# Overview of the NHLBI Trans-Omics for Precision Medicine (TOPMed) program: whole genome sequencing of >100,000 deeply phenotyped individuals

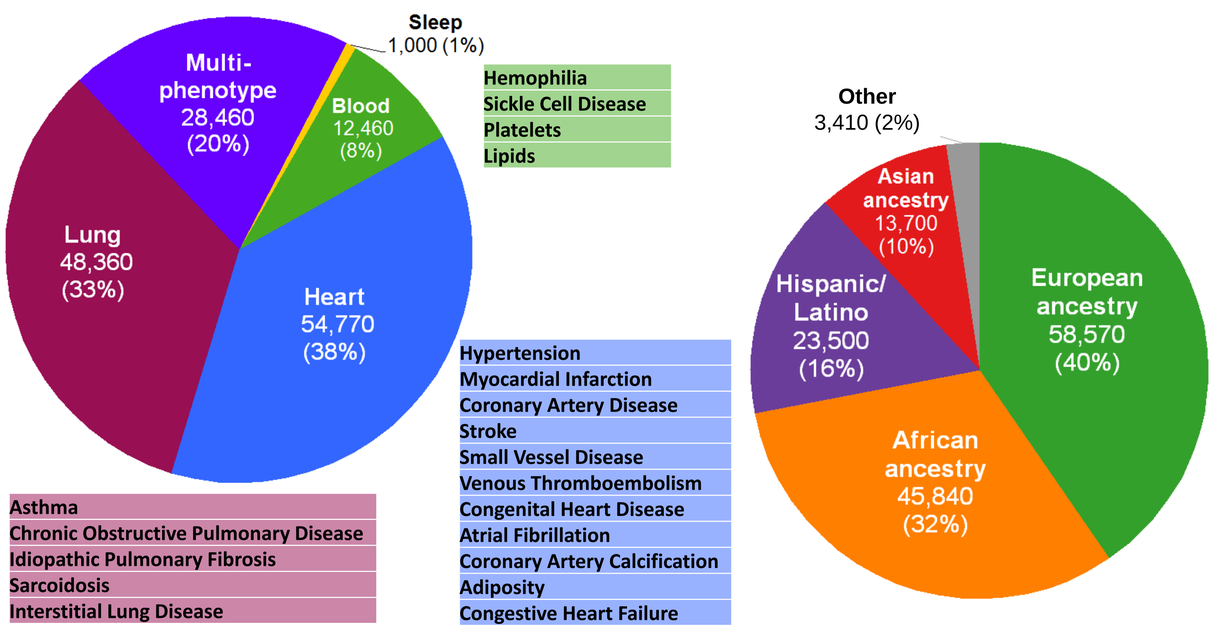
Cathy Laurie[1], Tom Blackwell[2], Goncalo Abecasis[2], Ken Rice[1], James Wilson[3], Deborah Nickerson[4], Stacey Gabriel[5], Richard Gibbs[6], Susan Dutcher[7], Soren Germer[8], Donna Arnett[9], Allison Ashley-Koch[10], Kathleen Barnes[11], Eric Boerwinkle[12], Stephen Rich[13], Edwin Silverman[14], Rebecca Beer[15], Julie Mikulla[15], Pothur Srinivas[15], Weiniu Gan[15], George Papanicolaou[15], Cashell Jaquish[15], and the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

1) Department of Biostatistics, University of Washington, 2) Department of Biostatistics, University of Michigan, 3) Department of Physiology, University of Mississippi, 4) Department of Genome Sciences, University of Washington, 5) Broad Institute, Massachusetts Institute of Technology, 6) Baylor College of Medicine Human Genome Sequencing Center, 7) McDonnell Genome Institute, Washington University in St. Louis, 8) New York Genome Center, 9) Department of Epidemiology, University of Kentucky, 10) Department of Medicine, Duke University Medical Center, 11) Department of Medicine, University of Colorado at Denver, 12) Department of Epidemiology, University of Texas Health at Houston, 13) Center for Public Health Genomics, University of Virginia, 14) Channing Division of Network Medicine, Brigham & Women's Hospital, 15) National Heart, Lung and Blood Institute, NIH

## Study Characteristics

A primary goal of the NHLBI TOPMed program is to improve scientific understanding of the fundamental biological processes that underlie heart, lung, blood, and sleep (HLBS) disorders. TOPMed is providing deep whole genome sequencing (WGS) and other omics data to pre-existing 'parent' studies having large samples of human subjects with rich phenotypic characterization and environmental exposure data.

Sample numbers by phenotype area and ancestry/ethnicity (N=144k total)

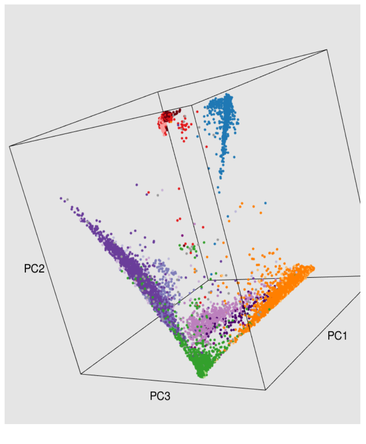


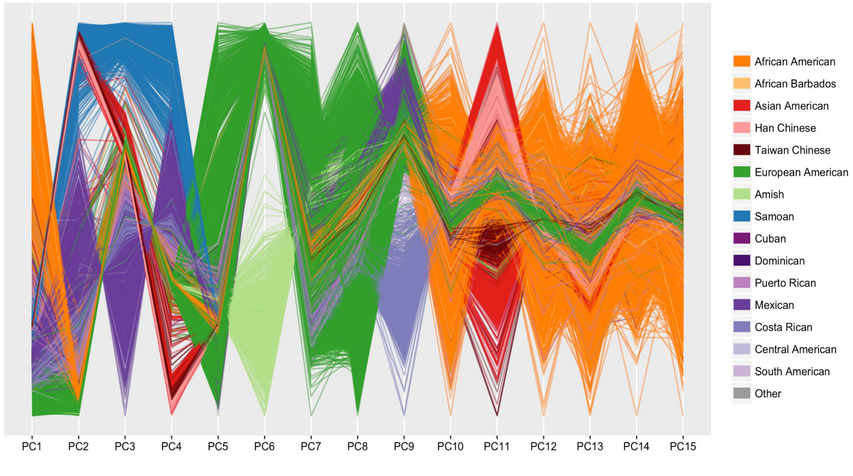

### Study Designs

Currently, TOPMed consists of ~144k participants from >80 different studies with varying designs. Prospective cohorts provide large numbers of disease risk factors, subclinical disease measures, and incident disease cases; case-control studies provide large numbers of prevalent disease cases; extended family structures and population isolates provide improved power to detect rare variant effects. The phenotype pie chart above shows the numbers and percentages of participants in studies with a focus on HLBS, as well as the percentage belonging to cohort studies that have collected many different phenotypes.

### Participant Diversity

Achieving ancestral and ethnic diversity was a priority in selecting contributing studies**.** Currently, the 144k participants consist of approximately **60% with substantial non-European ancestry** (see pie chart above, based on participant self-identification and study inclusion criteria). This diversity is also illustrated by the principal components analysis of ~55k participants (see 3D PCA plot and parallel coordinates plot below). Discovery of genotype-phenotype associations frequently includes pooled analysis across ancestry groups and studies, using statistical models that account for population structure and relatedness.




### Acknowledgments

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Please see https://www.nhlbiwgs.org/acknowledgements for full acknowledgments and funding sources. We acknowledge the support of Daniel Taliun, Quenna Wong, Sarah Nelson, Stephanie Gogarten, and Deepti Jain for providing data summaries and graphics in this poster.

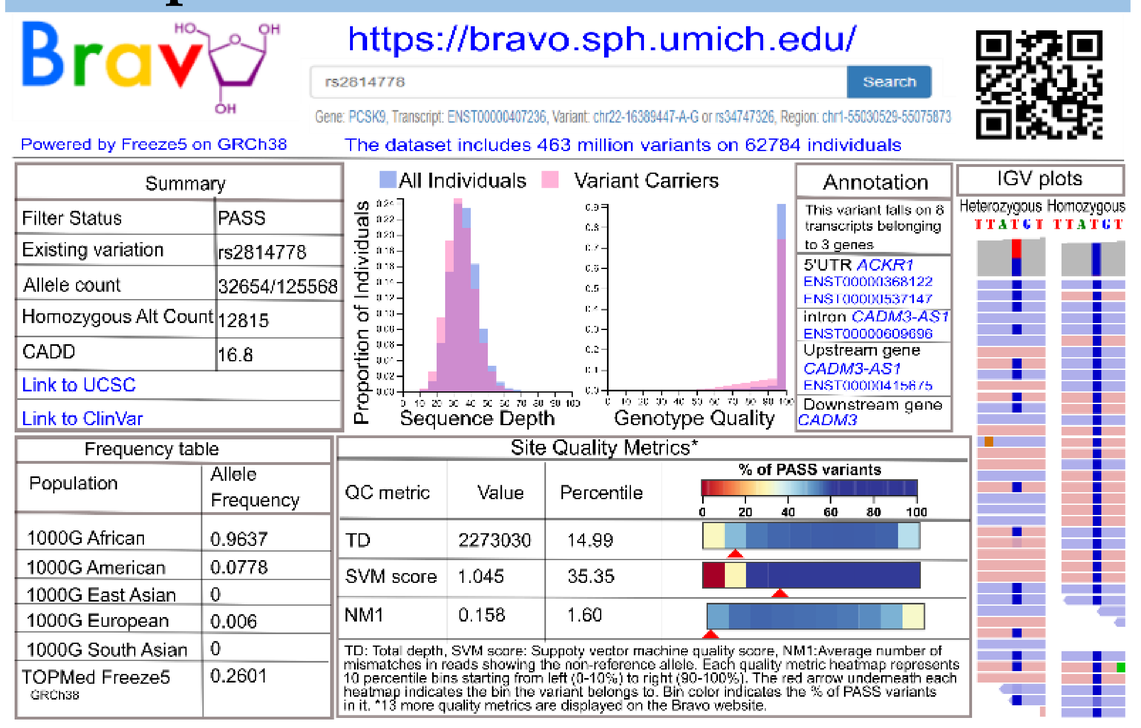
## Whole Genome Sequencing

**Coverage, sequencing depth and number of variants in TOPMed freeze 5 (~65k samples)**

| | Total Number | % Singletons | Per Sample 5th percentile | Per Sample median | Per Sample 95th percentile |
|---|---|---|---|---|---|
| **Bases (Gb)** | 8,523,992 | | 107 | 131 | 156 |
| **Depth (x)** | | | 32 | 39 | 47 |
| **Genome covered %** | | | | | |
| **All** | | | 98.12 | 99.85 | 99.90 |
| **Depth > 10x** | | | 97.80 | 99.18 | 99.62 |
| **Total variants** | 471,006,757 | 46 | 3,514,025 | 3,567,972 | 4,363,040 |
| **SNVs** | 437,606,741 | 46 | 3,333,422 | 3,384,184 | 4,128,810 |
| **Indels** | 33,400,016 | 47 | 180,486 | 183,885 | 234,188 |
| **Coding variation** | | | | | |
| **Synonymous** | 1,734,456 | 43 | 10,837 | 11,078 | 13,688 |
| **Non-synonymous** | 3,644,788 | 48 | 10,632 | 10,880 | 13,231 |
| **Stop/essential splice** | 122,804 | 54 | 427 | 458 | 567 |
| **Indels** | | | | | |
| **Frameshift** | 134,489 | 60 | 111 | 128 | 167 |
| **In-frame** | 64,356 | 48 | 85 | 100 | 129 |

WGS was performed by several sequencing centers to a median depth of 39X using DNA from blood, PCR-free library construction and Illumina HiSeq X technology. A Support Vector Machine quality filter was trained with known variants and Mendelian-inconsistent variants. Genotypes were called jointly across all samples available to produce genotype data "freezes." For freeze 5, the median pair-wise non-reference discordance rates are: $4 \times 10^{-4}$ for SNVs and $6 \times 10^{-3}$ for indels passing the quality filter (N=378 duplicate sample pairs).

## Explore TOPMed variants in Bravo




## TOPMed Resources for the Community

TOPMed data are being made available to the scientific community as a series of "data freezes": **genotypes and phenotypes via dbGaP**; read alignments via the SRA; and variants via the Bravo variant server and dbSNP. Genotypes for a set of 18.5k samples have been released on dbGaP, another freeze of 55k samples is currently being released, and a freeze of >100k samples is planned starting early 2019. TOPMed WGS data are contained in study-specific accessions with names containing "NHLBI TOPMed", while most phenotypic data are in parent study accessions. **The TOPMed web site has much more information about the study, as well as how to access the data through dbGaP — see www.nhlbiwgs.org.**

## Other Omics

TOPMed is currently adding other omic assays to samples that have been whole genome sequenced; these include RNAseq, metabolomics, proteomics and epigenomics. These data will become available via dbGaP starting in 2019.