HGGA, Volume *4*

Supplemental information

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

Scott Mastromatteo, Angela Chen, Jiafen Gong, Fan Lin, Bhooma Thiruvahindrapuram, Wilson W.L. Sung, Joe Whitney, Zhuozhi Wang, Rohan V. Patel, Katherine Keenan, Anat Halevy, Naim Panjwani, Julie Avolio, Cheng Wang, Guillaume Côté-Maurais, Stéphanie Bégin, Damien Adam, Emmanuelle Brochiero, Candice Bjornson, Mark Chilvers, April Price, Michael Parkins, Richard van Wylick, Dimas Mateos-Corral, Daniel Hughes, Mary Jane Smith, Nancy Morrison, Elizabeth Tullis, Anne L. Stephenson, Pearce Wilcox, Bradley S. Quon, Winnie M. Leung, Melinda Solomon, Lei Sun, Felix Ratjen, and Lisa J. Strug

Supplementary Figures

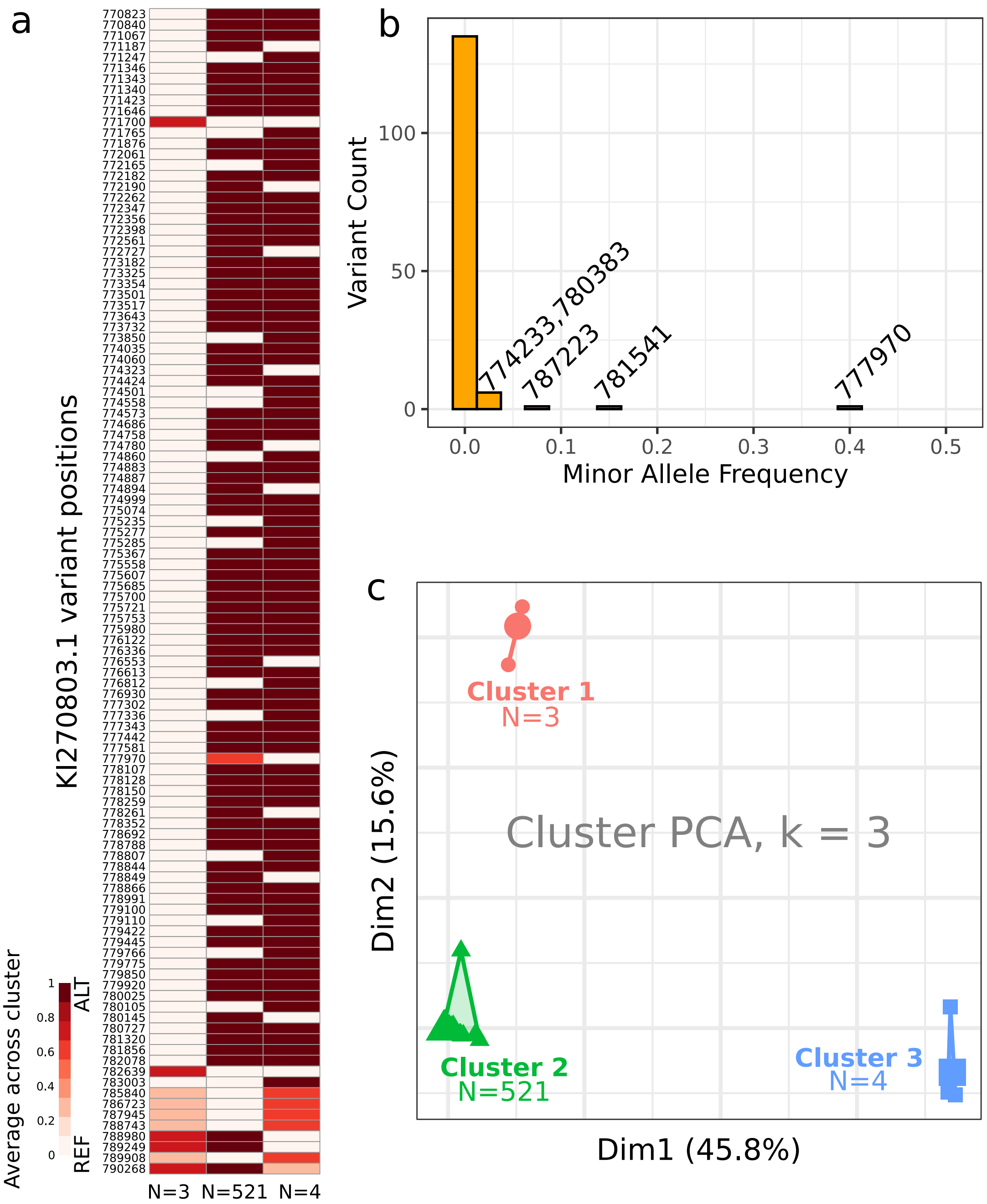


Figure S 1. Clustering non-deleted haplotypes. 144 variants called within the deletion boundary (KI270803.1:770437-790564) were used to cluster 528 non-deleted haplotypes into three groups via k-means clustering. **a** 108 informative variants are plotted; each row is a variant position and columns represents the average haplotype within each cluster. A light cell implies a reference base and a dark cell implies alternative; reference sequence KI270803.1 appears to be most similar to Cluster 3. **b** Distribution of allele frequency for variants in the largest non-deleted haplotype cluster. Allele frequency was calculated for 521 haplotypes. Little variation is observed in this haplotype group, only five variants had minor allele frequency >2.5%. KI270803.1 coordinate positions of these five variants are denoted. **c** The first two principal components are plotted with three distinct clusters. The smaller two clusters (cluster 1, n=3; cluster 3, n=4) correspond to six individuals with African ancestry.

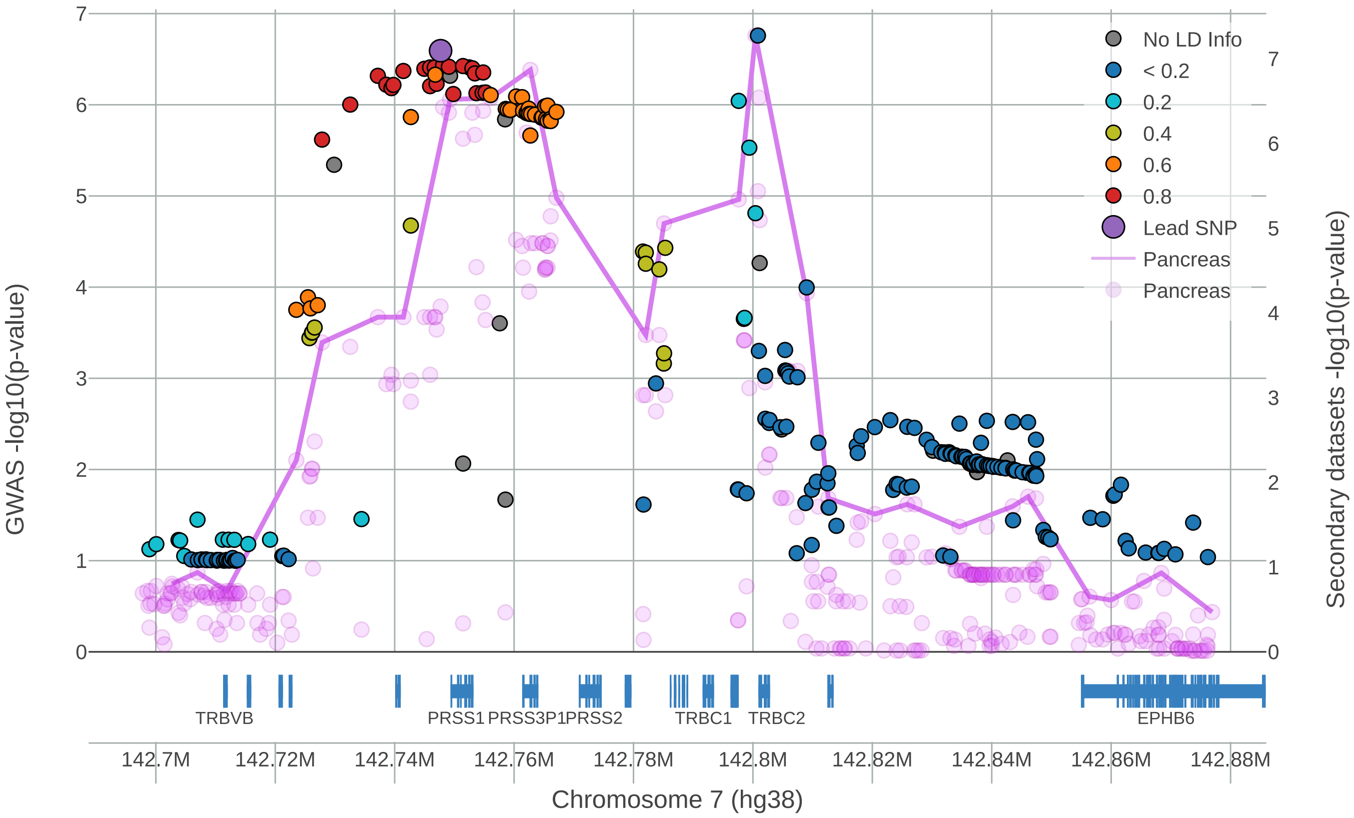
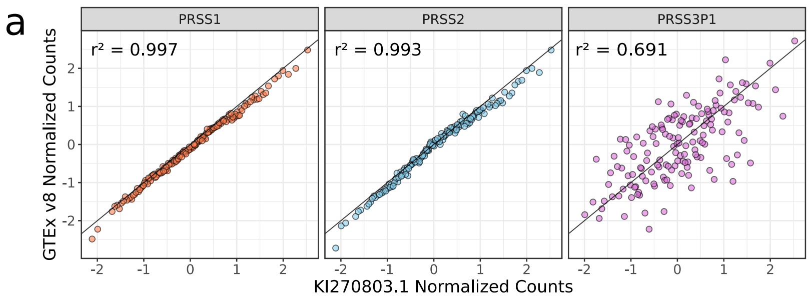


Figure S 2. Colocalization of GWAS and eQTL SNPs. Colocalization of meconium ileus GWAS summary statistics and GTEx v8 (1) pancreas *PRSS2* eQTLs using LocusFocus (2) GWAS summary statistics from (3) lifted to GRCh38. Linkage is shown with respect to rs3757377. Purple line follows the most significant pancreas eQTL in a sliding window. Simple sum colocalization *p*-value=7.1e-8.



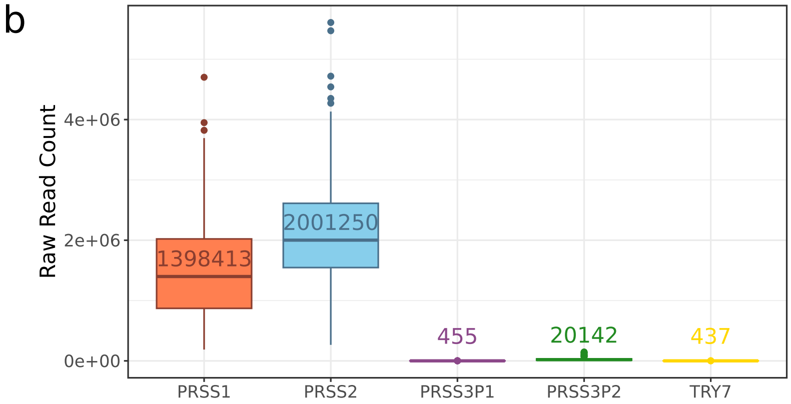


Figure S 3. GTEx v8 read counts using difference reference genomes. GTEx pancreas RNA-seq samples were aligned to a custom reference that replaces the sequence content from GRCh38 chr7 with KI270803.1. **a** Normalized read counts were computed from alignments to this reference (n=152) and compared to GTEx v8 normalized counts that were produced from alignments against GRCh38. For *PRSS1* and *PRSS2*, there is a strong concordance (r² correlation >0.99) between normalized read counts from KI270803.1 and GRCh38. In contrast, *PRSS3P1* displays much less concordance and alignments to *PRSS3P1* are susceptible to the reference sequence used. This is because *PRSS3P1* is not expressed and therefore only receives spurious alignments. Overall, the presence of the extra 20 kb sequence does not significantly shift the normalized gene expression counts for *PRSS1* or *PRSS2* when compared with GTEx v8 counts. **b** Unnormalized RNA-seq read counts to five trypsinogen paralogs after alignment to KI270803.1. Median value for each gene is annotated. *PRSS3P1* and *TRY7* are not expressed but receive a small number of spurious alignments (<1% of total from all five paralogs).

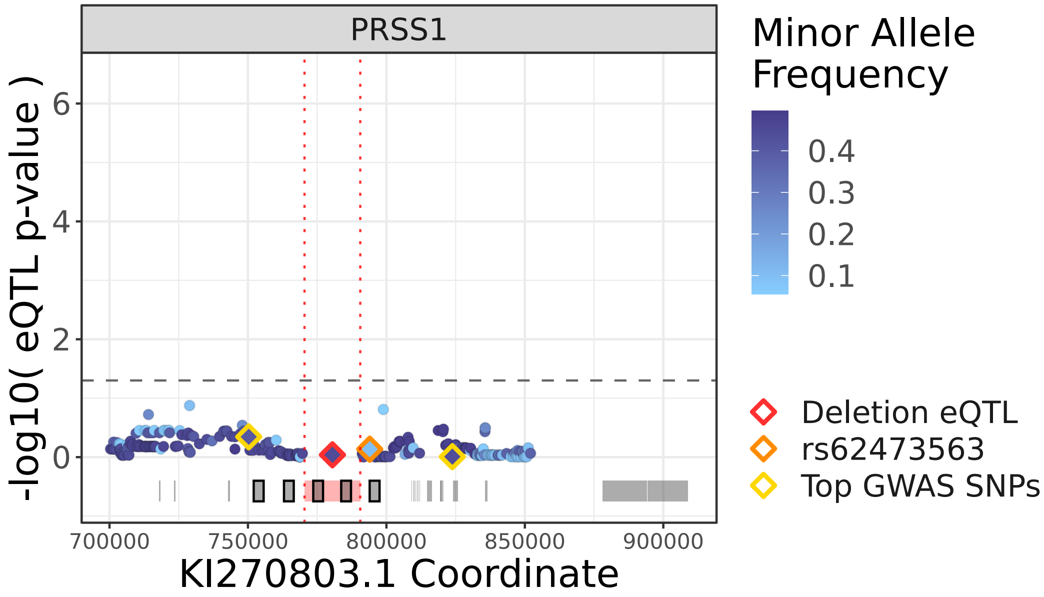


Figure S 4. *PRSS1* pancreas eQTLs. Recalculating *PRSS1* eQTLs from 252 GTEx samples after lifting variants to chr7 alternative contig KI270803.1. No variant passes the significance threshold p<0.05 (dotted line).

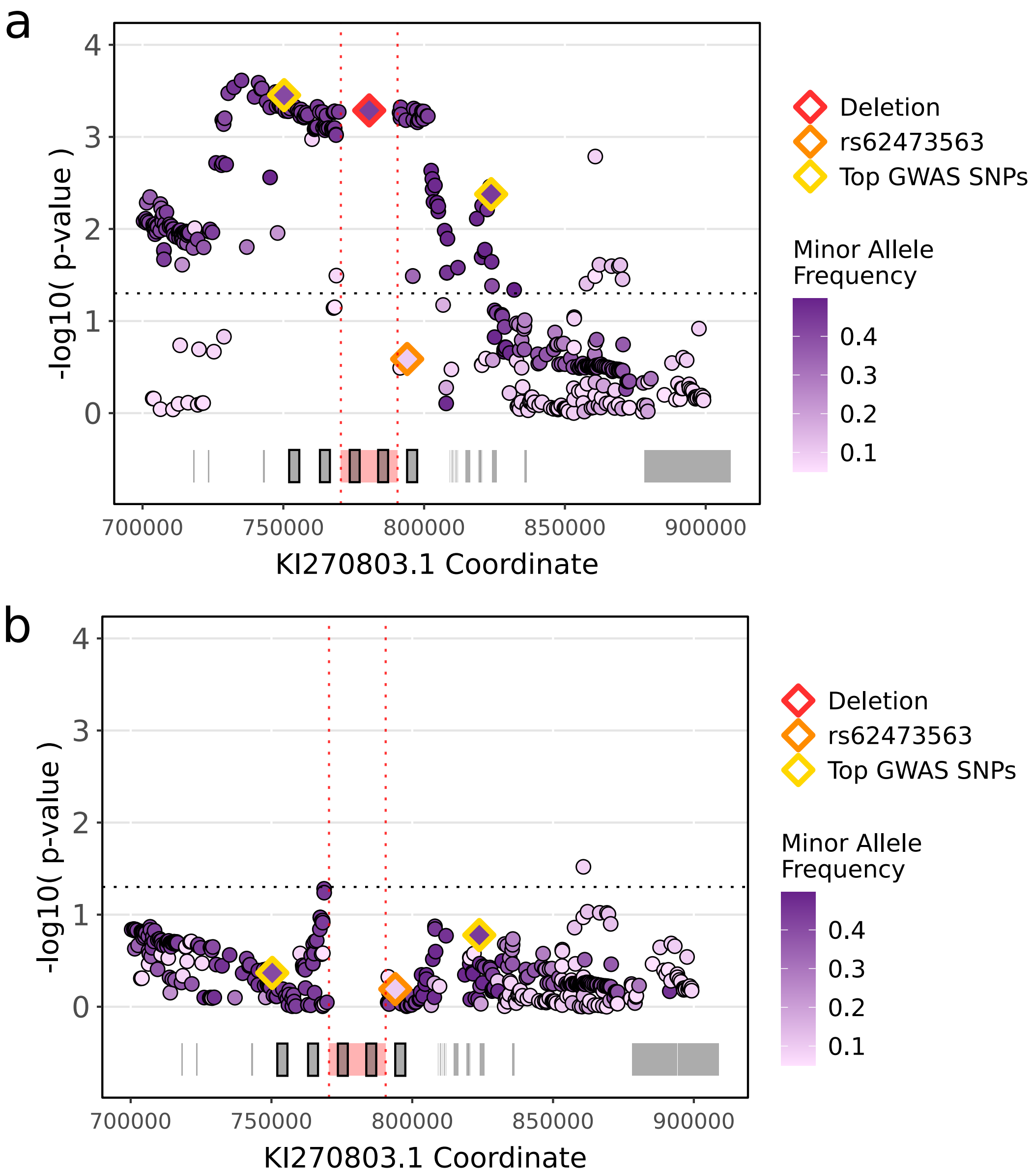


Figure S 5. *PRSS1* pancreas eQTLs. Association to meconium ileus using 10XG-imputed array data. Genotype data for 2635 CCGMS individuals with European ancestry lifted over to KI270803.1. Both the phase and the deletion polymorphism imputed using the 10XG CCGMS samples as a reference panel. Association between SNPs with imputed deletion polymorphism and meconium ileus is plotted. **a** Haplotype that includes the deletion polymorphism (red diamond) demonstrates significant association with an additive increased risk of disease (beta=0.29, p=5.2e-4). top GWAS-identified variants (yellow diamond) are replicated as significant. **b** Recalculating association with the deletion genotype included as a covariate, association signal is attenuated. Imputation quality for rs62473563 is poor. A masked VCF was created by filtering the 10XG sequencing calls to match the information available in each genotyping array. The masked VCF was imputed and compared back to the original sequencing calls. This process reveals that more imputation errors are made than correct calls for rs62473563. This problem is present for the 610Q and 660W array platforms which account for 74% of the samples.

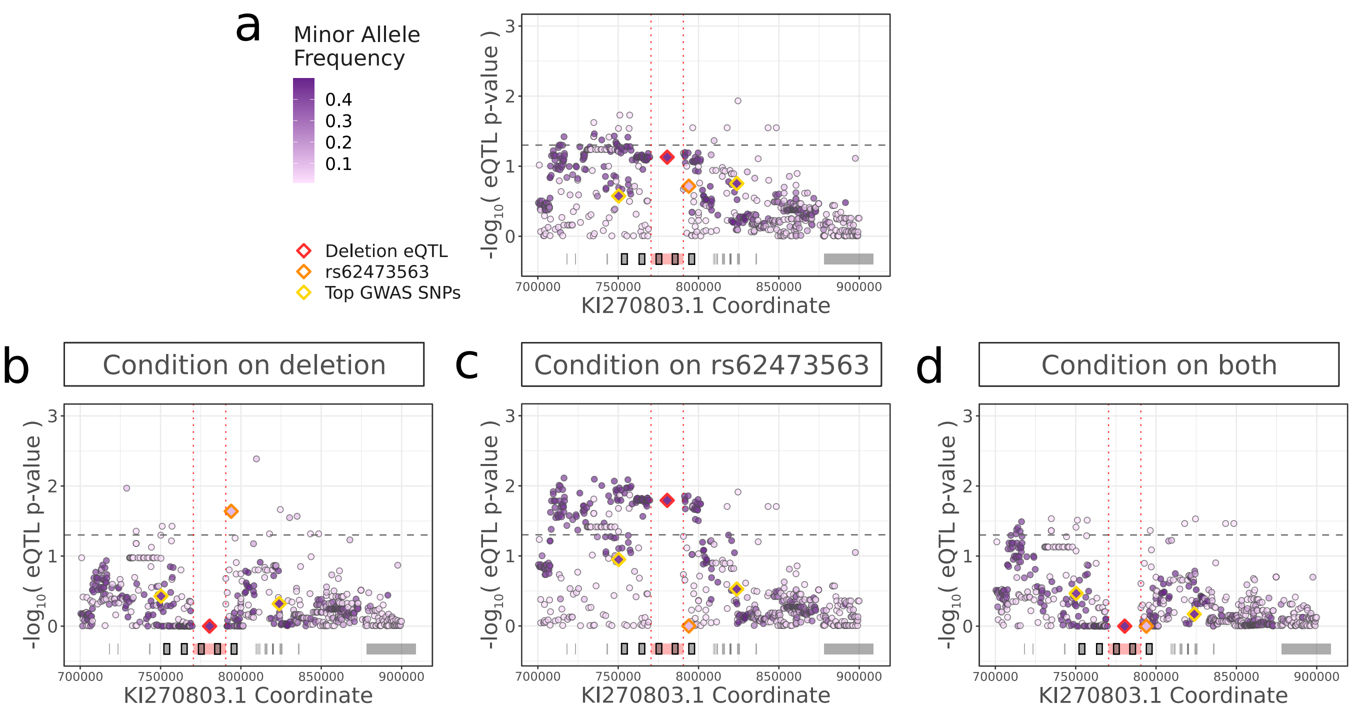


Figure S 6. Conditional association analysis for meconium ileus risk using all available samples. a Association with meconium ileus was similarly performed for 337 10XG samples where yellow diamonds are the top GWAS SNP reported in (3), red diamond is the deletion polymorphism and orange diamond represents rs62473563. **b** Meconium ileus risk conditioning on deletion polymorphism. **c** Meconium ileus risk conditioning on rs62473563. **d** Meconium ileus risk conditioning on both rs62473563 and deletion polymorphism.

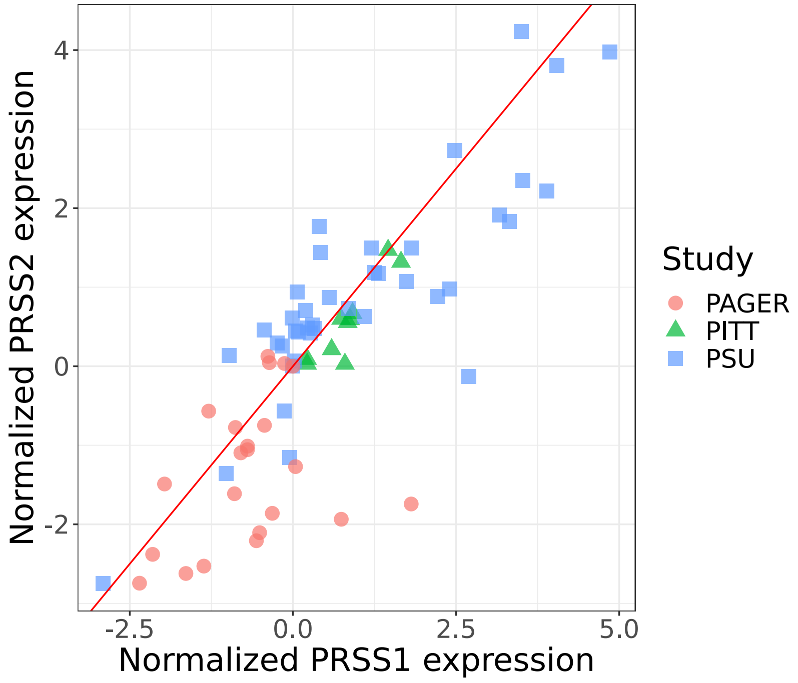


Figure S 7. *PRSS1* and *PRSS2* expression from 69 pancreas tissue samples. Raw data sourced from (4) which uses data from three studies (red circles are from "PAGER" study, green triangle from "PITT" and blue square from "PSU" study). RNA for *PRSS1* and *PRSS2* was quantified and association was demonstrated between rs1027369 and *PRSS1* expression (p=0.01). *PRSS2* expression shows correlation with *PRSS1* expression (r^2=0.83) and regression between rs1027369 and *PRSS2* expression is in same direction as *PRSS1* (p=0.053).

Supplementary Tables

Site Name	Province	City	Clinic	Patients
The Hospital for Sick Children	Ontario	Toronto	Pediatric	190
St. Michael's Hospital	Ontario	Toronto	Adult	61
Children's Hospital of Western Ontario	Ontario	London	Pediatric	24
Centre de recherche du CHUM	Québec	Montréal	Adult	55
Québec (IUCPQ-UL)	Québec	Laval	Adult	45
St. Paul's Hospital	British Columbia	Vancouver	Adult	52
BC Children's Hospital	British Columbia	Vancouver	Pediatric	23
Foothills Medical Centre	Alberta	Calgary	Adult	9
Alberta Children's Hospital	Alberta	Calgary	Pediatric	2
University of Alberta Hospital	Alberta	Edmonton	Adult	2
IWK Health Centre	Nova Scotia	Halifax	Pediatric	10
Royal University Hospital	Saskatchewan	Saskatoon	Pediatric	2
Janeway Children's Health & Rehabilitation Centre	Newfoundland	St. John's	Pediatric	2

Table S 1. Recruitment sites of the Canadian participants with cystic fibrosis. Participants were recruited into the study from 13 sites spanning seven Canadian provinces.

Metric	MagAttract mean (*min-max*)	Other methods mean (*min-max*)	NA12878	Source
Linked-reads per molecule	29.5 (*15.0-115.0*)	14.9 (*9.0-25.0*)	45.0	Long Ranger
10XG gems detected (million)	1.6 (*1.3-1.8*)	1.7 (*1.4-1.8*)	1.6	Long Ranger
Mean DNA per gem (kb)	562.6 (*173.8-726.9*)	423.2 (*321.8-535.6*)	414.5	Long Ranger
Mean molecule length (kb)	58.7 (*32.6-95.4*)	18.3 (*11.0-28.6*)	73.4	Long Ranger
Total number of reads (million)	734.1 (*631.4-1272.0*)	890.8 (*743.5-1126.0*)	695.4	Long Ranger
Mapped reads (%)	96.1 (*92.8-98.0*)	94.7 (*92.6-96.2*)	96.1	Long Ranger
Mean coverage	31.3 (*26.4-55.8*)	37.0 (*31.4-47.6*)	29.7	Long Ranger
Zero coverage (%)	0.54 (*0.15-0.98*)	0.49 (*0.16-0.92*)	0.82	Long Ranger
Median insert size (bp)	369.4 (*298.0-452.0*)	388.9 (*375.0-414.0*)	325.0	Long Ranger
PCR duplication (%)	3.1 (*1.9-6.4*)	5.0 (*3.4-7.1*)	2.3	Long Ranger
Genes >100kb phased (%)	98.8 (*95.8-99.5*)	91.3 (*81.1-96.5*)	98.9	Long Ranger
Phased blocks	2445 (*927-5958*)	17072 (*8530-30415*)	2237	WhatsHap
Longest phase block (Mb)	21.4 (*7.5-87.8*)	4.1 (*1.6-6.9*)	33.6	Long Ranger
Phase block N50 (Mb)	4.4 (*1.3-19.3*)	0.5 (*0.2-0.9*)	5.0	Long Ranger
Mean variants per block	1426.3 (*522.3-4415.7*)	215.3 (*93.3-358.1*)	1522.8	WhatsHap
Variants called (millions)	5.7 (*5.3-7.3*)	5.5 (*5.4-5.6*)	6.0	WhatsHap
Heterozygous SNPs (millions)	2.9 (*2.5-3.9*)	2.8 (*2.7-2.9*)	3.0	WhatsHap
Phased SNPs (millions)	2.5 (*2.2-3.5*)	2.5 (*2.2-3.5*)	2.6	WhatsHap
Short deletion calls	4659 (*4172-5319*)	4817 (*4624-5103*)	4528	Long Ranger

Table S 2. Genome-wide metrics for 10XG phasing. Comparison of metrics between CCGMS samples extracted using MagAttract (n=463), other DNA extraction methods (n=14) and publicly available sample NA12878 (5). Values were calculated and reported by either WhatsHap (6) or Long Ranger (7) as specified

Gene Symbol	Variant ID	rsID	p-value	NES
PRSS2	chr7_142770582_A_G_b38	rs2855983	8.80E-08	0.29
PRSS2	chr7_142776167_A_AT_b38	rs1426115328	9.20E-08	0.29
PRSS2	chr7_142776421_A_T_b38	rs2014445	9.20E-08	0.29
PRSS2	chr7_142778093_C_A_b38	rs2734218	9.20E-08	0.29
PRSS2	chr7_142778351_G_A_b38	rs2734219	9.20E-08	0.29
PRSS2	chr7_142762725_G_A_b38	rs3752404	1.40E-07	0.29
PRSS2	chr7_142800425_T_C_b38	rs1800907	5.30E-08	0.28
PRSS2	chr7_142756070_C_A_b38	rs2855972	2.90E-07	0.27
PRSS2	chr7_142748102_A_G_b38	rs9969188	3.70E-07	0.27
PRSS2	chr7_142762093_C_G_b38	rs12534595	7.40E-07	0.27
PRSS2	chr7_142753427_T_C_b38	rs10231771	7.90E-07	0.27
PRSS2	chr7_142779536_A_G_b38	rs2734221	8.50E-07	0.27
PRSS2	chr7_142775307_A_G_b38	rs151340166	1.00E-06	0.27
PRSS2	chr7_142749281_A_C_b38	rs4726576	3.00E-07	0.26
PRSS2	chr7_142754822_C_G_b38	rs4726577	4.10E-07	0.26
PRSS2	chr7_142749077_T_C_b38	rs10273639	4.30E-07	0.26
PRSS2	chr7_142753014_T_C_b38	rs6667	4.30E-07	0.26
PRSS2	chr7_142751439_C_T_b38	rs3857776	8.80E-07	0.26
PRSS2	chr7_142776921_T_C_b38	rs13242405	3.10E-06	0.26
PRSS2	chr7_142801003_T_C_b38	rs1799887	2.90E-07	0.25
PRSS2	chr7_142796922_C_G_b38	rs2071361	1.90E-06	0.25
PRSS2	chr7_142795717_A_G_b38	rs6961499	2.90E-06	0.25
PRSS2	chr7_142767091_C_T_b38	rs10952532	4.40E-06	0.25
PRSS2	chr7_142797624_G_A_b38	rs56352733	4.60E-06	0.25
PRSS2	chr7_142773135_A_G_b38	rs2075544	4.70E-06	0.25
PRSS2	chr7_142789018_T_C_b38	rs2367487	7.10E-06	0.25
PRSS2	chr7_142766134_C_T_b38	rs2886990	7.20E-06	0.25
PRSS2	chr7_142785121_T_C_b38	rs3114486	8.80E-06	0.25
PRSS2	chr7_142768299_C_G_b38	rs2734213	1.00E-05	0.25
PRSS2	chr7_142774654_A_C_b38	rs2855985	9.30E-06	0.24
PRSS2	chr7_142775024_G_A_b38	rs151339640	9.30E-06	0.24
PRSS2	chr7_142776778_A_C_b38	rs2734217	9.30E-06	0.24
PRSS2	chr7_142766103_G_A_b38	rs2367484	1.40E-05	0.24
PRSS2	chr7_142762842_G_C_b38	rs10952531	1.50E-05	0.24
PRSS2	chr7_142763495_G_A_b38	rs56225909	1.50E-05	0.24
PRSS2	chr7_142764751_T_A_b38	rs4726582	1.50E-05	0.24
PRSS2	chr7_142764755_T_A_b38	rs4726583	1.50E-05	0.24
PRSS2	chr7_142800839_T_C_b38	rs1799886	3.70E-06	0.23
PRSS2	chr7_142753685_G_C_b38	rs1811090	1.20E-05	0.23
PRSS2	chr7_142760340_G_C_b38	rs1969595	1.40E-05	0.23
PRSS2	chr7_142761342_A_G_b38	rs13225332	1.60E-05	0.23
PRSS2	chr7_142765588_C_G_b38	rs13229600	1.60E-05	0.23
PRSS2	chr7_142765617_G_A_b38	rs13228878	1.60E-05	0.23
PRSS2	chr7_142768623_A_C_b38	rs2855981	2.20E-05	0.23
PRSS2	chr7_142761494_C_T_b38	rs11765409	2.90E-05	0.23
PRSS2	chr7_142765247_C_T_b38	rs34500324	2.90E-05	0.23
PRSS2	chr7_142765339_C_T_b38	rs4726588	2.90E-05	0.23
PRSS2	chr7_142765565_A_AT_b38	rs71522195	2.90E-05	0.23
PRSS2	chr7_142765615_C_T_b38	rs13229701	2.90E-05	0.23
PRSS2	chr7_142762515_T_C_b38	rs11770572	5.60E-05	0.23
PRSS2	chr7_142801129_G_A_b38	rs1042955	8.00E-06	0.22
PRSS2	chr7_142765189_G_T_b38	rs4726585	3.10E-05	0.22
PRSS2	chr7_142765191_A_T_b38	rs4726586	3.10E-05	0.22
PRSS2	chr7_142779935_C_T_b38	rs2855990	6.00E-05	0.22
PRSS2	chr7_142780026_T_C_b38	rs2734222	6.00E-05	0.22
PRSS2	chr7_142753697_A_G_b38	rs1811091	2.90E-05	0.21
PRSS2	chr7_142746804_A_C_b38	rs11761222	7.30E-05	0.21
PRSS2	chr7_142754721_A_G_b38	rs1985888	7.50E-05	0.21
PRSS2	chr7_142809001_T_C_b38	rs762691	5.80E-05	0.2
PRSS2	chr7_142747676_A_G_b38	rs3757378	8.40E-05	0.2
PRSS2	chr7_142747687_T_C_b38	rs3757377	8.40E-05	0.2

Table S 3. Significant pancreas eQTLs for *PRSS2* reported by GTEx v8 (1). NES=Normalized effect size

References

1. The GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369:1318–1330, 2020.
2. Naim Panjwani, Fan Wang, Scott Mastromatteo, Allen Bao, Cheng Wang, Gengming He, Jiafen Gong, Johanna M. Rommens, Lei Sun, and Lisa J. Strug. Locusfocus: Web-based colocalization for the annotation and functional follow-up of gwas. *PLOS Computational Biology*, 16:1–8, 2020.
3. Jiafen Gong, Fan Wang, Bowei Xiao, Naim Panjwani, Fan Lin, Katherine Keenan, Julie Avolio, Mohsen Esmaeili, Lin Zhang, Gengming He, David Soave, Scott Mastromatteo, Zeynep Baskurt, Sangook Kim, Wanda K. O'Neal, Deepika Polineni, Scott M. Blackman, Harriet Corvol, Garry R. Cutting, Mitchell Drumm, Michael R. Knowles, Johanna M. Rommens, Lei Sun, and Lisa J. Strug. Genetic association and transcriptome integration identify contributing genes and tissues at cystic fibrosis modifier loci. *PLOS Genetics*, 15, 2019.
4. David C. Whitcomb, Jessica LaRusch, Alyssa M. Krasinskas, Lambertus Klei, Jill P. Smith, Randall E. Brand, John P. Neoptolemos, Markus M. Lerch, Matt Tector, Bimaljit S. Sandhu, Nalini M. Guda, Lidiya Orlichenko, Samer Alkaade, Stephen T. Amann, Michelle A. Anderson, John Baillie, Peter A. Banks, Darwin Conwell, Gregory A. Coté, Peter B. Cotton, James DiSario, Lindsay A. Farrer, Chris E. Forsmark, Marianne Johnstone, Timothy B. Gardner, Andres Gelrud, William Greenhalf, Jonathan L. Haines, Douglas J. Hartman, Robert A. Hawes, Christopher Lawrence, Michele Lewis, Julia Mayerle, Richard Mayeux, Nadine M. Melhem, Mary E. Money, Thiruvengadam Muniraj, Georgios I. Papachristou, Margaret A. Pericak-Vance, Joseph Romagnuolo, Gerard D. Schellenberg, Stuart Sherman, Peter Simon, Vijay P. Singh, Adam Slivka, Donna Stolz, Robert Sutton, Frank Ulrich Weiss, C. Mel Wilcox, Narcis Octavian Zarnescu, Stephen R. Wisniewski, Michael R. O'Connell, Michelle L. Kienholz, Kathryn Roeder, M. Michael Barmada, Dhiraj Yadav, Bernie Devlin, Marilyn S. Albert, Roger L. Albin, Liana G. Apostolova, Steven E. Arnold, Clinton T. Baldwin, Robert Barber, Lisa L. Barnes, Thomas G. Beach, Gary W. Beecham, Duane Beekly, David A. Bennett, Eileen H. Bigio, Thomas D. Bird, Deborah Blacker, Adam Boxer, James R. Burke, Joseph D. Buxbaum, Nigel J. Cairns, Laura B. Cantwell, Chuanhai Cao, Regina M. Carney, Steven L. Carroll, Helena C. Chui, David G. Clark, David H. Cribbs, Elizabeth A. Crocco, Carlos Cruchaga, Charles DeCarli, F. Yesim Demirci, Malcolm Dick, Dennis W. Dickson, Ranjan Duara, Nilufer Ertekin-Taner, Kelley M. Faber, Kenneth B. Fallon, Martin R. Farlow, Steven Ferris, Tatiana M. Foroud, Matthew P. Frosch, Douglas R. Galasko, Mary Ganguli, Marla Gearing, Daniel H. Geschwind, Bernardino Ghetti, John R. Gilbert, Sid Gilman, Jonathan D. Glass, Alison M. Goate, Neill R. Graff-Radford, Robert C. Green, John H. Growdon, Hakon Hakonarson, Kara L. Hamilton-Nelson, Ronald L. Hamilton, Lindy E. Harrell, Elizabeth Head, Lawrence S. Honig, Christine M. Hulette, Bradley T. Hyman, Gregory A. Jicha, Lee Way Jin, Gyungah Jun, M. Ilyas Kamboh, Anna Karydas, Jeffrey A. Kaye, Ronald Kim, Edward H. Koo, Neil W. Kowall, Joel H. Kramer, Patricia Kramer, Walter A. Kukul, Frank M. LaFerla, James J. Lah, James B. Leverenz, Allan I. Levey, Ge Li, Chiao Feng Lin, Andrew P. Lieberman, Oscar L. Lopez, Kathryn L. Lunetta, Constantine G. Lyketsos, Wendy J. MacK, Daniel C. Marson, Eden R. Martin, Frank Martiniuk, Deborah C. Mash, Eliezer Masliah, Ann C. McKee, Marsel Mesulam, Bruce L. Miller, Carol A. Miller, Joshua W. Miller, Thomas J. Montine, John C. Morris, Jill R. Murrel, Adam C. Naj, John M. Olichney, Joseph E. Parisi, Elaine Peskind, Ronald C. Petersen, Aimee Pierce, Wayne W. Poon, Huntington Potter, Joseph F. Quinn, Ashok Raj, Murray Raskind, Eric M. Reiman, Barry Reisberg, Christiane Reitz, John M. Ringman, Erik D. Roberson, Howard J. Rosen, Roger N. Rosenberg, Mary Sano, Andrew J. Saykin, Julie A. Schneider, Lon S. Schneider, William W. Seeley, Amanda G. Smith, Joshua A. Sonnen, Salvatore Spina, Robert A. Stern, Rudolph E. Tanzi, John Q. Trojanowski, Juan C. Troncoso, Debby W. Tsuang, Otto Valladares, Vivianna M. Van Deerlin, Linda J. Van Eldik, Badri N. Vardarajan, Harry V. Vinters, Jean Paul Vonsatte, Li San Wang, Sandra Weintraub, Kathleen A. Welsh-Bohmer, Jennifer Williamson, Randall L. Woltjer, Clinton B. Wright, Steven G. Younkin, Chang En Yu, and Lei Yu. Common genetic variants in the cldn2 and prss1-prss2 loci alter risk for alcohol-related and sporadic pancreatitis. *Nature Genetics*, 44:1349–1354, 2012.
5. Justin M. Zook, Jennifer McDaniel, Nathan D. Olson, Justin M. Wagner, Hemang Parikh, Haynes Heaton, Sean A. Irvine, Len Trigg, Rebecca Truty, Cory Y. McLean, Francisco M. De La Vega, Chunlin Xiao, Stephen Sherry, and Marc Salit. An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*, 37:561–566, 2019.
6. Marcel Martin, Murray Patterson, Shilpa Garg, Sarah O. Fischer, Nadia Pisanti, Gunnar W. Klau, Alexander Schöenhuth, and Tobias Marschall. Whatshap: fast and accurate read-based phasing. *bioRxiv*, page 85050, 2016.
7. Patrick Marks, Sarah Garcia, Alvaro Martinez Barrio, Kamila Belhocine, Jorge Bernate, Rajiv Bharadwaj, Keith Bjornson, Claudia Catalanotti, Josh Delaney, Adrian Fehr, Ian T. Fiddes, Brendan Galvin, Haynes Heaton, Jill Herschleb, Christopher Hindson, Esty Holt, Cassandra B. Jabara, Susanna Jett, Nikka Keivanfar, Sofia Kyriazopoulou-Panagiotopoulou, Monkol Lek, Bill Lin, Adam Lowe, Shazia Mahamdallie, Shamoni Maheshwari, Tony Makarewicz, Jamie Marshall, Francesca Meschi, Christopher J. O'Keefe, Heather Ordonez, Pranav Patel, Andrew Price, Ariel Royall, Elise Ruark, Sheila Seal, Michael Schnall-Levin, Preyas Shah, David Stafford, Stephen Williams, Indira Wu, Andrew Wei Xu, Nazneen Rahman, Daniel MacArthur, and Deanna M. Church. Resolving the full spectrum of human genome variation using linked-reads. *Genome Research*, 29:635–645, 2019.