Exome Aggregation Consortium (ExAC) Data Set (>60,000 exomes)

IBG 2015

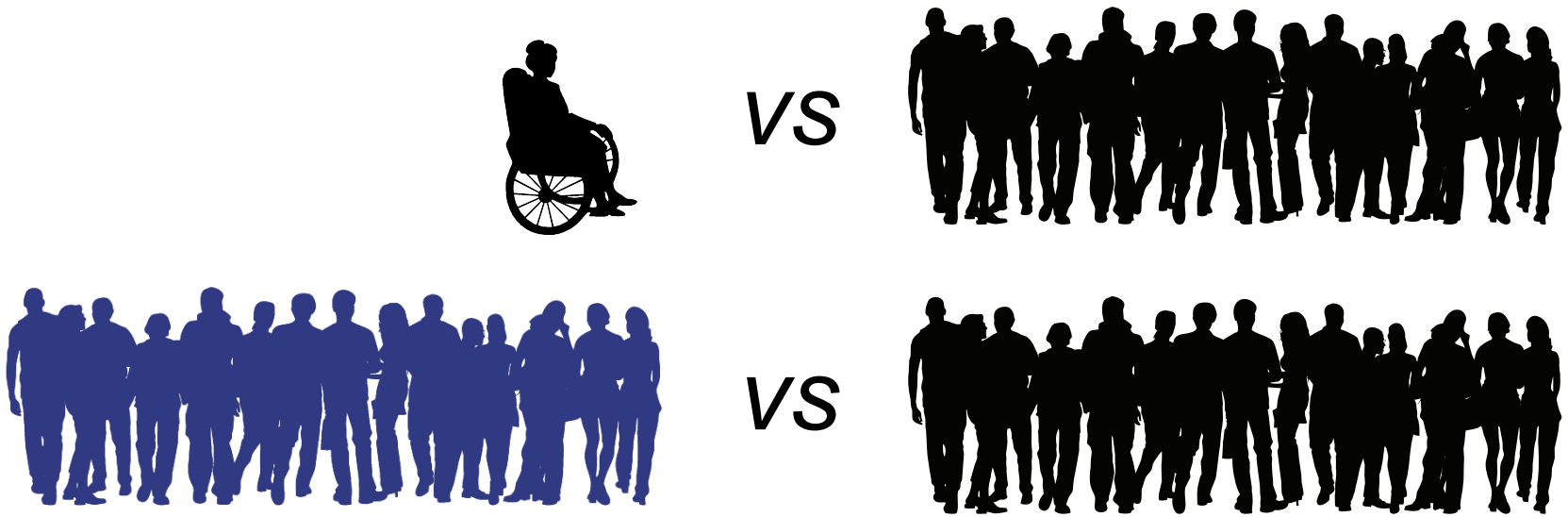
Monkol Lek
20150304

BROAD INSTITUTE

Rationale:

Fully understanding any human's variants requires placing them in the context of **tens of thousands** of other genomes

We have sequenced >200,000 human exomes, but data are siloed



ExAC Principal Investigators

Daniel MacArthur
David Altshuler
Diego Ardissino
Michael Boehnke
Mark Daly
John Danesh
Roberto Elosua
Gad Getz
Christina Hultman
Sekar Kathiresan
Markku Laakso
Steven McCarroll
Mark McCarthy
Ruth McPherson
Benjamin Neale
Aarno Palotie
Shaun Purcell
Danish Saleheen
Jeremiah Scharf
Pamela Sklar
Patrick Sullivan
Jaakko Tuomilehto
Hugh Watkins

Analysis

Monkol Lek
Eric Minikel
Kaitlin Samocha
Menachem Fromer
Doug Ruderfer
Pradeep Natarajan
Ron Do
Andrew Hill
James Ware
Adam Kiezun

Production

Monkol Lek
Fengmei Zhao
Ryan Poplin
Eric Banks
Timothy Fennell

Website

Konrad Karczewski
Brett Thomas
Ben Weisburd

Broad Genomics Platform

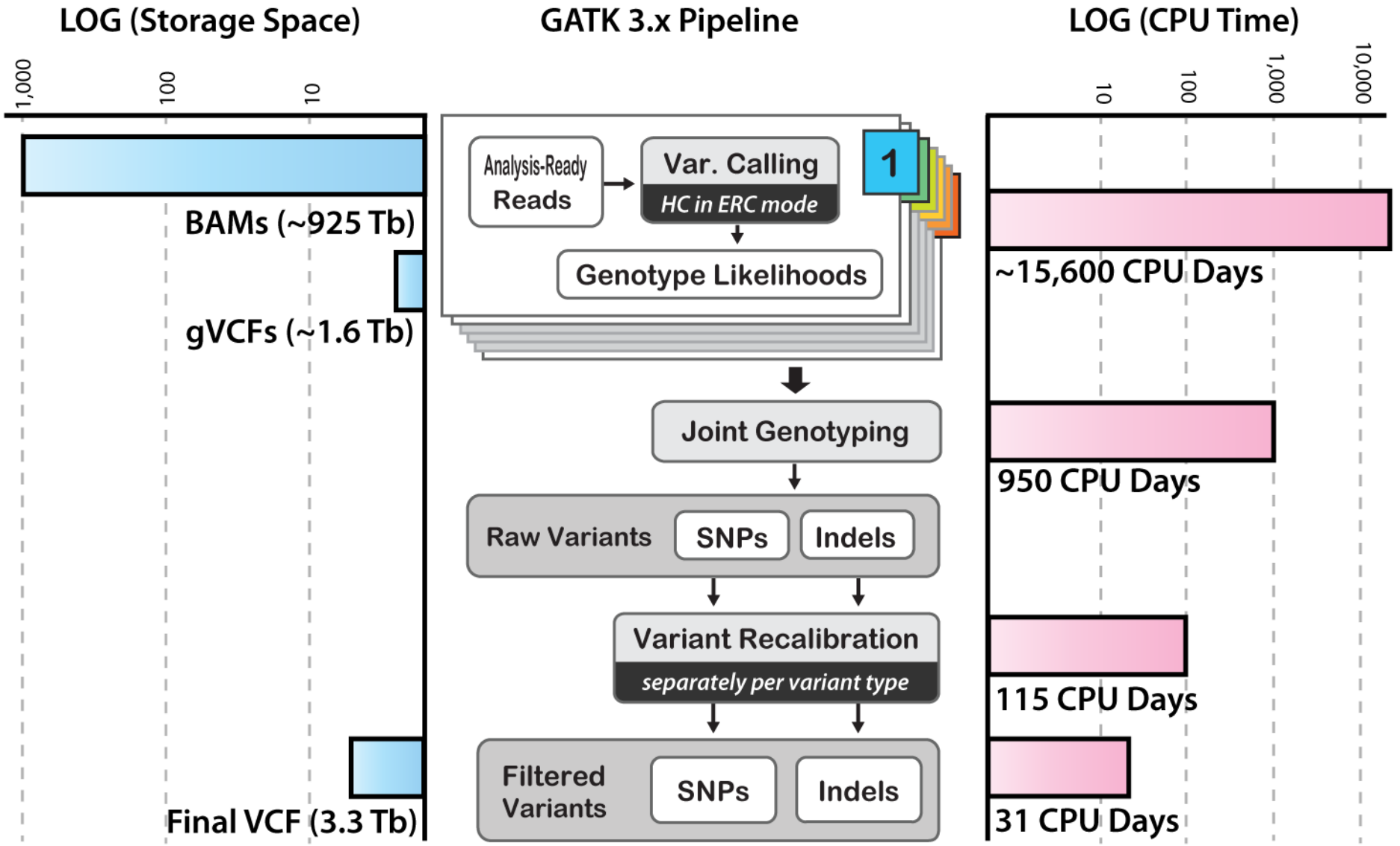
Contributing cohorts

1000 Genomes
Bulgarian Trios
Finland-United States Investigation of NIDDM Genetics (FUSION)
GoT2D
Inflammatory Bowel Disease
METabolic Syndrome In Men (METSIM)
Myocardial Infarction Genetics Consortium:

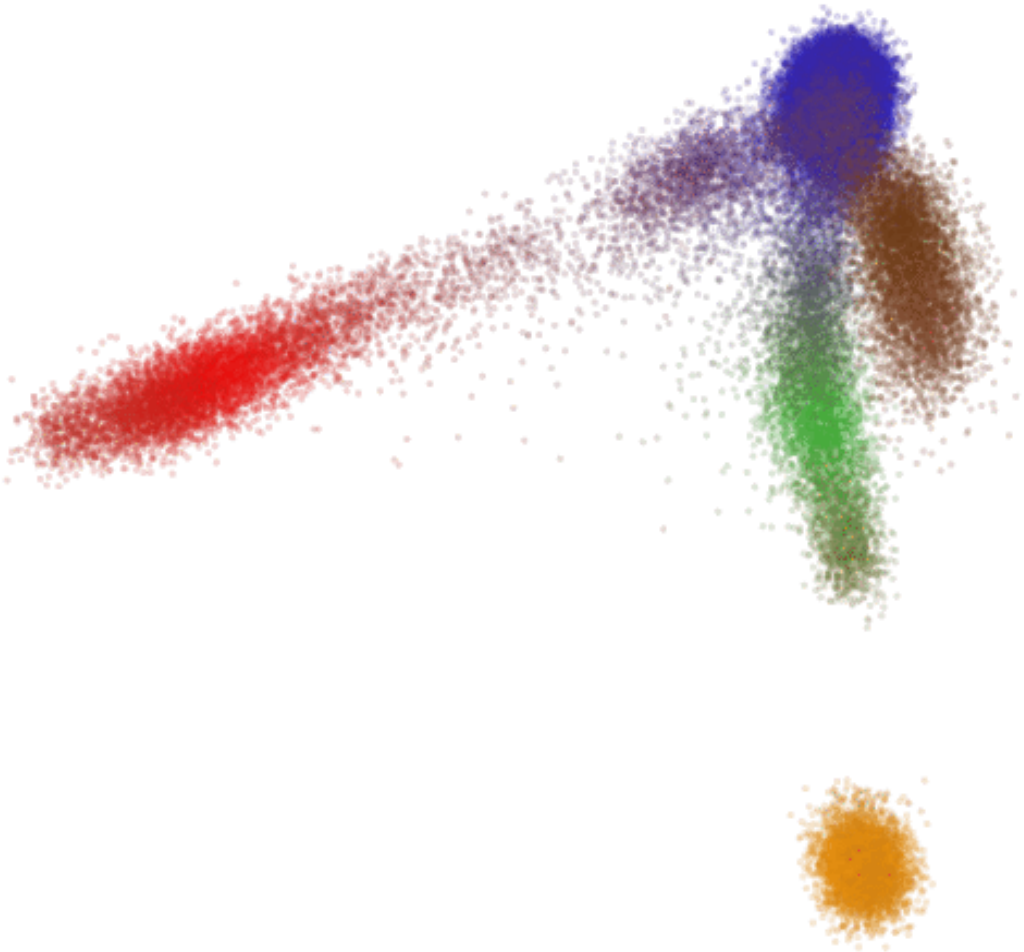
- Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group
- Ottawa Genomics Heart Study
- Pakistan Risk of Myocardial Infarction Study (PROMIS)
- Precocious Coronary Artery Disease Study (PROCARDIS)
- Registre Gironi del COR (REGICOR)

NHLBI-GO Exome Sequencing Project (ESP)
National Institute of Mental Health (NIMH) Controls
SIGMA-T2D
Sequencing in Suomi (SISu)
Swedish Schizophrenia & Bipolar Studies
T2D-GENES
Taiwanese Trios
The Cancer Genome Atlas (TCGA)
Tourette Syndrome Association International Consortium for Genomics (TSAICG)

Scalable Variant Calling Pipeline



European
Latino
South Asian
East Asian
African
Other



Principal component analysis
using 5,800 common SNPs

Catalogue of protein-coding variation

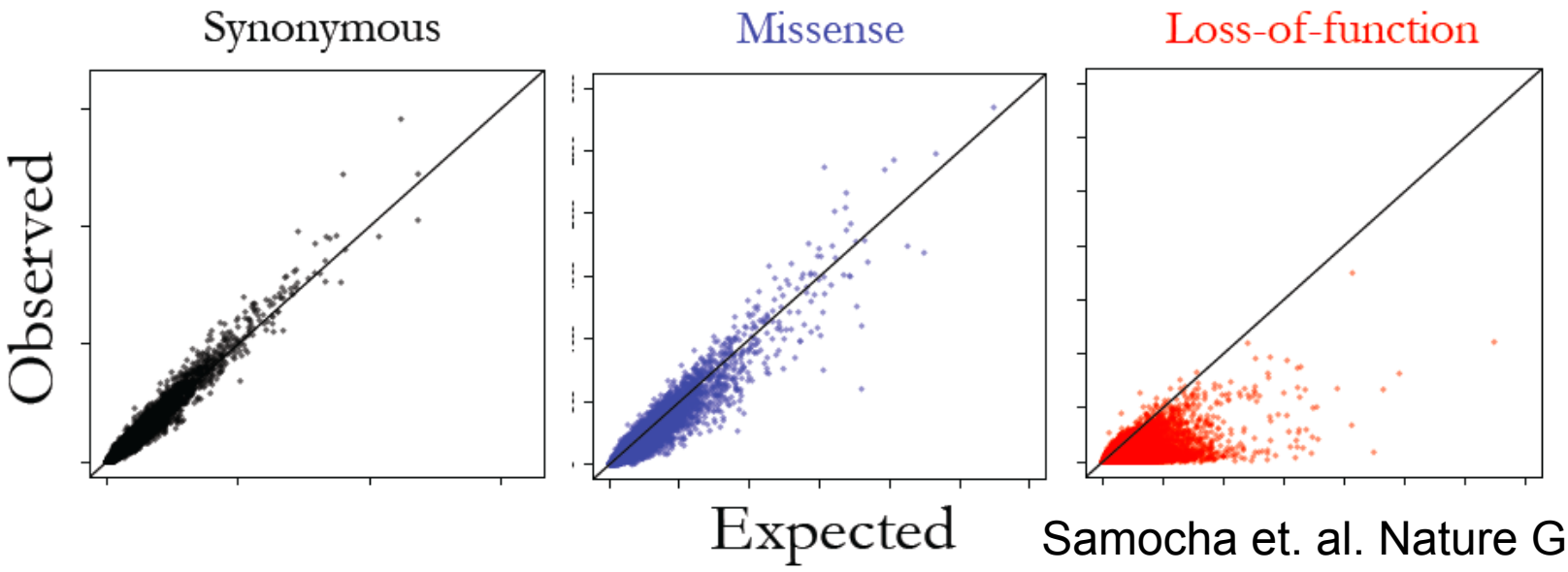
- Largest ever collection of human protein-coding genetic variants
  - **3,296,522** missense alleles: 99% with frequency <1%, 52% seen only once!
  - **232,094** candidate loss-of-function alleles affecting **16,007** protein-coding genes
  - **21,204** previously reported severe disease-causing mutations from HGMD

Gene Constraint

Model accurately predicts synonymous variation

Kaitlin Samocha

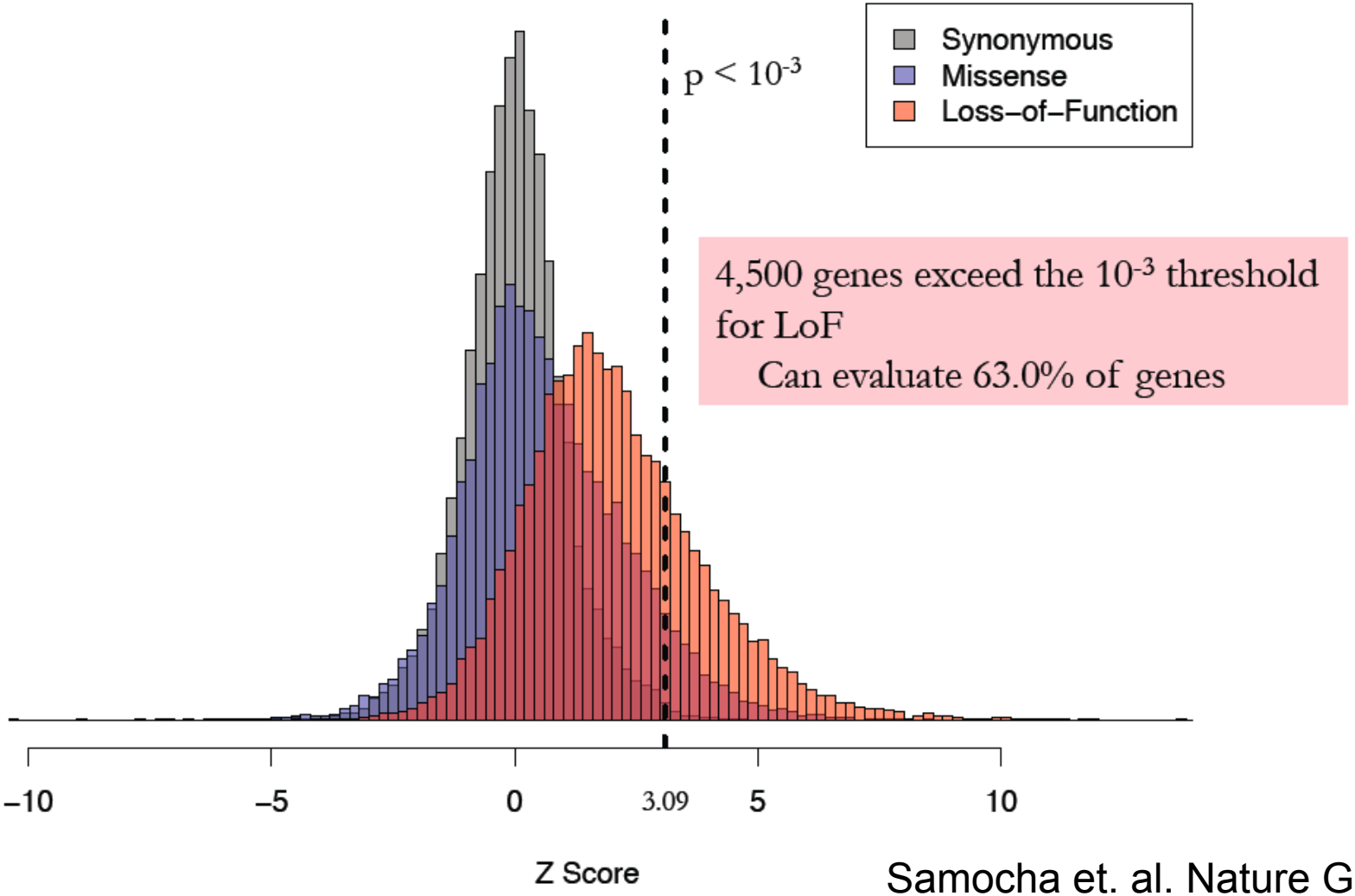
- Extracted rare (MAF < 0.1%) variants from the ~61,000 individuals in the ExAC dataset
- Correlation of number of rare synonymous variants in a gene with:
  - The gene's length 0.9199
  - The expected number of synonymous variants **0.9796**



Samocha et. al. Nature Genetics 2014

Gene Constraint

Extreme loss-of-function constraint in ExAC



Kaitlin Samocha

Samocha et. al. Nature Genetics 2014

Public Release

- All variants and population frequencies are publicly available:

exac.broadinstitute.org

Konrad Karczewski

Brett Thomas

Ben Weisburd

ExAC Principal Investigators

Daniel MacArthur
David Altshuler
Diego Ardissino
Michael Boehnke
Mark Daly
John Danesh
Roberto Elosua
Gad Getz
Christina Hultman
Sekar Kathiresan
Markku Laakso
Steven McCarroll
Mark McCarthy
Ruth McPherson
Benjamin Neale
Aarno Palotie
Shaun Purcell
Danish Saleheen
Jeremiah Scharf
Pamela Sklar
Patrick Sullivan
Jaakko Tuomilehto
Hugh Watkins

Analysis

Monkol Lek
Eric Minikel
Kaitlin Samocha
Menachem Fromer
Doug Ruderfer
Pradeep Natarajan
Ron Do
Andrew Hill
James Ware
Adam Kiezun

Production

Monkol Lek
Fengmei Zhao
Ryan Poplin
Eric Banks
Timothy Fennell

Website

Konrad Karczewski
Brett Thomas
Ben Weisburd

Broad Genomics Platform

Contributing cohorts

1000 Genomes
Bulgarian Trios
Finland-United States Investigation of NIDDM Genetics (FUSION)
GoT2D
Inflammatory Bowel Disease
METabolic Syndrome In Men (METSIM)
Myocardial Infarction Genetics Consortium:

- Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group
- Ottawa Genomics Heart Study
- Pakistan Risk of Myocardial Infarction Study (PROMIS)
- Precocious Coronary Artery Disease Study (PROCARDIS)
- Registre Gironi del COR (REGICOR)

NHLBI-GO Exome Sequencing Project (ESP)
National Institute of Mental Health (NIMH) Controls
SIGMA-T2D
Sequencing in Suomi (SISu)
Swedish Schizophrenia & Bipolar Studies
T2D-GENES
Taiwanese Trios
The Cancer Genome Atlas (TCGA)
Tourette Syndrome Association International Consortium for Genomics (TSAICG)