Appendix 1 (as supplied by the authors): The Personal Genome Project Canada – findings from whole genome sequences of the inaugural 56 participants

Miriam S Reuter, MD, Susan Walker, PhD, Bhooma Thiruvahindrapuram, MSc, Joe Whitney, MSc, Iris Cohn, MSc, Neal Sondheimer, MD, PhD, Ryan K C Yuen, PhD, Brett Trost, PhD, Tara A Paton, PhD, Sergio L Pereira, PhD, Jo-Anne Herbrick, BSc, Richard F Wintle, PhD, Daniele Merico, PhD, Jennifer Howe, Jeffrey R MacDonald, BSc, Chao Lu, PhD, Thomas Nalpathamkalam, BSc, Wilson W L Sung, MSc, Zhuozhi Wang, PhD, Rohan V Patel, MSc, Giovanna Pellecchia, PhD, John Wei, PhD, Lisa J. Strug, PhD, Sherilyn Bell, BSc, Barbara Kellam, BSc, Melanie M Mahtani, PhD, Anne S Bassett, MD, Yvonne Bombard, PhD, Rosanna Weksberg, MD, PhD, Cheryl Shuman, MS, Ronald D Cohn, MD, Dimitri J Stavropoulos, PhD, Sarah Bowdin, MD, MSc, Matthew R Hildebrandt, PhD, Wei Wei, MSc, Asli Romm, MSc, Peter Pasceri, BSc, James Ellis, PhD, Peter Ray, PhD, M Stephen Meyn, MD, PhD, Nasim Monfared, MSc, S Mohsen Hosseini, MD, PhD, Ann M Joseph-George, PhD, Fred W Keeley, PhD, Ryan A Cook, MBA, BSc, Marc Fiume, PhD, Hin C Lee, PhD, Christian R Marshall, PhD, Jill Davies, MSc, Allison Hazell, MSc, Janet A Buchanan, PhD, Michael J Szego, PhD, Stephen W Scherer, PhD

Supplementary Methods

Dataset and software versions

All databases are referred to GRCh37/hg19 genome build.
Annovar: February 2016 version.
Annovar database for 1000g: August 2015 version.
Annovar database for NHLBI-ESP: December 2014 version.
Annovar database for Exac database: November 2015 version.
Annovar database for “Sorting Intolerant From Tolerant” (SIFT), PolyPhen-2, Mutation assessor, MutationTaster: October 2015 version (dbnsfp).
Annovar database for CADD: Downloaded June 2016.
dbSNP: version 147.
Clinvar: Downloaded from NCBI on July 2016.
Cosmic: version 70.
HGMD: Commercial version, downloaded July 2016.
Gene database (refGene and UCSC genes): January 2017 version.
OMIM: Morbidmap downloaded June 2016.
CGD, HPO, MPO: Downloaded and processed June 2016.
Phastcon: Downloaded from UCSC, April 2014.
SegDups: Downloaded from UCSC, October 2011.
Phylop placental mammal: Downloaded from UCSC, November 2009 version.
Phylop 100 vertebrates: Downloaded from UCSC December 2014 version.
Repeats: Downloaded from UCSC June 2013.
PFAM: Downloaded from UCSC, July 2013 version.

Whole-genome sequencing - pipeline category definitions

(i) Sequence quality

Variants were required to pass Illumina default quality filters (Filter = PASS, with low quality variants being tagged based on the following cut-offs: (a) <10 supporting reads, (b) heterozygous variants with Genotype Quality (GQ) < 100 or alternate allele fraction < 30%, (c) homozygous variants with GQ < 50 or alternate allele fraction < 0.8), excluding no-calls and half-calls (i.e. where only one allele could be called with sufficient confidence). We also performed a manual inspection of the quality of all Insertions/Deletions (indels), variants with <20x coverage, variants with skewed allele fractions (reference allele >67%), and all copy number variants, by inspecting reads from the BAM files for confirmation. Structural variants were validated by Sanger sequencing of the breakpoints (primer sequences are available upon request).

(ii) Allele frequency

Since our analyses on single nucleotide variants (SNVs) and indels focused on rare variants, we considered only variants found at less than 5% allele frequencies in control samples. Overall and population-specific allele frequencies were derived from 1000 Genomes (African, American, East Asian, European, South Asian; http://www.internationalgenome.org/), NHLBI-ESP (African American, European; http://evs.gs.washington.edu/EVS/), and ExAC (African, American, East Asian, Finnish, Non-Finnish Europeans, South Asians, Others; http://exac.broadinstitute.org/). We also used Complete Genomics allele frequencies from the “Wellderly” population (597 Caucasian subjects) (1), the 1000 Genomes (436 subjects), and 54 unrelated subjects from the multi-ethnic reference panel (2). Rare CNVs were defined as those found at less than 1% frequency in the parents of the MSSNG dataset (3), in less than 0.1% in the population from microarray data, and overlapping with a region that is at least 75% copy-number-stable according to the CNV map of the human genome (4).

(iii) Predicted null alleles

We classified as null (loss of function (LoF)) alleles: frameshift insertions, deletions or substitutions, substitutions creating a stop codon, and alterations of the intronic dinucleotide adjacent to a coding-exonic splice junction. We focused on those in genes from the Clinical Genomic Database (CGD; https://research.nhgri.nih.gov/CGD/).

(iv) Disease variant databases

We analyzed for variants with an entry either in the Human Gene Mutation Database (HGMD; http://www.hgmd.cf.ac.uk/ac/index.php) (5), or in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) (6). Categories considered for HGMD were: DM (disease-causing mutation), DM? (likely disease-causing mutation), DFP (disease-associated polymorphism with additional functional evidence), DP (disease-associated polymorphism), or FP (in vitro or in vivo functional polymorphism). Categories considered for ClinVar were: Pathogenic, likely pathogenic, or conflicting interpretations of pathogenicity.

Whole-genome sequencing

We used DNA extracted from whole blood for genomic library preparation and whole-genome sequencing. For Personal Genome Project Canada (PGPC) samples 1-19, 26, and 34-56, library preparation and sequencing were carried out at The Centre for Applied Genomics (TCAG), Toronto, Canada, using the Illumina TruSeq Nano DNA Library Prep Kit with sequencing on the Illumina HiSeq X instrument following the manufacturer's instructions. Briefly, we quantified DNA using the Qubit High Sensitivity Assay and checked sample purity using the Nanodrop OD260/280 ratio. We used either 100 or 500ng of DNA as input material for library preparation, and fragmented to an average of 350bp by sonication on a Covaris LE220 instrument. Fragmented DNA was end-repaired, and A-tailed and indexed Illumina TruSeq adapters added by ligation. For 11 samples (PGPC-03, 07, 08, 10, 34, 36-38, 40, 41, 45), we amplified libraries by PCR prior to sequencing and for 32 samples, we omitted the PCR amplification step to generate "PCR-free" libraries. We assessed and quantified libraries by using a Bioanalyzer DNA High Sensitivity chip and qPCR with Kapa Library Quantification Illumina/ABI Prism Kit protocol (KAPA Biosystems). We pooled validated libraries in equimolar quantities and performed paired-end sequencing on an Illumina HiSeq X platform following Illumina's recommended protocol to generate paired-end reads of 150-bases in length. Samples 1-19 were additionally analyzed by Complete Genomics as described previously (2). For PGPC-samples 20-25 and 27-33, library preparation and sequencing were carried out by Macrogen Inc. using the Illumina TruSeq DNA PCR Free library construction kit with sequencing on the Illumina HiSeq X instrument (at the time, the Illumina HiSeq X instrument was not available at TCAG).

Alignment, variant calling and annotation

We called bases and analyzed data using Illumina HiSeq Analysis Software (HAS) version 2-2.5.55.1311. We mapped reads to the GRCh37/hg19 reference sequence using Isaac alignment software (Isaac alignment software: SAAC00776.15.01.27) and called SNV and small indel variants using the Isaac variant caller (Isaac Variant Caller (Starling): 2.1.4.2). Resulting variant calls were annotated using a custom pipeline developed at TCAG (7, 8) based on ANNOVAR (9). We called CNVs using the read-depth method with the programs ERDS v1.1 (estimation by read depth with single-nucleotide variants) (10) and CNVnator v0.3.2 with a window size of 500 bp (11). We used MANTA v0.23.1 and Canvas v1.1.0.5 to detect structural variants.

Mitochondrial analyses

We used Samtools v0.1.19 to filter reads from the BAM files mapping to mitochondria, and used Picard-tools v2.5.0 to convert the reads to fastq format. We then aligned reads to the mitochondrial b37 genome using BWA v0.7.8. We realigned indels using GATK v3.4.0 and called variants using the mpileup command in SAMtools (12). Using customized scripts, we identified heteroplasmic variants with greater than 5% deviation from a consensus call (while ignoring heteroplasmies at common repeat length polymorphisms; mt.310, mt.514, mt.567, mt.16180).

We additionally assigned mtDNA haplogroups using PhyMer (13), and evaluated the mtDNA sequences for non-haplogroup typical mutations using MITOMASTER (14).

Pharmacogenomics

We selected 391 variants in 14 pharmacogenes, based on guidelines by the Clinical Pharmacogenetics Implementation Consortium, Dutch Pharmacogenetic Working Group, Canadian Pharmacogenomics Network for Drug Safety, and U.S. Food and Drug Administration label recommendations (Table S2) (15). We predicted effects of allelic variation on dosing guidelines and phenotype assignments, using the Human Cytochrome P450 Allele Nomenclature Database (http://www.cypalleles.ki.se/) and the Pharmacogenomics Knowledgebase website (https://www.pharmgkb.org/) (16).

Samples were additionally genotyped with the iPLEX PGx 74 Panel (Agena Biosciences, San Diego, CA, USA) using iPLEX Pro chemistry on the MassARRAY Analyzer 4 System. The panel interrogates 69 SNPs/INDELs in 20 genes, plus 5 CNV assays in CYP2D6. Diplotypes and CNV calling for MassArray data were generated using the Agena PGx Report 2.0 Reporter plugin for the Typer Analyzer software (Agena Bioscience).

Ancestry determination

We determined the ancestry of the probands using data from 1752 unrelated samples from the 1000 Genomes Project as the reference set. The reference samples had been genotyped on Illumina HumanOmni2.5-4v1-B and Illumina HumanOmni25M-8v1-1_B chips (http://www.tcag.ca/tools/1000genomes.html). We extracted the genotypes for a subset of 282,273 positions for case samples with GATK HaplotypeCaller (option –emitRefConfidence BP_RESOLUTION) using the BAM files as input. The variant and non-variant calls for individual samples in gvcf files were combined using GATK CombineGVCFs and the variants were re-genotyped with GATK GenotypeGVCFs using –includeNonVariantSites to retain the homozygous reference calls. The resulting vcf was formatted for analysis using PLINK v1.90b2. We removed the SNPs with genotyping rate <99% in the case samples, both from the case and the reference set, before merging the two sets. Linkage disequilibrium-based pruning of the autosomal SNPs with parameters 50 (window size), 10 (step) and 0.1 (r^2 threshold) yielded 38,309 SNPs for the analysis. We performed population stratification by computing principal component analysis using smartPCA implemented in EIGENSTRAT (17). We then plotted the top three principal components using a custom R script (Figure S2). We also performed model based ancestry estimation using the program ADMIXTURE (Figure S1) (18).

Relationship inference

We used the program KING (19) to confirm the relationships in the provided pedigree and also to detect undeclared relationships between probands. Plots were generated using a custom R script to visualize the results (Table S7).

Runs of homozygosity (ROH) analysis

For analysis of potential uniparental isodisomy, we screened samples for ROHs > 10 kb (20) using PLINK (v1.90) 'Runs of homozygosity' implementation. Genotypes for approximately 2 million positions were extracted, corresponding to the list of pruned SNPs from the Illumina Omni2.5 platform (Roslin NM, Li W, Paterson AD, Strug LJ. Quality control analysis of the 1000 Genomes Project Omni2.5 genotypes, Abstract/Program #576/F, presented at the 66th Annual Meeting of The American Society of Human Genetics, October 18-22, 2016, Vancouver, Canada) (21). A sliding window of 50 SNPs in 5000 kb length was used to scan the genome. To find long segments of homozygosity, we used a linkage-disequilibrium-pruned list of SNPs (generated using window size 50kb, step size 5 and $r^{\wedge 2}$ threshold of 0.2).

Chromosomal microarray analysis

Genomic DNA extracted from whole blood was genotyped on an Affymetrix CytoScan HD platform at The Centre for Applied Genomics. One sample (PGPC-39) failed the data quality control. We performed CNV detection as previously described (22). Briefly, we used four CNV detection algorithms (Affymetrix Chromosome Analysis Suite (ChAS), iPattern, BioDiscovery Nexus, and Partek Genomics Suite) and defined “stringent” CNVs as those spanning five or more probes with a minimum size of 10kb, and found by both ChAS and iPattern, or, if detected by only one of these, also by one of Nexus or Partek. We analyzed for rare CNVs, present at less than 0.1% in population control samples compiled from eight publicly available datasets, totaling 10,851 unrelated individuals: PoPGen (Population-Based Recruitment for Genetics Research; n=1,107) (23) and OHI (Ottawa Heart Institute controls; n=1,224) (24) samples had been genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0 platform. KORA (Cooperative Health Research in the Region of Augsburg; n=1,775) (25) and COGEND (Collaborative Genetic Study of Nicotine Dependence; n=1,109) (26) samples had been genotyped using the OMNI 2.5M quad array platform. SAGE consortium (Study of Addiction: Genetics and Environment; n=1,764) (27), ONC (Ontario Familial Colorectal Cancer Registry; n=433) (28), and HABC (Health, Aging, and Body Composition study; n=2,566) (29) had been genotyped using the Illumina Human1M microarray.

Data Availability

Genomic data can be accessed via https://personalgenomes.ca/.

Table S1 (Pathogenic and risk variants, rare coding CNVs >100 kb) and Table S2 (Pharmacogenomics analyses) are available as separate files in Appendices 2 and 3, respectively.

Table S3. Comparison of rare variant analyses in whole genome sequencing.

	PGPC 2017	**Vassy et al. 2017 (30)**	**McLaughlin et al. 2014 (31)**	**Dewey et al. 2014 (32)**	**Ball et al. 2012 (33)**
Cohort, n	Adult volunteers, 56	Healthy adults, 50	Healthy, 100 or cardiac disease, 100	Adult volunteers, 12	Adult volunteers, 10 (PGP-US)
Technology (coverage)	WGS (38x)	WGS (42x)	WGS (>30x)	WGS	WGS
Rare variant analysis	SNVs, indels, CNVs, inversions in disease genes, mito.	SNVs, indels in disease genes	SNVs, indels in Mendelian disease genes	SNVs, indels, SVs in disease genes	SNVs, indels in disease genes
Variant interpretation	Manual curation (ACMG)	Manual curation (ACMG)	Manual curation	Manual curation	Automated prioritization, review of literature
Yield[a]	25% (potential health impact), 100% (incl. PGx and recessive)	22% (potential monogenic disease risk)	5% (diagnostic), 2% (incidental monogenic)	100% (reportable personal disease risk)	60% (significant phenotypic consequences)
Follow-up	Research reports, clinical interpretation, genetic counselling	Genome reports, clinical interpretation	Genome reports	Genome reports, clinical interpretation, counselling	Genome reports, clinical interpretation

[a] Participants with relevant finding.

ACMG, American College of Medical Genetics and Genomics; indels, insertions/deletions; mito., mitochondrial; NHLBI-ESP, National Heart, Lung, and Blood Institute – Exome Sequencing Project; ND, no data; PGx, pharmacogenomics; SNVs, single nucleotide variants; SVs, structural variants.

Table S4. Whole-genome sequencing coverage statistics per individual.

ID	Median coverage	Coverage 10X	Coverage 20X	Coverage 30X
PGPC-01	40	0.987	0.980	0.912
PGPC-02	47	0.993	0.978	0.931
PGPC-03	40	0.992	0.962	0.888
PGPC-04	33	0.990	0.935	0.678
PGPC-05	31	0.987	0.920	0.565
PGPC-06	41	0.986	0.979	0.932
PGPC-07	45	0.993	0.974	0.924
PGPC-08	37	0.986	0.969	0.805
PGPC-09	39	0.993	0.962	0.869
PGPC-10	39	0.992	0.959	0.870
PGPC-11	31	0.985	0.949	0.564
PGPC-12	38	0.992	0.956	0.834
PGPC-13	41	0.987	0.981	0.927
PGPC-14	41	0.987	0.981	0.930
PGPC-15	29	0.984	0.935	0.487
PGPC-16	34	0.990	0.938	0.704
PGPC-17	35	0.991	0.943	0.740
PGPC-18	34	0.986	0.969	0.752
PGPC-19	34	0.991	0.943	0.741
PGPC-20	34	0.984	0.960	0.745
PGPC-21	33	0.988	0.925	0.676
PGPC-22	39	0.990	0.953	0.842
PGPC-23	35	0.984	0.963	0.772
PGPC-24	35	0.989	0.937	0.749
PGPC-25	37	0.989	0.945	0.794
PGPC-26	35	0.991	0.946	0.765
PGPC-27	36	0.984	0.959	0.793
PGPC-28	35	0.984	0.963	0.767
PGPC-29	36	0.989	0.942	0.778
PGPC-30	36	0.984	0.966	0.796
PGPC-31	34	0.984	0.962	0.748
PGPC-32	38	0.985	0.973	0.864
PGPC-33	32	0.987	0.920	0.641
PGPC-34	46	0.993	0.973	0.906
PGPC-35	31	0.988	0.921	0.589
PGPC-36	39	0.992	0.960	0.878
PGPC-37	46	0.994	0.978	0.929
PGPC-38	45	0.987	0.982	0.955
PGPC-39	38	0.985	0.971	0.850

PGPC-40	37	0.991	0.947	0.797
PGPC-41	38	0.992	0.958	0.862
PGPC-42	55	0.988	0.984	0.979
PGPC-43	49	0.994	0.981	0.936
PGPC-44	48	0.987	0.983	0.972
PGPC-45	41	0.986	0.980	0.932
PGPC-46	38	0.987	0.978	0.878
PGPC-47	41	0.993	0.966	0.891
PGPC-48	35	0.992	0.947	0.779
PGPC-49	34	0.986	0.971	0.749
PGPC-50	38	0.992	0.957	0.845
PGPC-51	49	0.994	0.982	0.936
PGPC-52	37	0.992	0.955	0.831
PGPC-53	40	0.987	0.980	0.909
PGPC-54	37	0.992	0.953	0.829
PGPC-55	39	0.986	0.980	0.906
PGPC-56	38	0.986	0.979	0.885
Average	**38.27**	**0.989**	**0.961**	**0.818**
Median	**38**	**0.988**	**0.962**	**0.838**
Range	**29-55**	**0.984-0.994**	**0.920-0.984**	**0.487-0.979**

From sequence data, on average for all samples, 99% of the genome had at least 10X coverage and 96% at least 20X. Median sequence depth was 38X (range 29-55). Average coverage for mitochondrial DNA ranged from 732X to 2,229X with 100% coverage at >200X for all individuals at all positions.

Table S5. Comparison of CNV detection by WGS and microarray.

Rare, coding CNVs >10 kb*		**Sum (1-38 and 40-56)**	**Median size (kb)**	**Size range (kb)**
All	Deletions	36	32	11 - 1,919
	Duplications	55	48	11 - 863
Not detected by WGS	Deletions	3 of 36	25	24 - 29
	Duplications	0 of 55	-	-
Not detected by microarray	Deletions	4 of 36	11	11 - 16
	Duplications	18 of 55	22	12 - 148

* Not including aneuploidies.

WGS and microarray data of 55 PGPC participants were compared regarding their detection yield of rare, coding CNVs >10 kb. Thirty-six rare deletions and 55 rare duplications were identified by either WGS or microarray. Of those, 3 deletions were identified by microarray only, and 4 deletions were identified by WGS only. No duplications were identified by microarray only, whereas 18 duplications were identified by WGS only.

Table S6. Mitochondrial heteroplasmies.

ID	Position	Ref	Var	Gene	Var level	Prot	Effect	Pat report	Genbank
PGPC-05	14566	A	G	*ND6*	93% A	p.36G	S	no	2.08%
PGPC-06	12875	T	C	*ND5*	71% T	p.162I>T	NS	no	NR
PGPC-13	15323	G	A	*CYTB*	71% G	p.193A>T	NS	no	0.41%
PGPC-14	1201	A	G	*12S*	84% A	NA	unknown	no	NR
PGPC-16	10785	T	A	*ND4*	84% T	p.9I>N	NS	no	NR
PGPC-16	15034	A	G	*CYTB*	58% G	p.96L	S	no	0.04%
PGPC-23	6239	G	A	*CO1*	57% A	p.112L	S	no	0.01%
PGPC-31	13722	A	G	*ND5*	82% G	p.462L	S	no	0.70%
PGPC-31	16286	C	T	*DL*	83% C	NC	NC	no	0.44%
PGPC-32	902	G	C	*12S*	92% G	NA	unknown	no	NR
PGPC-32	1659	T	C	*MTTV*	93% T	NA	T-stem	reported	NR
PGPC-32	3094	G	A	*16S*	69% G	NA	unknown	no	NR
PGPC-34	6932	A	G	*CO1*	84% A	p.343G	S	no	0.09%
PGPC-36	15773	G	A	*CYTB*	89% G	p.343V>M	NS	LHON modifier	0.08%
PGPC-39	16294	C	T	*DL*	66% T	NA	NC	no	common (>5%)
PGPC-42	13032	A	G	*ND5*	54% A	p.232W>C	NS	no	NR
PGPC-43	2129	G	A	*16S*	81% G	NA	unknown	no	NR
PGPC-44	13032	A	G	*ND5*	73% G	p.232W>C	NS	no	NR

All heteroplasmies >5% identified in the PGPC cohort, excluding common heteroplasmies at several known repeat-length polymorphisms in the D-loop. This identified 22 heteroplasmic positions.

A single likely pathogenic heteroplasmic variant was identified: PGPC-32 carried a 7% load of mt.1659C. This variant was previously identified at a much higher heteroplasmic load in a patient with ataxia, developmental delay and elevated CSF lactate (34). The variant impacts the Watson-Crick pairing of the mitochondria-specific tRNA for valine. It was not identified in 32,059 Genbank controls. Although the variant does not apparently cause disease in the individual in this cohort, heteroplasmic mutations can shift between generations, and maternally related-individuals may be at an increased risk for disease.

LHON, Leber's hereditary optic neuropathy; NA, not applicable; NC, non-coding; NR, not reported in Genbank database; NS, non-synonymous; Var, variant; Prot, protein; Pat, patient; Ref, reference; S, synonymous.

Table S7. Mitochondrial haplogroups.

ID	Haplogroup	Haplogroup letter	Confidence
PGPC-01	U4b1b1a	U	0.998
PGPC-02	H1-16189	H	0.996
PGPC-03	H1ar1	H	0.998
PGPC-04	HV-16311	H	0.996
PGPC-05	H7e	H	0.998
PGPC-06	N1b1b1	N	0.997
PGPC-07	H5a1b	H	0.999
PGPC-08	J2b1a2	J	0.996
PGPC-09	H3g1b	H	0.998
PGPC-10	T2f1a1	T	0.996
PGPC-11	H1	H	0.995
PGPC-12	U4b2	U	0.996
PGPC-13	H1c	H	0.995
PGPC-14	H1e1a	H	0.997
PGPC-15	K1c1	K	0.995
PGPC-16	T2b21	T	0.996
PGPC-17	H1e1a	H	0.996
PGPC-18	K1a4b	K	0.994
PGPC-19	V-16298	V	0.998
PGPC-20	H1c3b	H	0.995
PGPC-21	W5a1a	W	0.995
PGPC-22	H1-16239	H	0.997
PGPC-23	H6a1a	H	0.996
PGPC-24	K1a3a	K	0.995
PGPC-25	I2	I	0.997
PGPC-26	X2c1c	X	0.995
PGPC-27	H1i1	H	0.998
PGPC-28	H1e	H	0.996
PGPC-29	X2c1	X	0.998
PGPC-30	J1b4	J	0.994
PGPC-31	T2e	T	0.995
PGPC-32	H1c1	H	0.997
PGPC-33	I2d	I	0.998
PGPC-34	K1a1b1a	K	0.996
PGPC-35	H6a1a	H	0.999
PGPC-36	Y1	Y	0.989
PGPC-37	T2b4a	T	0.995
PGPC-38	K1b1a1a	K	0.997
PGPC-39	J2b1a3	J	0.997
PGPC-40	H5a3a2	H	0.997

PGPC-41	U4b1b1	U	0.995
PGPC-42	H3ap	H	0.995
PGPC-43	H41a	H	0.998
PGPC-44	H3ap	H	0.995
PGPC-45	T1a2	T	0.994
PGPC-46	X2f	X	0.995
PGPC-47	HV0d	H	0.995
PGPC-48	K2a3	K	0.997
PGPC-49	U5a2c3a	U	0.998
PGPC-50	H1by	H	0.996
PGPC-51	U4a2a	U	0.997
PGPC-52	N1b1a4	N	0.995
PGPC-53	J2a1a1	J	0.996
PGPC-54	H10a1b	H	0.998
PGPC-55	H3ag	H	0.995
PGPC-56	X2b4a	X	0.995

The PGPC cohort had predominantly European (H, I, J, K, T, U, V, W, X), West-Eurasian (N), and Northeastern-Asian (Y) haplogroups.

Table S8. Top kinship coefficients.

ID1	ID2	N_SNP	Z0	Phi	HetHet	IBS0	Kinship	Error
PGPC-42	PGPC-44	277212	1	0	0.22	0.00	0.25	1
PGPC-42	PGPC-43	277214	1	0	0.21	0.00	0.25	1
PGPC-09	PGPC-12	277209	1	0	0.20	0.09	0.01	0
PGPC-18	PGPC-26	277210	1	0	0.20	0.10	0.01	0
PGPC-09	PGPC-18	277210	1	0	0.20	0.09	0.01	0
PGPC-09	PGPC-35	277211	1	0	0.20	0.10	0.01	0

We used the program KING (19) to confirm the relationships in the provided pedigree, and to detect undeclared relationships between probands.

Each row above provides information for one pair of individuals. The columns are (http://people.virginia.edu/~wc9c/KING/manual.html):
ID1 and ID2: Individual ID for the first and second individual of the pair.
N_SNP: The number of SNPs that do not have missing genotypes in either individual.
Z0: Probability (identical by descent = 0) as specified by the provided pedigree data.
Phi: Kinship coefficient as specified by the provided pedigree data.
HetHet: Proportion of SNPs with double heterozygotes (e.g., AG and AG).
IBS0: Proportion of SNPs with zero IBS (identical-by-state) (e.g., AA and GG).
Kinship: Estimated kinship coefficient from the SNP data.
Error: Flag indicating differences between the estimated and specified kinship coefficients (1 for error, 0.5 for warning).

PGPC-42 is the daughter of PGPC-43 (father) and PGPC-44 (mother). No other kinship scores were significant. With parental sequences available for PGPC-44, we could designate 71 SNVs/indels as *de novo* (but none were interpreted as disease-causing).

Table S9. Runs of homozygosity >5 Mb (5,000 kb).

ID	Chromosome	Start	End	Size (kb)
PGPC-05	1	48002447	54751900	6,749
PGPC-06	2	178553183	185209227	6,656
PGPC-14	9	65993846	71175685	5,181
PGPC-32	9	65993846	71173293	5,179
PGPC-41	9	65993846	71146533	5,152
PGPC-42	1	95315153	101338482	6,023
PGPC-43	3	175513789	181443422	5,929
PGPC-43	17	15294812	21189598	5,894
PGPC-44	20	20382912	25969265	5,586
PGPC-44	20	29523235	36741604	7,218
PGPC-51	9	65993846	71241017	5,247
PGPC-52	9	65993846	71182471	5,188
PGPC-53	9	65993846	71175409	5,181
PGPC-54	5	17061064	26349520	9,288
PGPC-54	8	46886735	53011930	6,125

Runs of homozygosity (segments of the genome where both inherited copies are identical) were analyzed to identify potential uniparental isodisomy. No runs of homozygosity >10 Mb were identified (20).

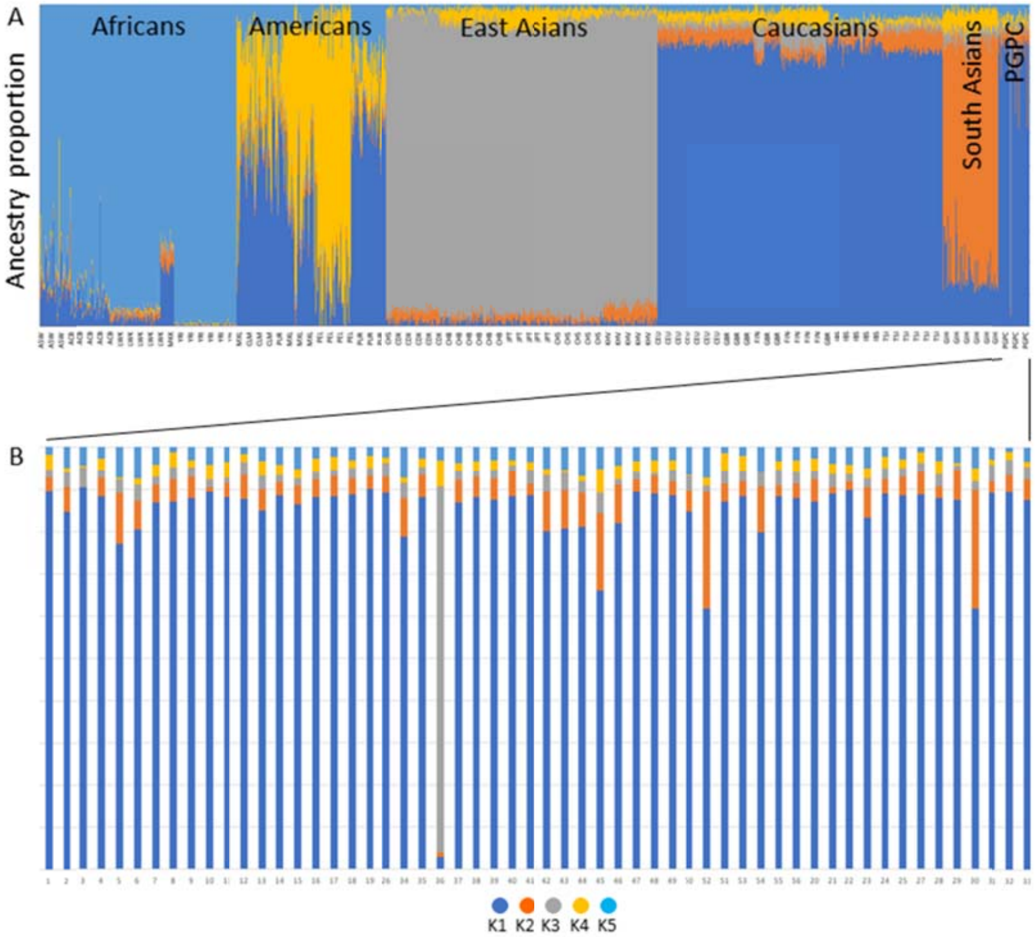

Figure S1. ADMIXTURE analysis. The analysis of WGS also included ancestry clustering (Figure S1 and S2), which enhances self-reported pedigree information (35), and can help to guide return of results and clinical management discussions (e.g. for recessive variants) (36, 37).

Ancestry estimation from genotype data for the 1000 Genomes Project and PGPC samples using the program ADMIXTURE: For K postulated ancestral populations, ADMIXTURE reports the fraction of an individual's genome that originates from each of the ancestral populations (18). The stacked-bar plot shows the ancestry coefficients computed for each sample for K=5. The 1000 Genome Project samples (A) are sorted by continental group (African, American, East Asian, European and South Asian). The bottom panel (B) is a zoomed in view of the PGPC samples.

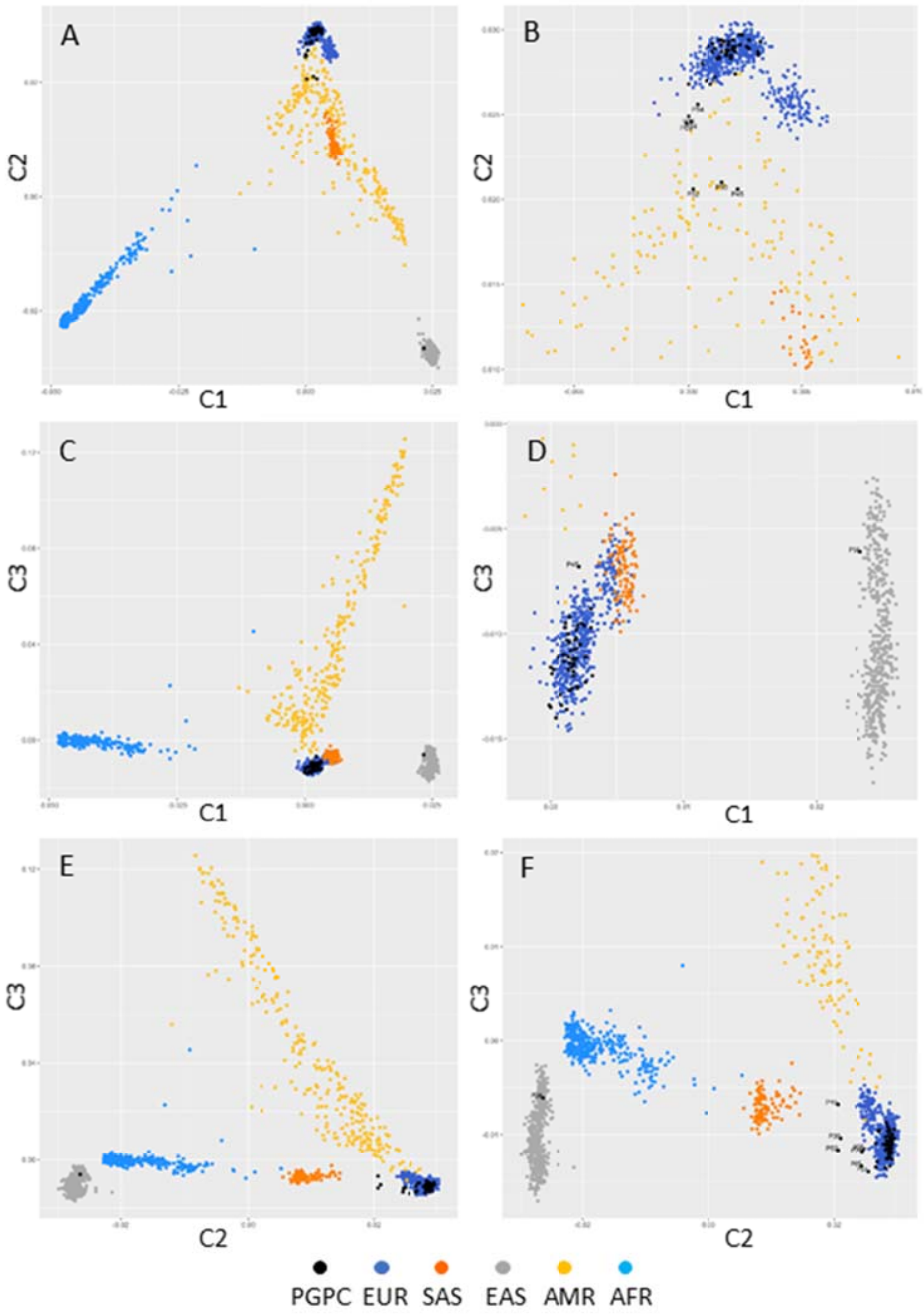

Figure S2. Principal Component Analysis. Principal Component Analysis to detect population structure using smartPCA from the EIGENSOFT package for the 1000 Genomes Project and PGPC samples (17). The value of C1/C2 is plotted along the x-axis and the value of C2/C3 is plotted along the y-axis. Each individual is represented by a dot in the plot. The 1000 Genomes Project samples are colored by the continental groups and the PGPC samples are coloured in black. Panels B, D and F are zoomed in views of A, C and E, respectively.

AFR, African; AMR, American; EAS, East Asian; EUR, European; SAS, South Asian.

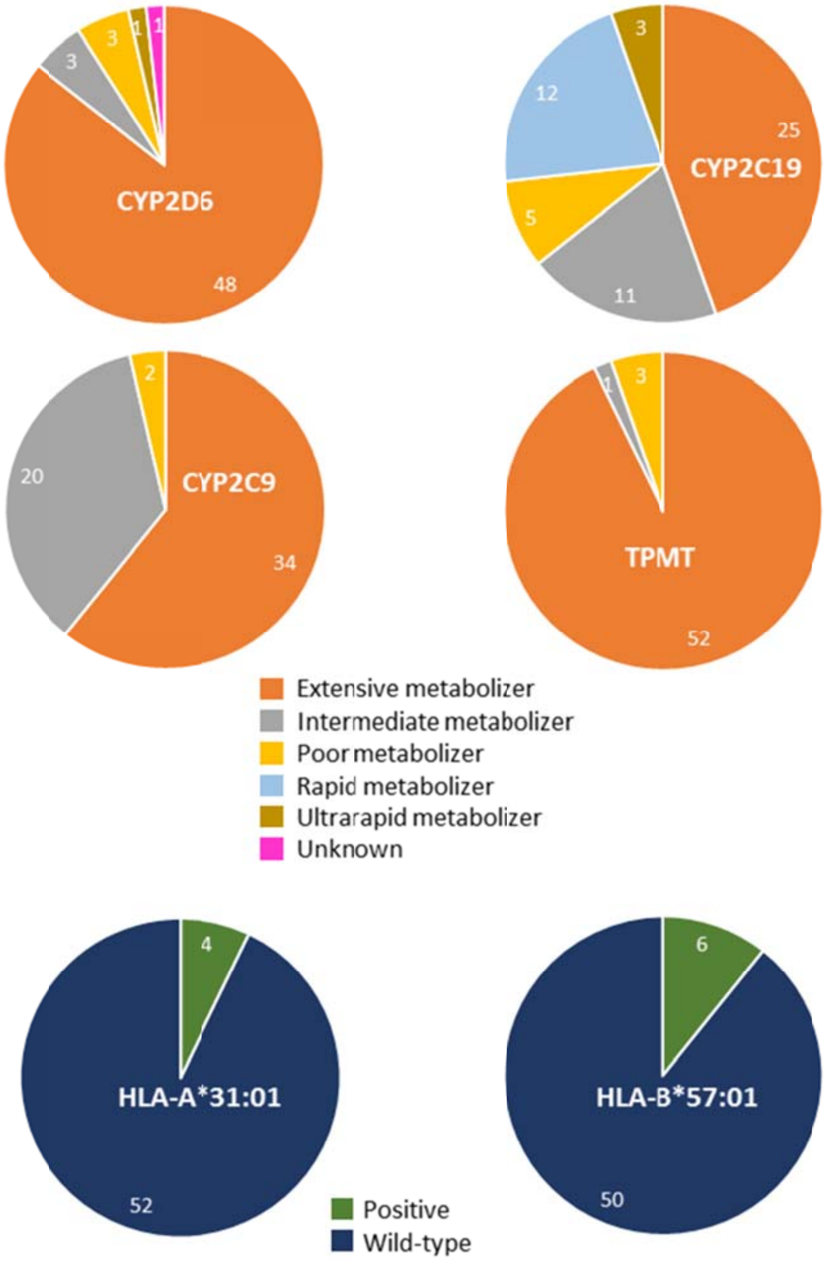

Figure S3. Selected pharmacogenomics results. Distribution of pharmacogenomics (PGx) results for selected pharmacogenes (number of PGPC participants with metabolizer type (orange coloured graphs) or genotype (blue-green coloured graphs), respectively). Additional PGx information is available in Table S2.

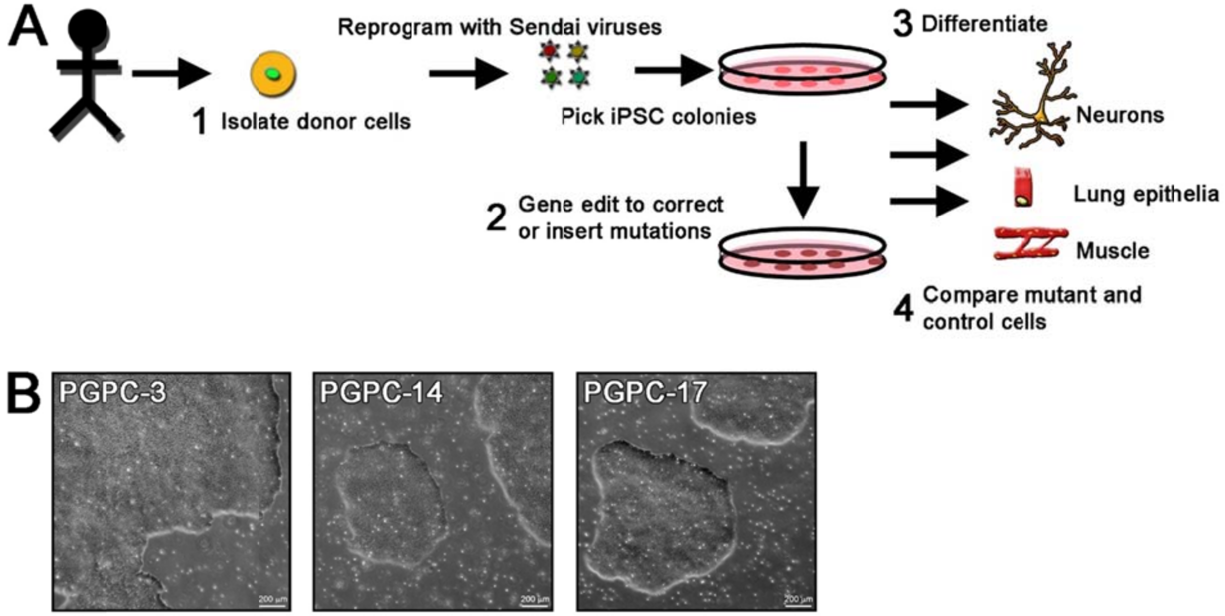

Figure S4. Derivation of induced pluripotent stem cell (iPSC) lines to model disease. For future functional or expression analyses, we generated iPSC lines for two males (PGPC-03, -17) and one female (PGPC-14). These stem cells are part of the data resource generated by the PGPC.

(A) Schematic representation of the generation and utility of donor iPSC lines. 1. Donor blood cells are infected with non-integrating Sendai viruses to deliver the four reprogramming factors OCT4, SOX2, KLF4 and c-MYC to generate iPSC colonies derived from three PGPC individuals. These colonies are picked and expanded as stem cells. 2. For analyses on the impact of genetic variants, stem cells can be gene edited, for example using CRISPR-Cas9, to correct or insert specific genetic variants. 3. IPSC can be differentiated into any cell type of the body including nerve, heart and lung epithelia cells. 4. Mutant and control cells can be compared to understand the impact of specific variants on cell shape and function. Those results in turn can be used as tests for personalized drug discovery.

(B) Representative brightfield microscopy images of iPSC colonies from each of the PGPC donors.

Supplementary Literature

1. Erikson GA, Bodian DL, Rueda M, Molparia B, Scott ER, Scott-Van Zeeland AA, et al. Whole-Genome Sequencing of a Healthy Aging Cohort. Cell. 2016;165(4):1002-11.

2. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010;327(5961):78-81.

3. Yuen RKC, Merico D, Bookman M, J LH, Thiruvahindrapuram B, Patel RV, et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci. 2017;20(4):602-11.

4. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16(3):172-83.

5. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003;21(6):577-81.

6. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44(D1):D862-8.

7. Stavropoulos DJ, Merico D, Jobling R, Bowdin S, Monfared N, Thiruvahindrapuram B, et al. Whole Genome Sequencing Expands Diagnostic Utility and Improves Clinical Management in Pediatric Medicine. NPJ Genom Med. 2016;1.

8. Yuen RK, Thiruvahindrapuram B, Merico D, Walker S, Tammimies K, Hoang N, et al. Whole-genome sequencing of quartet families with autism spectrum disorder. Nat Med. 2015;21(2):185-91.

9. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.

10. Zhu M, Need AC, Han Y, Ge D, Maia JM, Zhu Q, et al. Using ERDS to infer copy-number variants in high-coverage genomes. Am J Hum Genet. 2012;91(3):408-21.

11. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6):974-84.

12. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

13. Navarro-Gomez D, Leipzig J, Shen L, Lott M, Stassen AP, Wallace DC, et al. Phy-Mer: a novel alignment-free and reference-independent mitochondrial haplogroup classifier. Bioinformatics. 2015;31(8):1310-2.

14. Lott MT, Leipzig JN, Derbeneva O, Xie HM, Chalkia D, Sarmady M, et al. mtDNA Variation and Analysis Using Mitomap and Mitomaster. Curr Protoc Bioinformatics. 2013;44:1 23 1-6.

15. Cohn I, Paton TA, Marshall CR, Basran R, Stavropoulos DJ, Ray PN, et al. Genome sequencing as a platform for pharmacogenetic genotyping: a pediatric cohort study. npj Genomic Medicine. 2017;2(1):19.

16. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012;92(4):414-7.

17. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904-9.

18. Zhou H, Alexander D, Lange K. A quasi-Newton acceleration for high-dimensional optimization algorithms. Stat Comput. 2011;21(2):261-73.

19. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867-73.

20. King DA, Fitzgerald TW, Miller R, Canham N, Clayton-Smith J, Johnson D, et al. A novel method for detecting uniparental disomy from trio genotypes identifies a significant excess in children with developmental disorders. Genome Res. 2014;24(4):673-87.
21. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
22. Uddin M, Thiruvahindrapuram B, Walker S, Wang Z, Hu P, Lamoureux S, et al. A high-resolution copy-number variation resource for clinical and population genetics. Genet Med. 2015;17(9):747-52.
23. Krawczak M, Nikolaus S, von Eberstein H, Croucher PJ, El Mokhtari NE, Schreiber S. PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. Community Genet. 2006;9(1):55-61.
24. Stewart AF, Dandona S, Chen L, Assogba O, Belanger M, Ewart G, et al. Kinesin family member 6 variant Trp719Arg does not associate with angiographically defined coronary artery disease in the Ottawa Heart Genomics Study. J Am Coll Cardiol. 2009;53(16):1471-2.
25. Verhoeven VJ, Hysi PG, Wojciechowski R, Fan Q, Guggenheim JA, Hohn R, et al. Genome-wide meta-analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. Nat Genet. 2013;45(3):314-8.
26. Bierut LJ, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau OF, et al. Novel genes identified in a high-density genome wide association study for nicotine dependence. Hum Mol Genet. 2007;16(1):24-35.
27. Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E, et al. A genome-wide association study of alcohol dependence. Proc Natl Acad Sci U S A. 2010;107(11):5082-7.
28. Cotterchio M, Boucher BA, Manno M, Gallinger S, Okey AB, Harper PA. Red meat intake, doneness, polymorphisms in genes that encode carcinogen-metabolizing enzymes, and colorectal cancer risk. Cancer Epidemiol Biomarkers Prev. 2008;17(11):3098-107.
29. Coviello AD, Haring R, Wellons M, Vaidya D, Lehtimaki T, Keildson S, et al. A genome-wide association meta-analysis of circulating sex hormone-binding globulin reveals multiple Loci implicated in sex steroid hormone regulation. PLoS Genet. 2012;8(7):e1002805.
30. Vassy JL, Christensen KD, Schonman EF, Blout CL, Robinson JO, Krier JB, et al. The Impact of Whole-Genome Sequencing on the Primary Care and Outcomes of Healthy Adult Patients: A Pilot Randomized Trial. Ann Intern Med. 2017.
31. McLaughlin HM, Ceyhan-Birsoy O, Christensen KD, Kohane IS, Krier J, Lane WJ, et al. A systematic approach to the reporting of medically relevant findings from whole genome sequencing. BMC Med Genet. 2014;15:134.
32. Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H, et al. Clinical interpretation and implications of whole-genome sequencing. JAMA. 2014;311(10):1035-45.
33. Ball MP, Thakuria JV, Zaranek AW, Clegg T, Rosenbaum AM, Wu X, et al. A public resource facilitating clinical use of genomes. Proc Natl Acad Sci U S A. 2012;109(30):11920-7.
34. Blakely EL, Poulton J, Pike M, Wojnarowska F, Turnbull DM, McFarland R, et al. Childhood neurological presentation of a novel mitochondrial tRNA(Val) gene mutation. J Neurol Sci. 2004;225(1-2):99-103.
35. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007;5(10):e254.
36. Bowdin S, Gilbert A, Bedoukian E, Carew C, Adam MP, Belmont J, et al. Recommendations for the integration of genomics into clinical practice. Genet Med. 2016;18(11):1075-84.
37. Bowdin S, Ray PN, Cohn RD, Meyn MS. The genome clinic: a multidisciplinary approach to assessing the opportunities and challenges of integrating genomic analysis into clinical care. Hum Mutat. 2014;35(5):513-9.

38. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24.