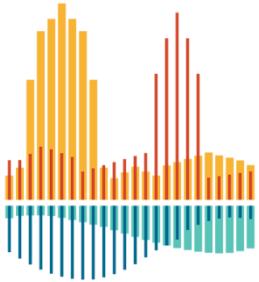
Covering All the Bases: A Primer on Today's Sequencing Technologies and Their Applications in Precision Medicine Research



INSTITUTE FOR COMPUTATIONAL BIOLOGY

September 7, 2018

Dana C. Crawford, PhD
Associate Professor
Population and Quantitative Health Sciences
Institute for Computational Biology

Genomes and Precision Medicine Research

Single nucleotide variants (SNVs) and disease

https://www.ebi.ac.uk/gwas/

CFH and age-related macular degeneration
rs10611790 (Y402H)
MAF ~30%

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Genomes and Precision Medicine Research

Single nucleotide variants (SNVs) and disease

3,541 publications
69,969 SNP-trait associations

https://www.ebi.ac.uk/gwas/

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Genomes and Precision Medicine Research

Genome-wide arrays

~700,000 -millions variants

~$40 - $140/sample

Limited to *known* variants

Mostly common variants

European bias*

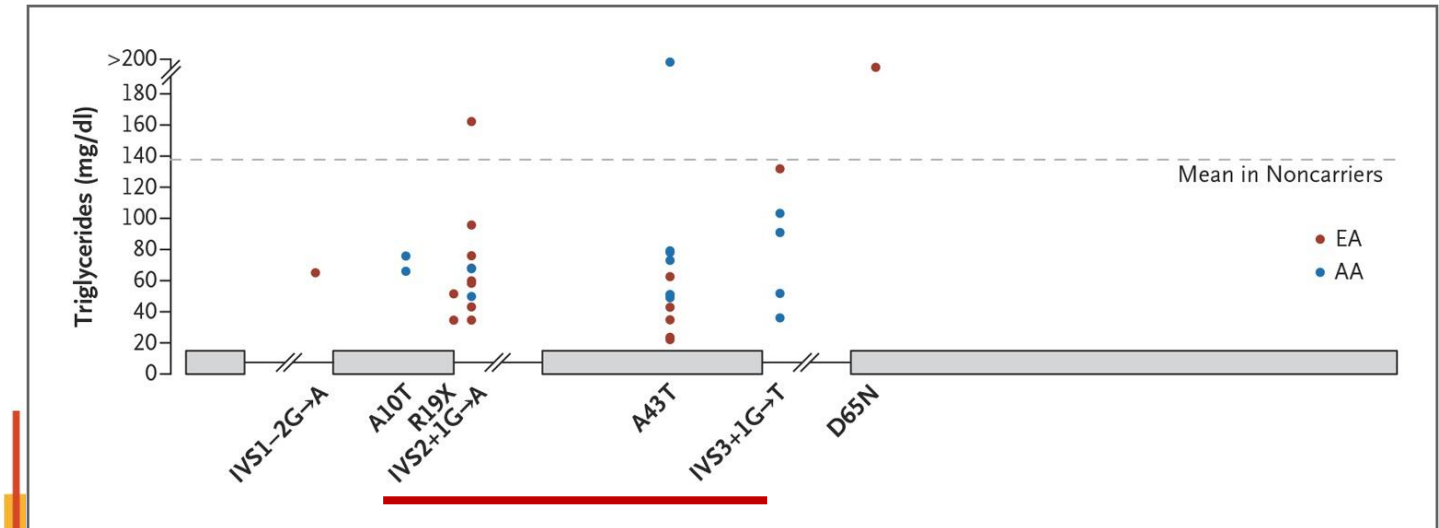
Illumina whole-genome arrays

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Genomes and Precision Medicine Research

Single nucleotide variants (SNVs) and disease

Cardioprotection and *APOC3* splice site, nonsense, and missense mutations
MAF <1%



NEJM (2014)
371:22-31

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Genomes and Precision Medicine Research

Sequencing

Millions to billions of bases

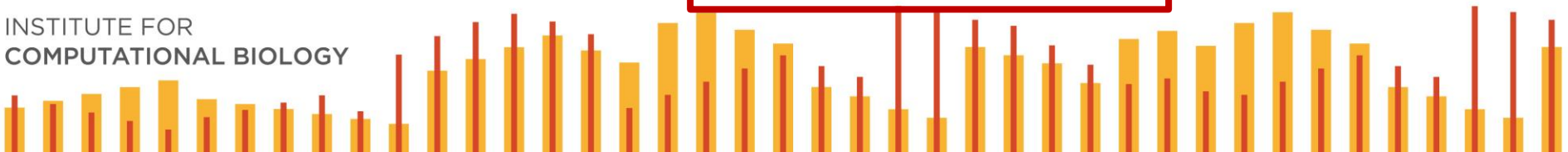
~$750 - $1,500/sample

Assays known and unknown variants

Common, rare, and everything in between

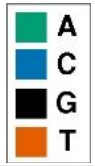
Lower throughput

$$$



INSTITUTE FOR
COMPUTATIONAL BIOLOGY

The Evolution of Sequencing

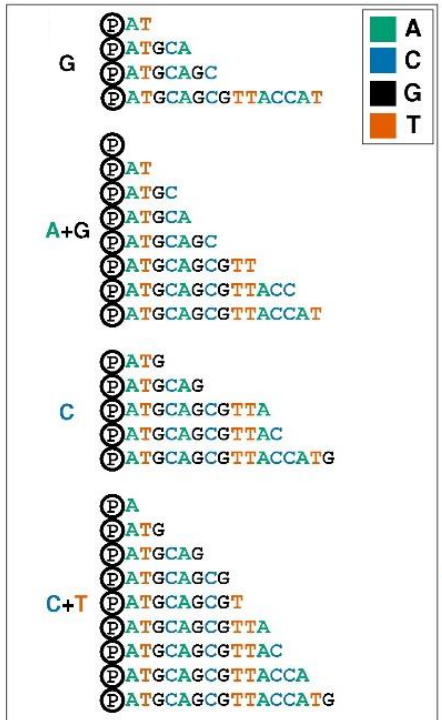


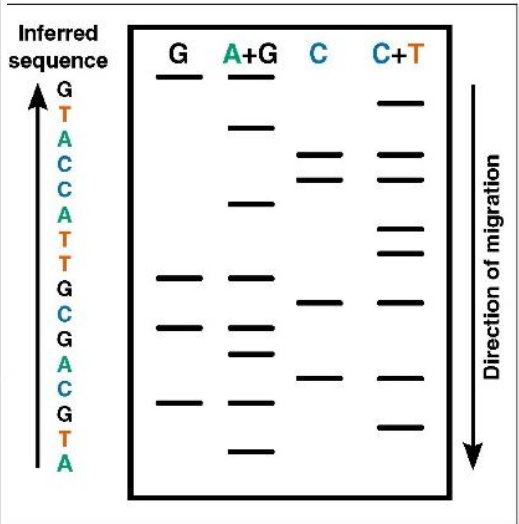
First generation sequencing (mid-1970s)

"Chemical sequencing"

- Add radioactive P^{32} to 5´ phosphate of DNA
- Add chemical treatments to selectively remove DNA bases
- Run on gel to infer position

Maxam-Gilbert

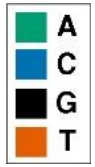




Heather and Chain (2016) *Genomics* 107:1-8

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

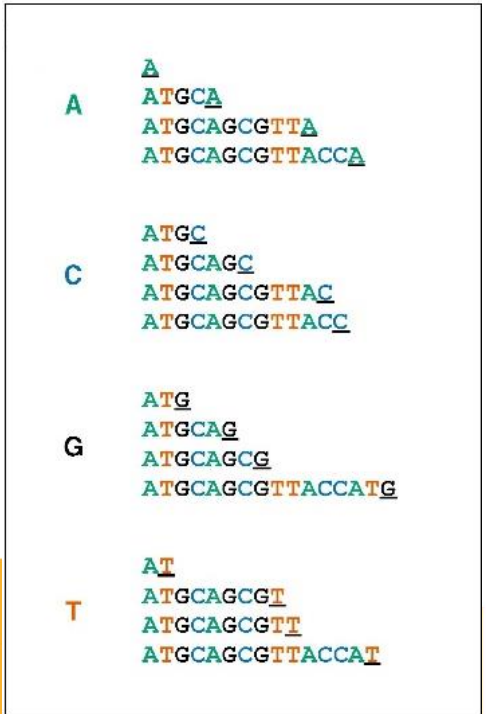
The Evolution of Sequencing



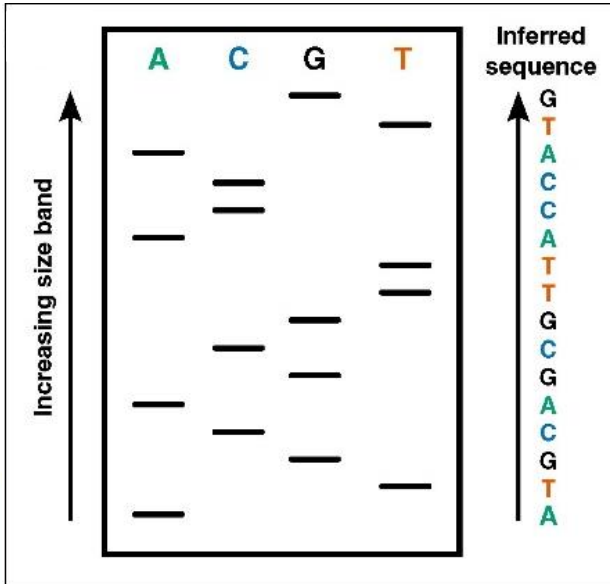
First generation sequencing (1977)

"Chain Termination"

- Incorporate radio- or fluorescently labeled dideoxynucleotide triphosphates (ddNTPs)
- ddNTPs terminate polymerization
- Run on gel to infer position



Sanger Sequencing

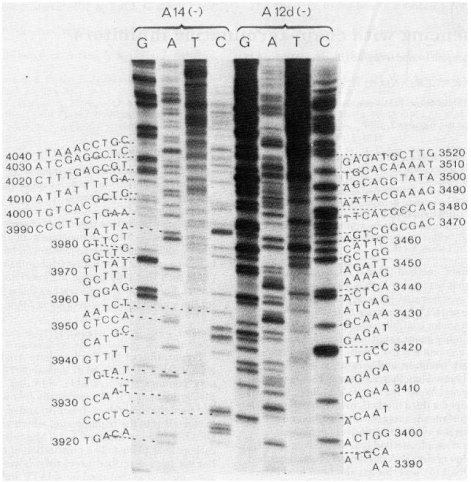


Heather and Chain (2016) *Genomics* 107:1-8

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

The Evolution of Sequencing

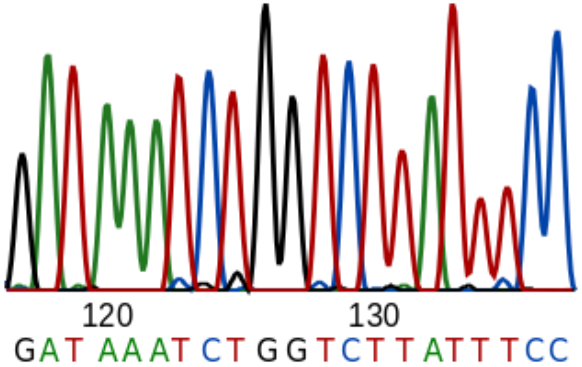
First generation sequencing (1977-present)



Sanger, Nicklen, Coulson (1977)
PNAS 74:5463-5467

Radioactive
Gel-based
One base per lane

Fluorescent
Gel-based
One sample per lane



https://en.wikipedia.org/wiki/Nucleic_acid_sequence

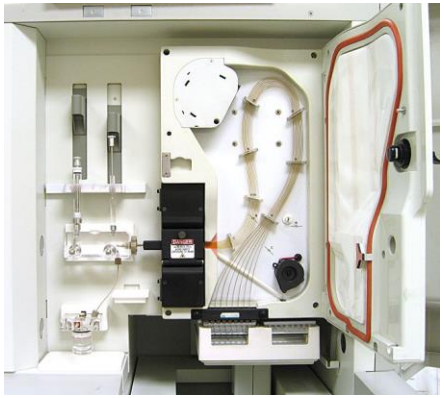
INSTITUTE FOR
COMPUTATIONAL BIOLOGY

The Evolution of Sequencing

First generation sequencing (1977-present)

ABI 370A (1986)

Photo credit: Giac83 on wikimedia



ABI 310
(1995)

Photo credit: Jacopo Werther on wikimedia

ABI 3700
(1999)

Springer (2006) American Laboratory News

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

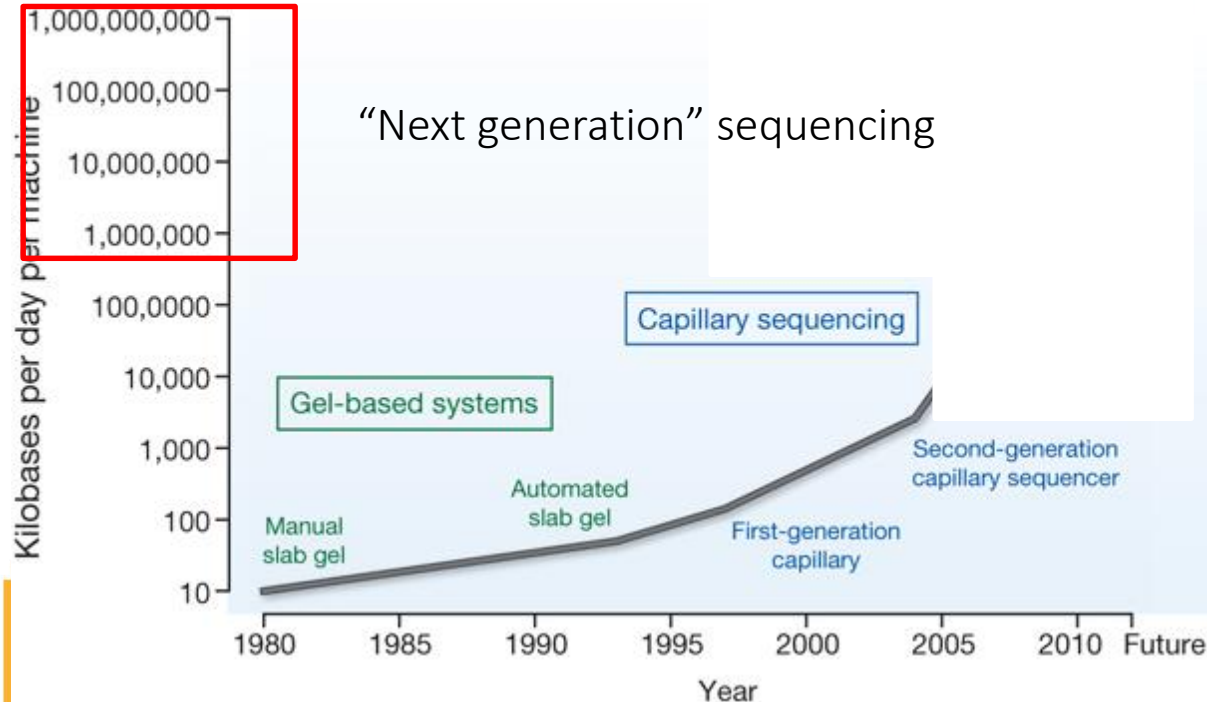
The Evolution of Sequencing

First generation sequencing (1977-present)

Throughput improved but limited by

- Separation methods
- Ability to parallel sequencing

Stratton, Campbell, Futreal (2009) *Nature* 458:719-724



"Next generation" sequencing

The Evolution of Sequencing

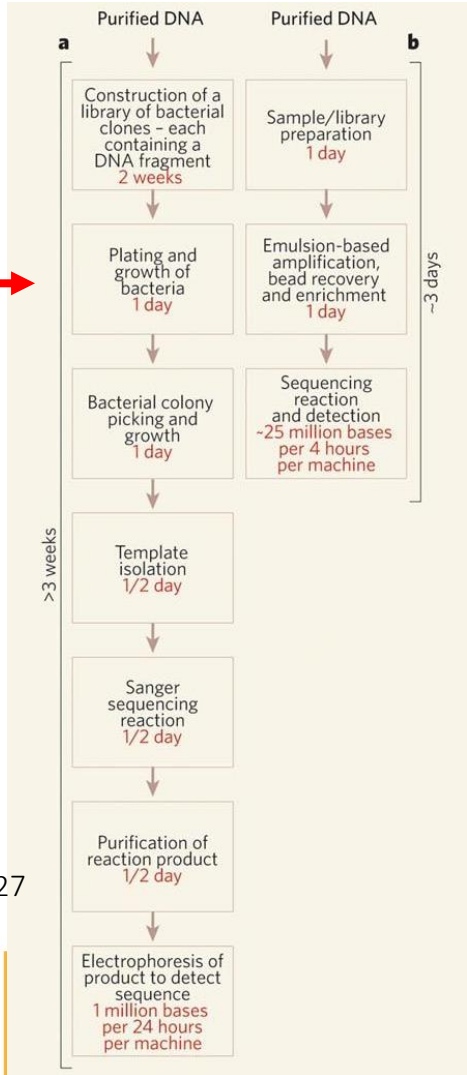
Next generation sequencing (2005-present)

Sequencing by synthesis (SBS):

- Tracks bases being added to growing DNA strand

Pyrosequencing (Roche 454)

- Uses natural nucleotides (dATP, dTTP, etc)
- Detection based on pyrophosphate released during polymerase reaction



Rogers and Venter (2005) *Nature* 437:326-327

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

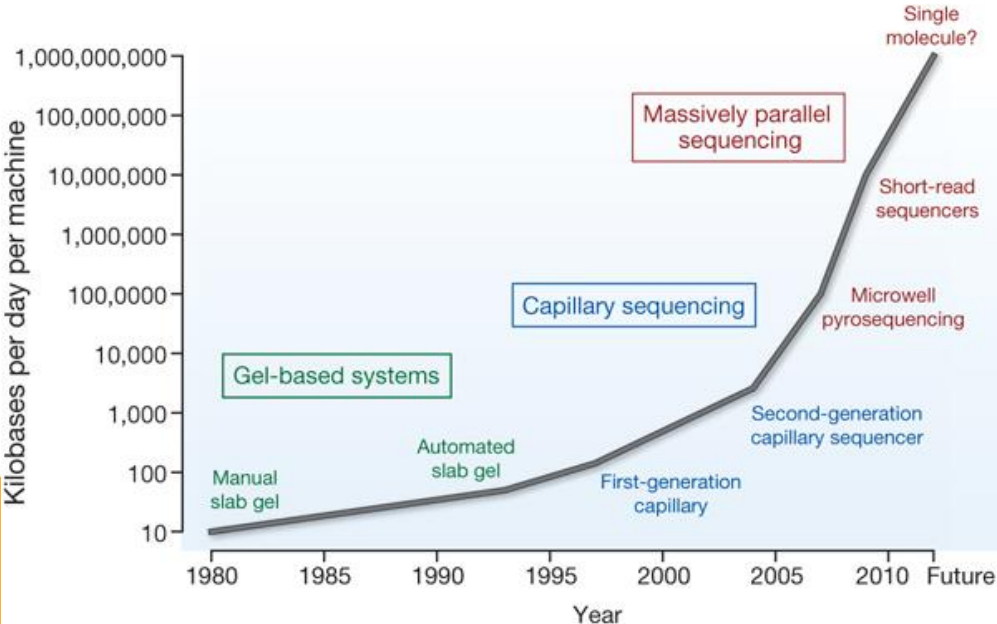
The Evolution of Sequencing

Next generation sequencing (2005-present)

Throughput improved greatly improved, but

- Reads were short (80-120 bases)
- Less accurate than Sanger (repeats, etc.)
- Need paired-end for *de novo* sequencing

Stratton, Campbell, Futreal (2009) *Nature* 458:719-724



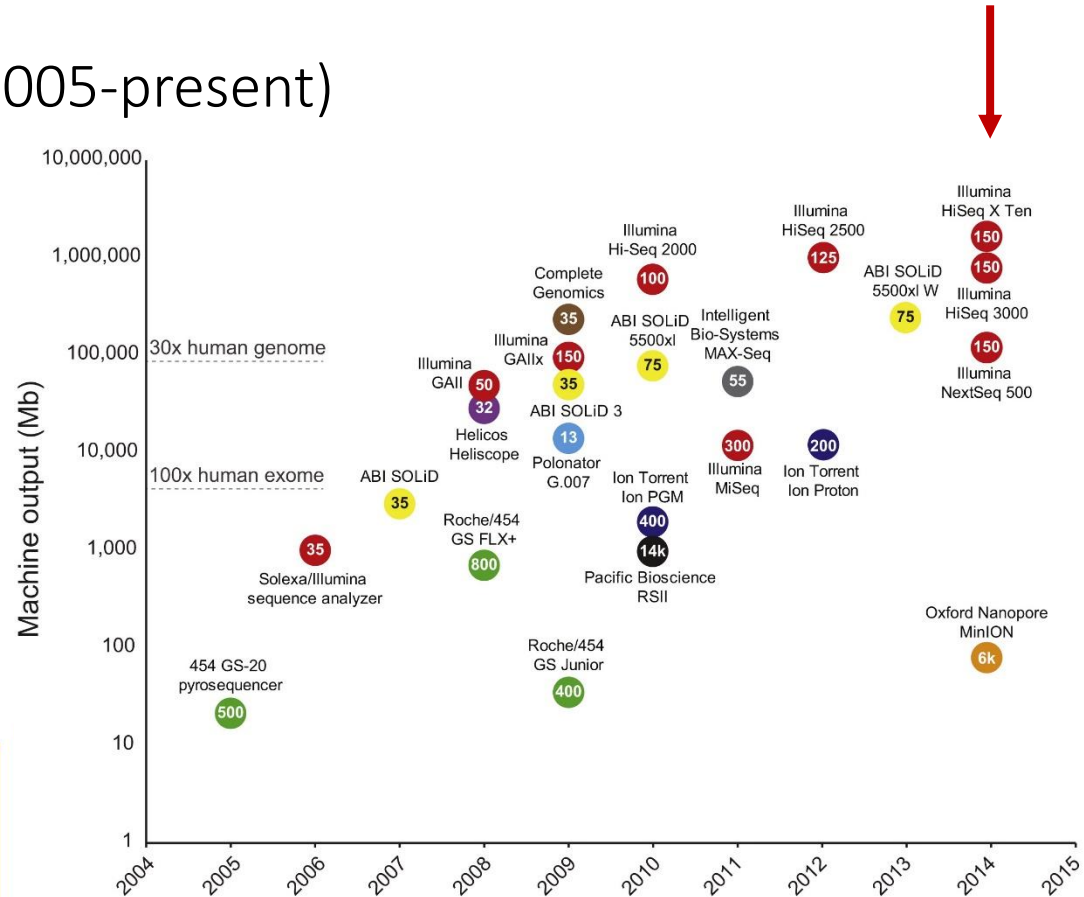
INSTITUTE FOR
COMPUTATIONAL BIOLOGY

The Evolution of Sequencing

Next generation sequencing (2005-present)

Trends in next-gen

- Longer reads
- Accuracy
- Price $$$$$$



Reuter, Spacek, Snyder (2015) *Molecular Cell* 58:586-597

INSTITUTE FOR COMPUTATIONAL BIOLOGY

The Evolution of Sequencing

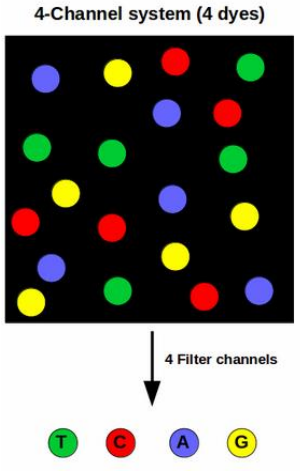
Illumina HiSeq X Ten (2014)

$1,000 genome

1.8 TB in <3 days

18K genomes/year

Genomes only



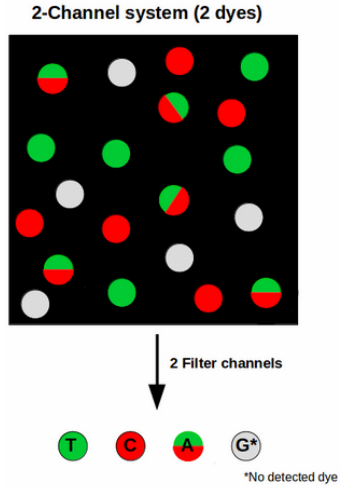
Illumina NovaSeq 6000 (2017)

$100 genome?

80 GB-6 TB in 1-2 days

3-48 whole genomes/run

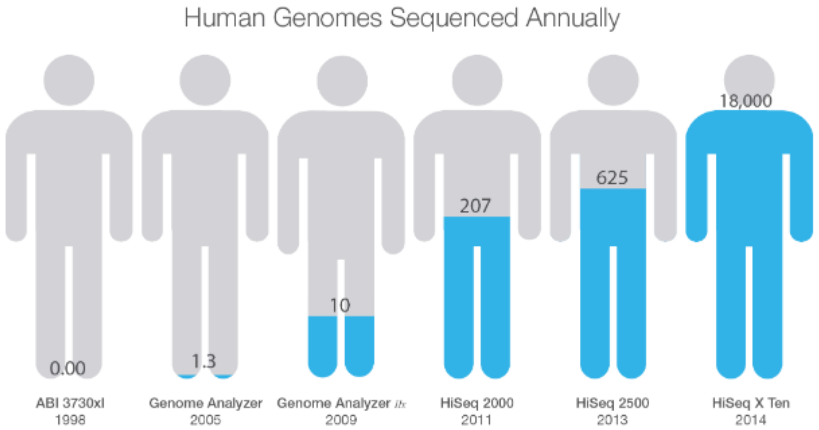
Genomes, exomes, RNA-Seq, ChIP-Seq, etc



https://www.ecseq.com/support/ngs/do_you_have_two_colors_or_four_colors_in_Illumina

Precision Medicine Research Sequencing

Enables population-scale whole genome sequencing for research

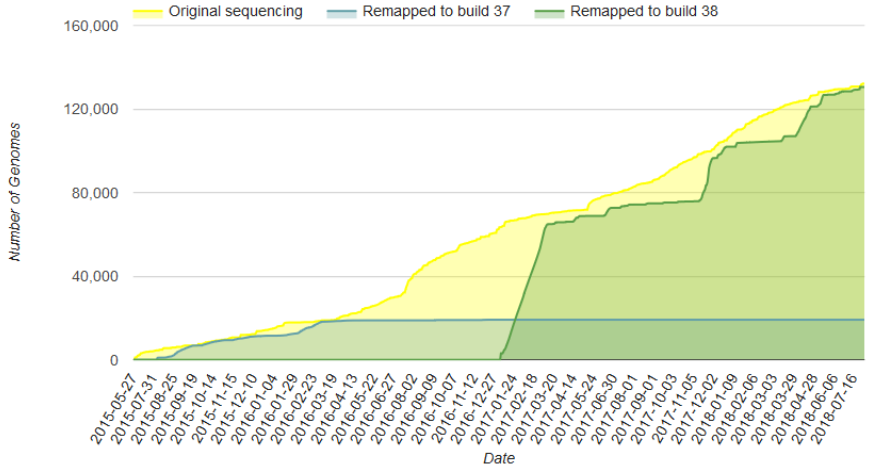


https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Precision Medicine Research

NHLBI Trans-Omics for Precision Medicine
Whole Genome Sequencing Program



http://nhlbi.sph.umich.edu/report/

biobank uk
Improving the health of future generations

50,000 (2018-2019)

MVP MILLION VETERAN PROGRAM

10,000 plus
>34,000 (2018-2019)

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Precision Medicine Research Sequencing

All of Us℠ The Precision Medicine Initiative®
THE FUTURE OF HEALTH BEGINS WITH YOU

Aims to ascertain
- At least 1 million US residents
- Representative of the population

Willing to provide
- Biospecimens
- Health data (EHRs)
- Lifestyle/behavior data

Longitudinally

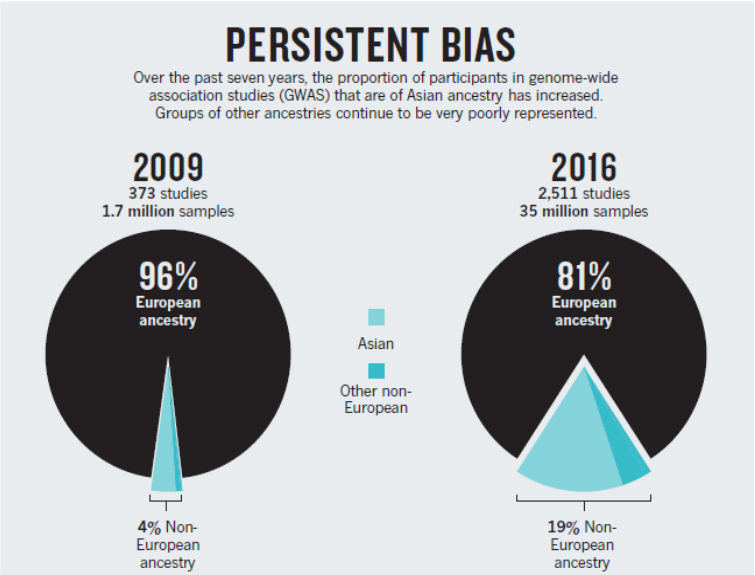
(Photo credit: Dr. Janina Jeff at the 2017 New Balance Bronx 10 Mile)

All of Us Research Program Newsletter September 2018

INSTITUTE FOR COMPUTATIONAL BIOLOGY

Data from Precision Medicine Research Sequencing

A catalog of single nucleotide variants (SNVs) in diverse populations



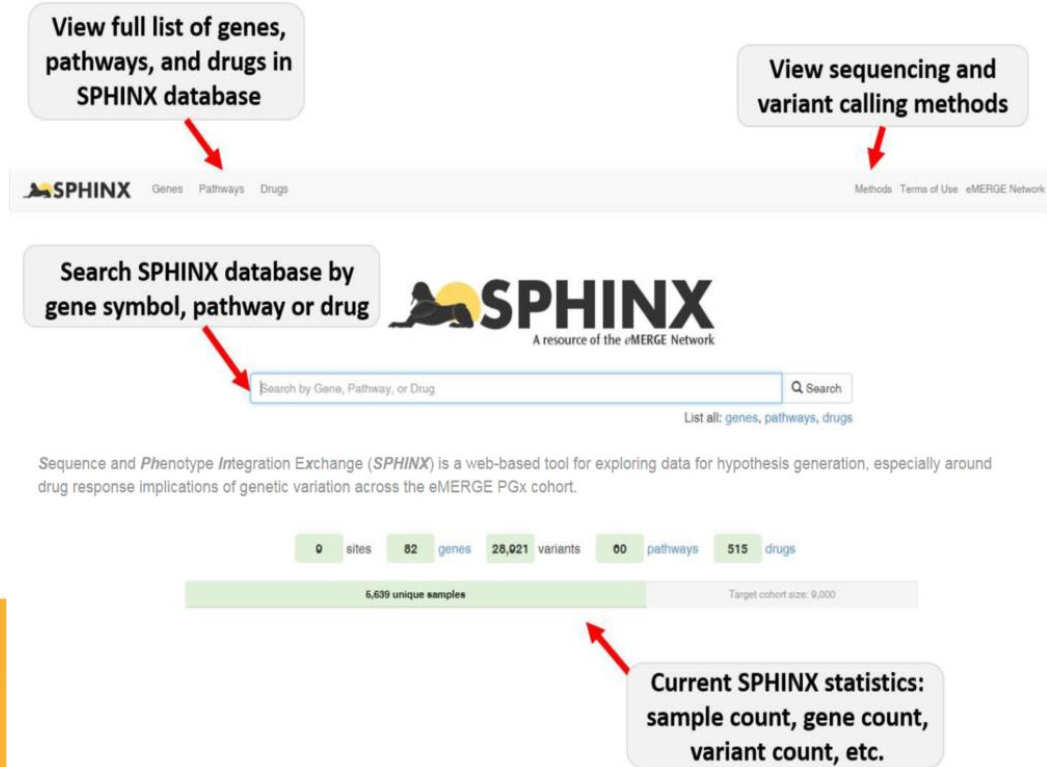
Popejoy and Fullerton (2016) *Nature* 538: 161-164

Data from Precision Medicine Research Sequencing

Variant servers with SNVs from diverse populations

Now includes eMERGEseq and PGRNseq with >24,000 samples and 1,138 genes

Bush et al (2016) *Clin Pharmacol Ther* 100:160-169



INSTITUTE FOR COMPUTATIONAL BIOLOGY

Data from Precision Medicine Research Sequencing

Variant servers with SNVs from diverse populations

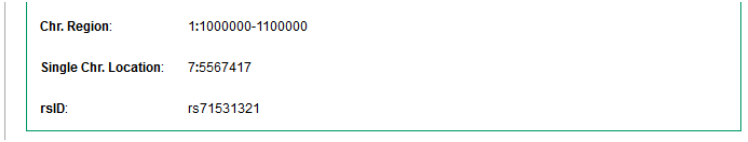
gnomAD browser beta | genome Aggregation Database

>6,000 sar
http://evs.

Search for a gene or variant or region

Example - Gene: PCSK9, Variant: 1-55516888-G-GA

>60,000 whole exome sequences
http://exac.broadinstitute.org/
>15,000 whole genome sequences
http://gnomad.broadinstitute.org/



INSTITUTE FOR
COMPUTATIONAL BIOLOGY

DATA FROM PRECISION MEDICINE

G

Bravo

Powered by TOPMed Freeze5 on GRCh38

(This dataset includes 463 million variants on 62784 individuals. Click here to switch to Freeze3a on GRCh37/hg19.)

Search for a gene or variant or region

Search

Gene: PCSK9, Transcript: ENST00000407236, Variant: chr22-16389447-A-G or rs34747326, Region: chr1-55030529-55075873

https://bravo.sph.umich.edu/freeze5/hg38/

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Data from Precision Medicine

Variant: 11:116830637 C / T (*APOC3*: p.Arg19Ter, p.Arg37Ter)

Summary | Annotations | Frequencies | Sequence Depth | Genotype Quality | Raw Sequences | Site Quality Metrics

Summary

Filter Status	PASS
Existing Variation	rs76353203
Allele Frequency	0.0008362
Allele Count	105 / 125568
Homozygous Alt Count	0
CADD	32.0
UCSC	11-116830637-C-T
ClinVar	11:116830637 is not in ClinVar Open rs76353203 in ClinVar

Annotations

This variant falls on 5 transcripts belonging to 1 gene:

stop gained

- *APOC3*
  - ENST00000227667 (p.Arg19Ter) LoF: High-confidence
  - ENST00000433777 (p.Arg19Ter) LoF: High-confidence
  - ENST00000375345 (p.Arg37Ter) LoF: High-confidence
  - ENST00000630701 (p.Arg37Ter) LoF: High-confidence

splice region

- *APOC3* - ENST00000470144 (n.87C>T)

Note: This list may not include additional transcripts in the same gene that the variant does not overlap.

Example TOPMed variant data

Frequency Table

Population	Allele Frequency
1000G African	Not available
1000G American	Not available
1000G East Asian	Not available
1000G European	Not available
1000G South Asian	Not available
TOPMed Freeze5 on GRCh38	0.0008362

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Sequencing and Precision Medicine Research Examples

General genotype-phenotype studies have been somewhat disappointing

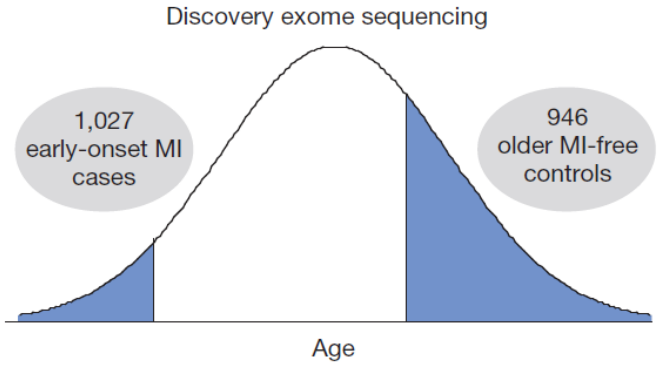
Rare variant analysis will require

- Large (>10K) sample size
- "Careful" statistical analyses

LETTER

doi:10.1038/nature13917

Exome sequencing identifies rare *LDLR* and *APOA5* alleles conferring risk for myocardial infarction



Do et al (2014) *Nature* 518:102-106

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Sequencing and Precision Medicine Research Examples

General genotype-phenotype studies have been somewhat disappointing

"At these sample sizes and for these phenotypes, the incremental value of WGS for discovery is limited but WGS permits simultaneous assessment of monogenic and polygenic models to severe hypercholesterolemia."

Natarajan et al (2018) *Nat Comm* 9:3391

ARTICLE

DOI: 10.1038/s41467-018-05747-8 OPEN

Deep-coverage whole genome sequences and blood lipids among 16,324 individuals

Precision Medicine Research Sequencing

??????

(Photo credit: Dr. Janina Jeff at the 2017 New Balance Bronx 10 Mile)

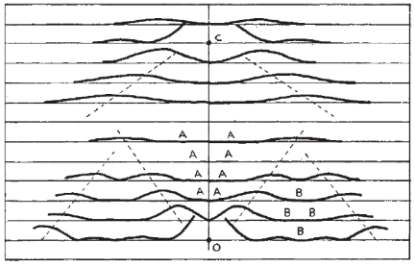
INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Sequencing and Precision Medicine Research Challenges

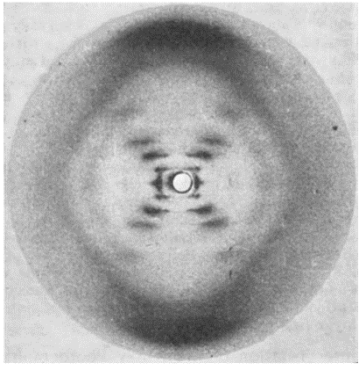
- Pipelines and expertise
- Functional annotation

} Bioinformatics

- Statistical methods *****
- Storage, secure access, and retrieval



Wilkins, Stokes, Wilson
(1953) *Nature* 171:738-740



Franklin and Gosling (1953)
Nature 171:740-741

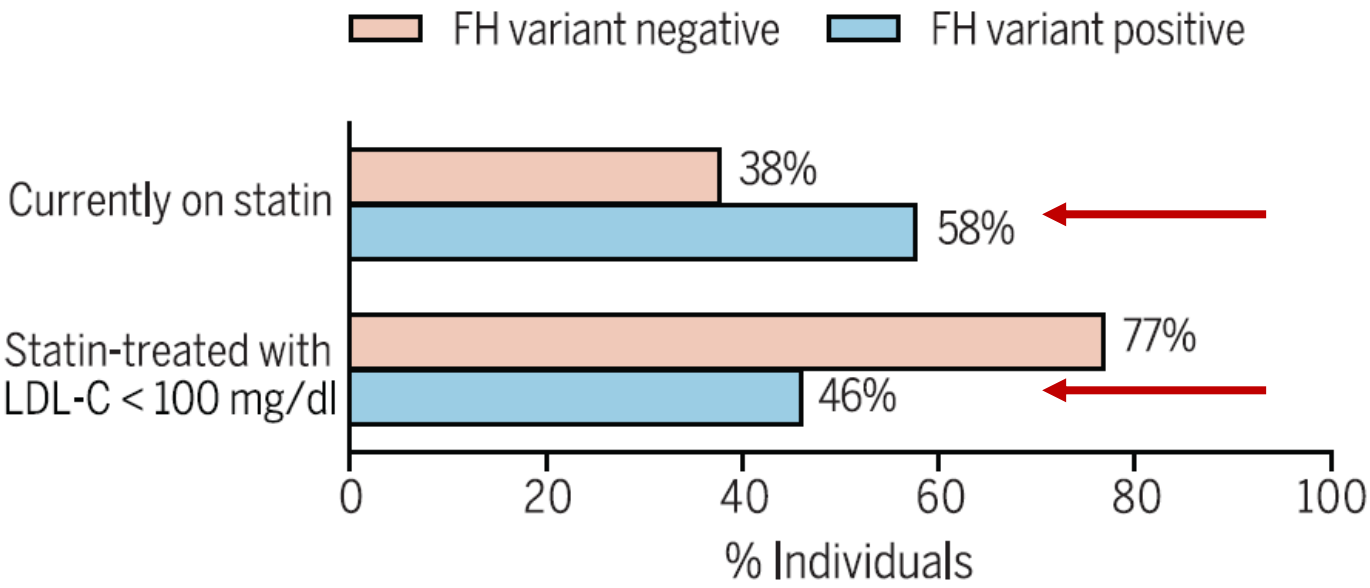
Watson and Crick
(1953) *Nature*
171:737-738

QUESTIONS?

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Sequencing and Precision Medicine Research Examples

Population studies and candidate genes



Abul-Husn et al (2016)
Science 354:1550

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

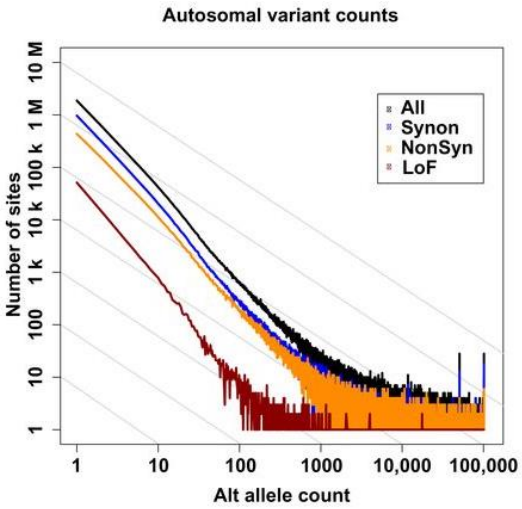
Data from Precision Medicine Research Sequencing

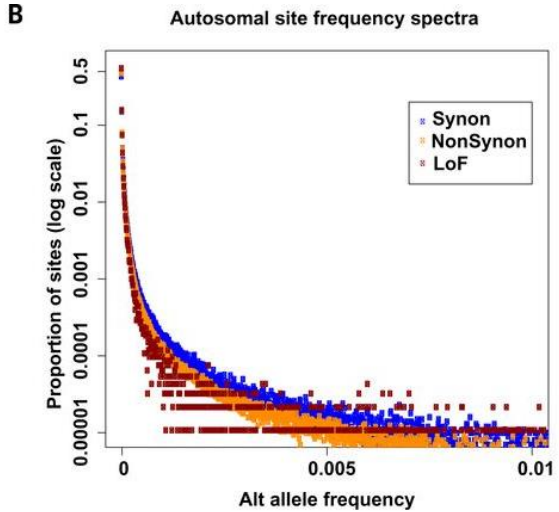
A catalog of single nucleotide variants (SNVs)

RESEARCH ARTICLE

HUMAN GENETICS

Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study





Dewey et al (2016) Science 354: aaf6814

INSTITUTE FOR COMPUTATIONAL BIOLOGY

Data from Precision Medicine Research Sequencing

SPECIAL ARTICLE

Genetic Misdiagnoses and the Potential for Health Disparities

Arjun K. Manrai, Ph.D., Birgit H. Funke, Ph.D., Heidi L. Rehm, Ph.D., Morten S. Olesen, Ph.D., Bradley A. Maron, M.D., Peter Szolovits, Ph.D., David M. Margulies, M.D., Joseph Loscalzo, M.D., Ph.D., and Isaac S. Kohane, M.D., Ph.D.

N Engl J Med 2016; 375:655-665 | August 18, 2016 | DOI: 10.1056/NEJMsa1507092

REPORT

Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans

Giulio Genovese[1,2,*], David J. Friedman[1,3,*], Michael D. Ross[4], Laurence Lecordier[5], Pierrick Uzureau[5], Barry I. Freedman[6], Donald W. Bowden[7,8], Carl D. Langefeld[8,9], Taras K. Oleksyk[10], Andrea L. Uscinski Knob[4], Andrea J. Bernhardy[1], Pamela J. Hicks[7,8], George W. Nelson[11], Benoit Vanhollebeke[5], Cheryl A. Winkler[12], Jeffrey B. Kopp[11], Etienne Pays[5,†], Martin R. Pollak[1,13,†]

+ Author Affiliations

†To whom correspondence should be addressed. E-mail: mpollak@bidmc.harvard.edu (M.R.P.); epays@ulb.ac.be (E.P.)

* These authors contributed equally to this work.

Science 13 Aug 2010:
Vol. 329, Issue 5993, pp. 841-845
DOI: 10.1126/science.1193032

INSTITUTE FOR
COMPUTATIONAL BIOLOGY