The American Journal of Human Genetics, Volume *102*

Supplemental Data

Profiling and Leveraging Relatedness in a Precision Medicine Cohort of 92,455 Exomes

Jeffrey Staples, Evan K. Maxwell, Nehal Gosalia, Claudia Gonzaga-Jauregui, Christopher Snyder, Alicia Hawes, John Penn, Ricardo Ulloa, Xiaodong Bai, Alexander E. Lopez, Cristopher V. Van Hout, Colm O'Dushlaine, Tanya M. Teslovich, Shane E. McCarthy, Suganthi Balasubramanian, H. Lester Kirchner, Joseph B. Leader, Michael F. Murray, David H. Ledbetter, Alan R. Shuldiner, George D. Yancoupolos, Frederick E. Dewey, David J. Carey, John D. Overton, Aris Baras, Lukas Habegger, and Jeffrey G. Reid

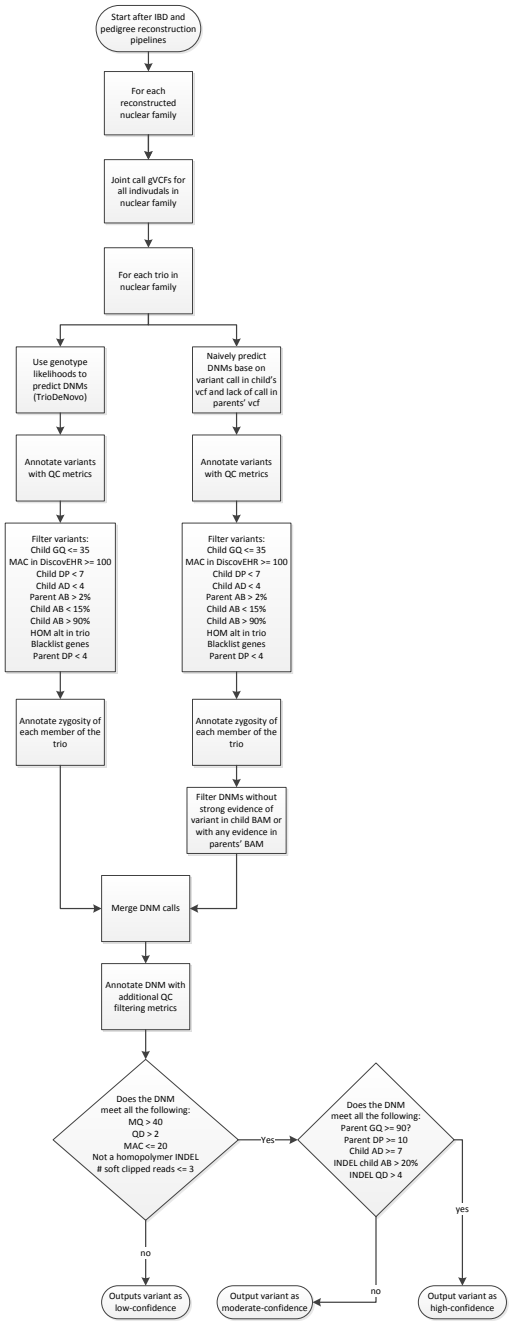

Figure S1. DNM calling, filtering, and confidence ranking workflow. GQ = genotype quality; MAC is minor allele count in DiscovEHR; DP = read depth at the DNM site; AD = the alternate allele depth; AB = alternate allele balance; MQ = mapping quality; QD = quality by depth for the DNM site in the joint called DiscovEHR pVCF; Homopolymer indel is an indel with more than 4 consecutive base pairs of the same nucleotide. Blacklisted genes include PDE4DIP, PRAMEF1, PABPC3, NBPF10, NBPF14, olfactory genes (OR), MUC genes (MUC*), and HLA genes (HLA-*). DNMs were excluded if either parent had a DP < 4, which was effective at filtering out potential DNMs in a child whose parent(s) had no or little read coverage at that DNM site due to being processed with a different capture (e.g. child captured with VCRome and the parent with xGen).*

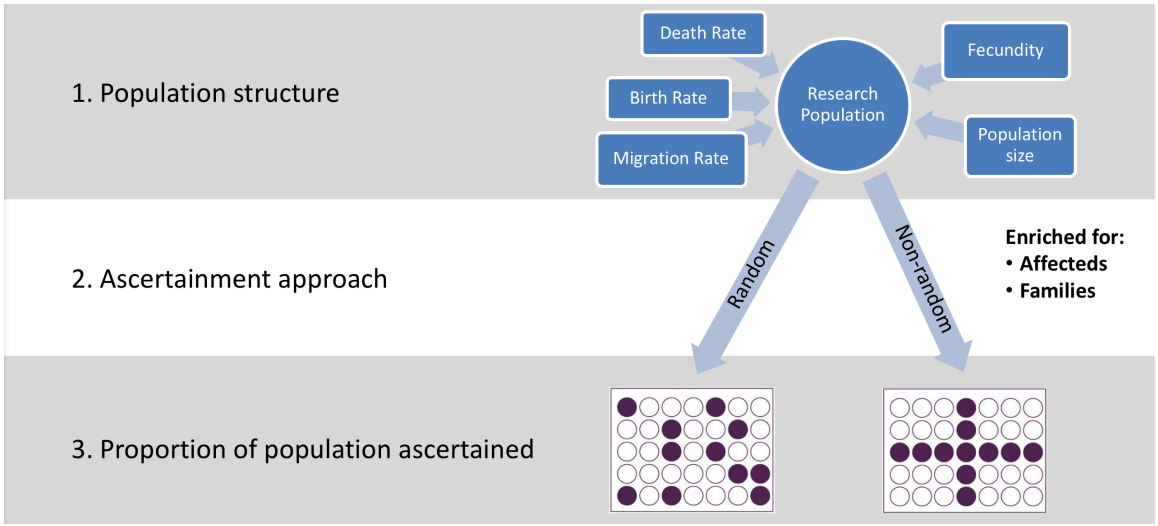

Figure S2. Some of the factors that drive the amount of relatedness in an ascertained dataset modeled by SimProgeny. 1) The populations structure is determined by several parameters that are modeled by SimProgeny. 2) SimProgeny simulates both random and clustered (non-random) ascertainment of the simulated populations. 3) SimProgeny simulates the ascertainment of up to 50% of the simulated population.

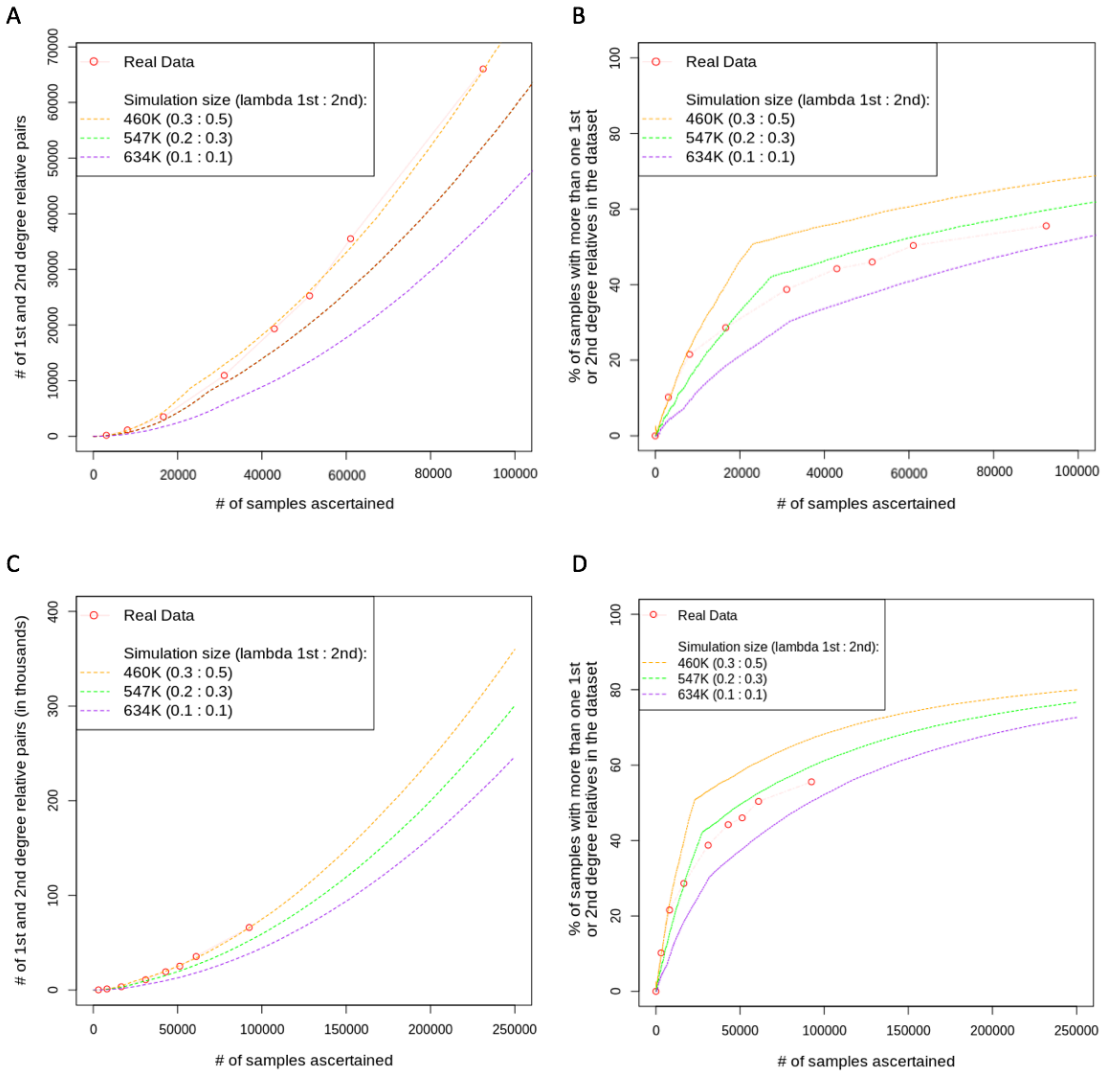

Figure S3. Simulated population and ascertainment fit to the accumulation of first- and second-degree relatedness in the DiscovEHR cohort. The real data was calculated at periodic "freezes" indicated with the punctuation points connected by the faint line. Most simulation parameters were set based on information about the real population demographics and the DiscovEHR ascertainment approach. However, two parameters were unknown and selected based on fit to the real data: 1. the effective population size from which samples were ascertained and 2. the increased chance that someone is ascertained given a first- or second-degree relative previously ascertained, which we call "clustered ascertainment". All panels show the same three simulated population sizes. We simulated clustered ascertainment by randomly ascertaining an individual along with a Poisson distributed random number of 1st degree relatives and a separate random number of 2nd degree relatives. Both Poisson distributions have a lambda indicated in the figure legends. (A) The accumulation of pairs of first- and second-degree relatives as additional samples are ascertained. (B) The proportion of the ascertained participants that have one or more first- and second-degree relatives that have also been ascertained. (C) Simulated ascertainment projections with upper and lower bounds of the number of first- and second-degree relationships we expect with our current DiscovEHR ascertainment approach as we scale to our goal of 250K participants. (D) Simulated projection with upper and lower bounds of the proportion of the ascertained participants that have 1 or more first- or second-degree relatives that have also been ascertained.

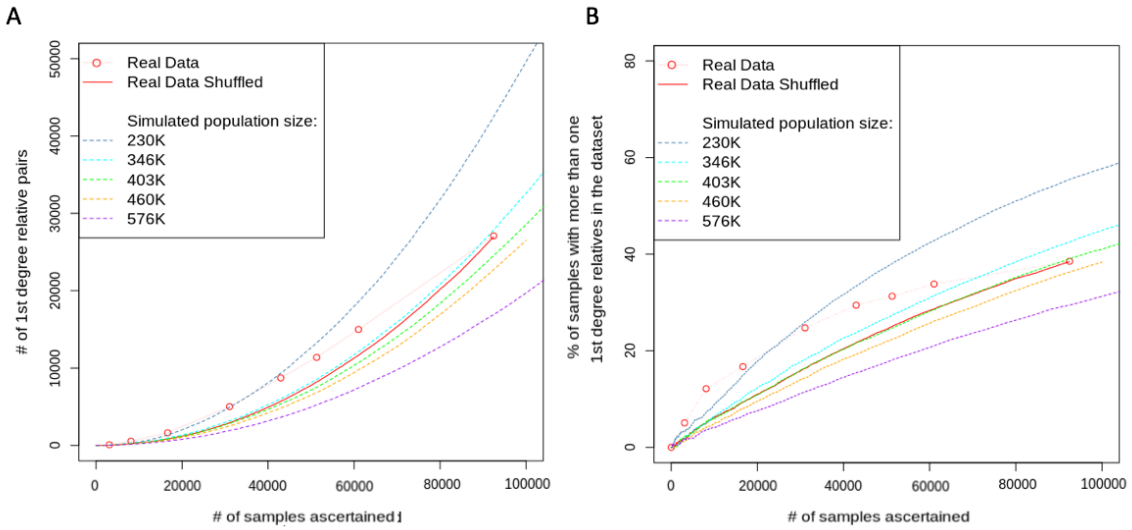

Figure S4. Comparison of the ascertainment of first-degree relatives among 92K DiscovEHR participants compared to random ascertainment of simulated populations. The real data was calculated at periodic "freezes" indicated with the punctuation points connected by the faint line. We also took the samples and relationships identified in the 92K-person freeze and then shuffled the ascertainment order to demonstrate that the first half of the 92K DiscovEHR participants were enriched for first-degree relationships relative the second half. We simulated populations of various sizes using parameters similar to the real population from which DiscovEHR was ascertained. We then perform random ascertainment from each of these populations to see which population size most closely fit the real data. The key takeaway is that none of these population sizes fit the real data and the random ascertainment approach is a poor fit. A different ascertainment approach that enriches for first-degree relatives compared to random ascertainment could produce a better fit. (A) Ascertainment of first-degree relative pairs in an effective sampling population of size 403K closely fit the shuffled version of the real data, but underestimate the # of relative pairs below 92K ascertained participants and dramatically over estimates the number of relative pairs above 92K participants. (B). Similarly, a population of 403K most closely fits the shuffled real data with respect to the number of individuals with one or more first-degree relatives, but is a poor fit to the real data.

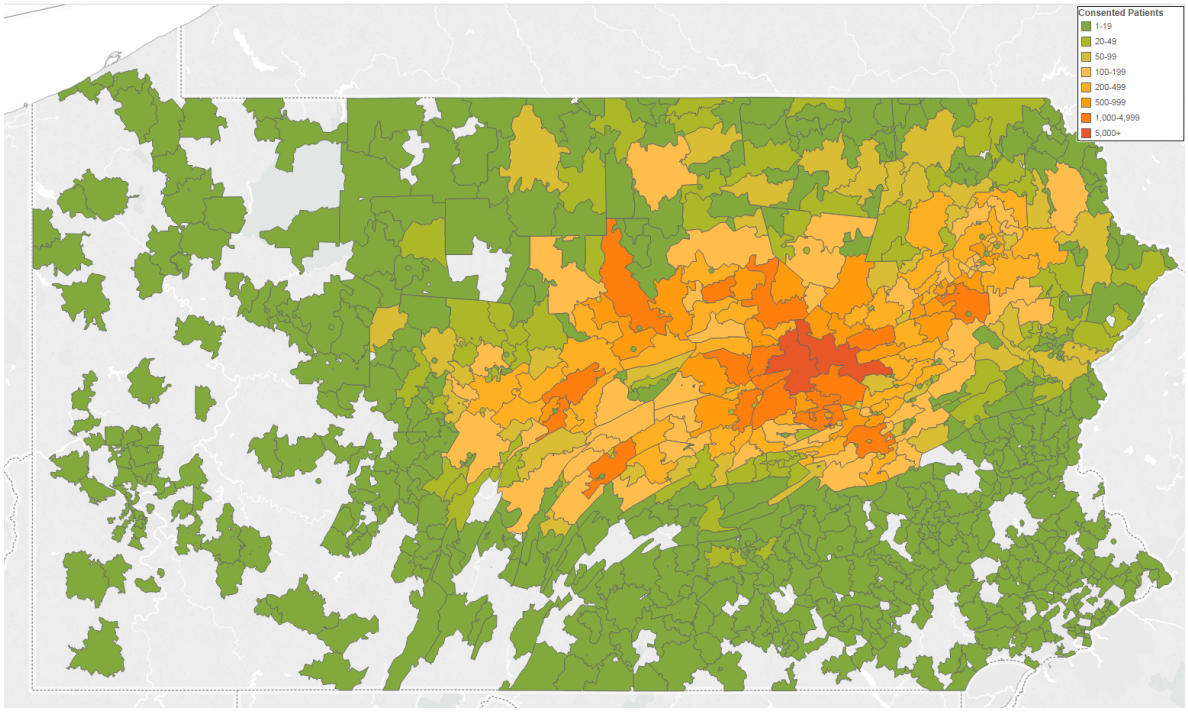

Figure S5. Heat map showing the concentration of where My Code participants live base on zip code. Although the highest concentration is in the areas around Danville (in the middle), there are also pockets in State College (west of Danville), Lewistown (southwest), Wilkes-Barre (northeast) and Shamokin (southeast), and areas in between.

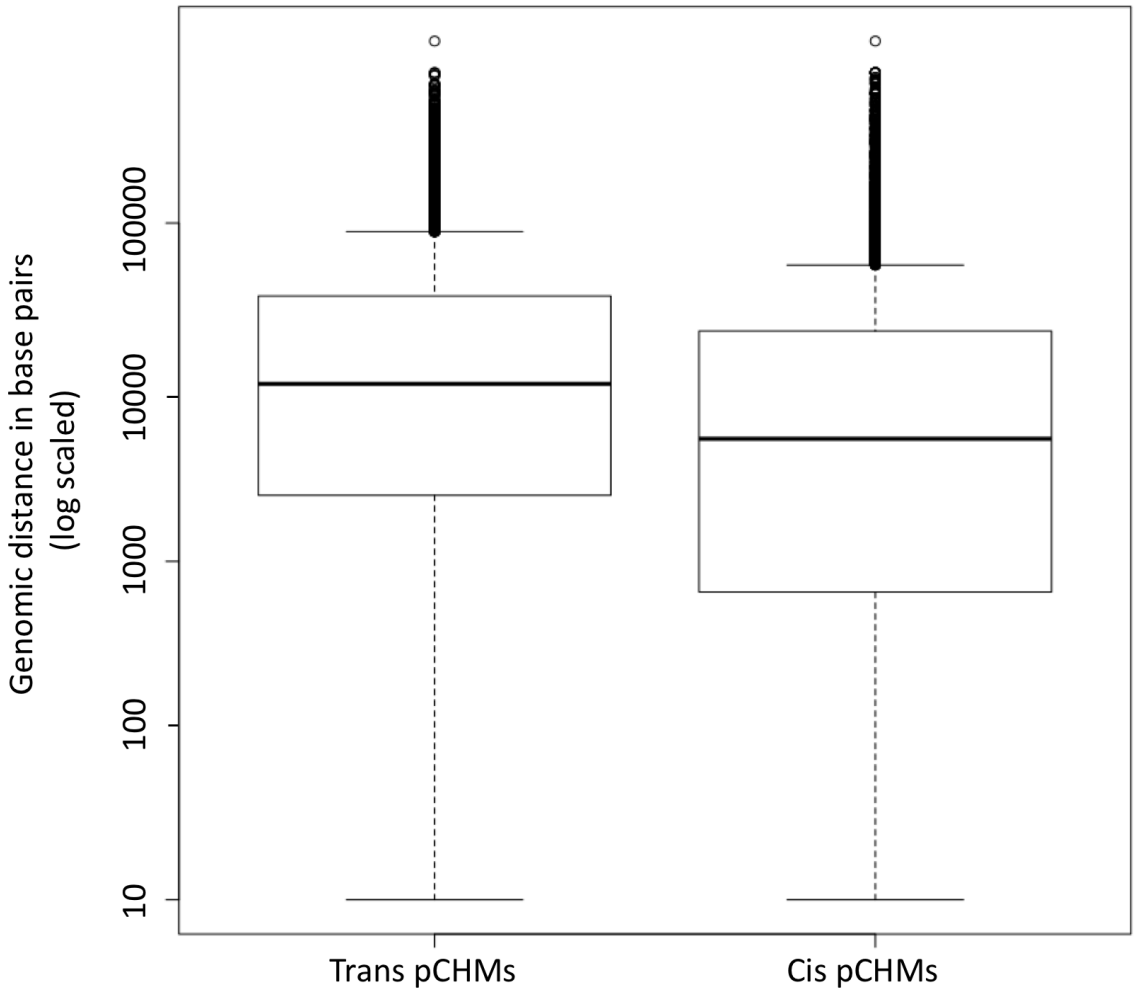

Figure S6. Range of genomic distance between phased pCHM variants showing that both trans and cis pCHMs span the same genomic distance range, but on average, cis pCHMs are closer.

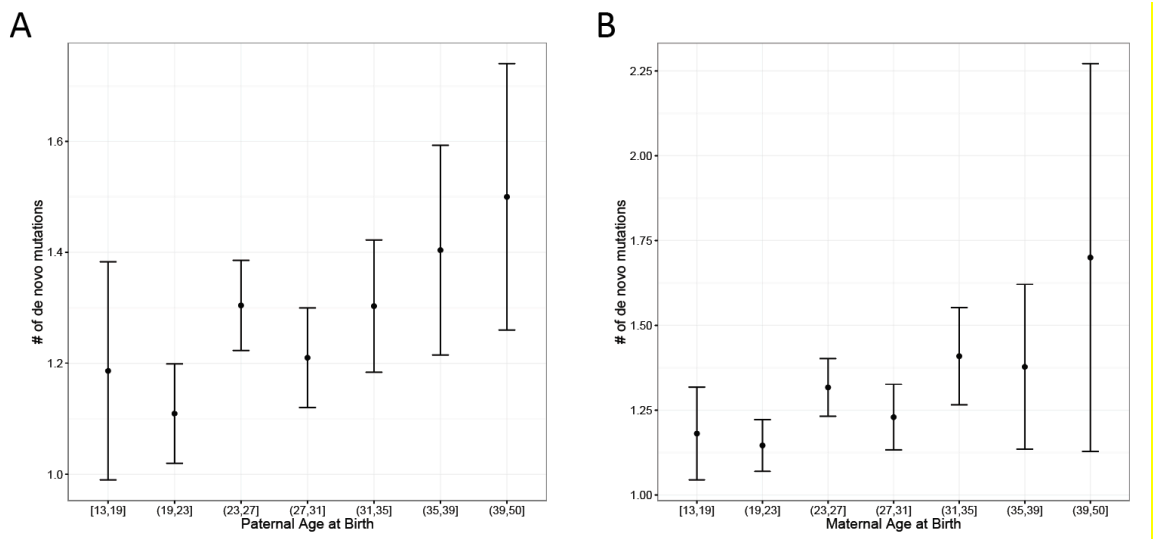

Figure S7. The expected number of exonic DNMs in the child given paternal (A) and maternal (B) age at birth with 95% confidence intervals for each age bin. There is a significant correlation between the number of DNMs in the child and both paternal (0.010 DNMs/year; $p=5.6x10^{-4}$) and maternal (0.011 DNMs/year, $p=7.3x10^{-4}$) age at birth, respectively. Testing for a correlation between parent age at conception and # of DNMs in the child. For this analysis, we excluded 16 samples where proband and parental ages could not be confidently assigned or where more than 10 DNMs were identified, likely indicating technical artifacts or somatic variation. Maternal and paternal age are highly correlated (rho=0.79); when modelled jointly, neither were significant due to collinearity (0.0059 maternal DNMs/year, p=0.29; 0.0063 paternal DNMs/year, p=0.21; Poisson regression). We then tested parental age difference (paternal-maternal age) alongside either maternal or paternal age at birth and, still, both paternal and maternal age were equally predictive of number of DNMs (i.e. age difference was not significantly associated with number of DNMs given maternal or paternal age).

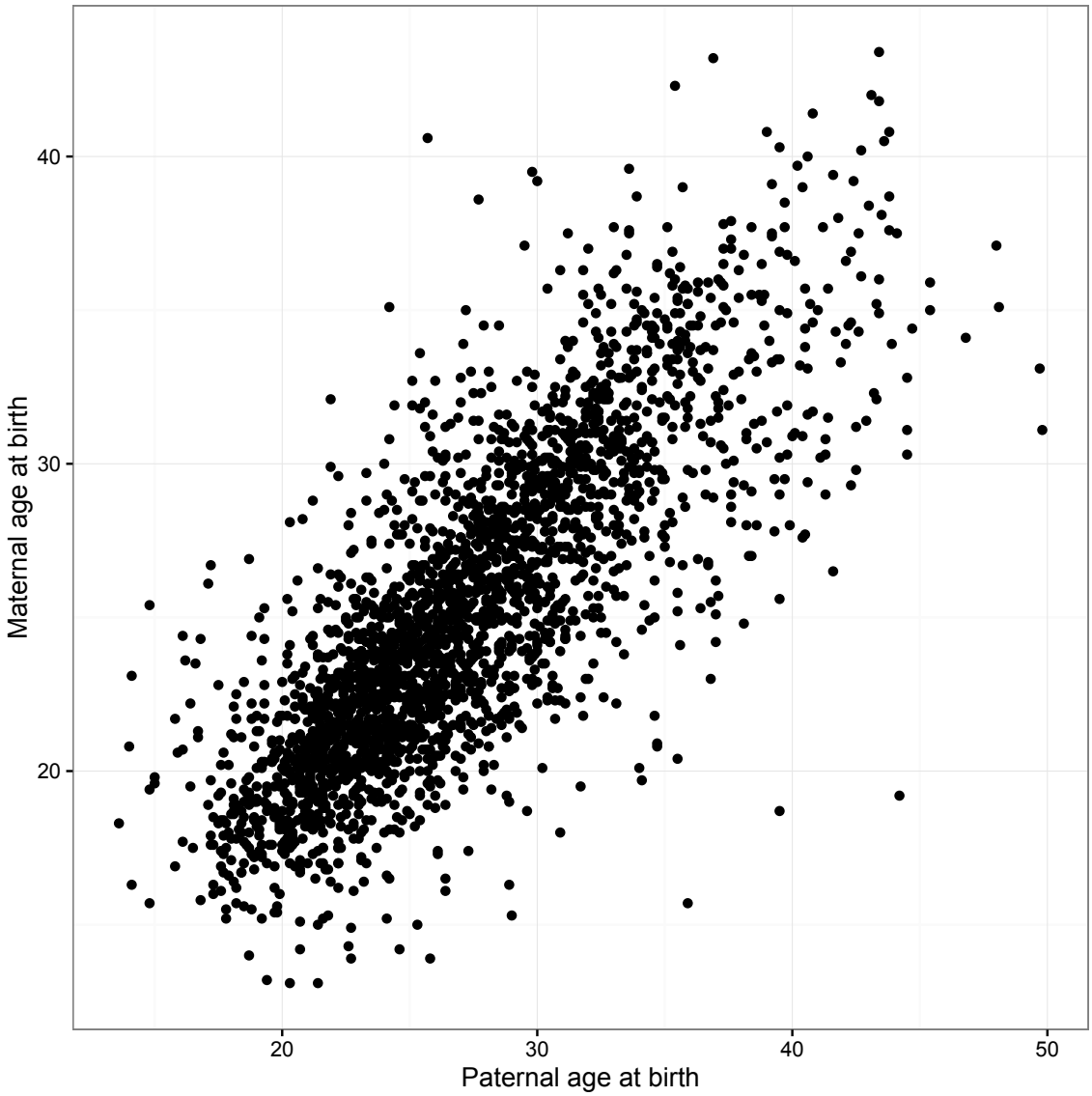

Figure S8. Maternal and Paternal age at birth of child are highly correlated (rho = 0.79). For this analysis, we excluded 16 samples where proband and parental ages could not be confidently assigned or where more than 10 DNMs were identified, likely indicating technical artifacts or somatic variation. Maternal and paternal age are highly correlated (rho=0.79)