Phenotype-driven exome sequencing

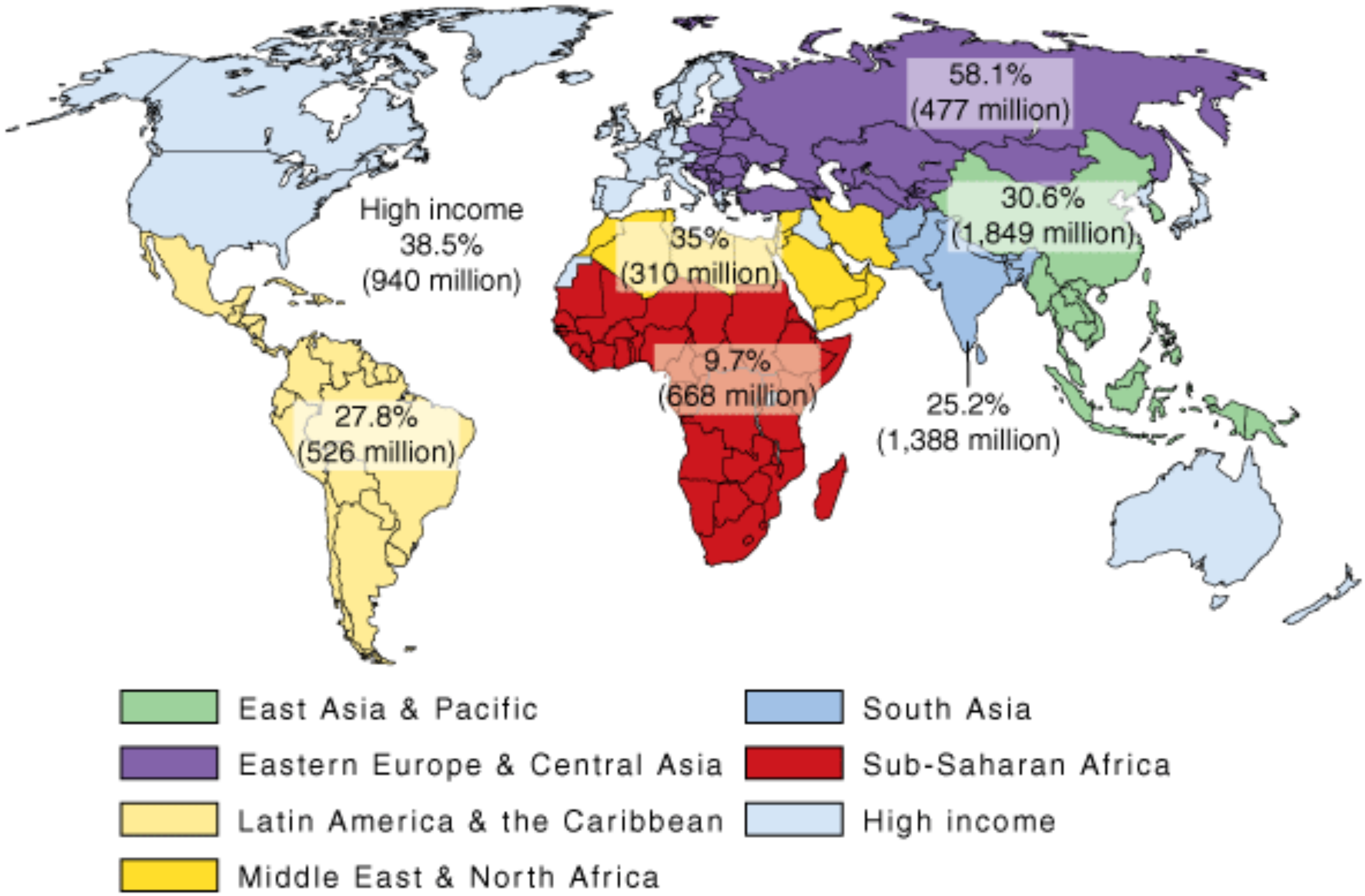
James Pirruccello

Kathiresan Lab
Massachusetts General Hospital
& Broad Institute

So you've got a phenotype…

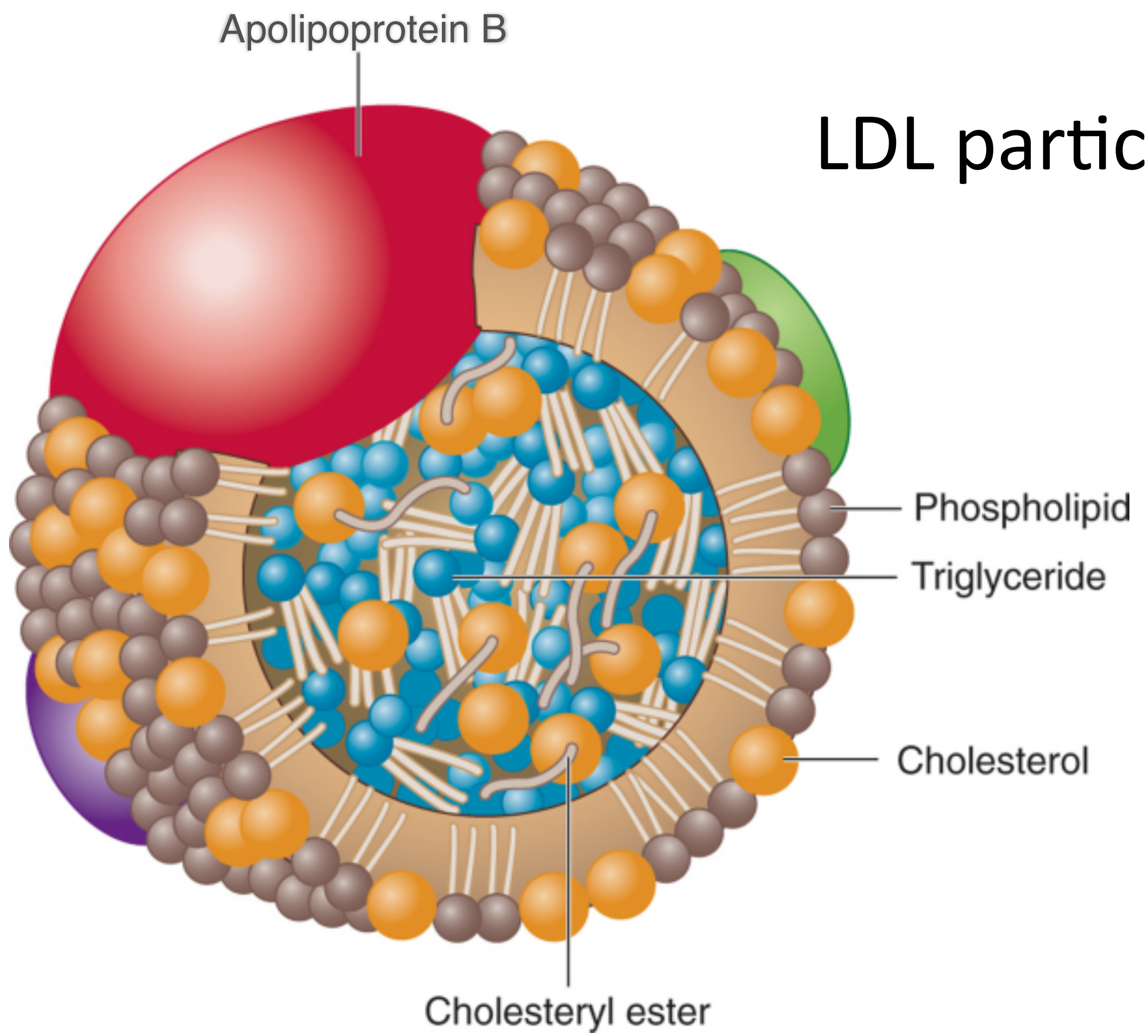
- A readily available procedure for identifying its genetic basis:
  - Select the right samples
  - Discover variants (sequencing)
  - Confirm genotype / phenotype association
  - Confirm that this association replicates
- In this talk, **lipids** are our phenotype

Cardiovascular disease is the leading cause of death



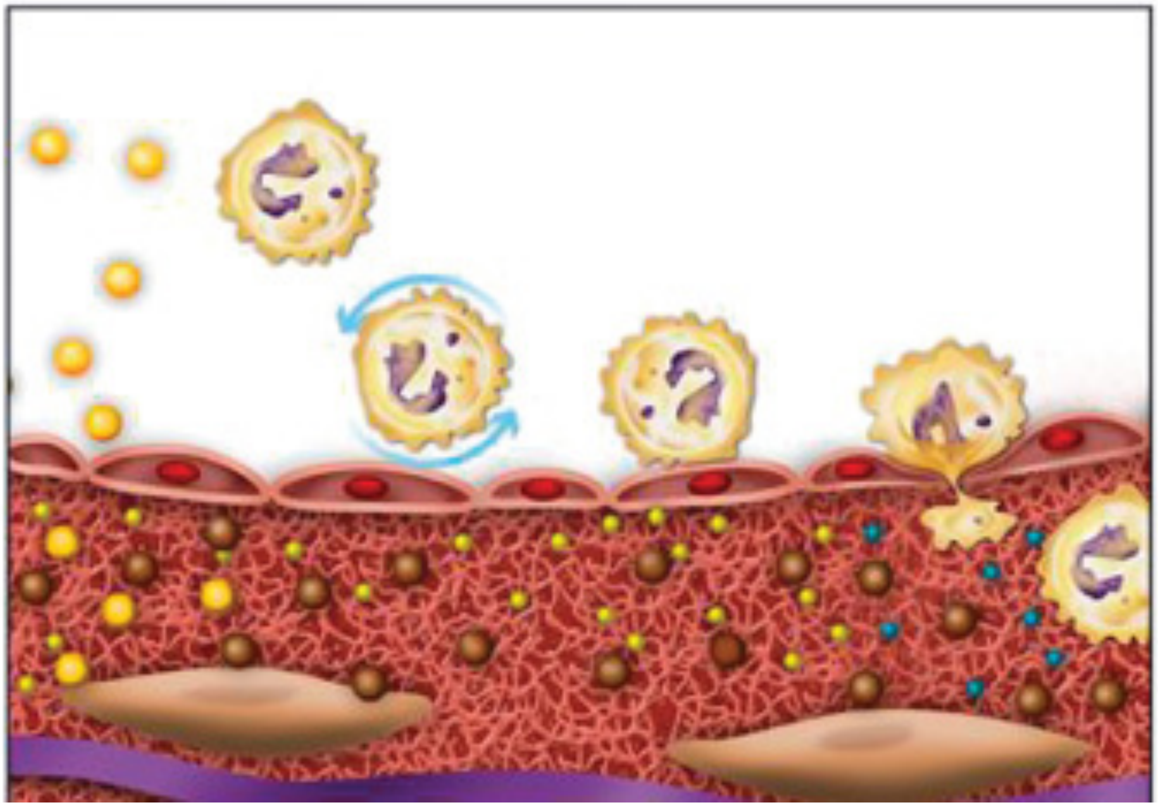
Fauci, Anthony et.al. Harrison's Principles of Internal Medicine. New York: McGraw-Hill, 2008.

LDL particle



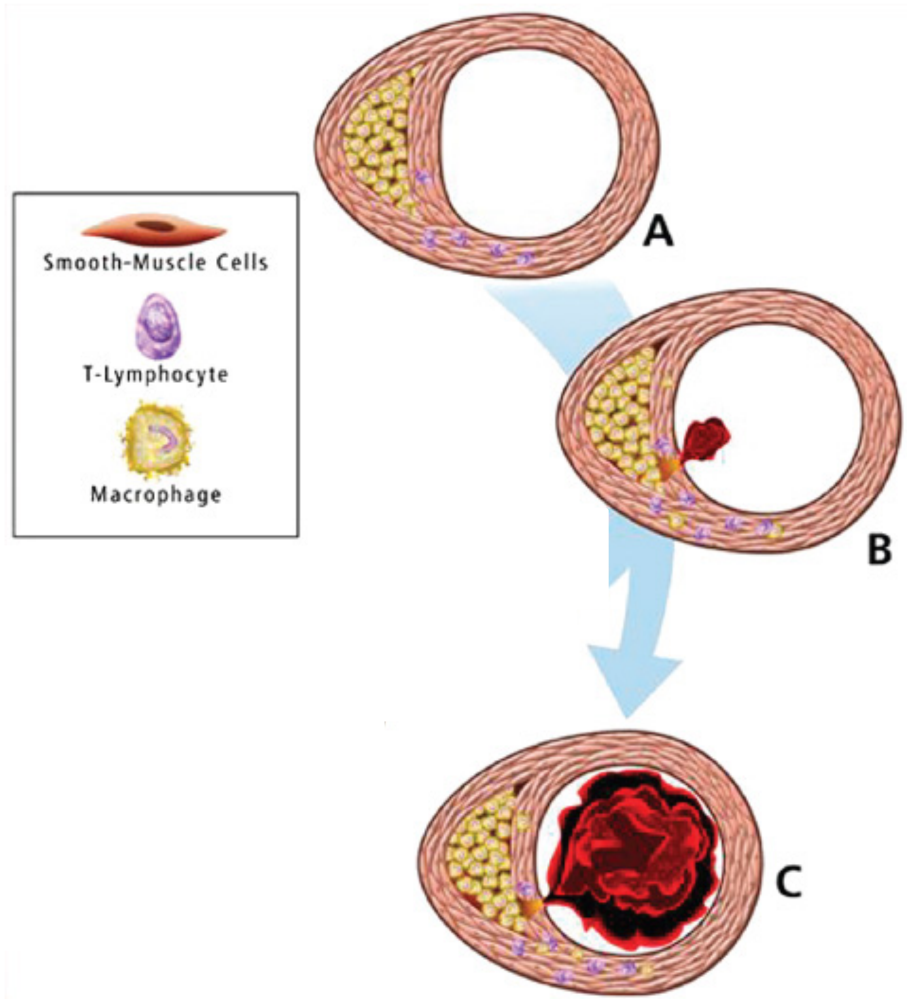
Mann, Douglas L.; Libby, Peter; Braunwald, Eugene; Robert Bonow; Zipes, Douglas P. (2008). *Braunwald's heart disease: a textbook of cardiovascular medicine*. Saunders/Elsevier. ISBN 1-4160-4107-9.

LDL deposition initiates atherosclerosis



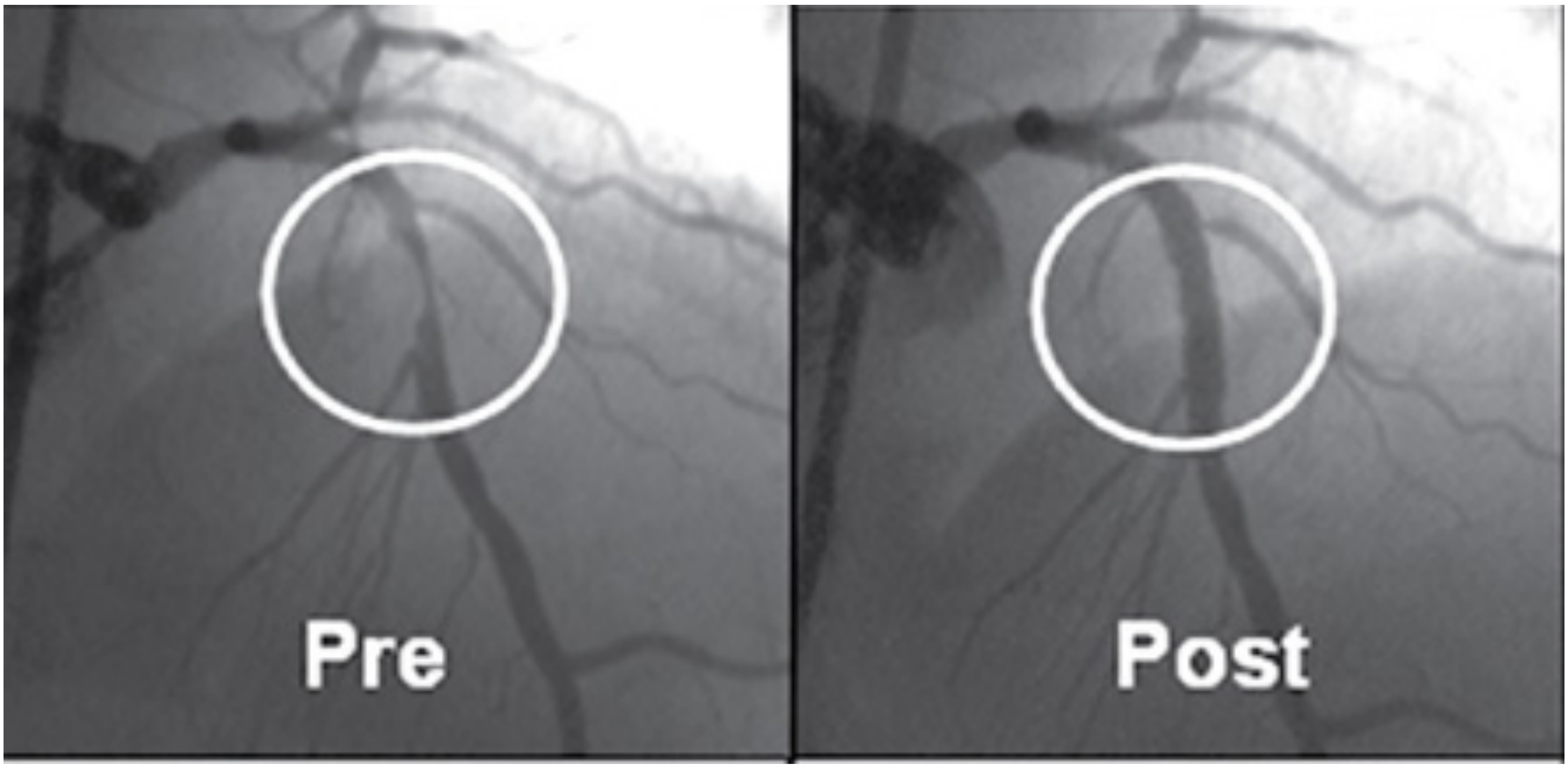
Fauci, Anthony et.al. Harrison's Principles of Internal Medicine. New York: McGraw-Hill, 2008.

Plaque rupture can be devastating



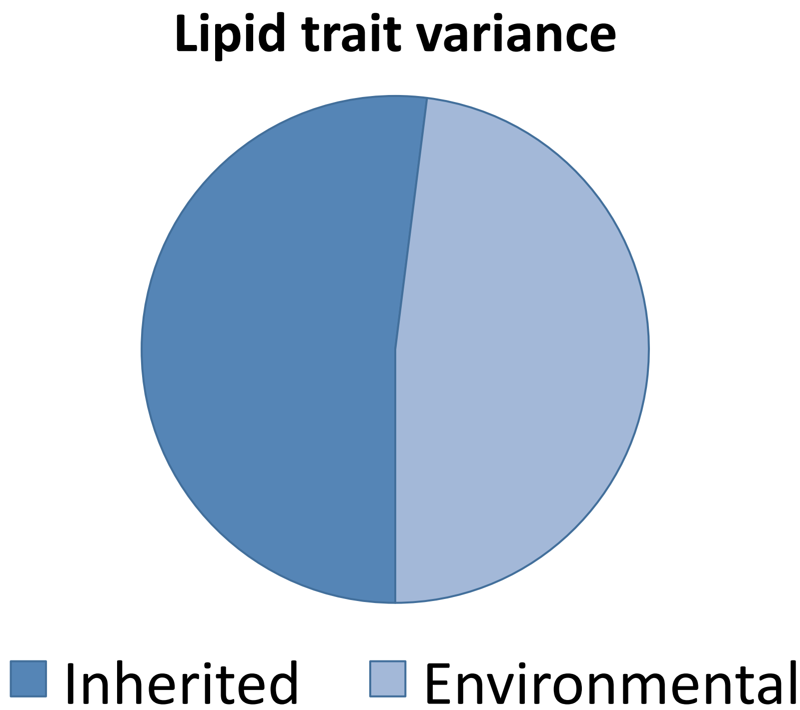
Fauci, Anthony et.al. Harrison's Principles of Internal Medicine. New York: McGraw-Hill, 2008.

Plaque rupture in the coronary vessels: myocardial infarction



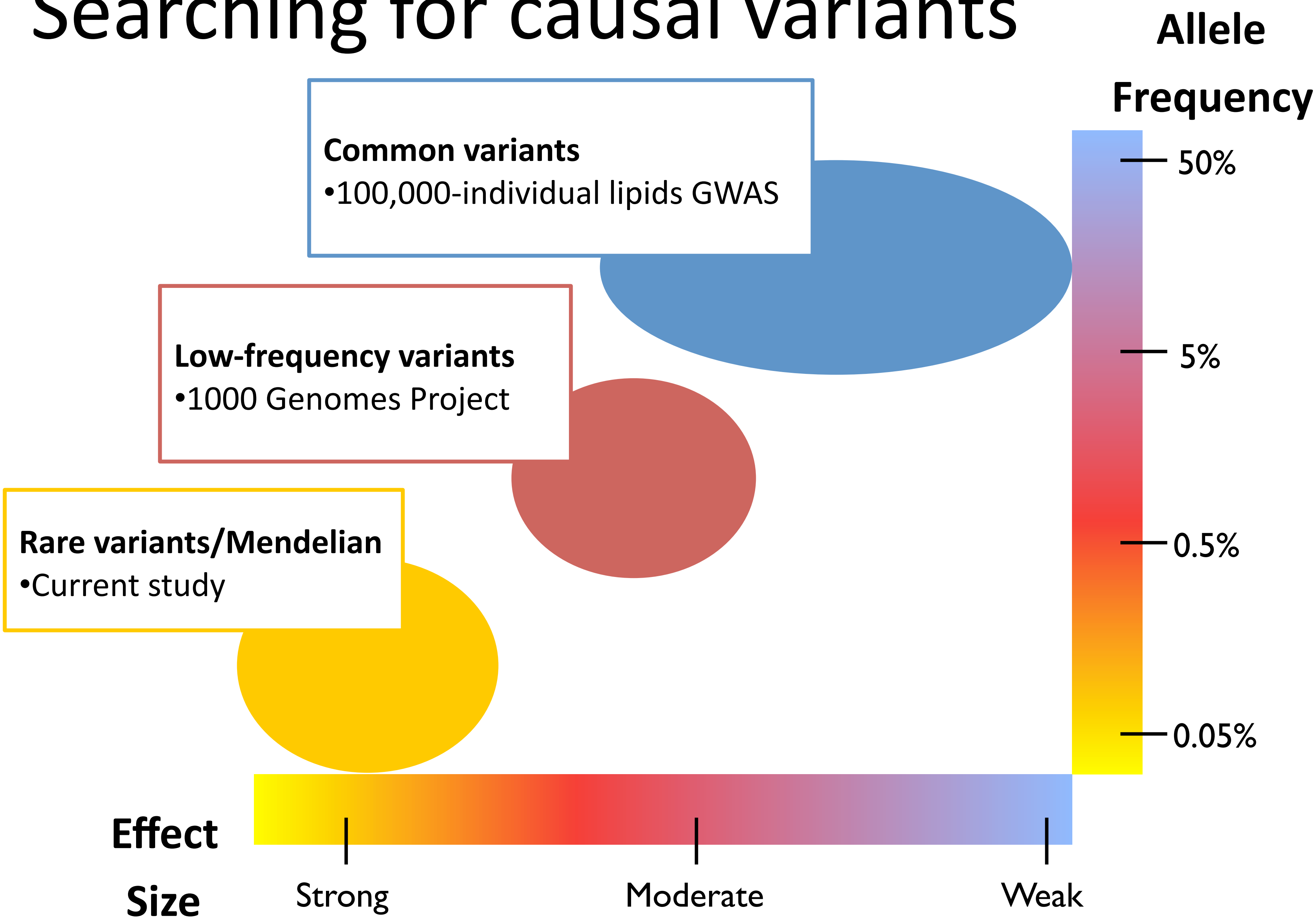
Fauci, Anthony et.al. Harrison's Principles of Internal Medicine. New York: McGraw-Hill, 2008.

Lipid traits are heritable



How do we identify which genetic variants are causal?

Searching for causal variants



Monogenic lipid disorders

Gene	Locus	GWAS SNP	Disorder and lipid phenotype
ABCA1	9q31.1	rs1883025	Tangier disease: low HDL
ABCG5	2p21	rs4299376	Sitosterolemia: high LDL
ABCG8	2p21	rs4299376	Sitosterolemia: high LDL
APOA1	11q23-q24	rs964184	ApoA-I deficiency: low HDL
APOA5	11q23	rs964184	ApoA-V deficiency: high VLDL and chylomicrons
APOB	2p24	rs515135	Familial hypobetalipoproteinemia: low LDL
			Familial defective ApoB-100: high LDL
APOC2	19q13	rs4420638	Familial ApoC-II deficiency: high chylomicrons
APOE	19q13	rs4420638	Familial dysbetalipoproteinemia: high VLDL remnants and chylomicrons
CETP	16q13	rs173539	Cholesteryl ester transfer protein deficiency: high HDL
LCAT	16q22	rs2271293	Lecithin-cholesterol acyltransferase deficiency (fish-eye disease): low HDL
LDLR	19p13	rs6511720	Familial hypercholesterolemia: high LDL
LDLRAP1	1p36-p35	rs12027135	Autosomal recessive hypercholesterolemia: high LDL
LIPC	15q22	rs10468017	Familial hepatic lipase deficiency: high VLDL remnants
LPL	8p21	rs12678919	Lipoprotein lipase deficiency: high chylomicrons
MTTP	4q24	N/A	Abetalipoproteinemia: low LDL
PCSK9	1p32	rs11206510	Autosomal-dominant hypercholesterolemia: high LDL
			PCSK9 deficiency: low LDL
SAR1B	5q31.1	N/A	Chylomicron retention disease: low chylomicrons

Edmondson AC, Rader DJ (October 2008). "Genome-wide approaches to finding novel genes for lipid traits: the start of a long road". *Circ Cardiovasc Genet* **1** (1): 3–6.

(Motivating description of the monogenic lipid disorder of interest)

Variant discovery framework

- **Select the right samples**
- Discover variants (sequencing)
- Confirm genotype / phenotype association
- Confirm that this association replicates

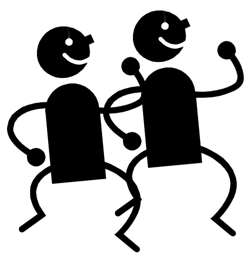
- (Pre-publication pedigree redacted.)

- (Pre-publication zoomed-in pedigree redacted.)
- Two individuals from the pedigree were selected for exome sequencing

Variant discovery framework

- Select the right samples
- **Discover variants (sequencing)**
- Confirm genotype / phenotype association
- Confirm that this association replicates

A schematic sequencing overview



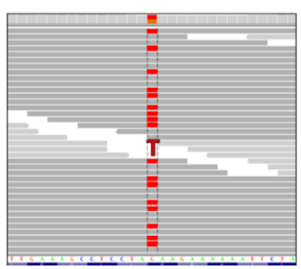
Obtain DNA

Sequence
(Illumina GA)

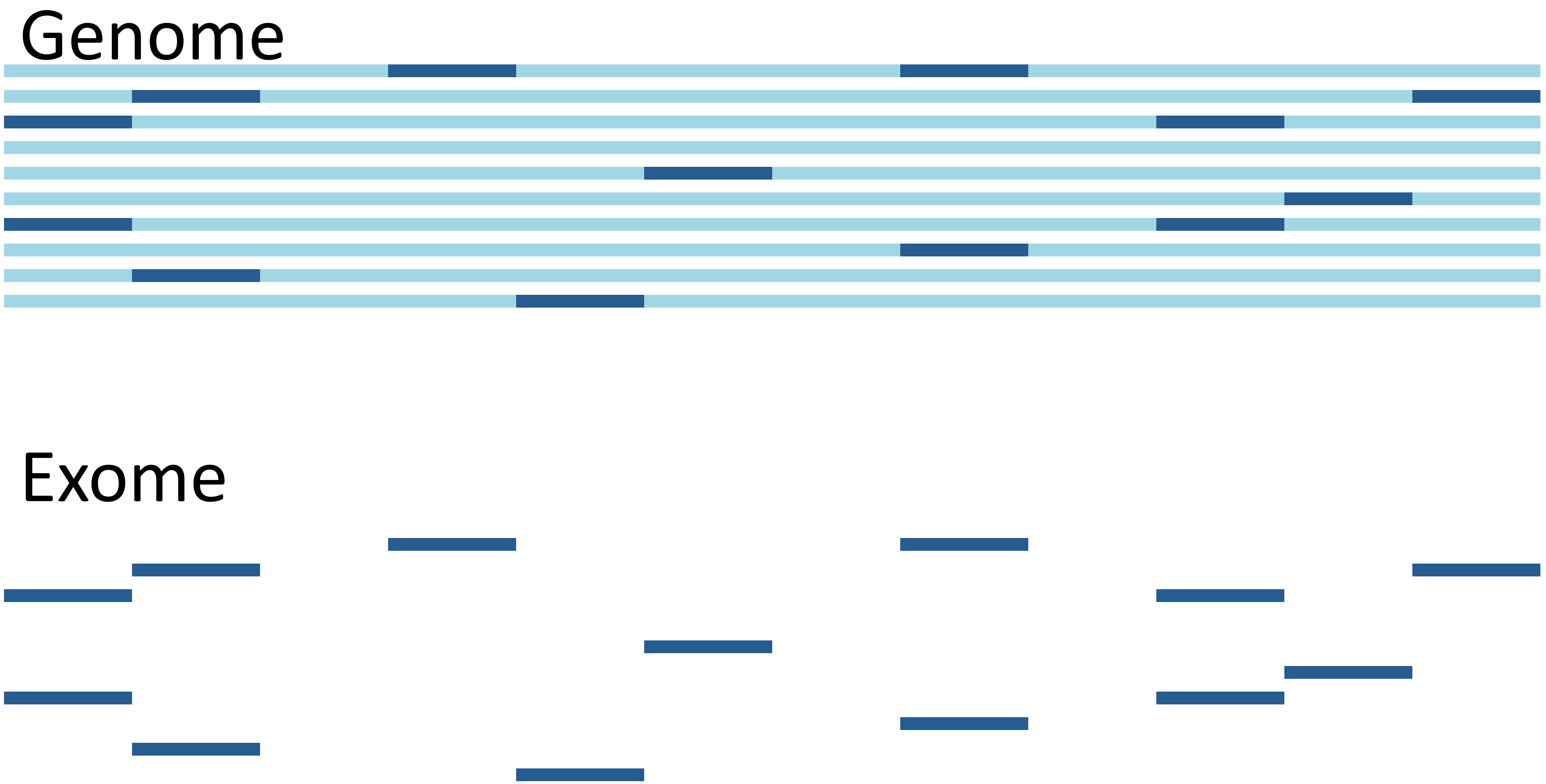
Recover variants
(SAMtools, GATK,
Picard)

Filter out low-
quality variants
(GATK, Firehose)

Interpret results
(R, IGV, Plink, intuition)



Exome sequencing is unbiased and efficient



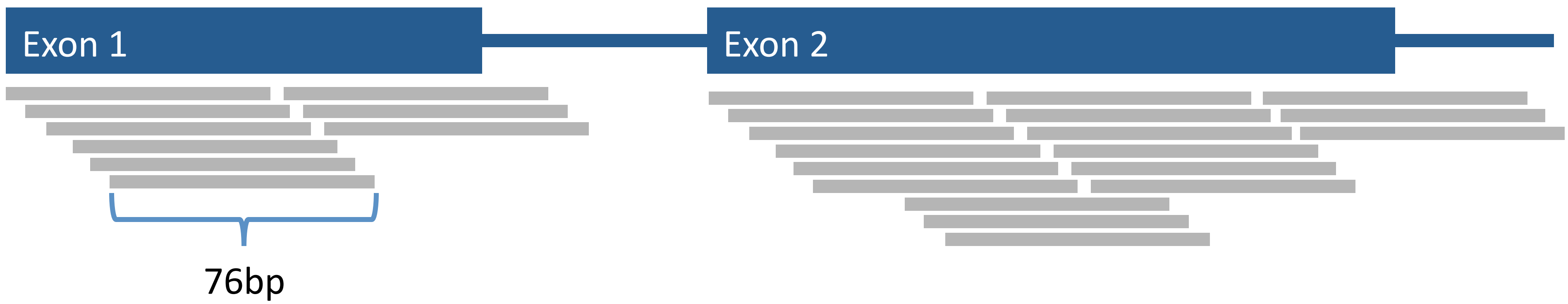
Why sequence exomes?

- Why not just look at lipid-related GWAS genes?
  - Not clear *a priori* that the causal variant will be in a GWAS gene
  - Samples are rare enough that taking a thorough look is a good value proposition

- Why not just look at genes in a linkage peak?
  - Possibly hundreds of genes! Exome is faster, easier.

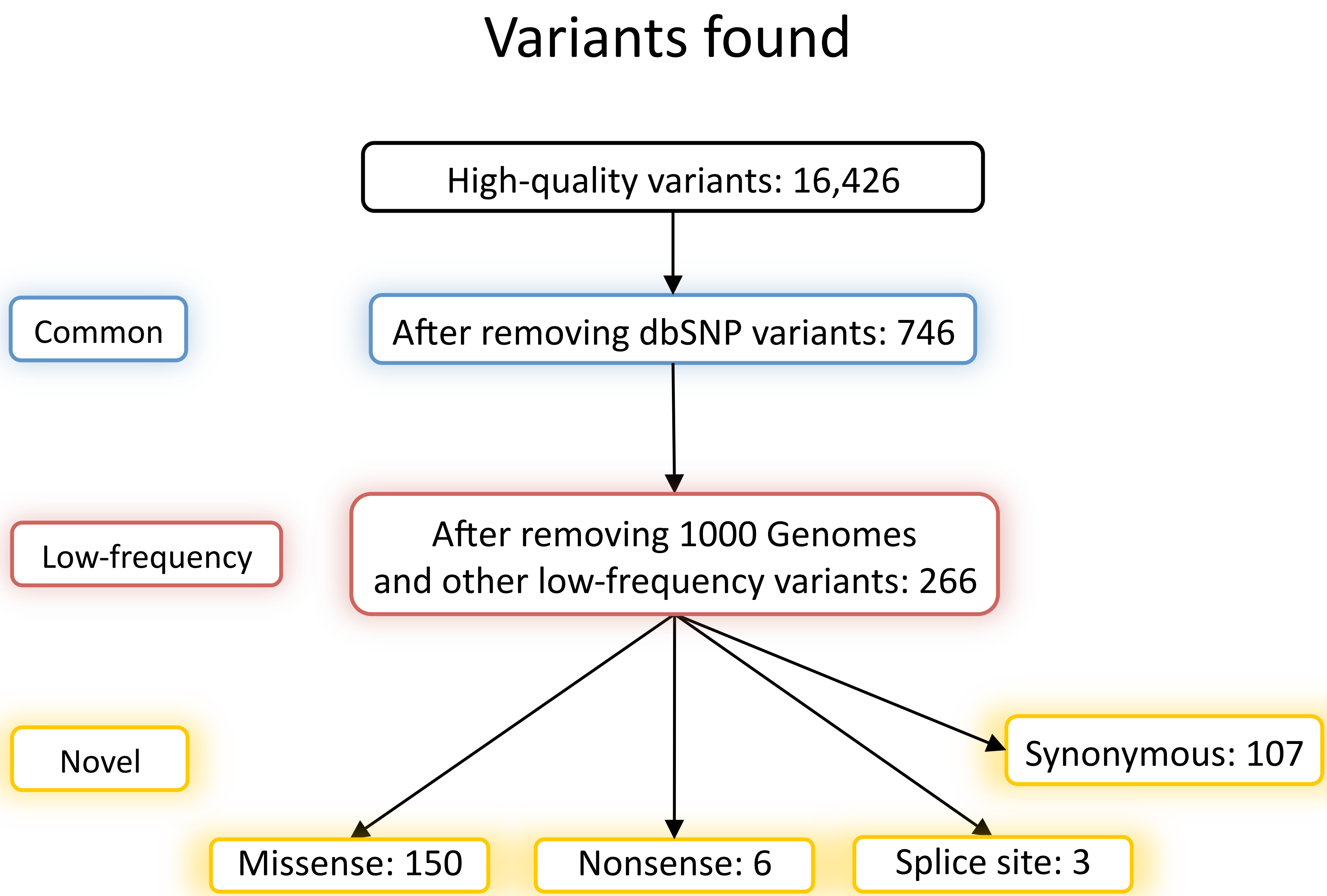
Exome sequencing yields high-quality coverage of targeted bases

Target: Exons of 16,000 genes (28.6 MB)



Sequencing yield: 6,000 MB (mean ~210x coverage)

90% of bases >= 20x

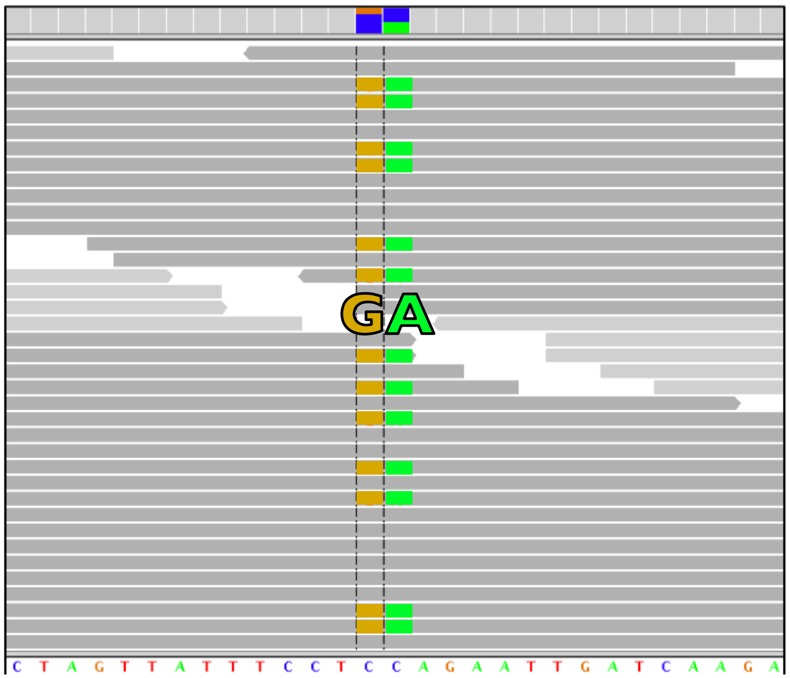

Variants found
High-quality variants: 16,426
Common
After removing dbSNP variants: 746
Low-frequency
After removing 1000 Genomes
and other low-frequency variants: 266
Novel
Synonymous: 107
Missense: 150
Nonsense: 6
Splice site: 3

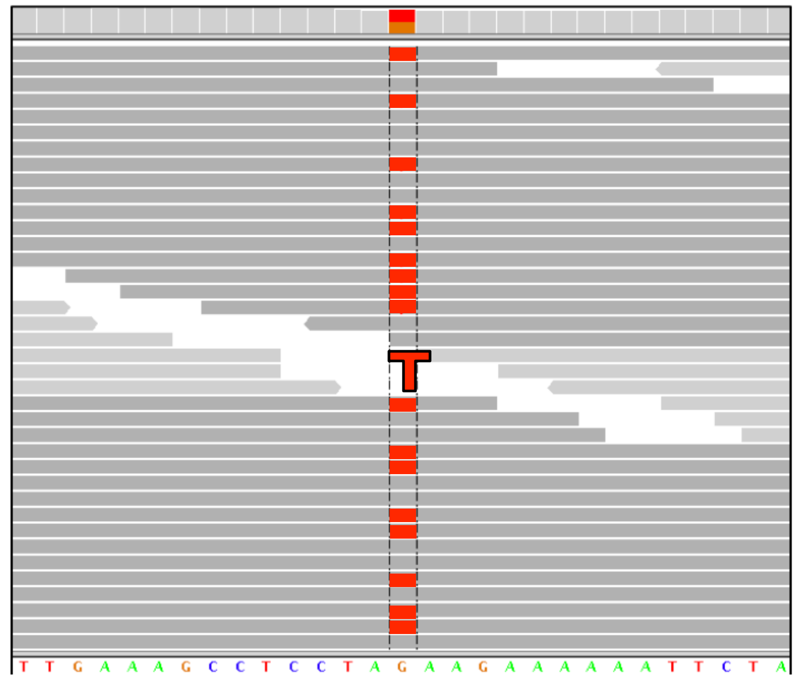
Both probands share two nonsense variants

Only one gene contained novel homozygous or compound heterozygous variants shared by both probands.

(Pre-publication gene name redacted.)

Next-generation sequencing confirms that the dinucleotide variant produces a stop codon





Variant discovery framework

- Select the right samples
- Discover variants (sequencing)
- **Confirm genotype / phenotype association**
- Confirm that this association replicates

(Graphs showing that a lipid trait in this family varies by genotype)

Variant discovery framework

- Select the right samples
- Discover variants (sequencing)
- Confirm genotype / phenotype association
- **Confirm that this association replicates**

Gene variants and lipid levels in XXX Study

- (Population-level testing between *gene of interest* and lipid levels showed association.)

Conclusions

- *Gene* appears to cause a monogenic lipid syndrome

- Exome sequencing can be used to identify the causes of inherited disorders

- Get to know your data

Acknowledgments

- Sekar Kathiresan
- Kiran Musunuru
- Ron Do
- Candace Guiducci
- Elena Gonzalez
- Carrie Sougnez
- Kiran Garimella
- Sheila Fisher
- Justin Abreu
- Andrew Barry
- Tim Fennell
- Eric Banks
- Lauren Ambrogio
- Kristian Cibulskis
- Andrew Kernytsky
- Mark DePristo
- Mark Daly
- David Altshuler
- Stacey Gabriel

Funding Sources

•NHGRI Medical Sequencing Program, U54 HG003067 (PI, Lander)

•Donovan Family Foundation (S. Kathiresan)

•Sarnoff Foundation (J. Pirruccello)