sample NA12878 were downloaded from the Genome in a Bottle Consortium.

Generating CFTR haplotypes

10XG variant calls for each sample were filtered to chr7: 117379963–117768665. Variants without "`PASS`" in the VCF "`FILTER`" column had the genotype set to missing. A multi-sample VCF was generated using BCFtools 1.12[28] `merge`, and variants with allele count less than three and variants without an rsID were filtered out. The intronic poly-T tract polymorphisms were manually called and phased using the 10XG sequencing reads. A graphical representation of haplotypes was generated by restricting variants to 50 bp around exonic *CFTR* variants. The vg toolkit 1.33.0[29] was used to generate a graph and a plot was produced using Sequence Tube Map[30]. The CFTbaRcodes Web application uses the complete multi-sample VCF but only displays summary-level information. Rare variants (allele count less than three) and rare haplotypes (observed in less than three individuals) are filtered and hidden from display. The code for the application can be found at https://github.com/strug-hub/CFTbaRcodes.

10XG realignment and deletion polymorphism calling

10XG sequencing reads aligned to the PRSS1-PRSS2 locus (GRCh38 chr7:142500000–143000000) and a region spanning *PRSS3* (GRCh38 chr9:33700000–33900000) were extracted from the Long Ranger BAM file using SAMtools v1.9.[31] The extracted reads were realigned using Long Ranger 2.2.2 to a custom reference containing KI270803.1 and the *PRSS3* locus (GRCh38 chr9:33500000–34100000). The *PRSS3* locus was included because it shares a high base pair identity to the PRSS1-PRSS2 locus, and we observed some reads aligned to *PRSS3* map better to the chromosome 7 locus.

To call the large deletion polymorphism observed on KI270803.1, a custom Python script was used to determine the presence of the deletion by comparing the coverage of the deleted region (KI270803.1:771000–790000) with a flanking region of the same size (KI270803.1:760500–770000 and KI270803.1: 791000–800500) on both sides of the deleted region. Deletion calls were also visually validated using IGV. A dummy SNP was added to the VCF to encode the genotype of the deletion. An additional step was required to phase heterozygous deletion calls with respect to the other variants called by Long Ranger. Using haplotype-tagged 10XG linked reads, all heterozygous deletion calls were manually phased using IGV with respect to rs3757377, which lies upstream of the deletion. In the case where the deletion was heterozygous and rs3757377 was homozygous, the deletion was instead phased with respect to rs6666. Phase of the deletion calls in the VCF were updated using a custom script to reflect the phase relationship observed in the linked reads.

Each 10XG VCF was filtered for variants with `PASS` in the `FILTER` column. Using BCFtools 1.12[28] `merge`, a multi-sample VCF was created by combining all the individual VCFs (`-missing-to-ref`). Variants in the multi-sample VCF called outside of KI270803.1 were removed. Variants with allele counts less than three, multi-allelic variants, and indels longer than five bases (other than the 20-kb deletion that was coded as a SNP) were removed. SHAPEIT version 4.1.2[32] was used to impute the missing variants and completely phase the multi-sample VCF to enable use as a reference panel (`-use-PS 0.0001 -sequencing`). LD blocks were computed from this VCF using LDBlockShow version 1.36[33] (`-BlockType 2 -SeleVar 1`).

Illumina genotype arrays and quality control

CGMS data are genotyped on four different Illumina platform: 610Quad, 660W, Omni2.5, and Omni5. Genotype calling was performed using GenomeStudio V2011.1. Quality control steps were performed separately for each platform and described in detail in Gong *et al.*[10] Briefly, PLINK[34] was used for most quality control steps, while KING[35] identified any cryptic familial relationships among all individuals and PC-AiR[36] calculated PCs. Parents in six parent-offspring pairs, 19 samples clustered with Hapmap3[37]

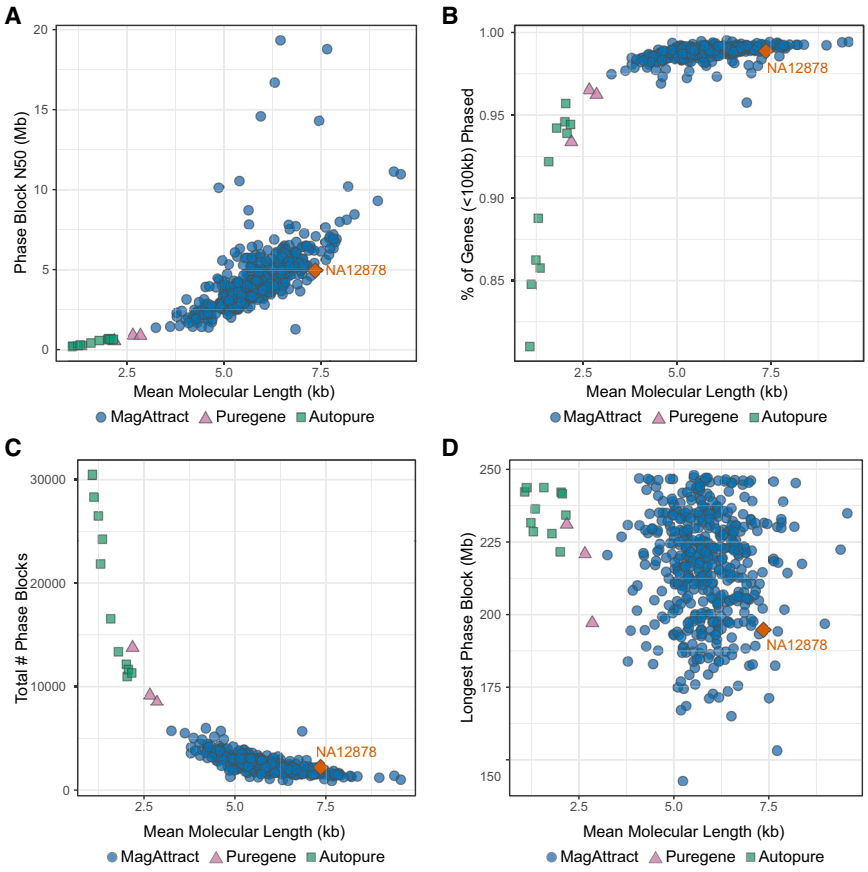

Figure 1. Genome-wide phasing statistics versus mean molecular length for CGMS samples and NA12878 sequenced by 10XG

DNA from CGMS cohort extracted using either MagAttract (blue circle), Autopure (green square), or Puregene (purple triangle). Public sample NA12878 (orange diamond) was down-sampled to a comparable coverage (30×). Statistics are compared against mean molecular length reported by Long Ranger.[26]

(A)Phase block N50.

(B)Proportion of phased genes with length less than 100 kb.

(C)Total number of phase blocks.

(D)Size of longest phase block in base pairs.

African and East Asian ethnicity, and 10 samples with sex mismatch were excluded. Significant PCs were selected to be included in the association based on the Tracy-Widom test result using the function `twtable` in POPGEN of Eigensoft.[38] For colocalization of meconium ileus association with GTEx eQTLs, GWAS summary statistics[10] were reformatted as BED file and lifted to GRCh38 by LiftOver[39] for colocalization analysis against GTEx v8 in LocusFocus.[40]

Imputation of genotype data using 10XG

Genotype array data were generated against GRCh37 and required lifting to alternative contig KI270803.1 before imputation. A two-step lift-over was performed using Picard LiftoverVcf,[41] first from GRCh37 to GRCh38 using a chain file available from the University of California Santa Cruz (UCSC) genome browser and then from GRCh38 to alternative contig KI270803.1. The chain file from GRCh38 to KI270803.1 was created by downloading a PSL file for alternative haplotypes using the UCSC table browser and converting to a chain file using axtChain. Genotype array calls were organized by array platform into separate multi-sample VCF files and imputed by BEAGLE v5.1[25] using the 10XG reference panel and default parameters.

Association with meconium ileus

Variants from 2,635 pancreatic insufficient individuals with BEAGLE imputation quality DR2 $>$ 0.3 were kept for association analysis with meconium ileus using imputation dosage of each variant, which was performed using the `geeglm` function from the R geepack package,[42] with exchangeable correlation structure and binomial family. Sex, array platform, and 11 PCs were included in the model. For conditional analysis, the dosage of the deletion was added as a covariate. For association testing with the 10XG data, only pancreatic-insufficient individuals with available meconium ileus status were considered. 10XG variant calls within the range KI270803.1:700000–900000 were regressed against meconium ileus status (n = 337 samples) using logistic regression. For conditioning on deletion genotype or rs62473563, the respective dosage was included as a covariate in the model. A subsequent regression was conducted where 28 individuals with the highest *CFTR* severity score were excluded.

Re-processing of GTEx RNA-seq data

A custom reference genome was generated by adding the alternative contig KI270803.1 to a GRCh38 reference FASTA file. To remove sequence redundancy, the region on the chromosome 7 main contig corresponding to KI270803.1 (chr7:142038121–143088503) was masked with the ambiguous base "`N`". Then 172 RNA sequencing (RNA-seq) GTEx samples from pancreas were downloaded and reads were aligned to our custom reference using the scripts from the GTEx pipeline.[43] First, GENCODE v26[44] annotations were retrieved from the GTEx Portal and annotations within chr7:142038121–143088503 were removed. GENCODE v35 annotations for KI270803.1 were downloaded and collapsed using `collapse_annotation.py`, available from the GTEx pipeline. The two resulting GTF files were combined into a single annotation file. We indexed our custom reference assembly with this annotation file using STAR v2.7.0[45] (`--sjdbOverhang 75`). For each sample, we aligned RNA-seq reads using the `run_STAR.py` script from the GTEx pipeline. Transcript quantification was performed by mmquant[46] (`-l 20`) and read counts were normalized by conversion to transcripts per million (TPM).

Recalculating GTEx pancreas eQTL data

Calculation of eQTLs was performed following the GTEx pipeline.[43] GTEx v8 variant calls were filtered to chr7:142038121–143088503 and only included 252 pancreas samples with race

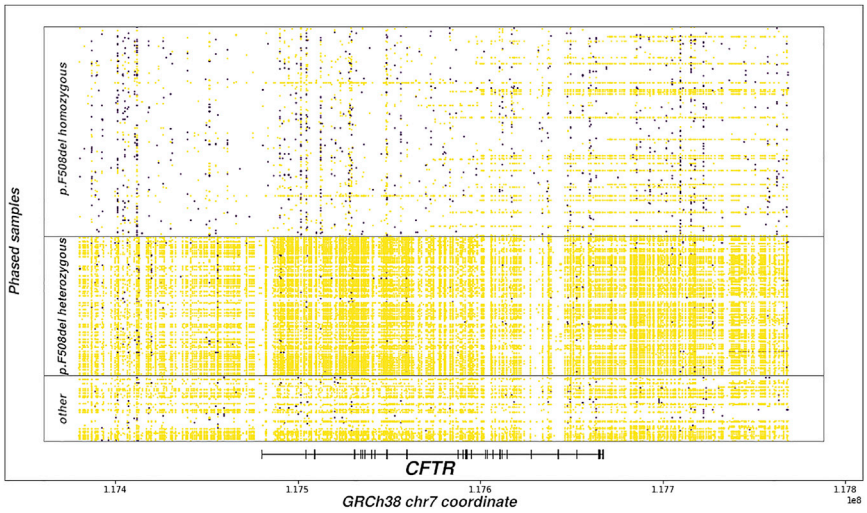

Figure 2. Phased heterozygous variants across chr7:117379963-117768665 grouped by *CFTR* mutation

10XG phased variant calls around *CFTR*. Each row represents a single individual, and rows are grouped by the number of p.F508del mutations present. Heterozygous variants are plotted by position as points where yellow points belong to the largest phase block and purple ones do not. Individuals with two different CFTR mutations have fewer runs of homozygosity and more complete phase blocks.

labeled as "`white`." Using the previously generated chain file, the GTEx multi-sample VCF and annotation BED file was lifted over from GRCh38 to KI270803.1. BEAGLE v5.1 was then used to impute the deletion from the 10XG reference panel into the GTEx VCF. Matching GTEx v8 read counts were normalized between samples using TMM.[47] PEER factors were calculated from the normalized gene expression values using `run_PEER.R` from the GTEx pipeline. In addition to 15 PEER factors, the covariates used by GTEx v8 were included (five PCs, sex, PCR status, and platform). FastQTL v2.184[48] performed the eQTL analysis restricted to gene annotations on KI270803.1. For conditioning on deletion genotype or rs62473563, the respective dosage was included as a covariate in the model.

Results

Phasing 477 Canadians with CF

The 477 Canadians with CF were recruited from 13 CF centers across the country (Table S1), and their DNA was sequenced using the 10XG linked-read technology at 30× depth (25× after trimming the 10XG barcode). The phasing distance of 10XG linked reads is limited by the size of DNA molecules extracted. We investigated different extraction methods and found MagAttract produces the best results, consistent with publicly available NA12878 10XG data[49] (Figure 1). Mean molecular length averages 58.7 kb (range, 32.6–95.4 kb) across 463 MagAttract-extracted samples and is a strong predictor of the quality of the phasing. The average MagAttract-extracted sample is phased in 2,444 blocks, with N50 of 4.39 Mb and a mean of 1,428 variants per block. The largest phase block across all samples is 247.97 Mb, and all but two samples have >97% of all genes shorter than 100 kb phased in a single block. Additional statistics can be found in Table S2.

To complement genome-wide statistics and investigate phasing of complex alleles, compound heterozygosity and homozygous causal genotypes in a recessive mendelian disease, we assess the local phasing of a 389-kb region encompassing the CF causal gene, *CFTR* (GRCh38 chr7:117379963–117768665; *CFTR* plus 100-kb on both sides). The most common CF-causing variant is p.Phe508del; 241 individuals homozygous for this variant comprise about half of the sequenced samples. Due to a conserved haplotype, individuals homozygous for p.Phe508del possess high levels of homozygosity along the entire *CFTR* gene, which makes it difficult to phase. The median p.Phe508del homozygous individual has 10 heterozygous variant calls within the assessed region (one per ∼40 kb) compared with 236 heterozygous variants (one per ∼1.6 kb) for the median individual with heterozygous CF-causing variants (Figure 2). Consequently, 152 of the 199 individuals with heterozygous CF-causing variants have a single phase block spanning the complete 389-kb region. This demonstrates how the phasing of causal loci in disease cohorts with a recessive mode of inheritance could pose unique challenges for read-based phasing techniques but also highlights the potential to identify complex alleles that may explain disease variation.[7]

We construct a graph representation of the phased sequence at the *CFTR* locus to provide a visual understanding of the 10XG-derived haplotypes (Figure 3). The graph includes the multiallelic poly-T tract polymorphism to highlight how a graph representation of haplotypes can inform disease phenotypes. Variation at the poly-T tract results in altered splicing and can cause CF if in *cis* with specific *CFTR* mutations[50]; p.R117H in phase with a short poly-T is CF causing, while the clinical manifestations for those with longer poly-T sequence is less certain. Nine different poly-T alleles are visualized and their phase is shown with respect to downstream variants including p.Phe508del. We have created a Web application called CFTbaRcodes, available at https://cftbarcode.research.sickkids.ca, that enables exploration of the CFTR haplotypes found within the CGMS cohort. The tool allows selection of CFTR variants of interest and visualizes haplotypes using the textile plot[51] (Figure 4).

Structure of the chr7q35 trypsinogen CF modifier locus

The meconium ileus GWAS-suggestive locus on chr7q35 has five duplicated trypsinogen paralogs (*PRSS1*, *PRSS3P1*, *PRSS3P2*, *TRY7*, and *PRSS2*) but is structurally variable across

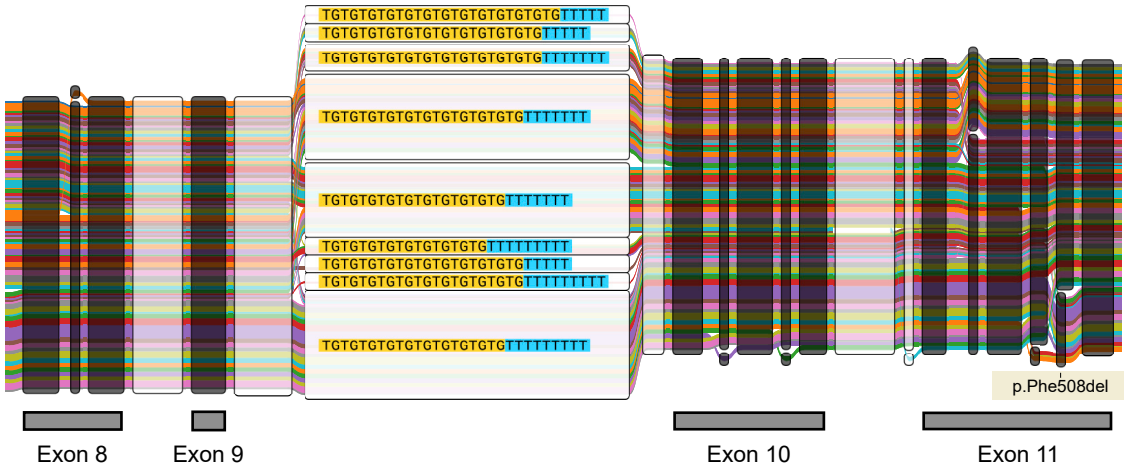

Figure 3. Graph representation of exonic variants for 898 *CFTR* haplotypes
The graph is composed of nodes representing sequence and haplotype groups as colored edges. The haplotype sequence can be reconstructed by concatenating the nodes along a path. The thickness of each edge denotes the haplotype frequency in the dataset. Nodes belonging to exons are annotated and colored black. The intronic poly-T tract is included in the graph representation. Nine different poly-T alleles are visualized here and shown with respect to p.Phe508del downstream from the poly-T tract.

reference assemblies (Figure 5). The GRCh38 chr7 primary sequence includes a large deletion polymorphism that removes *PRSS3P2* and *TRY7*. While this causes no issues for individuals homozygous for the deletion, short reads from individuals who carry a non-deleted haplotype align spuriously to GRCh38, resulting in false variant calls (Figure 6). Realignment of 10XG reads to alternative contig KI270803.1 improves the calling and phasing of variation and enables the large deletion polymorphism to be unambiguously called (Figure 7).

Of 477 10XG samples, 424 are completely phased in a single block across a conservative 200-kb region surrounding the PRSS1-PRSS2 locus (KI270803.1:700000–900000). With respect to the 20-kb sequence spanning the deletion boundary, we identify three distinct haplotype groups (Figure S1). Two of these groups are composed of seven haplotypes derived from six individuals with admixed African ancestry. These individuals are excluded from subsequent analysis. Among individuals with European ancestry, we find almost no variation within the deletion boundary on haplotypes lacking the deletion (Figure S1B). A simple genotype coding of the deletion sufficiently captures the genetic variation contained in this subregion and is used for all subsequent analysis.

The 10XG phase information elucidates the LD structure of this locus for the CGMS cohort and is shown alongside CF meconium ileus GWAS summary statistics in Figure 8A. Two association peaks centered at rs3757377 (KI270803.1:750284C>T) and rs1799886 (KI270803.1:823812T>C) are present in different LD blocks. The rs3757377 risk allele "T" has a frequency of 41% in the 10XG calls. We phase this SNP with respect to other variants of interest within the same LD block; the two major haplotypes account for 94.7% of the observed data (Figure 8B).

The second peak centered at rs1799886 has a similar minor allele frequency of 43.5% but is not in strong LD with the deletion polymorphism ($D' = -0.55$, $r^2 = 0.19$). A search for variants in *cis* with rs1799886 reveals a nonsynonymous *PRSS2* variant (p.Thr8Ile), rs62473563 (KI270803.1:793978C>T), with 10.7% minor allele frequency and a high D′ with rs1799886 ($D' = -0.98$, $r^2 = 0.09$). The rs1799886 T allele is in *cis* with p.Thr8Ile for 100 out of 101 haplotypes. The GWAS signal is tagging this protein-coding SNP; this relationship was not uncovered in the original analysis of the GWAS results due to the absence of *PRSS2* from the GRCh37 reference.

Analyzing the chr7q35 trypsinogen CF modifier locus

A query of the Genotype-Tissue Expression (GTEx) v8 data [52] was conducted to search for pancreas eQTLs with respect to the five trypsinogen paralogs. *PRSS3P2* and *TRY7* are not reported by GTEx v8 due to their absence from GRCh38. *PRSS3P1* does not have significant pancreas eQTLs, but this is expected as it is not transcribed. Significant pancreas eQTLs are reported for *PRSS2* (Table S3) but not for *PRSS1*. This result is surprising because there is a common SNP in the promoter region that is reported to alter *PRSS1* expression [19] but did not appear as a significant eQTL. LocusFocus[40] identifies colocalization between meconium ileus genotype association p values and GTEx v8 *PRSS2* pancreas eQTLs (colocalization p value = 7.1×10^{-8}; Figure S2). This suggests that meconium ileus risk could be modulated by altered *PRSS2* expression. The reliability of GTEx results depends on accurate accounting of the 20-kb deletion polymorphism during read alignment to GRCh38. GTEx v8 variant calls were made from an alt-aware alignment pipeline and avoids mapping issues. However, GTEx v8 RNA-seq reads are aligned to the GRCh38 main chromosomes and therefore any reads derived from *PRSS3P2* and *TRY7* expression are potentially

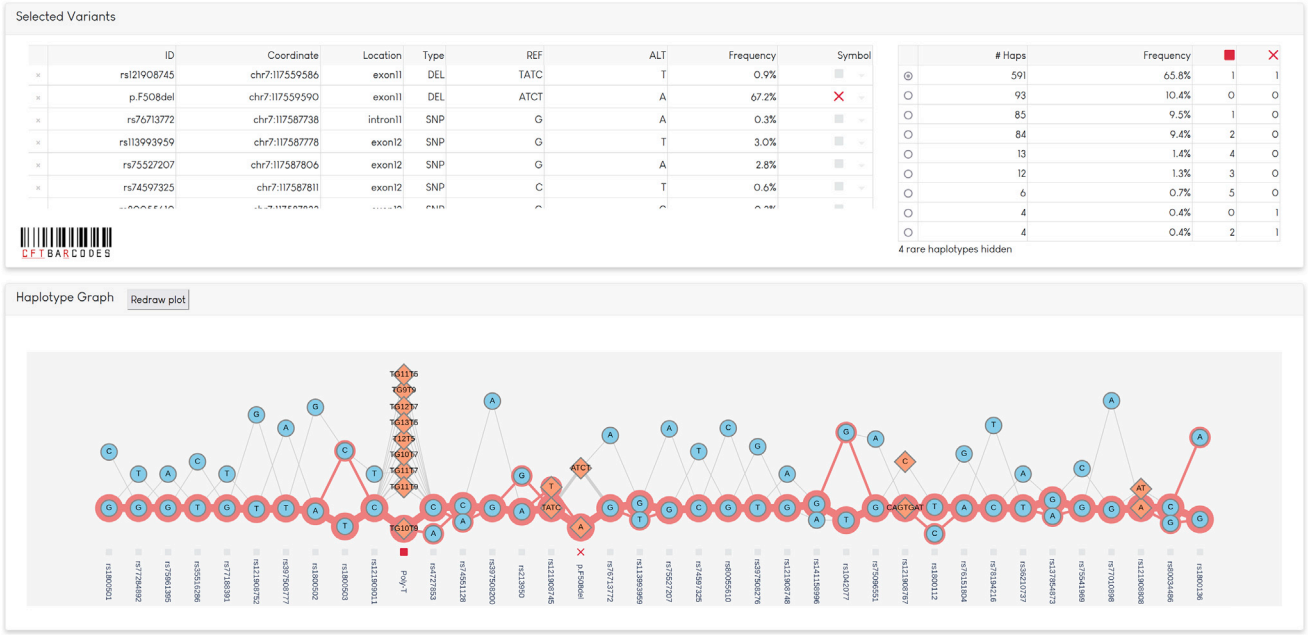

Figure 4. Screenshot of CFTbaRcodes interface

CFTR variants are selected through the application interface and can be annotated with a symbol. In the screenshot, p.F508del is annotated with a cross and the poly-T tract is annotated with a red square. The distribution of haplotypes is shown in a table, based on the variants assigned a symbol. A textile plot[51] of the variants is generated, which is optimized to place alleles in high LD on a horizontal line. SNPs are drawn as blue circles, indels as orange diamonds. Lines show the possible haplotype paths through different alleles, and the weight of the line correlates to frequency observed. If a haplotype is chosen from the table, the path of that haplotype is highlighted by a red line. In the screenshot, the most common p.F508del haplotype is highlighted in red.

misaligned to other paralogs. We find that accounting for the presence of the extra 20-kb sequence does not significantly alter the normalized RNA-seq gene expression counts for *PRSS1* or *PRSS2* compared with GTEx v8 counts (r^2 correlation of the two datasets >0.99; Figure S3).

To improve comparison with the predominantly European CGMS data, 252 GTEx samples with the race labeled as white were used to recalculate pancreas eQTLs. The GTEx v8 GRCh38 variant calls were lifted to KI270803.1 coordinates, which is a precise procedure outside of the 20-kb deletion polymorphism given the 99.9% shared base pair identity between the two contigs. The deletion polymorphism was imputed using the 10XG CGMS samples as a reference panel. Similar to the GTEx v8 results, there are no significant ($p < 0.05$) eQTLs for *PRSS1* (Figure S4), but *PRSS2* pancreas eQTLs are identified (Figure 9A). The imputed deletion polymorphism appears as a strong *PRSS2* pancreas eQTL ($p = 7.8 \times 10^{-5}$). Conditioning on the deletion polymorphism reveals that rs62473563 (*PRSS2* missense variant, p.Thr8Ile) acts as an independent eQTL (Figure 9B). Conditioning on rs62473563 increases the significance of the deletion polymorphism (Figure 9C), and conditioning on both eliminates the *PRSS2* eQTL signal (Figure 9D). The presence

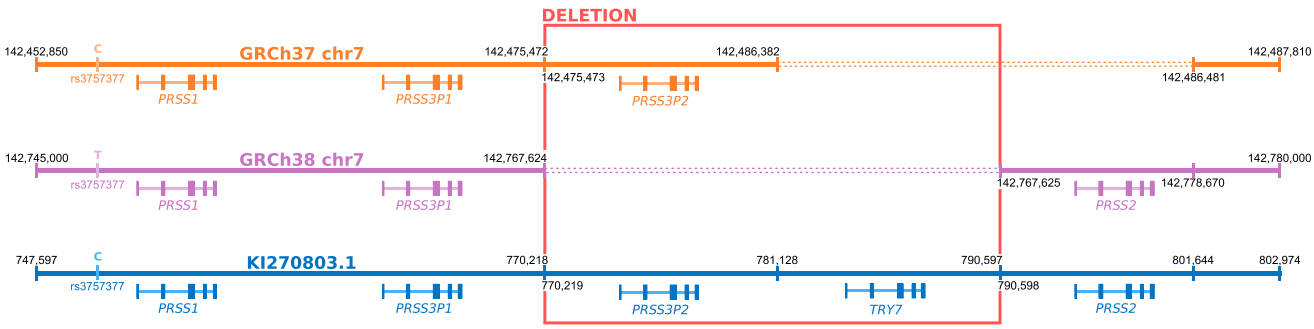

Figure 5. Characterizing the chr7q35 trypsinogen locus

Differences between chromosome 7 reference assemblies for GRCh37, GRCh38, and alternative contig KI270803.1. In the GRCh37 assembly, *TRY7* and *PRSS2* are absent. The GRCh38 assembly does not include *PRSS3P2* and *TRY7* because it accurately represents a haplotype with a common ~20-kb deletion polymorphism (highlighted in red). KI270803.1 represents a haplotype without the deletion polymorphism.

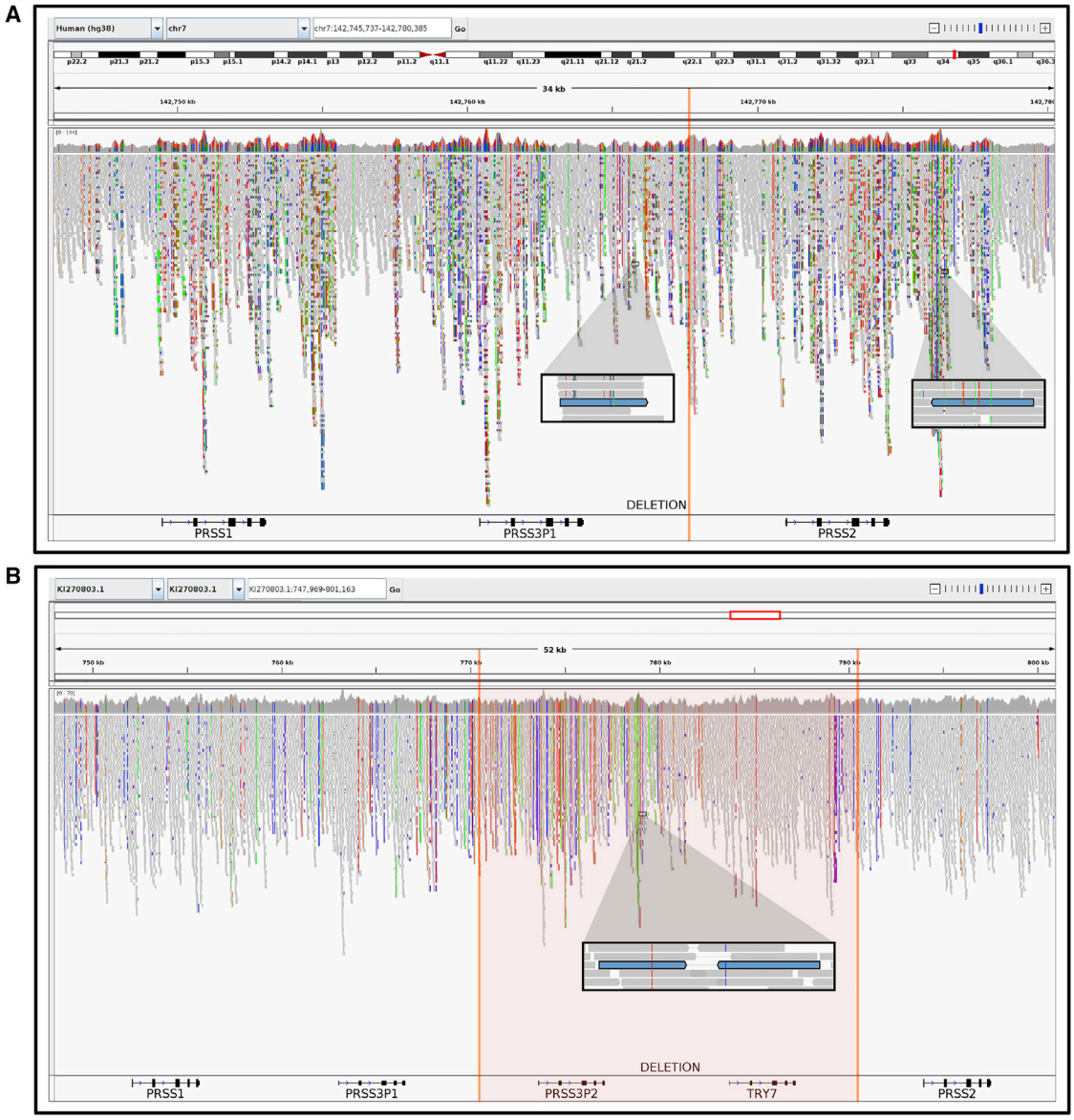

Figure 6. Demonstration of reference bias induced by 20-kb deletion polymorphism and correction using alternative contig
GRCh38 chr7 reference sequence excludes 20-kb present in the sequenced individual (orange vertical lines); reads that would align in this region generate spurious mappings to nearby homologous positions.
(A)IGV visualization of 10XG reads aligned to the GRCh38 chr7 reference chromosome. Regions of higher coverage and an excessive variation are produced. A highlighted read pair maps kilobases apart, downstream of *PRSS3P1* and *PRSS2*.
(B)IGV visualization of 10XG reads aligned to the KI270803.1 alternative contig. Alignments produce more uniform coverage and less variation with respect to the reference. The same highlighted read pair maps with a reasonable insert size, downstream of *PRSS3P2*. The use of KI270803.1 instead of the GRCh38 chr7 reference sequence produces more accurate alignments for individuals carrying the 20-kb absent from the reference sequence.

of p.Thr8Ile and the deletion polymorphism are both associated with reduced *PRSS2* expression. This conditional analysis is summarized in Table 1.

To understand these eQTL results in the context of meconium ileus, we analyze genotyping array data from the CGMS cohort (n = 2,635; Figure S5). The deletion polymorphism is associated with an additive increased risk of disease (beta = 0.29, $p = 5.2 \times 10^{-4}$), but imputation for rs62473563 is poor. Instead, we performed fine-mapping using the 10XG sequencing calls for whom meconium ileus status and 10XG data were available. A similar association pattern observed for the pancreas eQTL was recapitulated for the meconium ileus phenotype in 337 individuals sequenced with the 10XG technology (Figure S6). Interestingly, the contribution of this locus in CF individuals with two minimal-function *CFTR* alleles appears attenuated, which is likely due to their already elevated risk due to *CFTR*.[8] Exclusion of 28 individuals with minimal-function CF alleles produces stronger evidence of association with meconium ileus despite the smaller sample

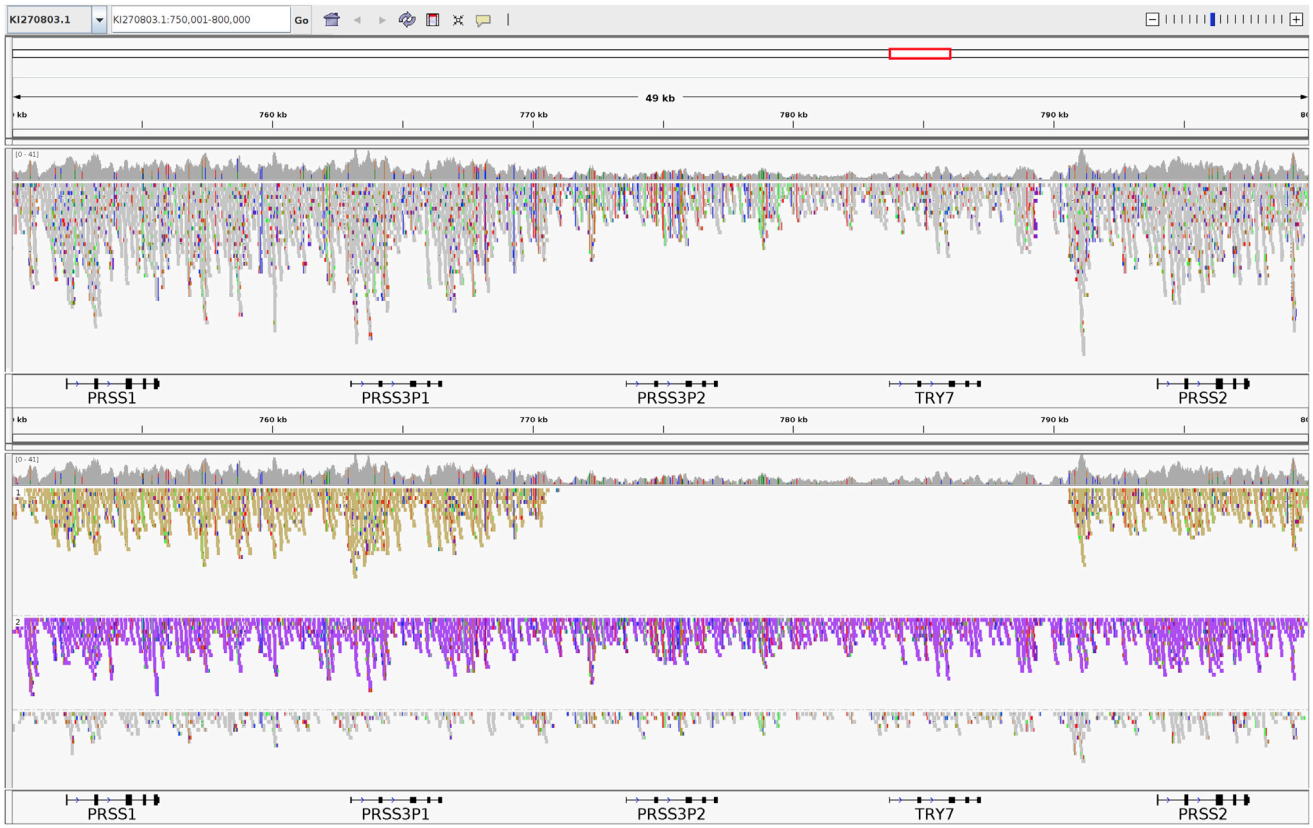

Figure 7. IGV visualizations of 10XG reads aligned to an individual with heterozygous 20-kb deletion polymorphism
10XG reads from an individual heterozygous for the deletion polymorphism aligned to alternate contig KI270803.1 using Long Ranger. Without using linked-read information, haplotype information is obscured (top). Clustering and coloring reads by haplotype tag (haplotype 1, yellow; haplotype 2, purple; unphased, grey) reveals the heterozygous deletion polymorphism (bottom). The deletion is visible on the yellow haplotype and can be phased with respect to the small variants outside of the deletion.

size (Figures 9E–9H and Table 1). Notably, the *PRSS2* variant p.Thr8Ile remained associated with meconium ileus after accounting for the deletion polymorphism in the model (beta = 0.93, p = 0.011). Both deletion and p.Thr8Ile are associated with a reduction in *PRSS2* expression and a higher risk of meconium ileus.

Discussion

Phasing of genetic sequence improves understanding of causal variation at GWAS-associated loci, especially in regions of complex genetic architecture and when allelic heterogeneity is present. However, haplotype reconstruction is typically not a priority when studying disease cohorts following up GWAS-identified loci. Here we demonstrate the benefits of access to phase information when performing epidemiological studies of complex loci. In particular, insights made available though LD structure, genome graphs, and reference panel construction are dependent on accurate phase information. It was therefore discouraging to receive news during this study that 10XG was discontinuing their linked-read sequencing with no intention to make it available through other providers. We hope analogous methods such as Universal Sequencing Technologies TELL-Seq and long-read technology such as PacBio HiFi sequencing and Oxford Nanopore continue to mature to allow the research community continued access to read-based phasing that is cost-effective for population studies.

We evaluated a 389-kb region around *CFTR* to assess the phasing of a gene with clinically relevant complex alleles. Due to the frequency and shared ancestry of the p.Phe508-del haplotype, we found that many individuals were difficult to phase in a single contiguous block. Fortunately, this issue was specific to individuals with homozygous CF causal variants. In contrast, 76% of individuals with heterozygous CF-causing alleles were phased in a single complete 389-kb block. We demonstrated the capability of phasing p.R117H with the nearby poly-T allele, an example of a complex allele with clinical consequence.[50] We have made a Web application available at https://cftbarcode.research.sickkids.ca to enable further investigation of allele relationships.

We also investigated the chr7q35 trypsinogen locus that did not reach genome-wide significance in our largest GWAS of meconium ileus in CF to date.[10] Nonetheless this locus was tantalizing due to the role trypsinogen plays in digestion and the specificity to the pancreas, one of the organs most significantly affected in CF. The architecture

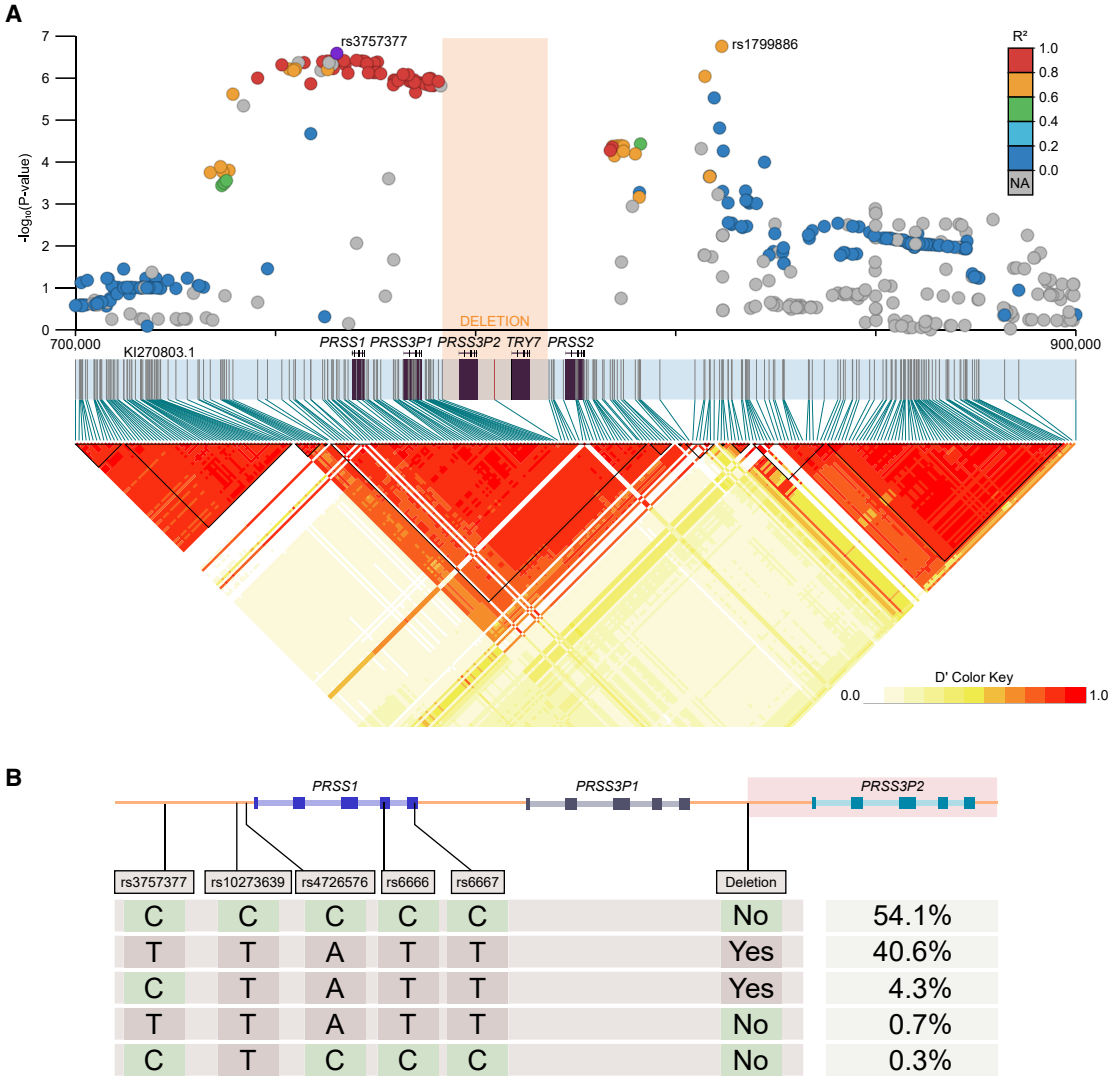

Figure 8. Linkage and haplotype structure at the chr7q35 trypsinogen locus

(A)LD matrix calculated from 10XG phased calls; deletion allele is denoted by orange rectangle. Haplotype blocks are drawn as black triangles, and all five trypsinogen paralogs are located within a single block (KI270803.1:737033–802909). Meconium ileus GWAS summary statistics lifted from GRCh37 to KI270803.1 are shown, R^2 with respect to rs3757377.

(B)Four SNPs in the same LD block as rs3757377, phased with the deletion polymorphism. SNPs include a common pancreatitis risk allele (rs10273639),[18] a *PRSS1* promoter SNP (rs4726576) that alters expression of a reporter gene in mice,[19] and two synonymous *PRSS1* variants (rs6666 and rs6667). Five unique haplotypes are observed in 10XG data, and the frequencies are shown as a percentage. The two major haplotypes account for 94.7% of the observed data. The two rarest haplotypes (1% of the observed data) were supported by manual inspection in IGV.

of the chr7q35 trypsinogen locus requires careful analytic consideration. The region is heavily susceptible to reference bias, where misleading results may be produced based on the reference assembly used. Reference bias at this locus has had documented clinical consequences, specifically the detection of a pathogenic *PRSS1* variant called from misaligned reads derived from trypsinogen pseudogenes.[53,54] We mitigated misalignments by using reference sequence KI270803.1, which provides a more complete representation of this locus. The reference bias issues here motivate the general need to transition from linear references to more comprehensive representations, such as graph-based references that can capture and accommodate the range of variation found within a population. The construction of these graphs can also benefit from the read-based phasing made available through technologies such as linked reads, as demonstrated by the *CFTR* graph we present here. Future areas of focus include using the phased data to study other CF modifier genes, such as *SLC26A9*, *SLC6A14*, and *SLC9A3*.[9,10]

The chr7q35 trypsinogen locus, and *PRSS1* in particular, is well studied in the context of non-CF pancreatitis. An amino acid substitution in *PRSS1* (p.R122H) is the most common cause of hereditary pancreatitis in Europeans.[16] This small change alters a trypsin cleavage site that is important for regulation of trypsin activity through

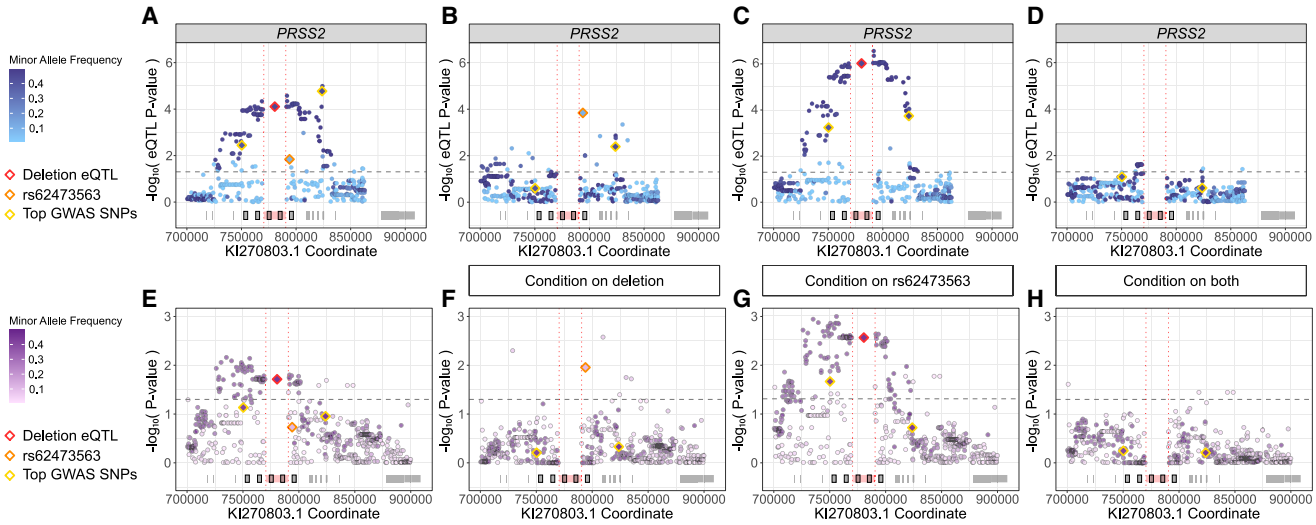

Figure 9. Conditional association analysis reveals common association pattern for *PRSS2* pancreas eQTLs and meconium ileus risk
GTEx v8 variant calls lifted to KI270803.1 and deletion allele (red diamond) imputed from 10XG calls.
(A)Recalculated *PRSS2* pancreas eQTLs.
(B)GTEx *PRSS2* pancreas eQTLs conditioning on deletion polymorphism.
(C)*PRSS2* pancreas eQTLs conditioning on rs62473563 (orange diamond).
(D)GTEx *PRSS2* pancreas eQTLs conditioning on both rs62473563 and deletion polymorphism.
(E)Association with meconium ileus was similarly performed for 309 10XG samples.
(F)Meconium ileus risk conditioning on deletion polymorphism.
(G)Meconium ileus risk conditioning on rs62473563.
(H)Meconium ileus risk conditioning on both rs62473563 and deletion polymorphism.

autoinactivation of trypsinogen.[17] Similarly, chronic pancreatitis has been shown to be associated with a common T>C variant (rs10273639) near *PRSS1*,[18] thought to be associated with altered risk by tagging a promotor SNP (rs4726576) that increases *PRSS1* expression.[19] Increased genetic risk of pancreatitis is typically manifested as increased trypsin activity, by the production of more functional trypsin or greater resistance to degradation via autoinactivation.[55] Despite the depth of evidence supporting a relationship between *PRSS1* and pancreatitis, there is not the same level of support for *PRSS2*. Transgenic human *PRSS2* in mice has been shown to aggravate pancreatitis,[56] and the *PRSS2* variant p.G191R promotes degradation and provides some protection against chronic pancreatitis.[57] This supports the hypothesis that PRSS2 activity may also contribute to pancreatitis risk.

The data presented here suggest a more relevant role for *PRSS2* over *PRSS1* in meconium ileus. We identify two putatively contributing polymorphisms that independently alter meconium ileus risk and *PRSS2* expression: a 20-kb deletion polymorphism and a nonsynonymous variant in exon 1 of *PRSS2*. These polymorphisms are in *cis* with risk variants in two independent meconium ileus-associated SNP clusters, confirming the evidence of allelic heterogeneity seen in our previous meconium ileus GWAS.[10] The deletion polymorphism is in *cis* with the common SNP rs10273639 found to alter non-CF pancreatitis risk.[18] While previous work has suggested a connection between this haplotype and *PRSS1* expression, the results presented in this current work do not implicate *PRSS1* expression as the mechanism. The association between rs10273639 and *PRSS1* expression was initially established using 69 pancreas tissue samples after removal of three outliers.[17] However, the raw data show positive correlation between *PRSS1* and *PRSS2* expression (r^2 = 0.83) and suggestive evidence of an association between rs10273639 and *PRSS2* (p = 0.053; Figure S7). While the data were interpreted to support *PRSS1* expression as a causal explanation, they do not exclude a *PRSS2* contribution. Given the extreme transcriptional activity of this locus in pancreatic cells, it would not be surprising that a structural change caused by the large 20-kb deletion polymorphism upstream of the *PRSS2* promoter could alter *PRSS2* transcription.

A second meconium ileus GWAS association signal is in high LD with the p.Thr8Ile variant in *PRSS2* (rs62473563). When restricted to a European subset, this variant is also the most significant *PRSS2* pancreas eQTL. When conditioning on the deletion polymorphism, p.Thr8Ile showed evidence of increased meconium ileus risk in the 10XG samples, highlighting its independent effect. PRSS2 trypsin operates extracellularly and therefore must be targeted for the endoplasmic reticulum (ER) during translation. The first 15 amino acids contain the sequence specific for binding of the signal recognition particle (SRP) targeting for the ER. An amino acid change here can alter SRP recognition efficiency, which triggers a translation quality control.[58] As p.Thr8Ile is a common variant found in healthy individuals, it does not seem consequential enough to cause a disease phenotype in

Table 1. GTEx *PRSS2* pancreas eQTL analysis and association to meconium ileus risk using 10XG data under conditional analysis.

		PRSS2 pancreas eQTL		Meconium ileus association	
		n = 252		n = 309	
Variant	Conditioned on	Slope (SE)	p value	Beta (SE)	p value
rs62473563	–	−0.24 (0.10)	0.014	0.42 (0.32)	0.19
rs62473563	-kb deletion	−0.38 (0.097)	1.4×10^{-4}	0.93 (0.37)	0.011
20-kb deletion	–	−0.24 (0.060)	7.8×10^{-5}	0.53 (0.22)	0.019
20-kb deletion	rs62473563	−0.31 (0.060)	9.5×10^{-7}	0.75 (0.25)	0.0028

isolation, but perhaps it is sufficient to modify severity of phenotypes when found in combination with disease states such as CF.

Non-CF pancreatitis is related to increased trypsin activity, typically attributed to *PRSS1*.[18] For meconium ileus we see the opposite relationship, where more trypsin activity reduces risk and our data suggest this is due to *PRSS2* expression variation. Although there is conflicting evidence of whether *PRSS1* or *PRSS2* is the relevant gene, in both contexts the haplotype with the common deletion polymorphism is associated with lower levels of trypsinogen. Similarly, the presence of p.Thr8Ile is associated with lower *PRSS2* expression and higher meconium ileus risk; the effect on non-CF pancreatitis—if any—has not been reported to our knowledge. As meconium ileus is a neonatal intestinal blockage caused by thick and adhesive consistency of the first stool, a simple explanation is that higher trypsin levels in the intestine break down and discourage the formation of this blockage-causing stool, thereby reducing risk. In fact, it is known that the meconium of individuals with CF contain high levels of protein[59] and more active trypsin could provide a protective effect against blockage.

Ethics approval and consent to participate

The Canadian CF Gene Modifier Study (CGMS) was approved by the Research Ethics Board of the Hospital for Sick Children (#0020020214 from 2002 to 2019 and #1000016662 from 2019 to the present) and all participating sub-sites. Written informed consent was obtained from all participants or parents/guardians/substitute decision makers prior to inclusion in the study. The CGMS is approved by the Research Ethics Board of the Hospital for Sick Children for the usage of public and external data.

Consents for publication

Not applicable.

Data and code availability

The CGMS genotype data and 10XG data reported in this study are available at the Canadian CF registry at https://www.cysticfibrosis.ca/our-programs/cf-registry/requesting-canadian-cf-registry-data.

The GTEx RNA-seq data and GTEx v8 variant calls are available at dbGaP: phs000424.v8.p and the GTEx Portal at https://www.gtexportal.org/home/datasets/, respectively. The source code for the CFTbaRcodes tool generated during this study is available at https://github.com/strug-hub/CFTbaRcodes with documentation available at https://cftbarcode.readthedocs.io/en/latest/.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2022.100156.

Acknowledgments

We thank the patients, care providers and clinic research assistants, collaborators, and principal investigators involved in CF Centers throughout Canada for their contributions to the CF Canada Patient Registry and Canadian Gene Modifier Study. The authors wish to acknowledge the staff supporting the High Performance Computing cluster and Research Helpdesk department and The Centre for Applied Genomics at the Hospital for Sick Children, Toronto. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Funding was provided by Cystic Fibrosis Foundation (STRUG17PO); Canadian Institutes of Health Research (FRN 167282), Cystic Fibrosis Canada (2626), and the CFIT Program funded by the SickKids Foundation and CF Canada; and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-03742, 250053-2013). This work was also funded by the Government of Canada through Genome Canada (OGI-148) and supported by a grant from the Government of Ontario. The funders of the study played no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

Conceptualization: L.J.S. Sample recruitment: K.K., J.A., A.H. G.C.M., D.A., S.B., C.B., M.C., A.P., M.P., R.V.W., D.M.C., D.H., M.J.S., N.M., E.T., A.L.S., B.S.Q., P.W., W.M.L., M.S., E.B. Sample processing: F.L., K.K. Data processing: S.M., A.C., J.G., B.T., W.W.L.S., J.W., Z.W., R.V.P., N.P., C.W. Formal analysis: S.M., A.C., J.G. Funding acquisition: L.J.S. Investigation: S.M., L.J.S. Methodology: S.M., A.C. Project administration: L.J.S. Supervision: L.J.S. Visualization: S.M., A.C. Writing-original draft: S.M., A.C., J.G., L.J.S. Writing, review, and editing: all authors.

Declaration of interests

D.M.-C. received an honorarium for teaching module development for Vertex Pharmaceuticals. N.M. is doing contract research trials for Vertex Phaemaceuticals and Abbvie. A.L.S. has received speaking fees for educational programs sponsored by Vertex Pharmaceuticals. B.S.Q. has received speaker fees from Vertex Pharmaceuticals and has served as site principal investigator for several Vertex-sponsored clinical trials. W.M.L. is a study investigator for Vertex Pharmaceuticals. E.T. and F.R. act as consultants for Vertex Pharmaceuticals. M.S. participated in Vertex clinical trials and received payment for education modules. S.M., A.C., J.G., F.L., B.T., W.W.L.S., J.W., Z.W., R.V.P., K.K., A.H., N.P., J.A., C.W., G.C.M., S.B., D.A., E.B., C.B., M.C., A.P., M.P., R.V.W., D.H., M.J.S., E.T., P.W., L.S., F.R., and L.J.S. declare no competing interests.

Received: July 8, 2022
Accepted: October 13, 2022

Web resources

Web application for CFTR haplotypes: https://cftbarcode.research.sickkids.ca
Source code for Web application: https://github.com/strug-hub/CFTbaRcodes
Documentation for Web application: https://cftbarcode.readthedocs.io/en/latest/
KI270803.1 alternative contig: https://www.ncbi.nlm.nih.gov/nuccore/KI270803.1
GTEx analysis pipeline: https://github.com/broadinstitute/gtex-pipeline
Picard tools: https://broadinstitute.github.io/picard
LiftOver tool: https://genome.ucsc.edu/cgi-bin/hgLiftOver
geepack R package: https://cran.r-project.org/web/packages/geepack/index.html
LocusFocus: https://locusfocus.research.sickkids.ca/

References

1. Tewhey, R., Bansal, V., Torkamani, A., Topol, E.J., and Schork, N.J. (2011). The importance of phase information for human genomics. Nat. Rev. Genet. *12*, 215–223.
2. Cystic Fibrosis Genotype-Phenotype Consortium (1993). Correlation between genotype and phenotype in patients with cystic fibrosis. N. Engl. J. Med. *329*, 1308–1313.
3. Rommens, J.M., Iannuzzi, M.C., Kerem, B.S., Drumm, M.L., Melmer, G., Dean, M., Rozmahel, R., Cole, J.L., Kennedy, D., Hidaka, N., et al. (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. Science *245*, 1059–1065.
4. Sosnay, P.R., Siklosi, K.R., Van Goor, F., Kaniecki, K., Yu, H., Sharma, N., Ramalho, A.S., Amaral, M.D., Dorfman, R., Zielenski, J., et al. (2013). Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. Nat. Genet. *45*, 1160–1167.
5. Massie, R.J., Poplawski, N., Wilcken, B., Goldblatt, J., Byrnes, C., and Robertson, C. (2001). Intron-8 polythymidine sequence in australasian individuals with cf mutations r117h and r117c. Eur. Respir. J. *17*, 1195–1200.
6. Strug, L.J., Stephenson, A.L., Panjwani, N., and Harris, A. (2018). Recent advances in developing therapeutics for cystic fibrosis. Hum. Mol. Genet. *27*, R173–R186.
7. Cutting, G.R. (2015). Cutting. Cystic fibrosis genetics: from molecular understanding to clinical application. Nat. Rev. Genet. *16*, 45–56.
8. Dupuis, A., Keenan, K., Ooi, C.Y., Dorfman, R., Sontag, M.K., Naehrlich, L., Castellani, C., Strug, L.J., Rommens, J.M., and Gonska, T. (2016). Prevalence of meconium ileus marks the severity of mutations of the cystic fibrosis transmembrane conductance regulator (cftr) gene. Genet. Med. *18*, 333–340.
9. Sun, L., Rommens, J.M., Corvol, H., Li, W., Li, X., Chiang, T.A., Lin, F., Dorfman, R., Busson, P.F., Parekh, R.V., et al. (2012). Multiple apical plasma membrane constituents are associated with susceptibility to meconium ileus in individuals with cystic fibrosis. Nat. Genet. *44*, 562–569.
10. Gong, J., Wang, F., Xiao, B., Panjwani, N., Lin, F., Keenan, K., Avolio, J., Esmaeili, M., Zhang, L., et al. (2019). Genetic association and transcriptome integration identify contributing genes and tissues at cystic fibrosis modifier loci. PLoS Genet. *15*, e1008007.
11. Ooi, C.Y., and Durie, P.R. (2012). Cystic fibrosis transmembrane conductance regulator (cftr) gene mutations in pancreatitis. J. Cyst. Fibros. *11*, 355–362.
12. Blackman, S.M., Commander, C.W., Watson, C., Arcara, K.M., Strug, L.J., Stonebraker, J.R., Wright, F.A., Rommens, J.M., Sun, L., Pace, R.G., et al. (2013). Genetic modifiers of cystic fibrosis-related diabetes. Diabetes *62*, 3627–3635.
13. Lin, Y.-C., Keenan, K., Gong, J., Panjwani, N., Avolio, J., Lin, F., Adam, D., Barrett, P., Bégin, S., Berthiaume, Y., Bilodeau, L., Bjornson, C., Brusky, J., Burgess, C., Chilvers, M., Consunji-, R., et al. (2021). Cystic fibrosis–related diabetes onset can be predicted using biomarkers measured at birth. Genet. Med. *23*, 927–933.
14. Gibson-Corley, K.N., Meyerholz, D.K., and Engelhardt, J.F. (2016). Pancreatic pathophysiology in cystic fibrosis. J. Pathol. *238*, 311–320.
15. Sontag, M.K., Corey, M., Hokanson, J.E., Marshall, J.A., Sommer, S.S., Zerbe, G.O., and Accurso, F.J. (2006). Genetic and physiologic correlates of longitudinal immunoreactive trypsinogen decline in infants with cystic fibrosis identified through newborn screening. J. Pediatr. *149*, 650–657.
16. Howes, N., Lerch, M.M., Greenhalf, W., Stocken, D.D., Ellis, I., Simon, P., Truninger, K., Ammann, R., Cavallini, G., et al.; European Registry of Hereditary Pancreatitis and Pancreatic Cancer EUROPAC (2004). Clinical and genetic characteristics of hereditary pancreatitis in europe. Clin. Gastroenterol. Hepatol. *2*, 252–261.
17. Whitcomb, D.C., Gorry, M.C., Preston, R.A., Furey, W., Sossenheimer, M.J., Ulrich, C.D., Martin, S.P., Gates, L.K., Amann, S.T., Toskes, P.P., et al. (1996). Hereditary pancreatitis is caused by a mutation in the cationic trypsinogen gene. Nat. Genet. *14*, 141–145.
18. Whitcomb, D.C., LaRusch, J., Krasinskas, A.M., Klei, L., Smith, J.P., Brand, R.E., Neoptolemos, J.P., Lerch, M.M., Tector, M., Sandhu, B.S., et al. (2012). Chang En Yu, and Lei Yu. Common genetic variants in the cldn2 and prss1-prss2 loci alter risk for alcohol-related and sporadic pancreatitis. Nat. Genet. *44*, 1349–1354.

19. Boulling, A., Sato, M., Masson, E., Génin, E., Chen, J.-M., and Férec, C. (2015). Identification of a functional prss1 promoter variant in linkage disequilibrium with the chronic pancreatitis-protecting rs10273639. Gut *64*, 1837–1838.
20. Lee, R., Koop, B.F., and Leroy, H. (1996). The complete 685-kilobase dna sequence of the human β t cell receptor locus. Science *272*, 1755–1762.
21. Chen, J.M., and Ferec, C. (2000). Genes, cloned cdnas, and proteins of human trypsinogens and pancreatitis-associated cationic trypsinogen mutations. Pancreas *21*, 57–62.
22. Wagner, K., Grzybowska, E., Butkiewicz, D., Pamula-Pilat, J., Pekala, W., Tecza, K., Hemminki, K., and Försti, A. (2007). High-throughput genotyping of a common deletion polymorphism disrupting the try6 gene and its association with breast cancer risk. BMC Genet. *8*, 41.
23. McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J., Altshuler, D.M.; and International HapMap Consortium (2006). Common deletion polymorphisms in the human genome. Nat. Genet. *38*, 86–92.
24. Ncbi - homo Sapiens Chromosome 7 Genomic Contig, Grch38 Reference Assembly Alternate Locus Group Alt_ref_loci_1. (2022).
25. Browning, S.R., and Browning, B.L. (2011). Haplotype phasing: existing methods and new developments. Nat. Rev. Genet. *12*, 703–714.
26. Marks, P., Garcia, S., Barrio, A.M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A., et al. (2019). Resolving the full spectrum of human genome variation using linked-reads. Genome Res. *29*, 635–645.
27. Chen, Z., Pham, L., Wu, T.-C., Mo, G., Xia, Y., Chang, P.L., Porter, D., Phan, T., Che, H., Tran, H., et al. (2020). Rob Knight, Pavel Pevzner, Son Pham, Yong Wang, and Ming Lei. Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. Genome Res. *30*, 898–909.
28. Li, H. (2011). A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.
29. Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Garg, S., Markello, C., Lin, M.F., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol. *36*, 875–879.
30. Beyer, W., Novak, A.M., Hickey, G., Chan, J., Tan, V., Paten, B., and Zerbino, D.R. (2019). Sequence tube maps: making graph genomes intuitive to commuters. Bioinformatics *35*, 5318–5320.
31. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.; and 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and samtools. Bioinformatics *25*, 2078–2079.
32. Delaneau, O., Zagury, J.F., Robinson, M.R., Marchini, J.L., and Dermitzakis, E.T. (2019). Accurate, scalable and integrative haplotype estimation. Nat. Commun. *10*, 5436.
33. Dong, S.-S., He, W.-M., Ji, J.-J., Zhang, C., Guo, Y., and Yang, T.-L. (2021). Ldblockshow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. Brief. Bioinform. *22*, bbaa227.
34. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
35. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.
36. Conomos, M.P., Miller, M.B., and Thornton, T.A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet. Epidemiol. *39*, 276–293.
37. International HapMap 3 Consortium, Altshuler, D.M., Yu, F., Bonnen, P.E., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Moutsianas, L., Whittaker, P., Gibbs, R.A., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
38. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.
39. Ucsc - Lift Genome Annotations. (2022).
40. Naim, P., Wang, F., Scott, M., Allen, B., Wang, C., He, G., Gong, J., Rommens, J.M., Sun, L., and Lisa, J.S. (2020). Locusfocus: Web-based colocalization for the annotation and functional follow-up of gwas. PLoS Comput. Biol. *16*.
41. Broad institute - picard. (2022).
42. geepack: Generalized estimating equation package (2022).
43. Github - Broadinstitute/gtex-Pipeline. (2022).
44. Adam, F., Diekhans, M., Anne-Maud, F., Rory Baldwin, J., Jungreis, I., Jane, L., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). Gencode reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*.
45. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). Star: ultrafast universal rna-seq aligner. Bioinformatics *29*, 15–21.
46. Zytnicki, M. (2017). mmquant: how to count multi-mapping reads? BMC Bioinf. *18*, 411.
47. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. Genome Biol. *11*, R25.
48. Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T., and Delaneau, O. (2016). Fast and efficient qtl mapper for thousands of molecular phenotypes. Bioinformatics *32*, 1479–1485.
49. Zook, J.M., McDaniel, J., Olson, N.D., Wagner, J., Parikh, H., Heaton, H., Irvine, S.A., Trigg, L., Truty, R., McLean, C.Y., et al. (2019). An open resource for accurately benchmarking small variant and reference calls. Nat. Biotechnol. *37*, 561–566.
50. Laselva, O., Moraes, T.J., He, G., Bartlett, C., Szàrics, I., Ouyang, H., Gunawardena, T.N.A., Strug, L., Bear, C.E., and Gonska, T. (2020). The cftr mutation c.3453g >c (d1152h) confers an anion selectivity defect in primary airway tissue that can be rescued by ivacaftor. J. Pers. Med. *10*, E40.
51. Kumasaka, N., Nakamura, Y., and Kamatani, N. (2010). The textile plot: a new linkage disequilibrium display of

multiple-single nucleotide polymorphism genotype data. PLoS One *5*, e10207.
52. GTEx Consortium (2020). The gtex consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.
53. Weiss, F.U., Laemmerhirt, F., and Lerch, M.M. (2021). Next generation sequencing pitfalls in diagnosing trypsinogen (prss1) mutations in chronic pancreatitis. Gut *70*, 1602–1604.
54. Génin, E., Cooper, D.N., Masson, E., Férec, C., and Chen, J.-M. (2021). Ngs mismapping confounds the clinical interpretation of the prss1 p.ala16val (c.47c>t) variant in chronic pancreatitis. Gut *71*, 841–842.
55. Hegyi, E., and Sahin-Tóth, M. (2017). Genetic risk in chronic pancreatitis: the trypsin-dependent pathway. Dig. Dis. Sci. *62*, 1692–1701.
56. Wan, J., Haddock, A., Edenfield, B., Ji, B., and Bi, Y. (2020). Transgenic expression of human prss2 exacerbates pancreatitis in mice. Gut *69*, 2051–2052.
57. Witt, H., Sahin-Tóth, M., Landt, O., Chen, J.M., Kähne, T., Drenth, J.P.H., Kukor, Z., Szepessy, E., Walter, H., Dahm, S., et al. (2006). A degradation-sensitive anionic trypsinogen (prss2) variant protects against chronic pancreatitis. Nat. Genet. *38*, 668–673.
58. Karamyshev, A.L., Patrick, A.E., Karamysheva, Z.N., Griesemer, D.S., Hudson, H., Tjon-Kon-Sang, S., Nilsson, I.M., Otto, H., Liu, Q., Rospert, S., et al. (2014). Inefficient srp interaction with a nascent chain triggers a mrna quality control pathway. Cell *156*, 146–157.
59. Brock, D.J., and Barron, L. (1986). Biochemical analysis of meconium in fetuses presumed to have cystic fibrosis. Prenat. Diagn. *6*, 291–298.