



Sequencing and Introduction to Hail: Overview

2021 Virtual Workshop on Statistical Genetics Methods for Human Complex Traits

June 16th, 2021 (practical)

Hosted by the Institute for Behavioral Genetics, University of Colorado, Boulder

Kumar Veerapen, PhD

Hail Support and Community Outreach Manager

Tim Poterba, Carolin Diaz, Dan Howrigan, John Compitello



<https://hail.is>
@mkveerapen / @hailgenetics
veerapen@broadinstitute.org
#scalableGenomics
#hailGenetics #ATGUstrong

Learning Objectives

- To understand the overview of DNA sequencing methods
- To capture the need for Hail in the analysis of genomic datasets
- To be able to use basic Hail functions
- To apply basic GWAS analysis techniques using Hail on their own datasets
- To describe the use of PCA in Hail to decipher ancestries
- To obtain resources to further explore the extent of Hail capabilities
- To learn how to use Hail on public compute clouds

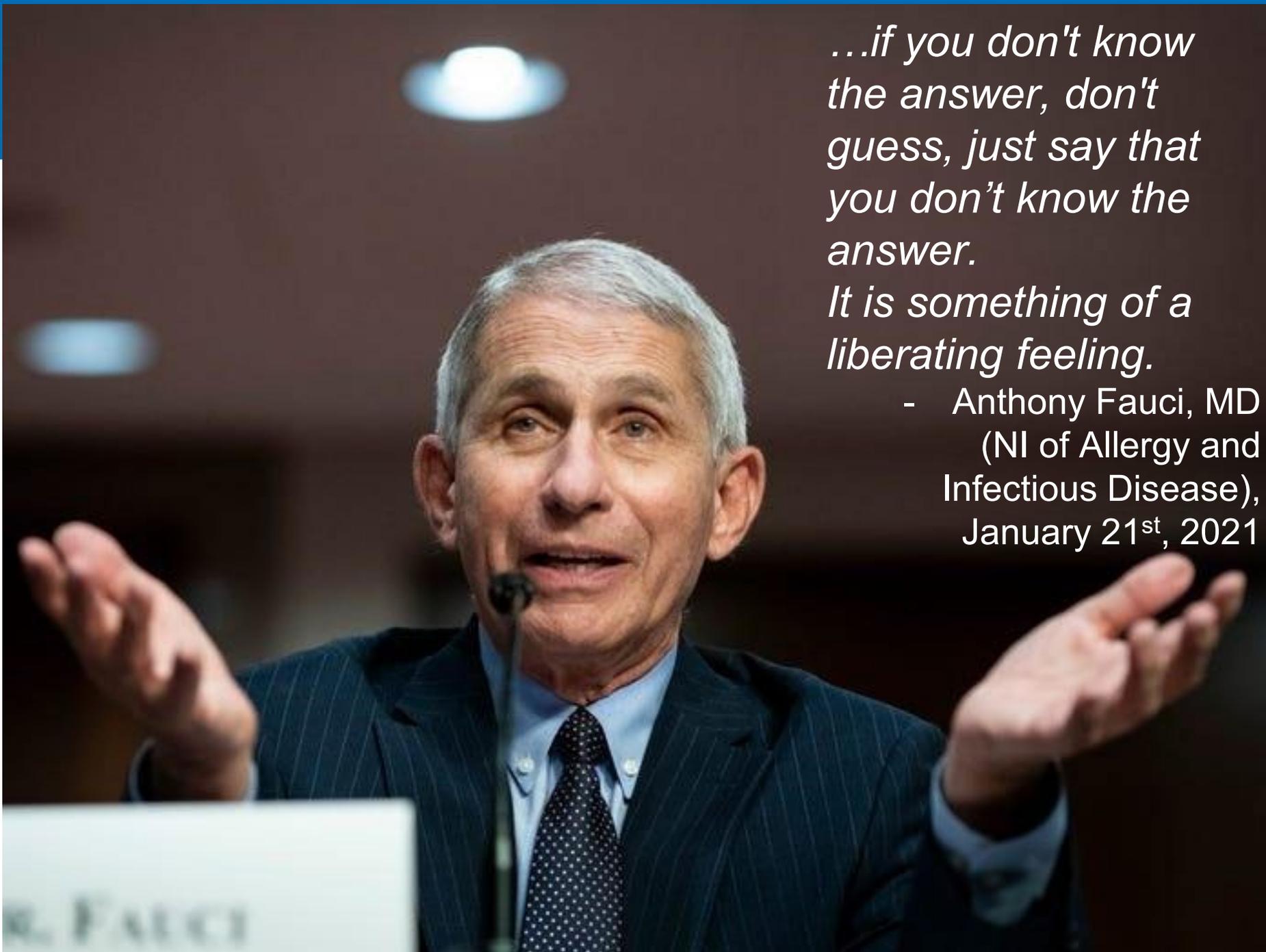
How we breakdown our sections:

Lectures (on demand)

- Traditional sequencing technology
- “Next-generation” sequencing technology
- “Next-generation” sequencing technology (informatics)
- Analysis of sequencing data using Hail
- How can I use Hail? (practicum)
- Unlocking the power of the cloud with Hail

Practicum (in “person” / real time)

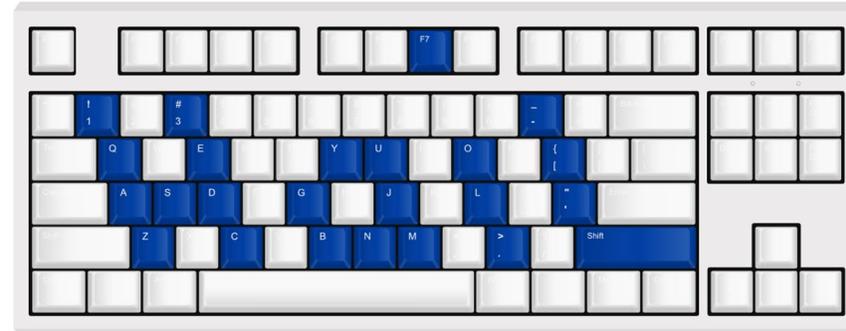
- **Practical 1:**
 - Import, joining data together, and quality control (QC)
- **Practical 2:**
 - Genome Wide Association Studies (GWAS)
- **Practical 3:**
 - Principal Component Analysis (PCA) and Deciphering Ancestry



*...if you don't know
the answer, don't
guess, just say that
you don't know the
answer.*

*It is something of a
liberating feeling.*

- Anthony Fauci, MD
(NI of Allergy and
Infectious Disease),
January 21st, 2021



Sequencing and Introduction to Hail: Overview

2021 Virtual Workshop on Statistical Genetics Methods for Human Complex Traits

June 16th, 2021 (practical)

Hosted by the Institute for Behavioral Genetics, University of Colorado, Boulder

Kumar Veerapen, PhD

Hail Support and Community Outreach Manager

Tim Poterba, Carolin Diaz, Dan Howrigan, John Compitello



<https://hail.is>
@mkveerapen / @hailgenetics
veerapen@broadinstitute.org

#scalableGenomics

#hailGenetics #ATGUstrong

hail

Sequencing and Introduction to Hail: Traditional Sequencing Technology

2021 Virtual Workshop on Statistical Genetics Methods for Human Complex Traits

June 16th, 2021 (practical)

Hosted by the Institute for Behavioral Genetics, University of Colorado, Boulder

Kumar Veerapen, PhD

Hail Support and Community Outreach Manager

Tim Poterba, Carolin Diaz, Dan Howrigan, John Compitello



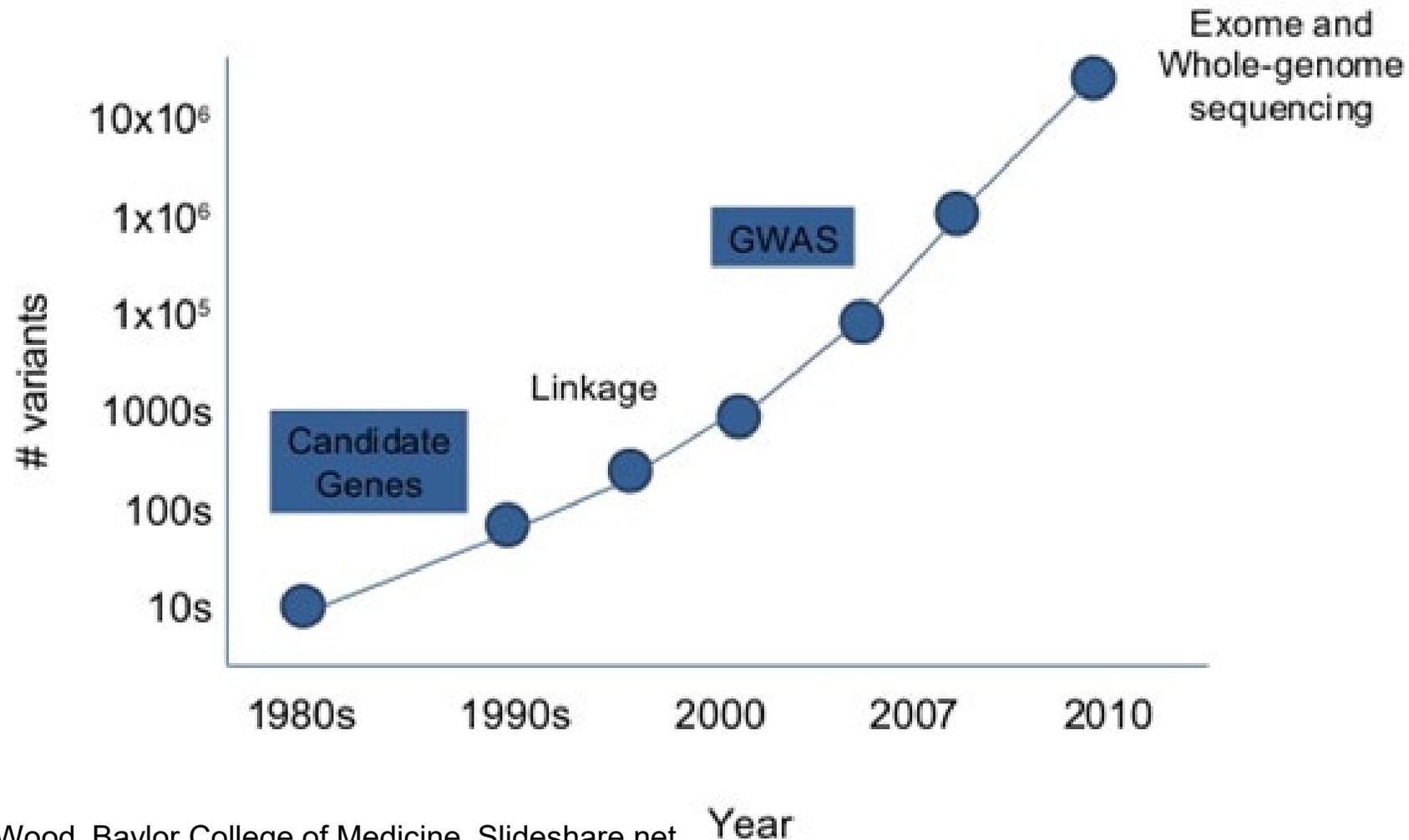
<https://hail.is>
@mkveerapen / @hailgenetics
veerapen@broadinstitute.org
#scalableGenomics
#hailGenetics #ATGUstrong

Learning Objectives

- To understand the overview of traditional DNA sequencing methods
 - To understand the basic concepts of PCR
 - To appreciate the concepts of PCR evolving to Sanger sequencing method
 - To understand the difference between sequencing and genotyping
 - To comprehend the limitations of traditional sequencing methods

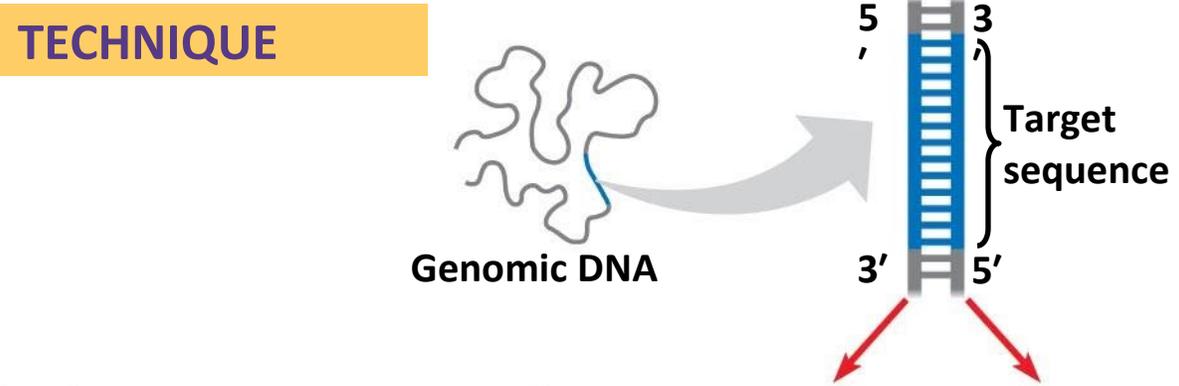
Technological Growth in Genetics and Genomics

Growth of the field of human genetics

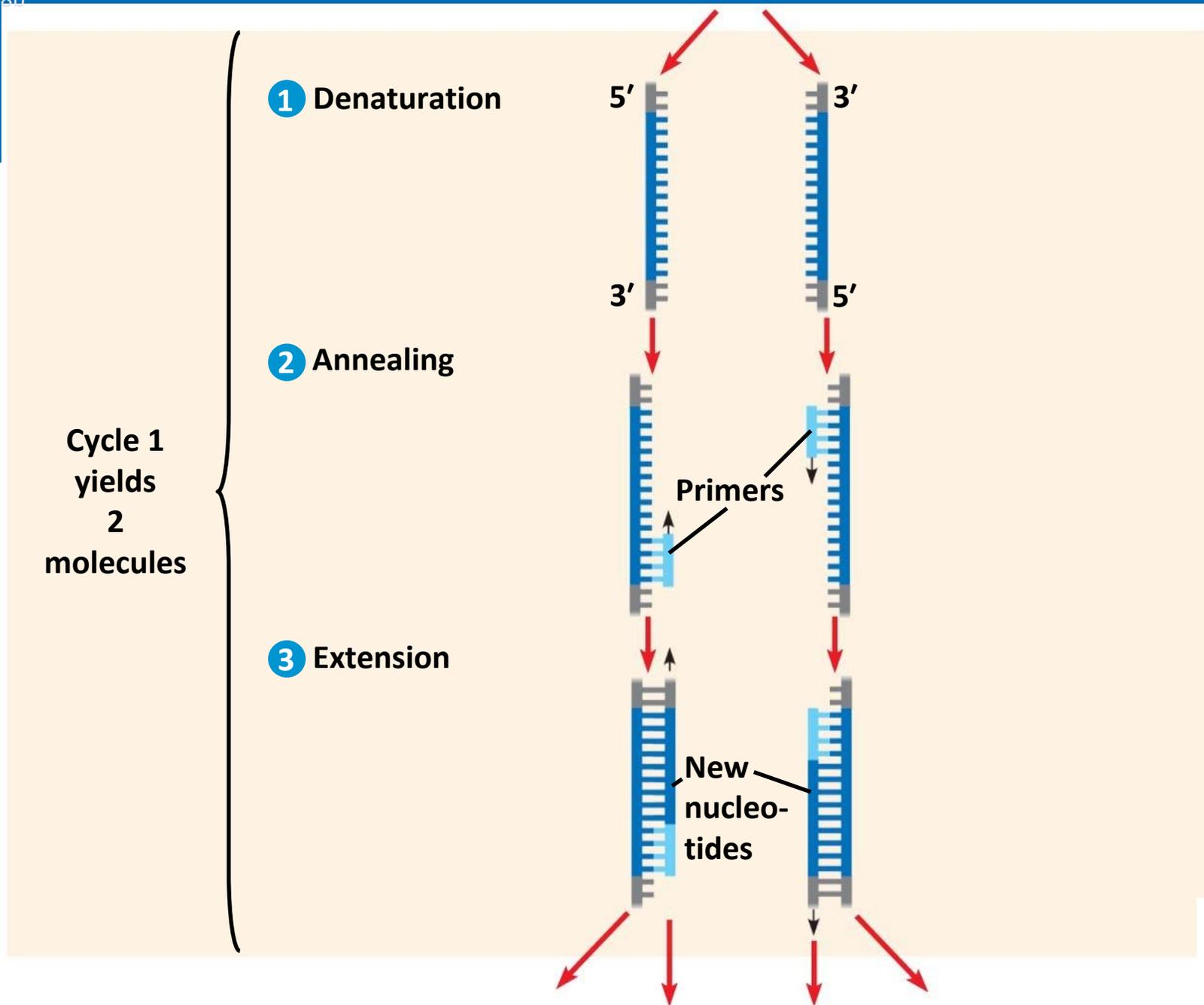


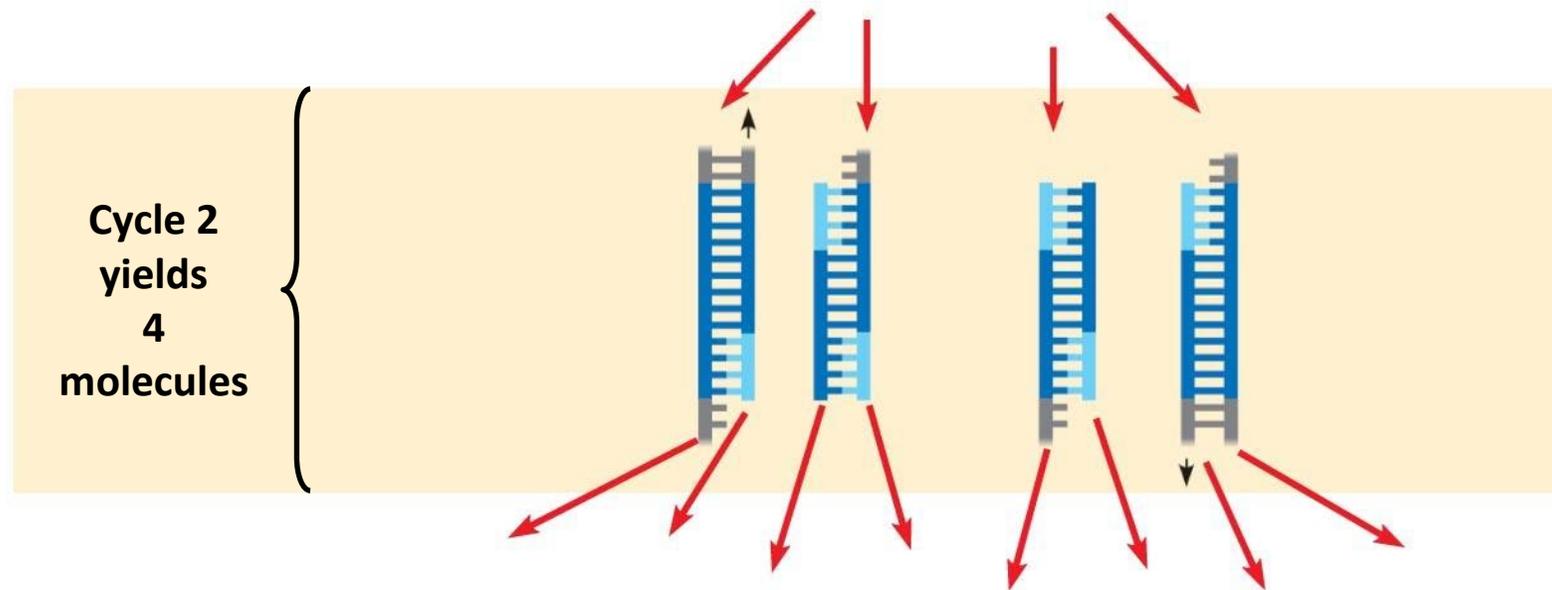
Amplifying DNA *in Vitro*: The Polymerase Chain Reaction (PCR)

- The **polymerase chain reaction, PCR**, can produce many copies of a specific target segment of DNA
- A three-step cycle—heating, cooling, and replication—brings about a chain reaction that produces an exponentially growing population of identical DNA molecules
- Kary Mullis -- December 16, 1983

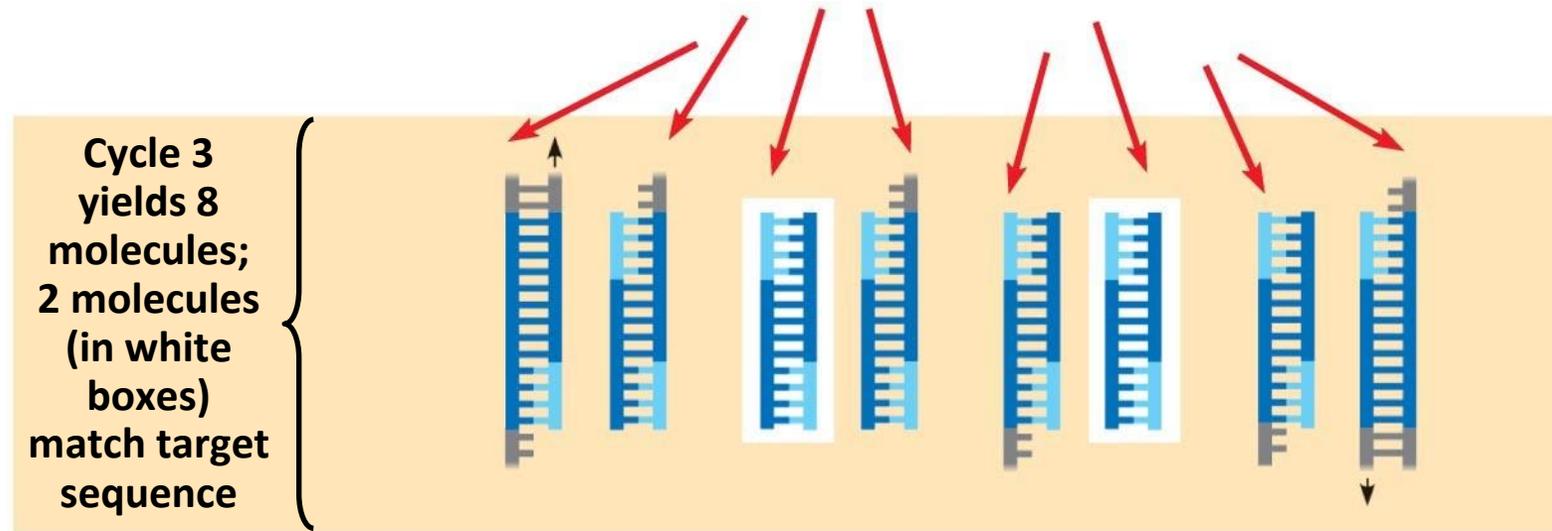


Copyright © 2008 Pearson Education, Inc., publishing as Pearson Benjamin Cummings.



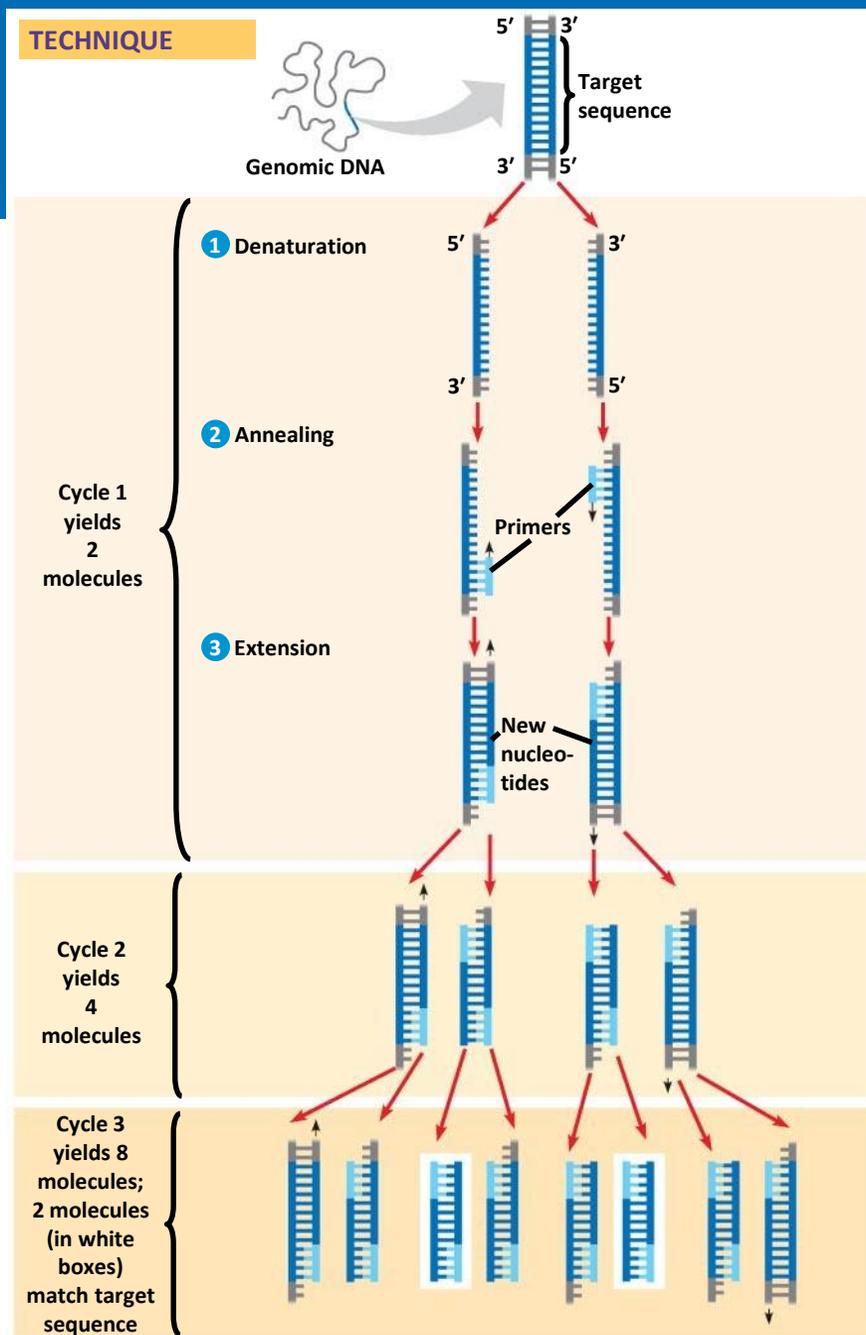


Copyright © 2008 Pearson Education, Inc., publishing as Pearson Benjamin Cummings.



Copyright © 2008 Pearson Education, Inc., publishing as Pearson Benjamin Cummings.

Fig. 20-8

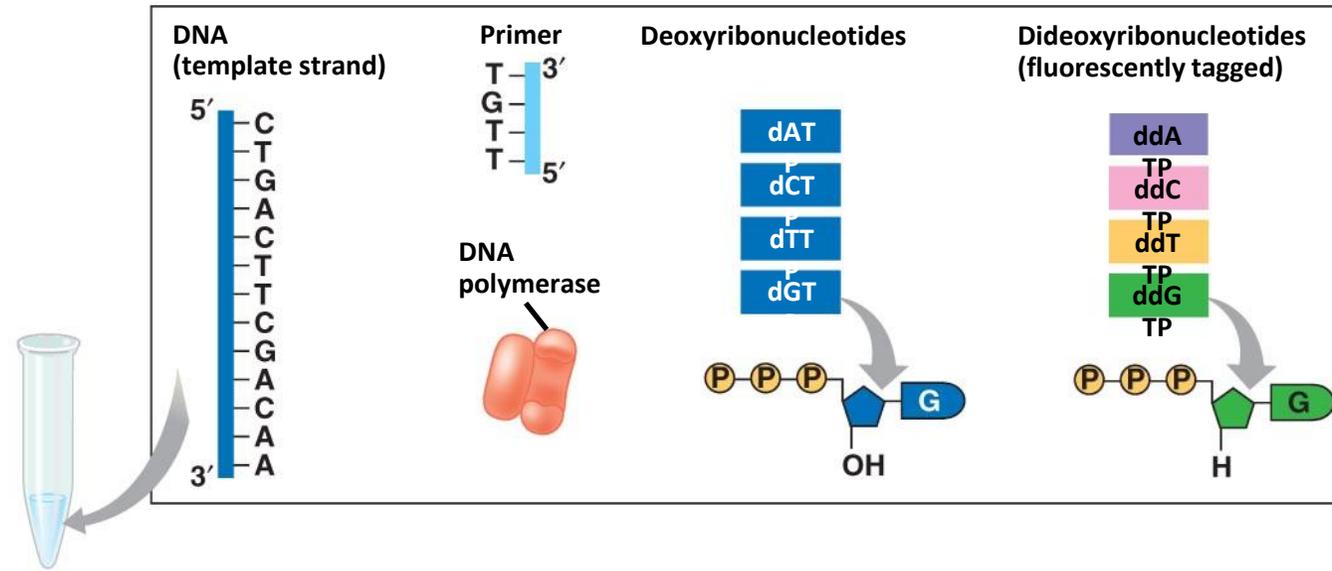


Copyright © 2008 Pearson Education, Inc., publishing as Pearson Benjamin Cummings.

DNA Sequencing

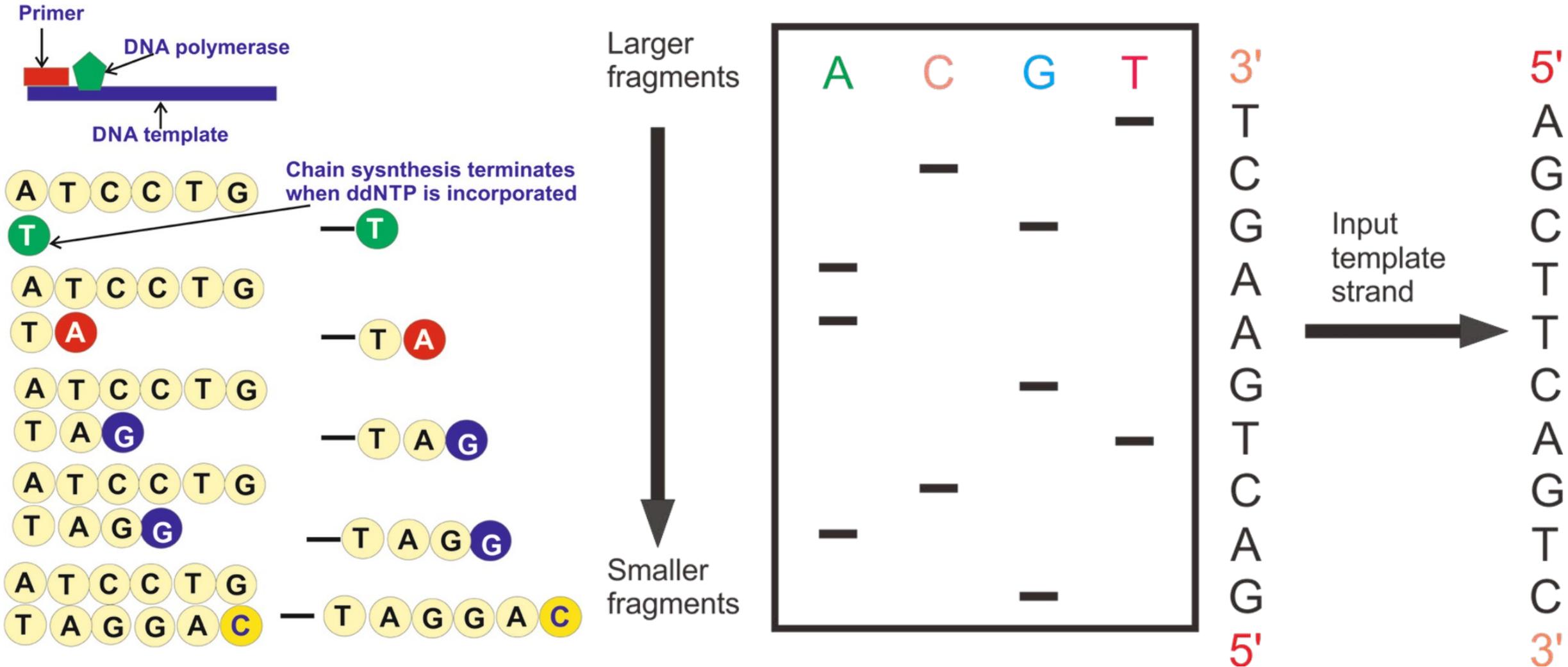
- Relatively short DNA fragments can be sequenced by the *dideoxy chain termination method*
- Modified nucleotides called dideoxynucleotides (ddNTP) attach to synthesized DNA strands of different lengths
- Each type of ddNTP is tagged with a distinct fluorescent label that identifies the nucleotide at the end of each DNA fragment
- The DNA sequence can be read from the resulting spectrogram

TECHNIQUE

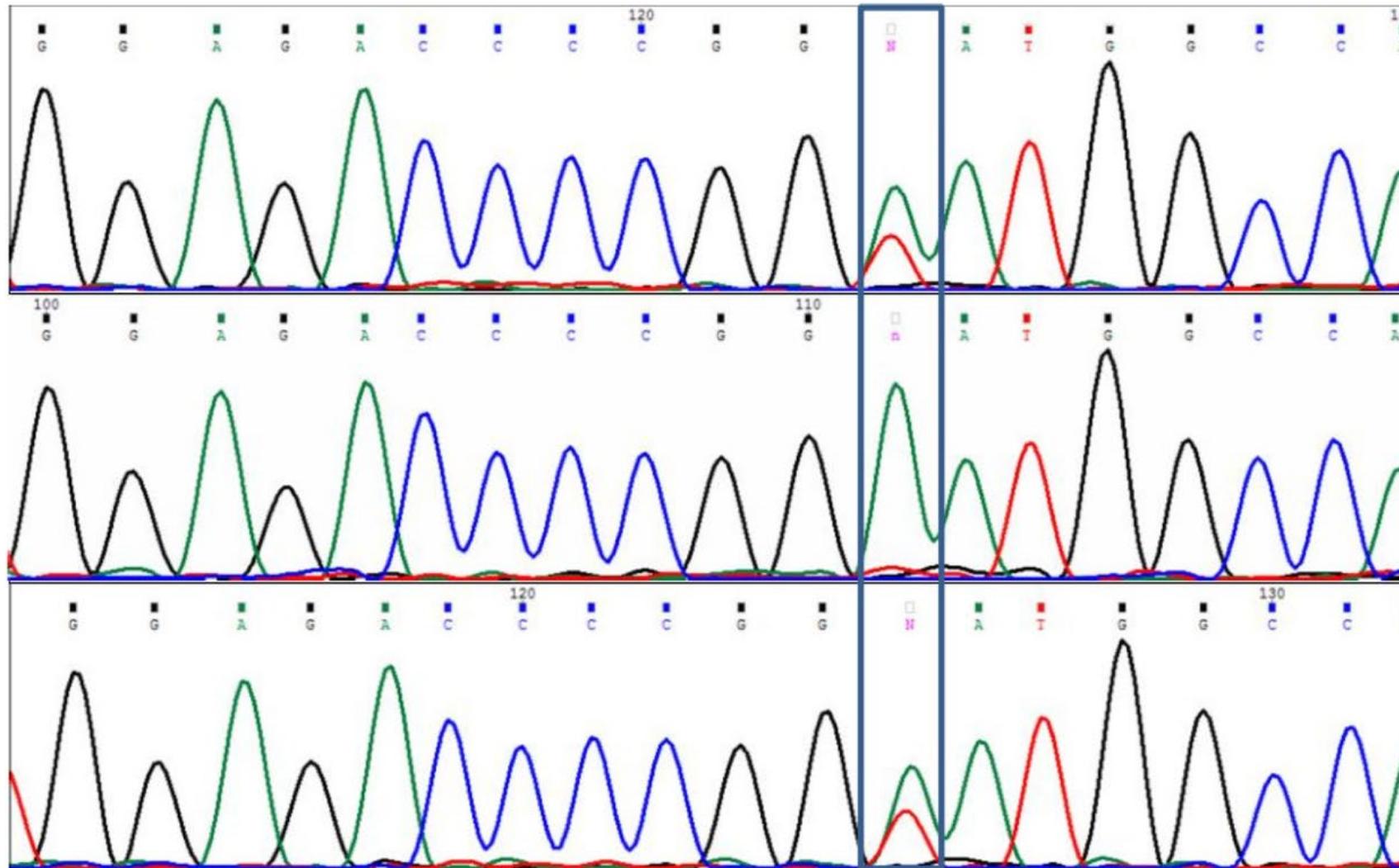


Copyright © 2008 Pearson Education, Inc., publishing as Pearson Benjamin Cummings.

Sanger sequencing



Sanger Sequencing Chromatogram

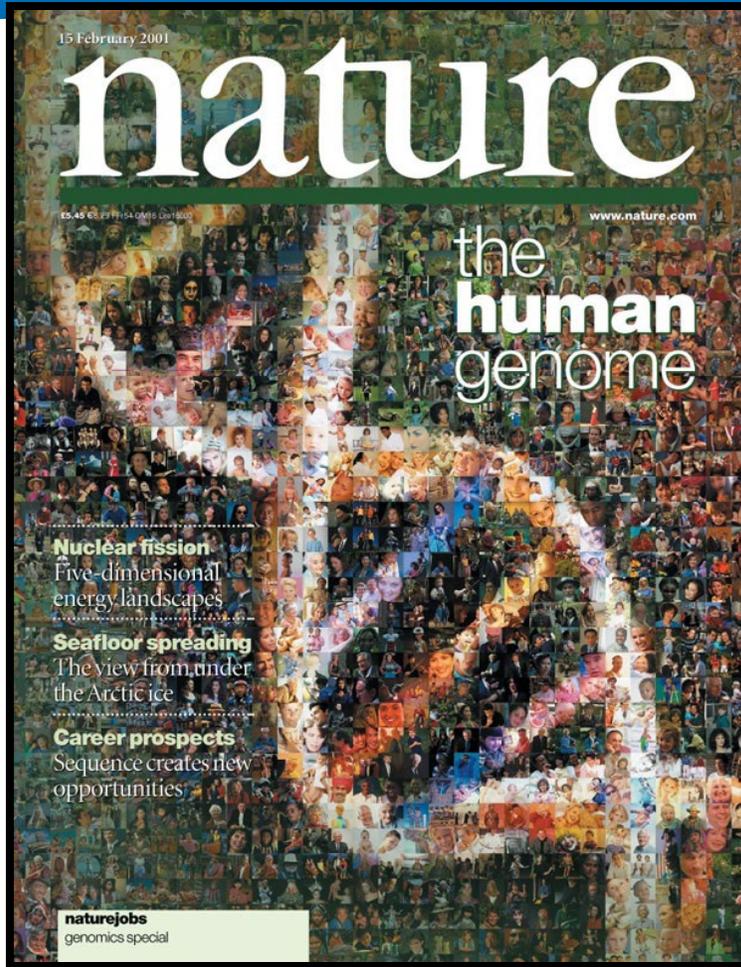


Mother

Proband

Father

The Reference Human Genome Sequence



15 February 2001



16 February 2001

Slide borrowed
from HSPH
GINGER program



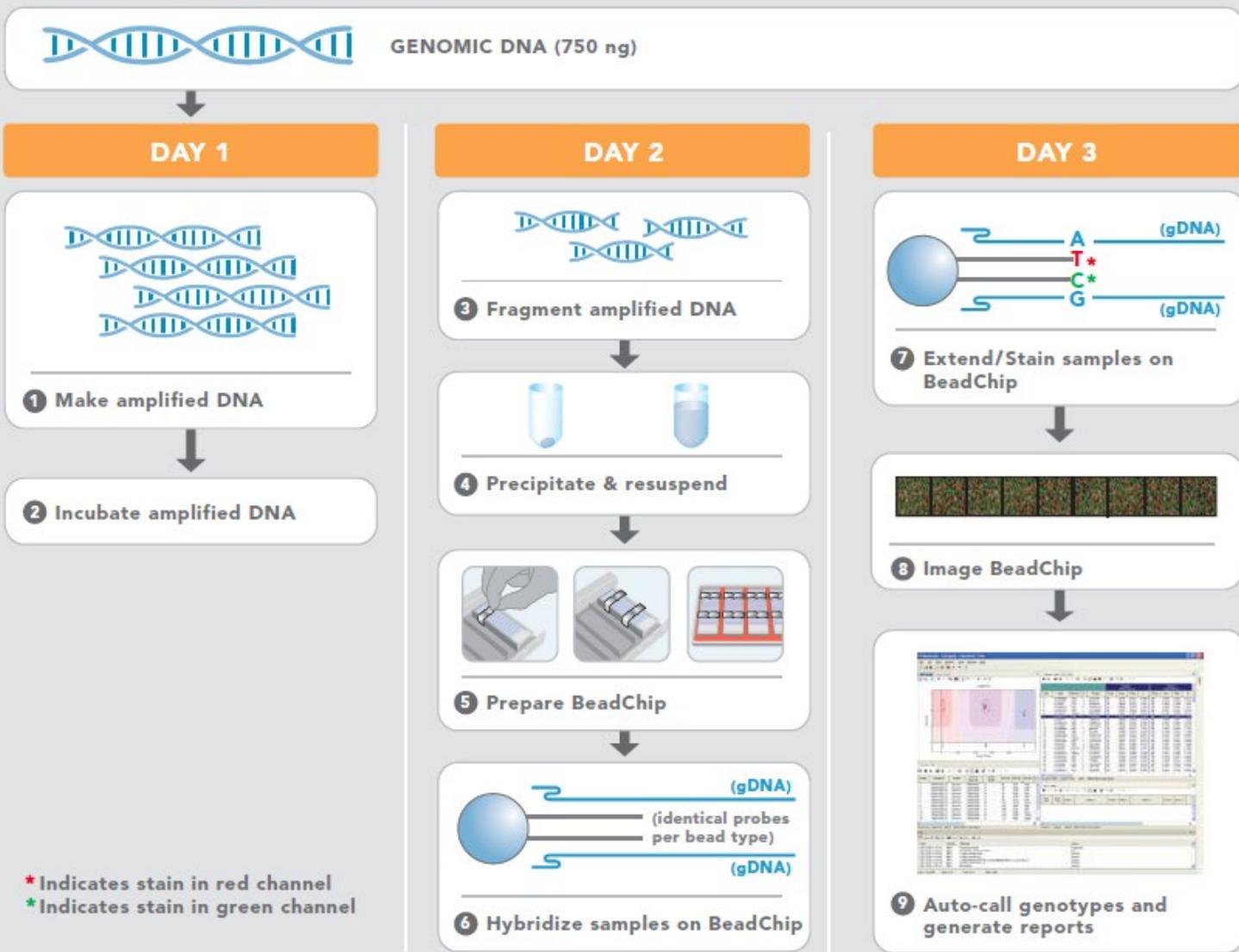
HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

<http://www.sciencemag.org/content/331/6019.toc>

<http://www.nature.com/nature/supplements/collections/humangenome/commentaries/>

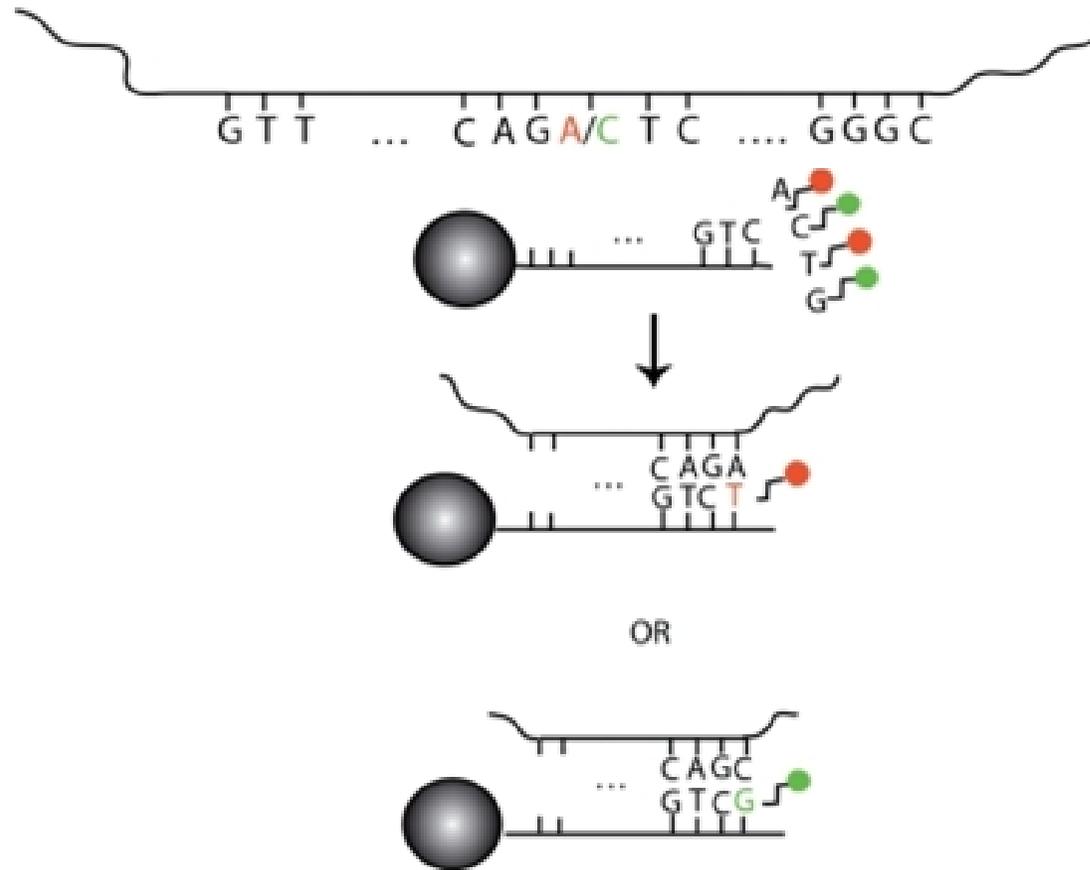
One marker/variant is not enough! There are 3 billion bp. Hence, high throughput genotyping

FIGURE 1: INFINIUM II ASSAY PROTOCOL



100,000 – 2.5 million markers
in the genome on a single
chip!

Genotyping chip



Single nucleotide polymorphism arrays:
a decade of biological, computational and technological advances

Sanger sequencing: Applications

- Targeting smaller genomic regions in a larger number of samples
- Sequencing of variable regions
- Validating results from next-generation sequencing (NGS) studies
- Verifying plasmid sequences, inserts, mutations
- HLA typing
- Genotyping of microsatellite markers
- Identifying single disease-causing genetic variants

Disadvantages

- Short sequence (~500-700 bp)
- Not great in the first 15 to 40 bases because that is where the primer binds.
- quality degrades after 700 to 900 bases.

Next Gen sequencing technologies

	Feature generation	Sequencing by synthesis	Cost per megabase	Cost per instrument	1° error modality	Read-length
454	Emulsion PCR	Polymerase (pyrosequencing)	~ \$60	\$500,000	Indel	250 bp
Solexa	Bridge PCR	Polymerase (reversible terminators)	~ \$2	\$430,000	Subst.	36 bp
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)	~ \$2	\$591,000	Subst.	35 bp
Polonator	Emulsion PCR	Ligase (nonamers)	~ \$1	\$155,000	Subst.	13 bp
HeliScope	Single molecule	Polymerase (asynchronous extensions)	~ \$1	\$1,350,000	Del	30 bp

Shendure J. & Hanlee J. (2008). *Nature*



Applied Biosystems
ABI 3730XL
1 Mb / day



Roche / 454
Genome Sequencer FLX
100 Mb / run



Illumina / Solexa
Genetic Analyzer
2000 Mb / run



Applied Biosystems
SOLiD
3000 Mb / run



Applied Biosystems
ABI 3730XL
1 Mb / day



Roche / 454
Genome Sequencer FLX
100 Mb / run



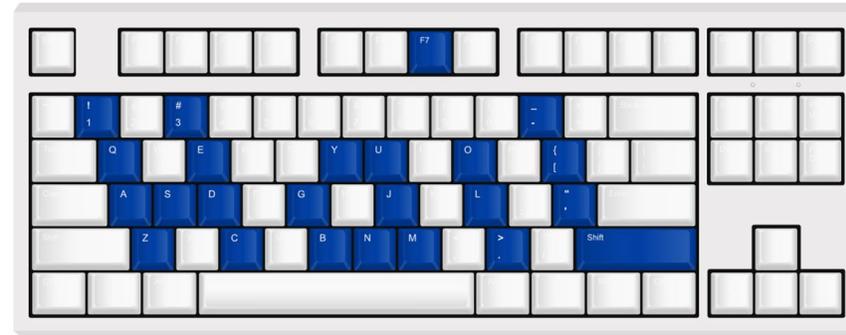
Illumina / Solexa
Genetic Analyzer
2000 Mb / run



Applied Biosystems
SOLiD
3000 Mb / run

Learning Outcomes

- You can describe the overview of traditional DNA sequencing methods
 - You understand the basic concepts of PCR
 - You appreciate the concepts of PCR evolving to Sanger sequencing method
 - You grasp the difference between genotyping and sequencing
 - You comprehend the limitations of traditional sequencing methods



Sequencing and Introduction to Hail: Traditional Sequencing Technology

2021 Virtual Workshop on Statistical Genetics Methods for Human Complex Traits

June 16th, 2021 (practical)

Hosted by the Institute for Behavioral Genetics, University of Colorado, Boulder

Kumar Veerapen, PhD

Hail Support and Community Outreach Manager

Tim Poterba, Carolin Diaz, Dan Howrigan, John Compitello



<https://hail.is>
@mkveerapen / @hailgenetics
veerapen@broadinstitute.org

#scalableGenomics

#hailGenetics #ATGUstrong

hail

Sequencing and Introduction to Hail: “Next-generation” Sequencing Technology

2021 Virtual Workshop on Statistical Genetics Methods for Human Complex Traits

June 16th, 2021 (practical)

Hosted by the Institute for Behavioral Genetics, University of Colorado, Boulder

Kumar Veerapen, PhD

Hail Support and Community Outreach Manager

Tim Poterba, Carolin Diaz, Dan Howrigan, John Compitello



<https://hail.is>
@mkveerapen / @hailgenetics
veerapen@broadinstitute.org

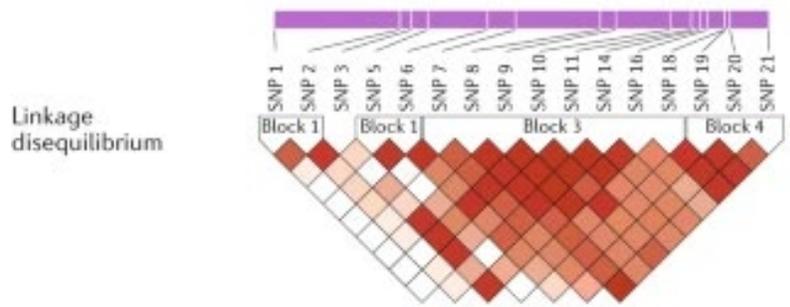
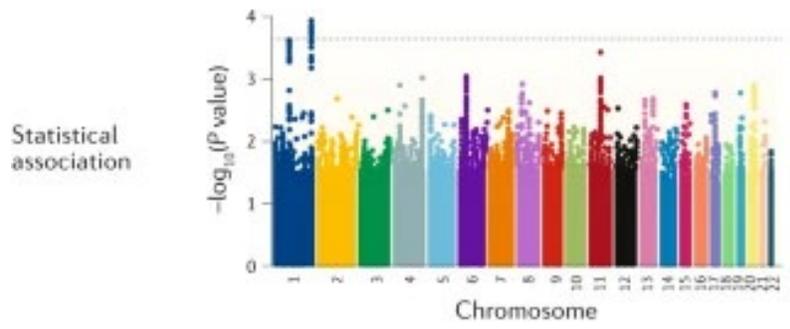
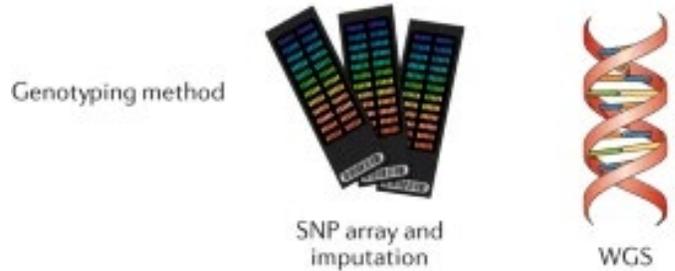
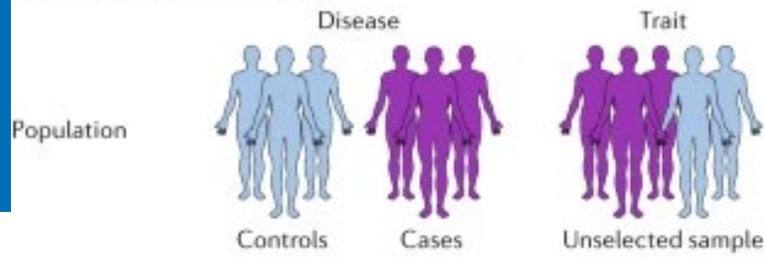
#scalableGenomics

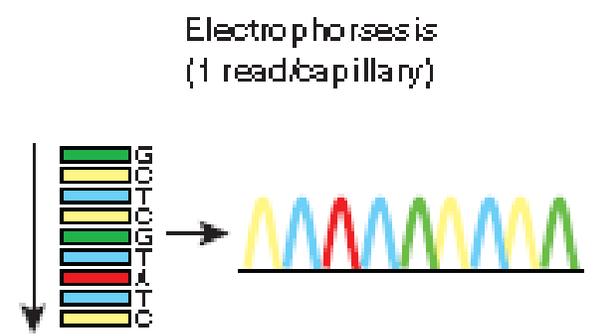
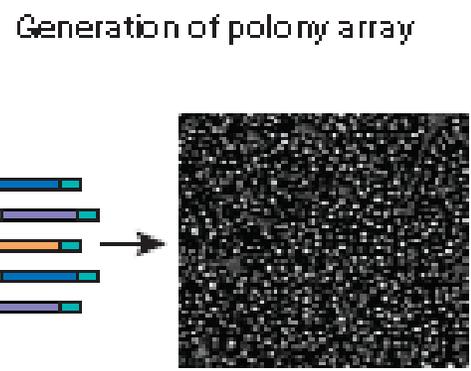
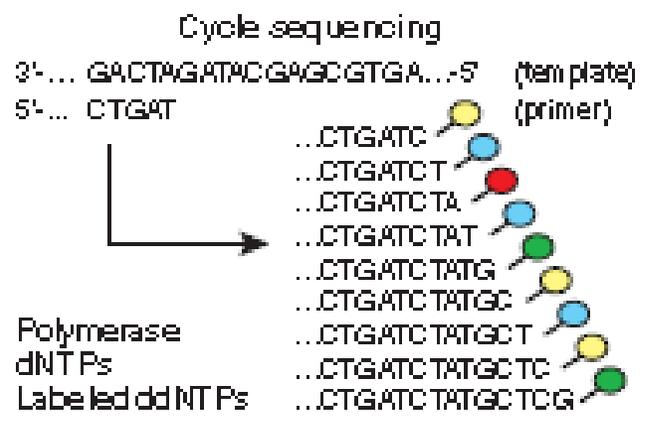
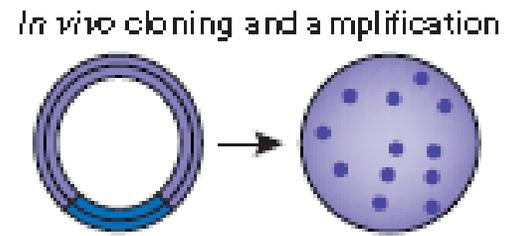
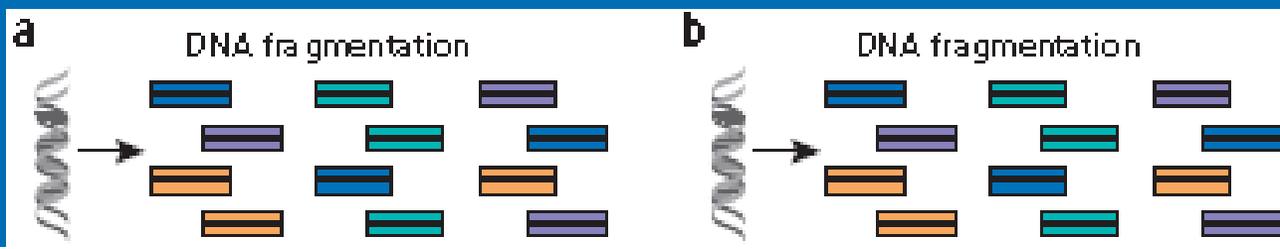
#hailGenetics #ATGUstrong

Learning Objectives

- To understand the overview of "next-generation" DNA sequencing methods
 - To appreciate the rationale in moving from Sanger sequencing and other "next-gen" methods for sequencing
 - To capture the concepts of sequencing-by-synthesis
 - To understand that there are a wide variety of applications for sequencing

a Genome-wide association



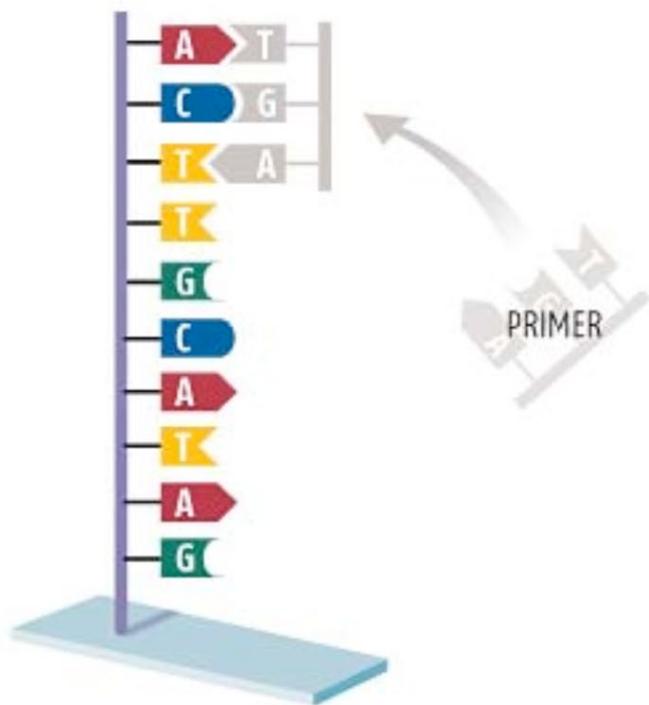


ng

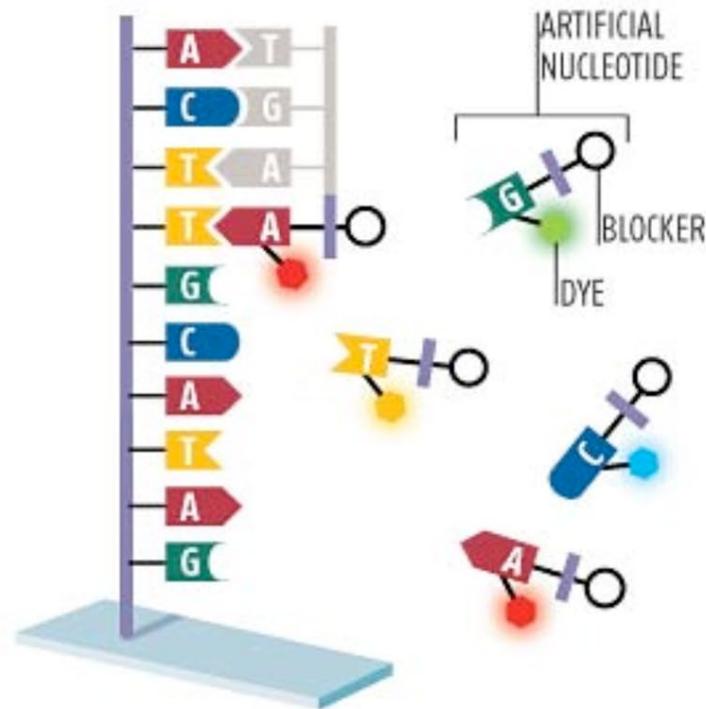
Sequence by Synthesis (SBS)

The sequence is read as a new strand is assembled

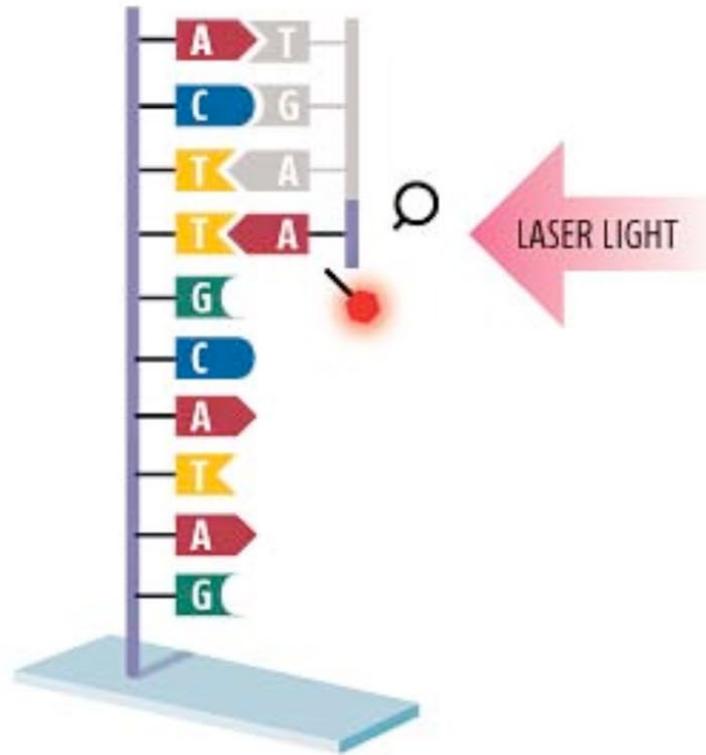
1: One strand of double-stranded DNA is firmly anchored at one end to a glass chip, and the other, complementary strand is separated and washed away. Next a "primer" is attached to the free end to allow synthesis of a new complementary strand to begin



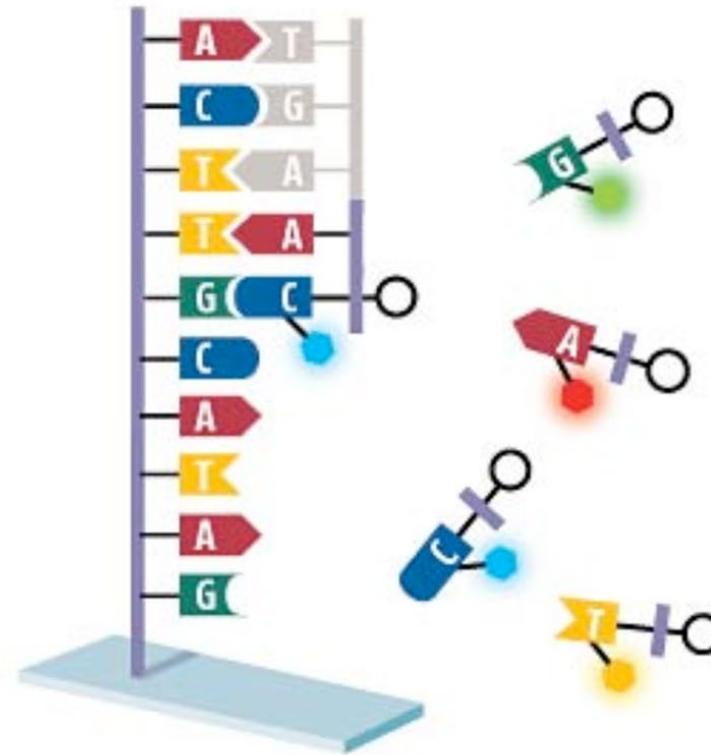
2: The polymerase enzyme starts building the new strand using artificial nucleotides. Normally it would continue building the strand, but the blocker on the artificial nucleotides halts synthesis after just one nucleotide is added. Once the excess nucleotides have been washed away, fluorescent dyes reveal which one has been added

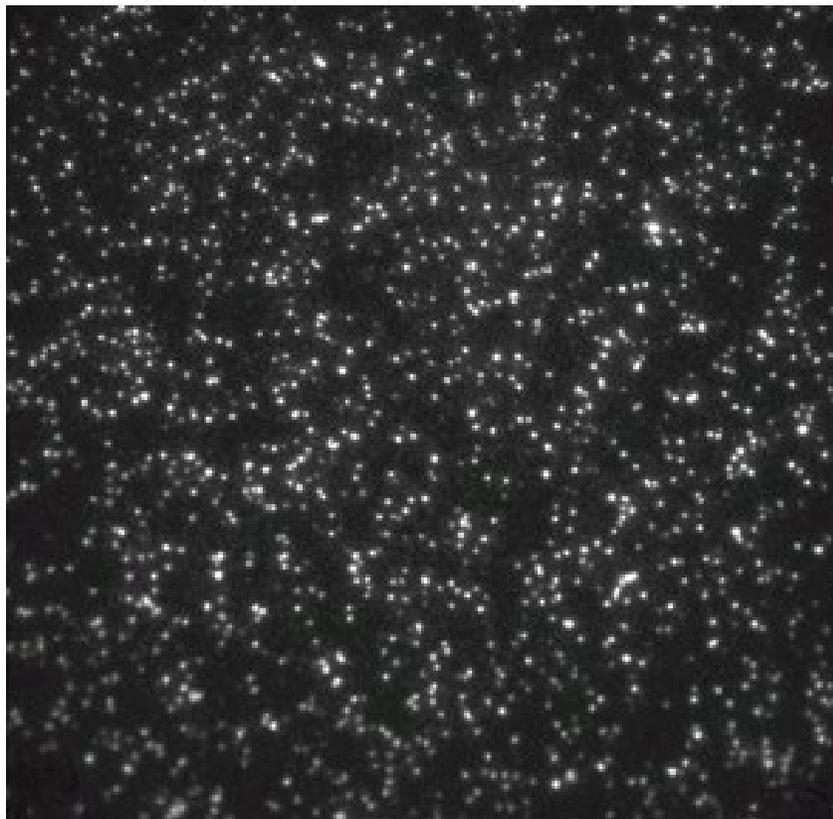


3: The fluorescent dye is removed by shining laser light on the chip, and the blocker group is removed by adding a palladium catalyst



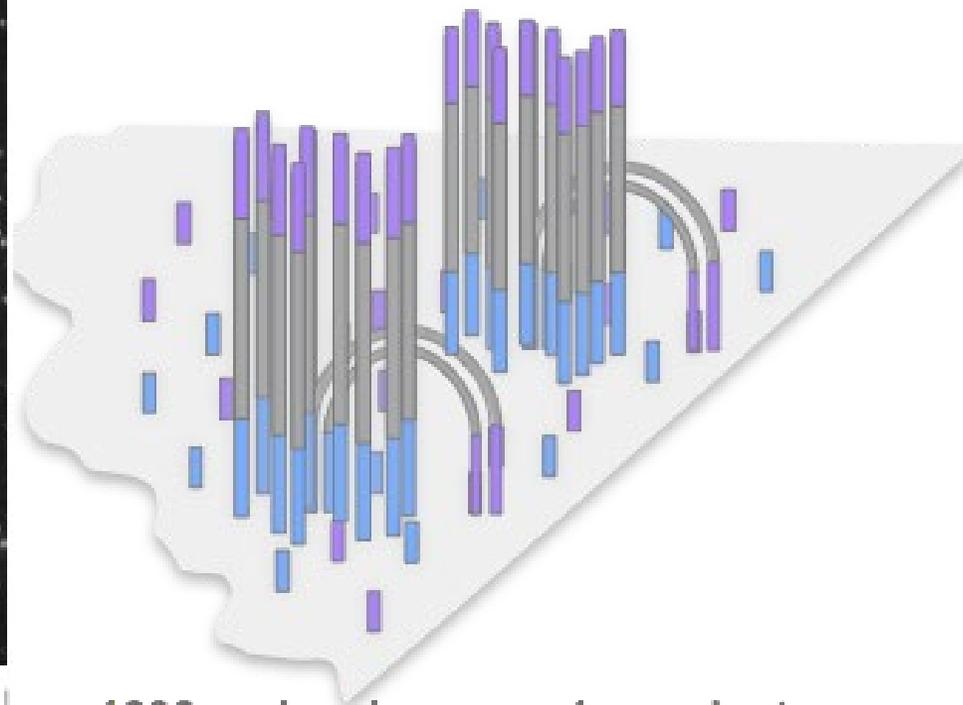
4: The process begins again with the next incoming nucleotide





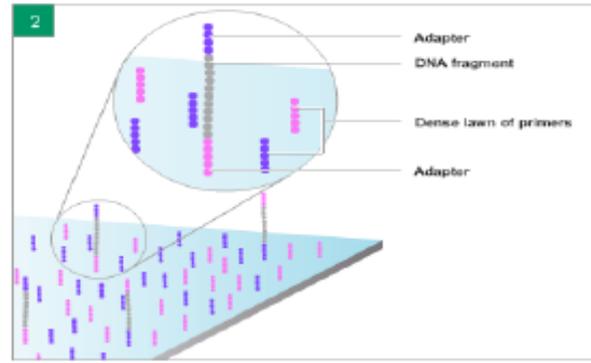
340um

Random array of clusters



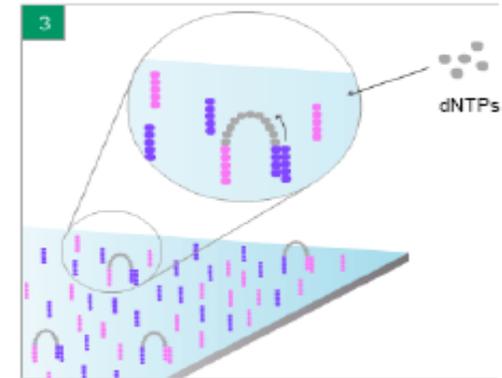
- ~1000 molecules per ~ 1 um cluster
- ~20,000 clusters per tile
- ~32 million clusters per experiment
- >1 Gb sequence per experiment

Attach primers and DNA molecule to the surface of flow cell channels



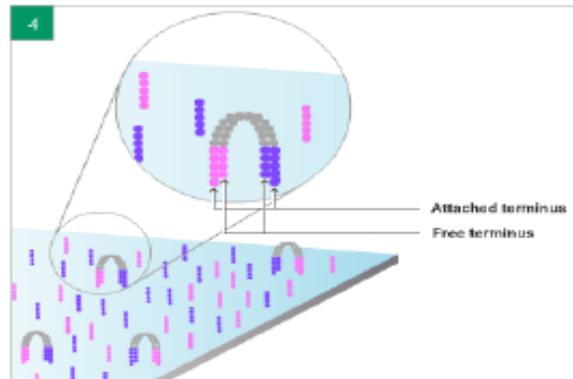
www.illumina.com

Cloning: "bridge" PCR amplification



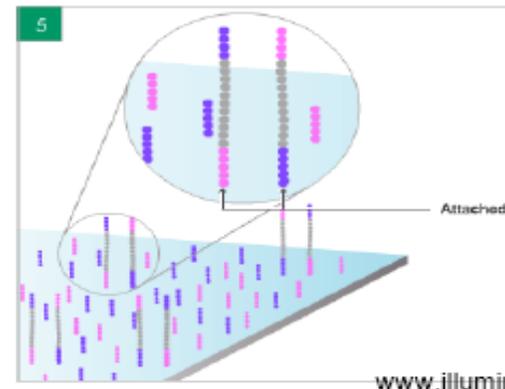
www.illumina.com

1st cycle of PCR amplification



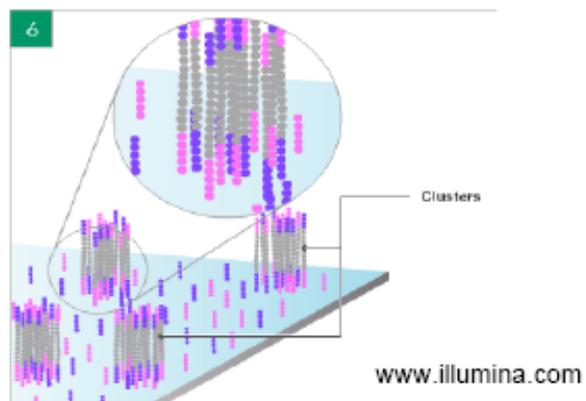
www.illumina.com

Denature the product



www.illumina.com

End result: PCR colonies ("Polonies")



First sequencing cycle: determine 1st base

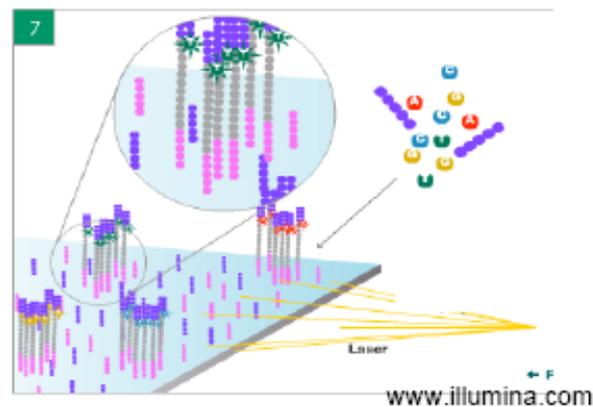
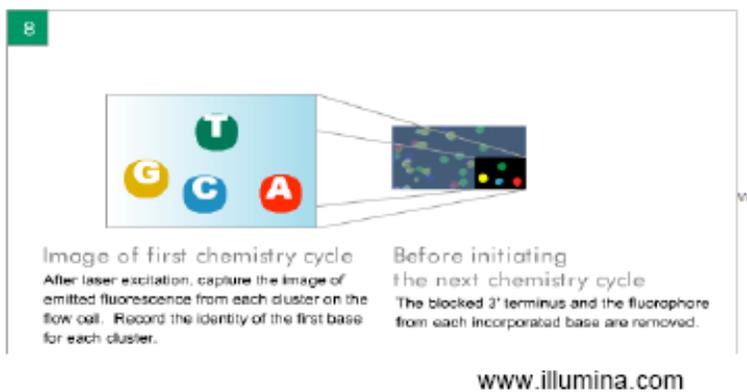


Image the base added at each "polony"

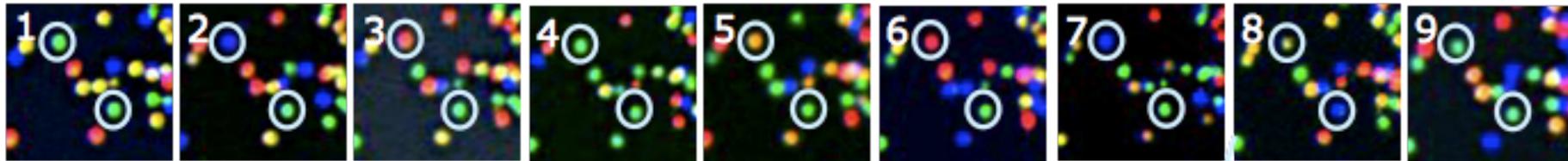


After additional sequencing cycles



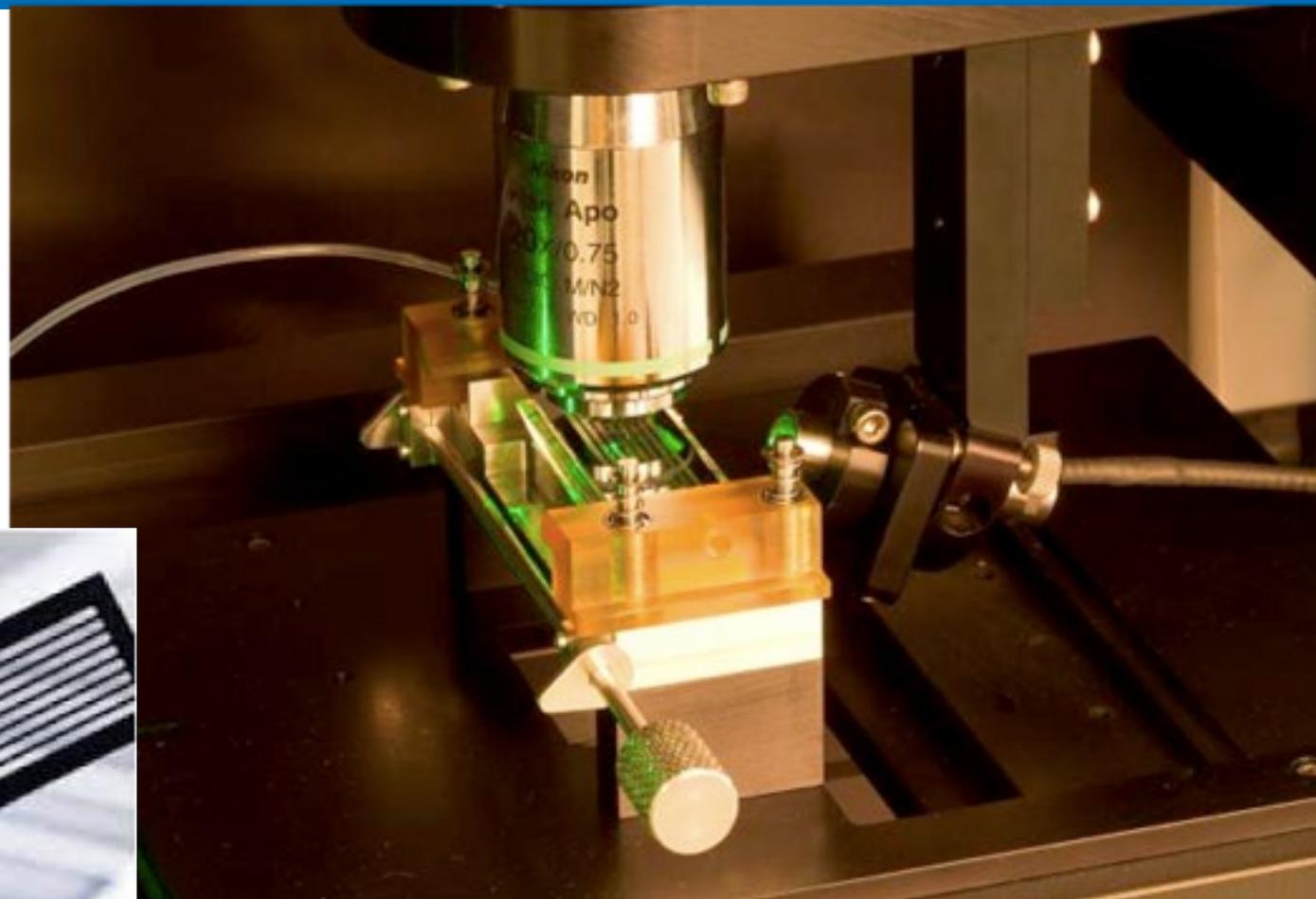
www.illumina.com

TGCTACGAT...

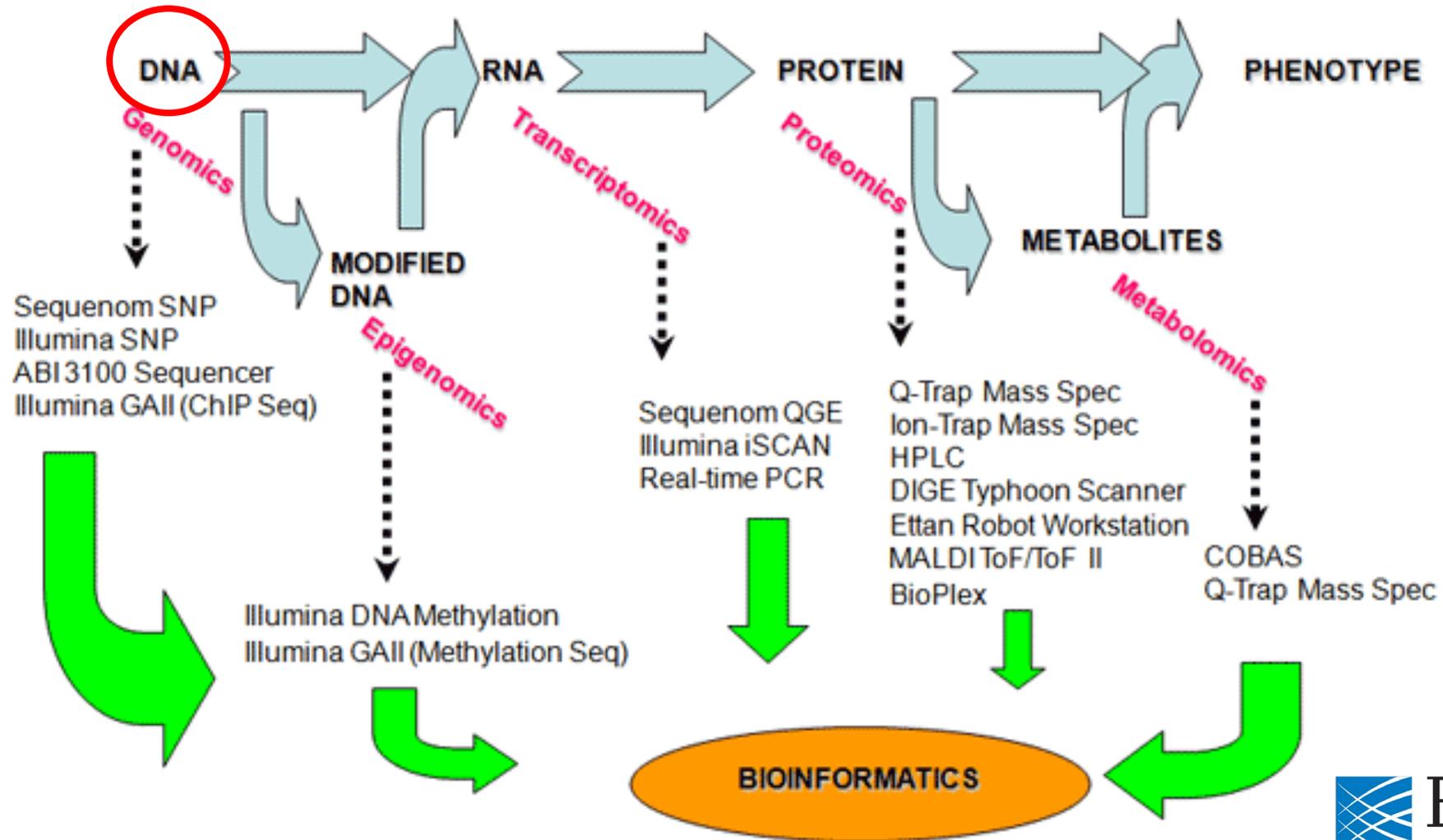


TTTTTTGT...

The identity of each base of a cluster is read off from sequential images



Overview: -Omics



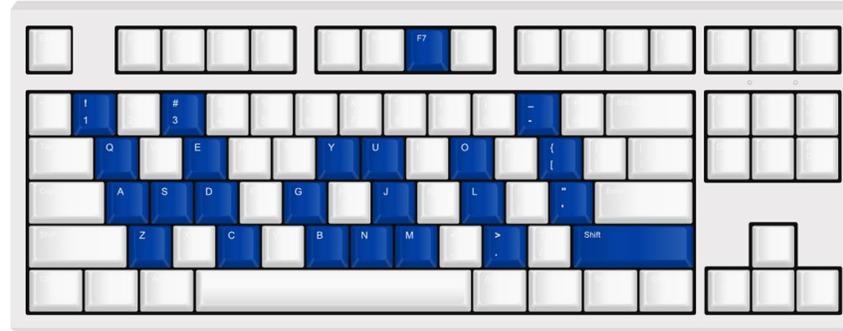
-Omics Applications

Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation-sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals

Shendure J. & Hanlee J. (2008) *Nature Biotechnology*

Learning Outcomes

- You understand the overview of "next-generation" DNA sequencing methods
 - You appreciate the rationale in moving from Sanger sequencing and other "next-gen" methods for sequencing
 - You can capture the concepts of sequencing-by-synthesis
 - You understand that there are a wide variety of applications for sequencing



Sequencing and Introduction to Hail: “Next-generation” Sequencing Technology

2021 Virtual Workshop on Statistical Genetics Methods for Human Complex Traits

June 16th, 2021 (practical)

Hosted by the Institute for Behavioral Genetics, University of Colorado, Boulder

Kumar Veerapen, PhD

Hail Support and Community Outreach Manager

Tim Poterba, Carolin Diaz, Dan Howrigan, John Compitello



<https://hail.is>
@mkveerapen / @hailgenetics
veerapen@broadinstitute.org

#scalableGenomics

#hailGenetics #ATGUstrong



Sequencing and Introduction to Hail: “Next-generation” Sequencing Technology (informatics)

2021 Virtual Workshop on Statistical Genetics Methods for Human Complex Traits

June 16th, 2021 (practical)

Hosted by the Institute for Behavioral Genetics, University of Colorado, Boulder

Kumar Veerapen, PhD

Hail Support and Community Outreach Manager

Tim Poterba, Carolin Diaz, Dan Howrigan, John Compitello

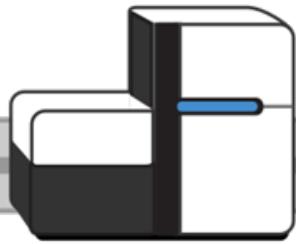


<https://hail.is>
@mkveerapen / @hailgenetics
veerapen@broadinstitute.org
#scalableGenomics
#hailGenetics #ATGUstrong

Learning Objectives

- To understand the overview of "next-generation" DNA sequencing methods
 - To obtain an overview of the bioinformatics involved after obtaining sequencing reads

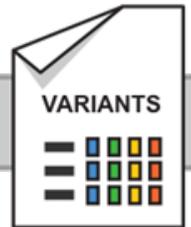
From samples to sequencing to analysis



Sequencing



.fastq



.bam /
.sam /
.cram

.vcf

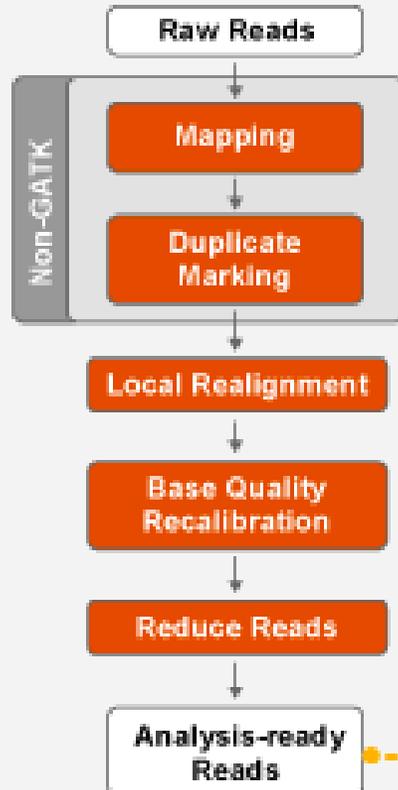
NextGen Alignment and Assembly

- BLAST or BLAT too expensive for huge numbers of short-reads
- New tools apply established alignment algorithms and new algorithms
- Some use quality values to align:
 - MAQ on Solexa or SOLiD data, SHRiMP on SOLiD
- *De novo* assembly is challenging but aided by paired reads in Illumina reads

Calling Variants with the GATK

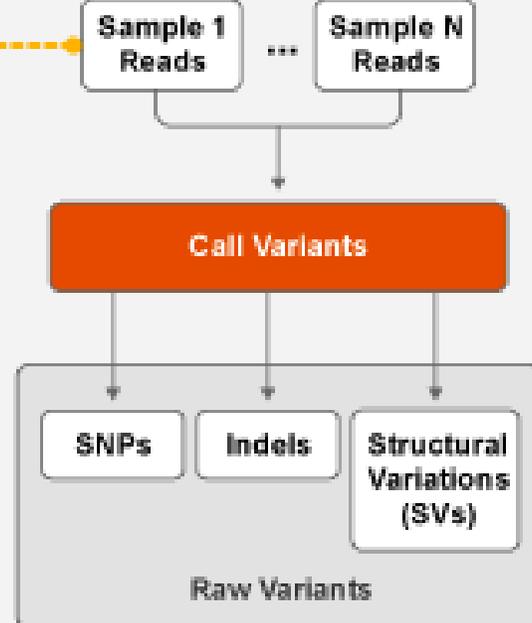
PHASE 1: NGS DATA PROCESSING

Typically by Lane

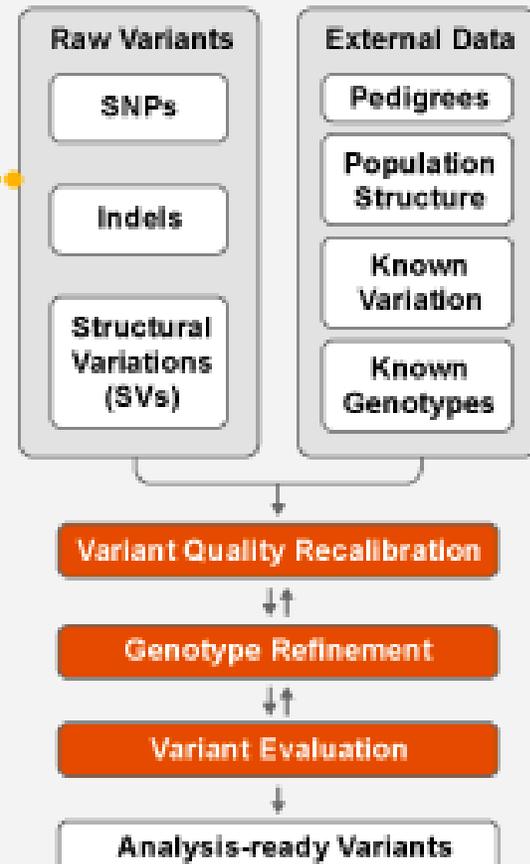


PHASE 2: VARIANT DISCOVERY AND GENOTYPING

Typically Multiple Samples
Simultaneously



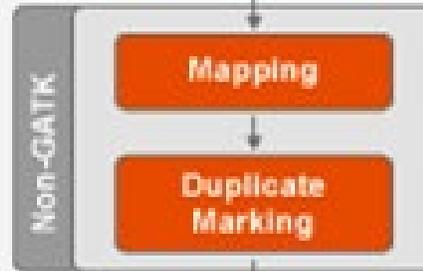
PHASE 3: INTEGRATIVE ANALYSIS



PHASE 1: NGS DATA PROCESSING

Typically by Lane

Raw Reads



CollectHsMetrics

BWA-MEM v0.7.15
(burrows-wheeler)

Picard v2.4.1

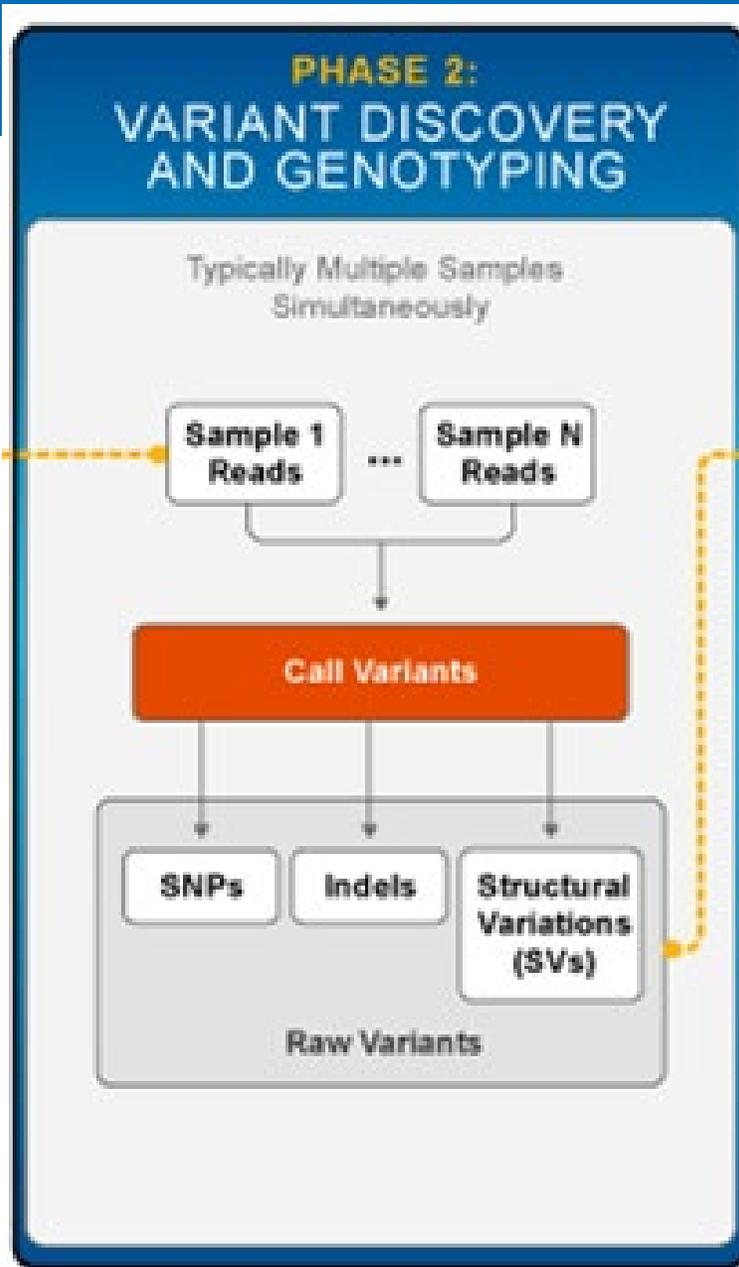
BaseRecalibrator
ApplyBQSR

For more information:
COVID-19 Host Genetics Initiative:
Whole Exome/Genome Sequencing Analysis Plan

https://docs.google.com/document/d/1X_qjplH8T4BJXSeMQ_sBfQUTiu_kAisicOqGb6B8hcM/edit#

GATK v4.1.4.1



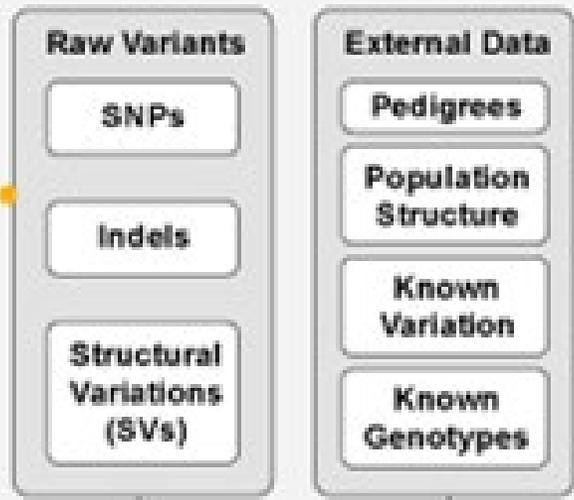


Haplotype Caller

For more information:
COVID-19 Host Genetics Initiative:
Whole Exome/Genome Sequencing Analysis Plan

https://docs.google.com/document/d/1X_qjplH8T4BJXSeMQ_sBfQUTiu_kAisicOqGb6B8hcM/edit#

PHASE 3: INTEGRATIVE ANALYSIS



ReblockGVCF
or
Hail VCF combiner

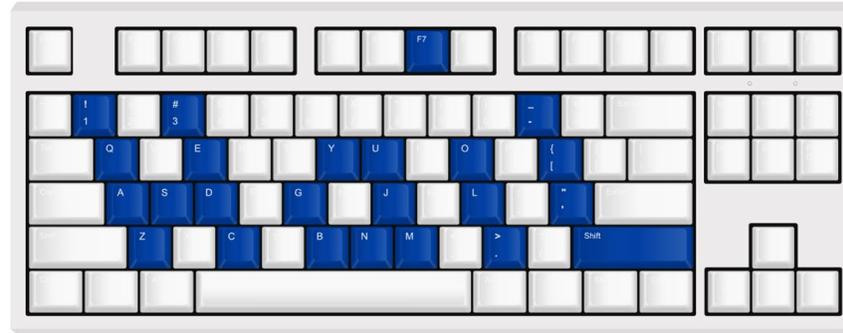
GenomicsDBImport
VariantRecalibrator
ApplyVQSR

For more information:
COVID-19 Host Genetics Initiative:
Whole Exome/Genome Sequencing Analysis Plan

https://docs.google.com/document/d/1X_qjplH8T4BJXSeMQ_sBfQUTiu_kAisicOqGb6B8hcM/edit#

Learning Objectives

- You understand the overview of "next-generation" DNA sequencing methods
 - You have an overview of the bioinformatics involved after obtaining sequencing reads



Sequencing and Introduction to Hail: “Next-generation” Sequencing Technology (informatics)

2021 Virtual Workshop on Statistical Genetics Methods for Human Complex Traits

June 16th, 2021 (practical)

Hosted by the Institute for Behavioral Genetics, University of Colorado, Boulder

Kumar Veerapen, PhD

Hail Support and Community Outreach Manager

Tim Poterba, Carolin Diaz, Dan Howrigan, John Compitello



<https://hail.is>
@mkveerapen / @hailgenetics
veerapen@broadinstitute.org

#scalableGenomics

#hailGenetics #ATGUstrong



Sequencing and Introduction to Hail: Analysis of sequencing data using Hail

2021 Virtual Workshop on Statistical Genetics Methods for Human Complex Traits

June 16th, 2021 (practical)

Hosted by the Institute for Behavioral Genetics, University of Colorado, Boulder

Kumar Veerapen, PhD

Hail Support and Community Outreach Manager

Tim Poterba, Carolin Diaz, Dan Howrigan, John Compitello



<https://hail.is>
@mkveerapen / @hailgenetics
veerapen@broadinstitute.org
#scalableGenomics
#hailGenetics #ATGUstrong

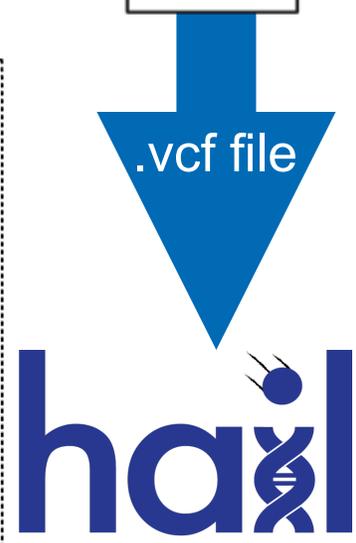
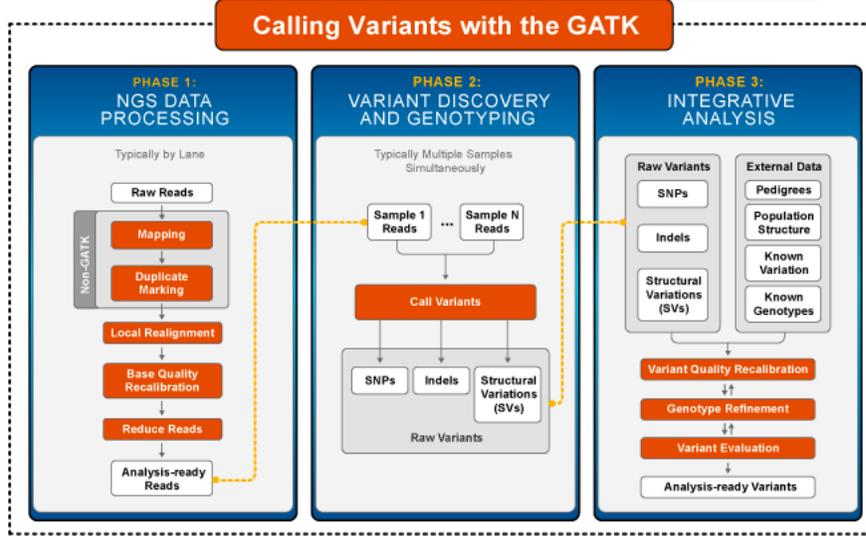
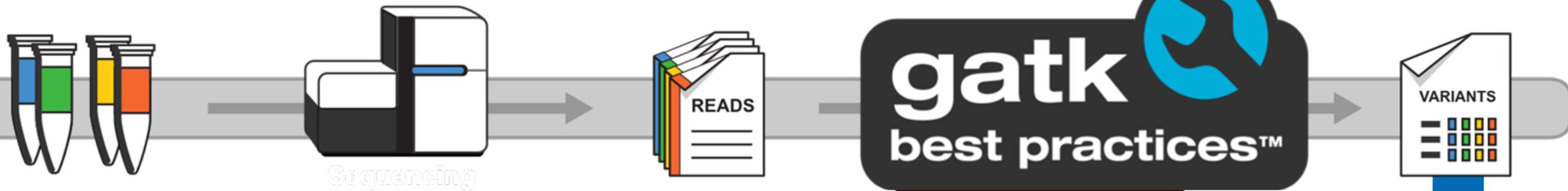
Learning Objectives

- To capture the need for Hail in the analysis of genomic datasets
 - Why Hail?
 - Who are in the Hail team?
 - What can you use Hail for?

Accelerating Genomic Data e.g. Call Sets, variant files etc



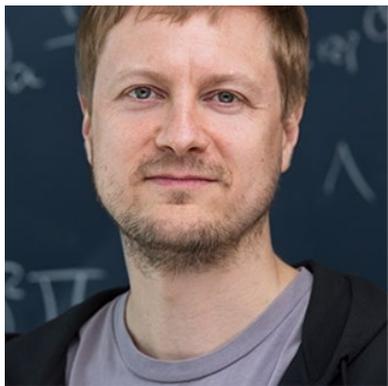
What is Hail's role in callset generation?





The Hail Team is a systems engineering team building tools to accelerate biological research.

Hail Team



*Cotton Seed, PhD
Team Leader*



Tim Poterba



Dan King



Jackie Goldstein



Daniel Goldstein



Patrick Schultz, PhD



*Whitney Wade
Operations*



*Kumar Veerapen, PhD
Support and Outreach*



John Compitello



Carolin Diaz



Chris Vittal



Patrick Cummings

What is Hail?

“On a scale from zero to dplyr, the Hail 0.2 interface scores an 8/10 for general-purpose data analysis.” - Konrad K., lead analyst, gnomAD

Open-Source Data
Science Library

Slice, dice, query,
and model any
kind of data

Scalability

Easy to use with
both small and
biobank-scale
genomic data

Unified Genomic
Data Representation

The MatrixTable is a
single interface for
working with all kinds
of genomic data

Community

Forum and chatroom
for people interested in
thinking + talking
about genomic data
analysis

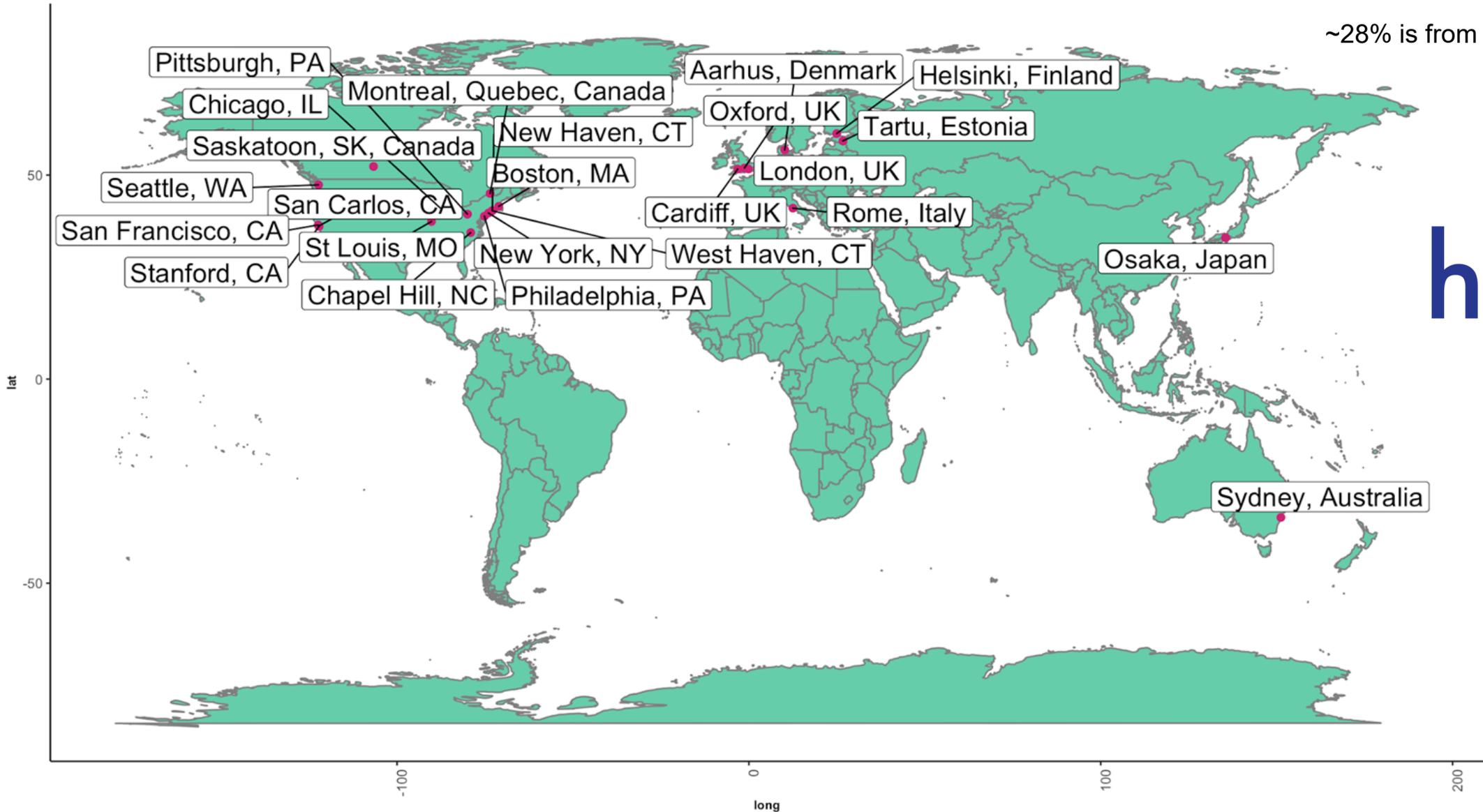


Learn more at [Hail.is](https://hail.is)

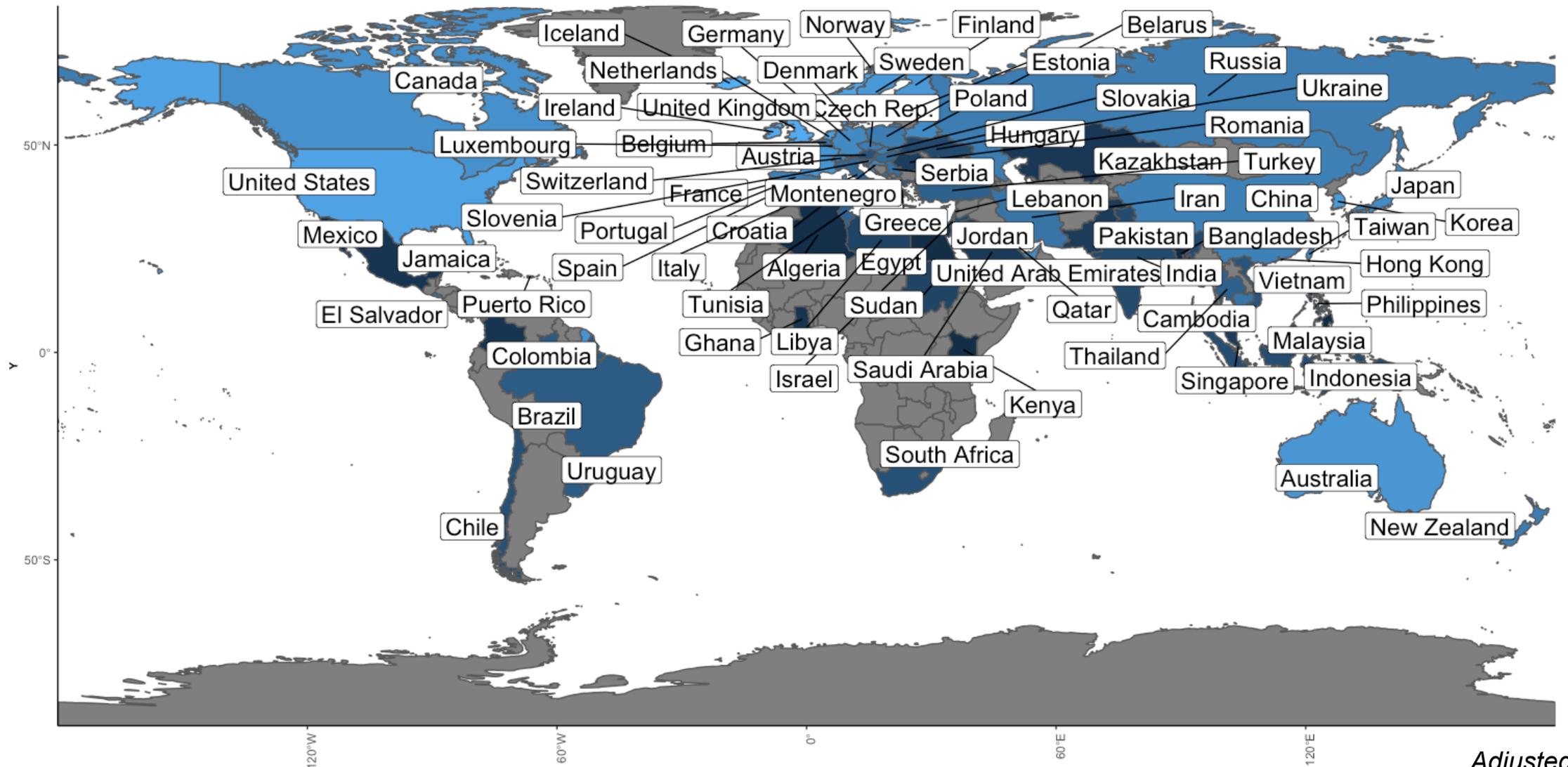
*We can't read your
minds, so talk to us
discuss.hail.is

Where has Hail been used?

~28% is from Boston, MA

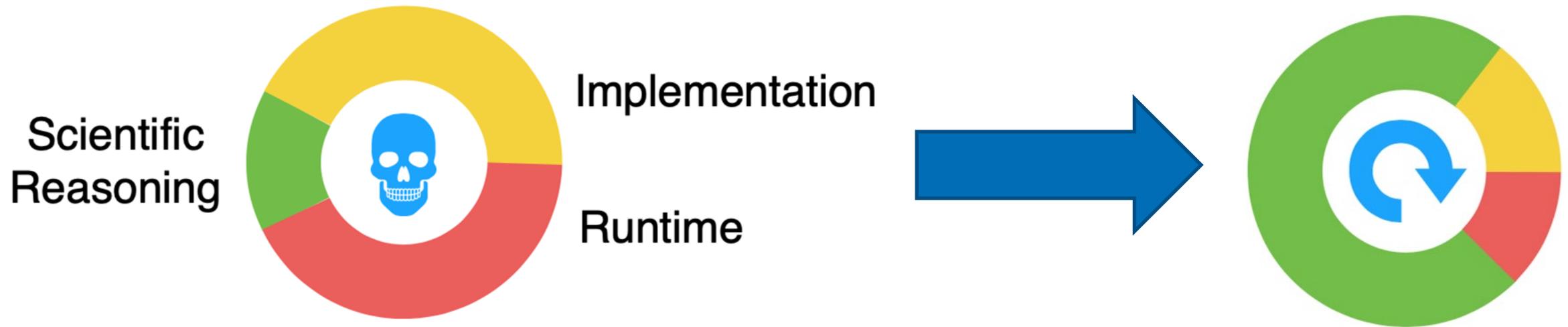


Where in the world has Hail been “pip”-ed a.k.a. downloaded?



Adjusted for total population

Why would you use Hail?



Hail as a data science library

Data slinging

Analytical toolbox

Hail as a data science library

Data slinging

Analytical toolbox

- **Read and write common formats**
- Filter, group, aggregate
- Annotation
- Visualization

VCF

TSV

BGEN

PLINK

JSON

GEN

BED

GTF

Hail as a data science library

Data slinging

- Read and write common formats
- **Filter, group, aggregate**
- Annotation
- Visualization

Analytical toolbox

- Compute mean depth per variant or per sample
 - Among heterozygotes
 - Grouped by ancestry labels & sex
- Count transitions & transversions called per sample

Hail as a data science library

Data slinging

- Read and write common formats
- Filter, group, aggregate
- **Annotation**
- Visualization

Analytical toolbox

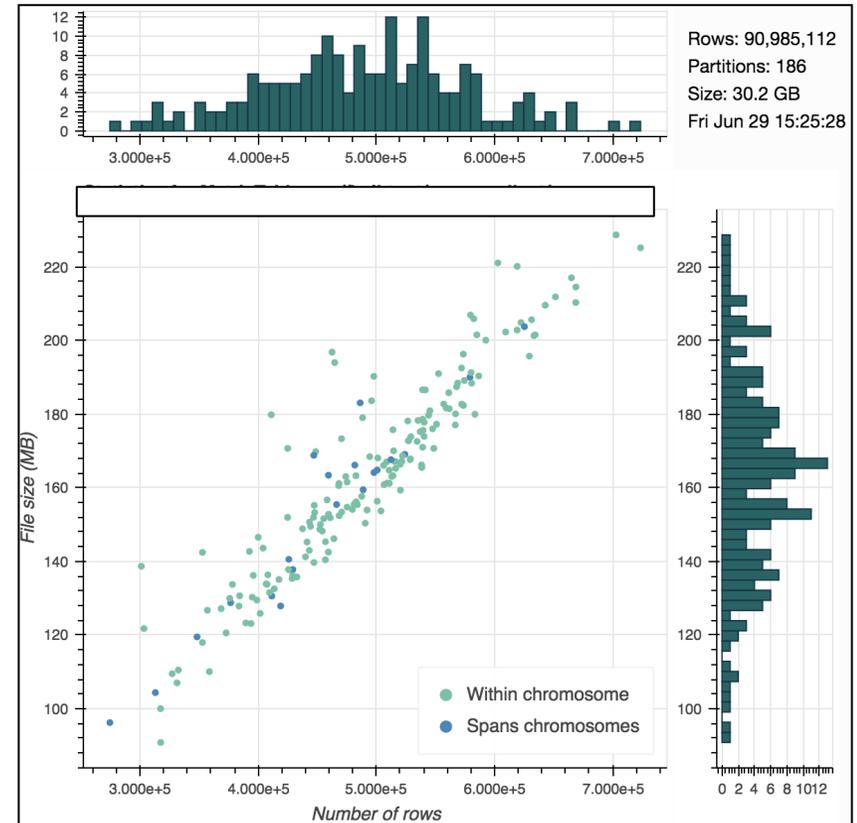
- Built-in wrapper for the Variant Effect Predictor (VEP). We did the setup so you don't have to!
- Join with annotations by variant, locus, interval, gene
- Annotation database

Hail as a data science library

Data slinging

Analytical toolbox

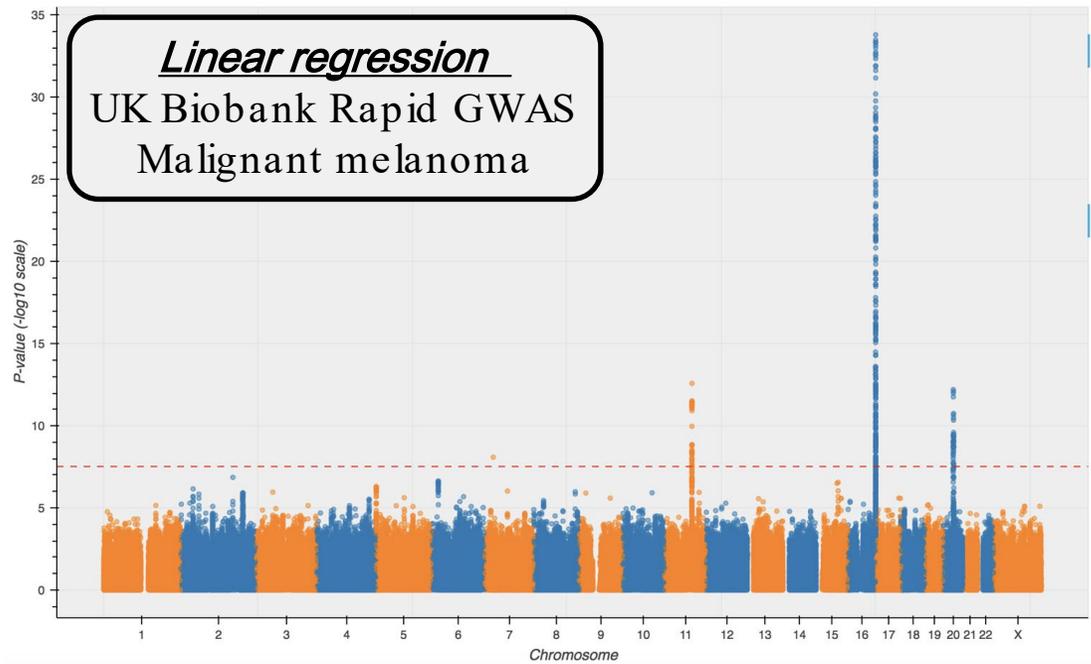
- Read and write common formats
- Filter, group, aggregate
- Annotation
- **Visualization**



Hail as a data science library

Data slinging

Analytical toolbox

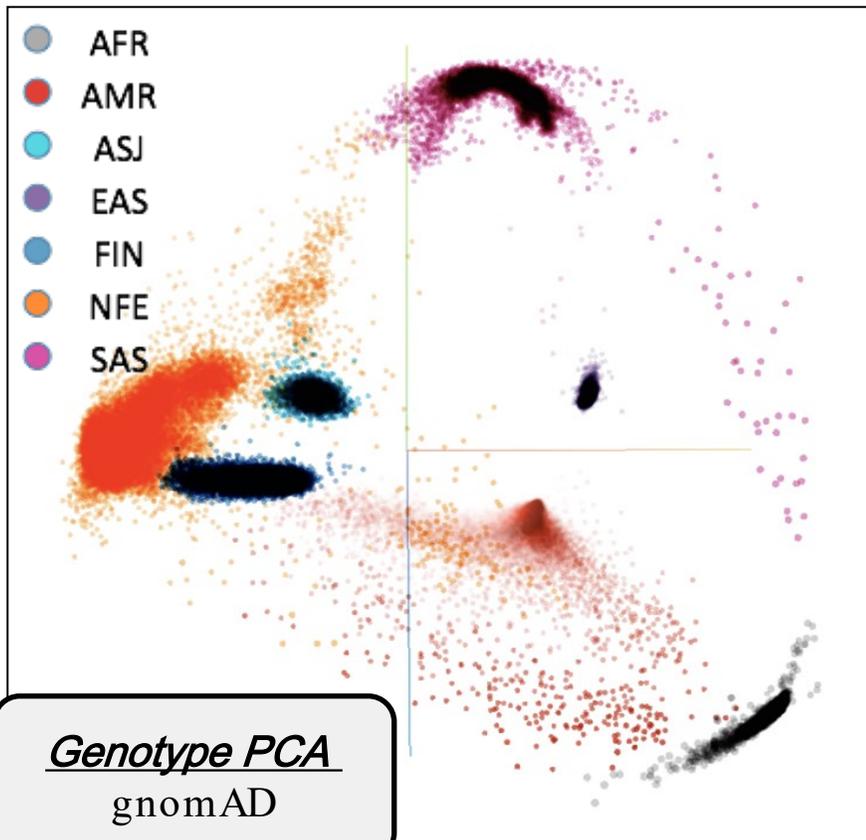


- **Statistical methods for genetics**
- Linear algebra

Hail as a data science library

Data slinging

Analytical toolbox

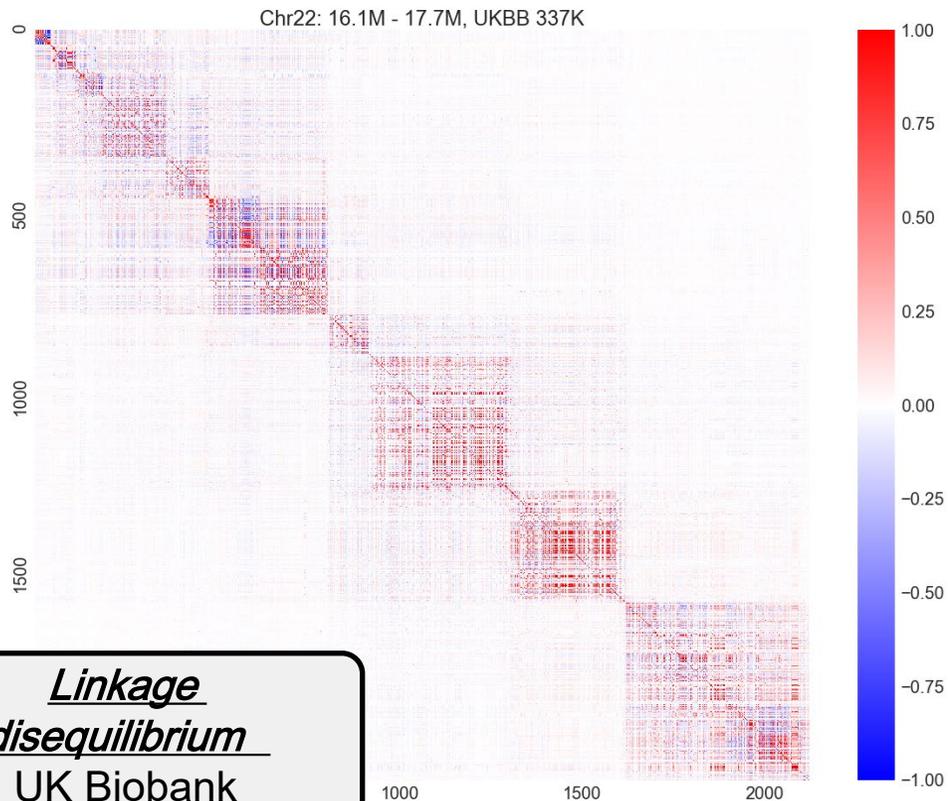


- **Statistical methods for genetics**
- Linear algebra

Hail as a data science library

Data slinging

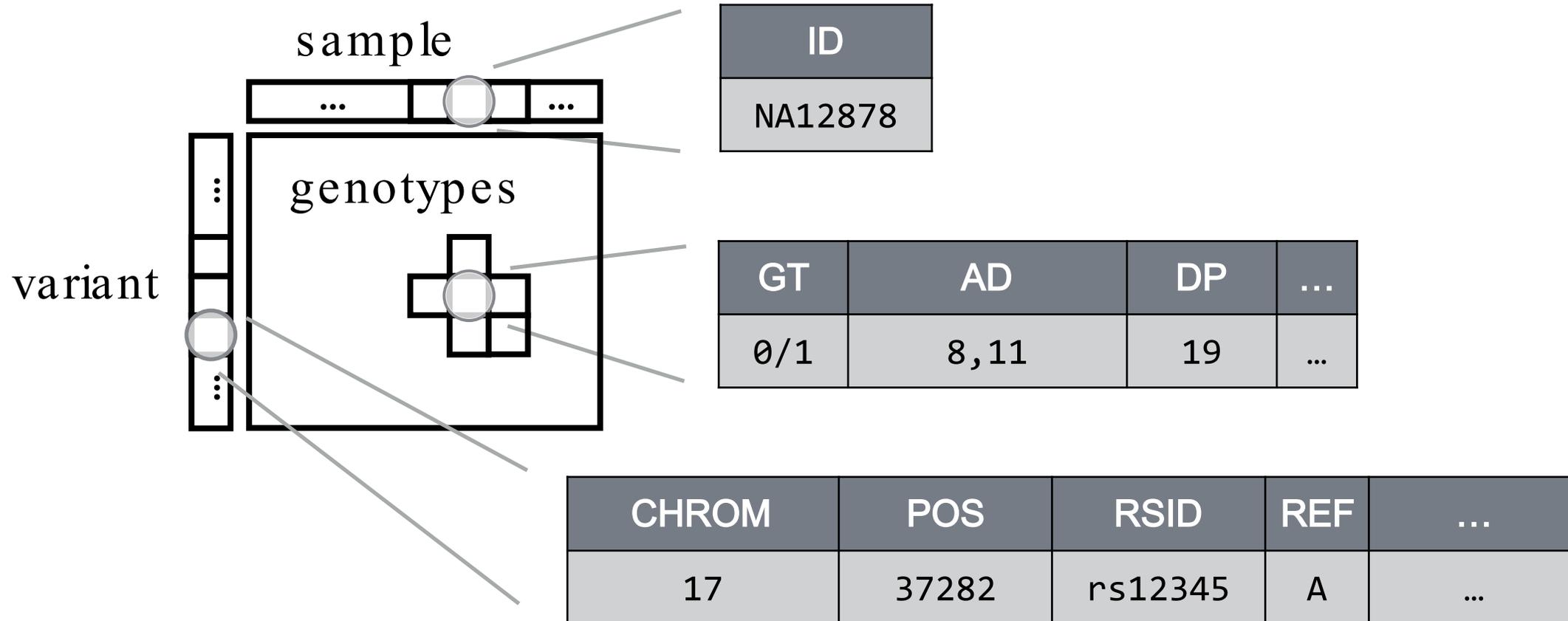
Analytical toolbox



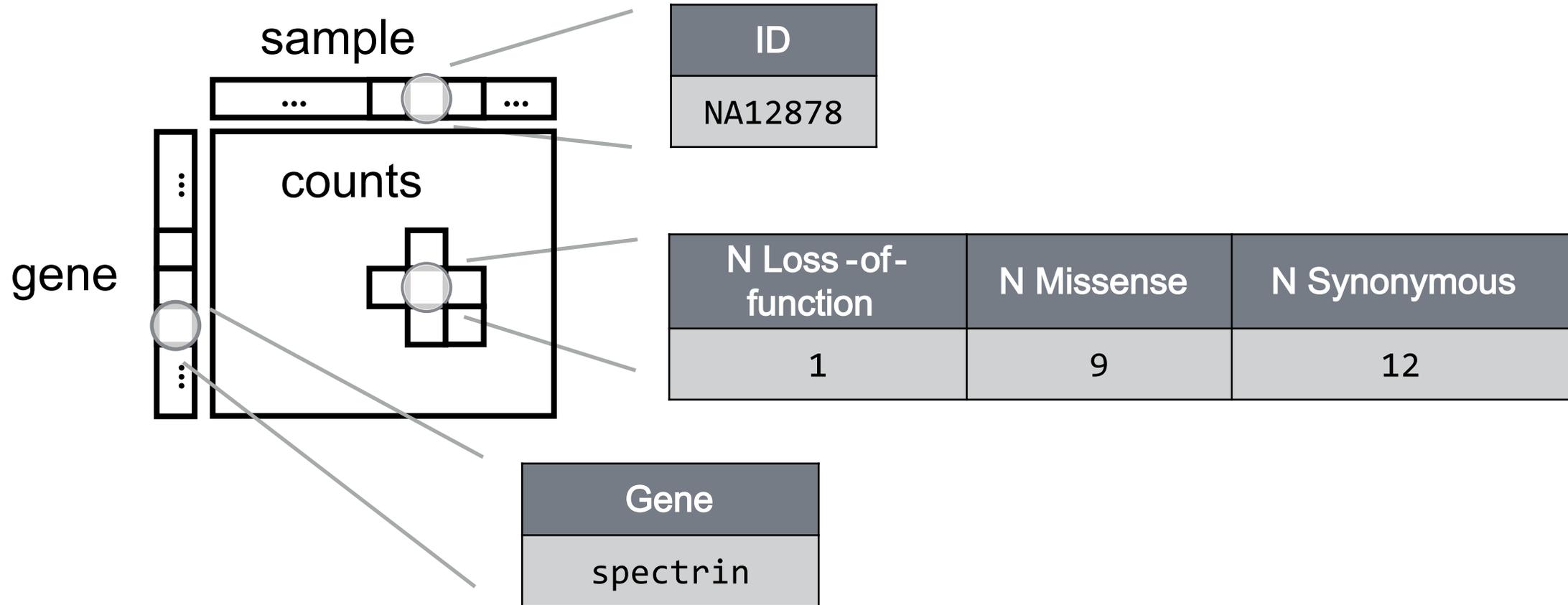
Linkage
disequilibrium
UK Biobank

- Statistical methods for genetics
- **Linear algebra (early stages)**

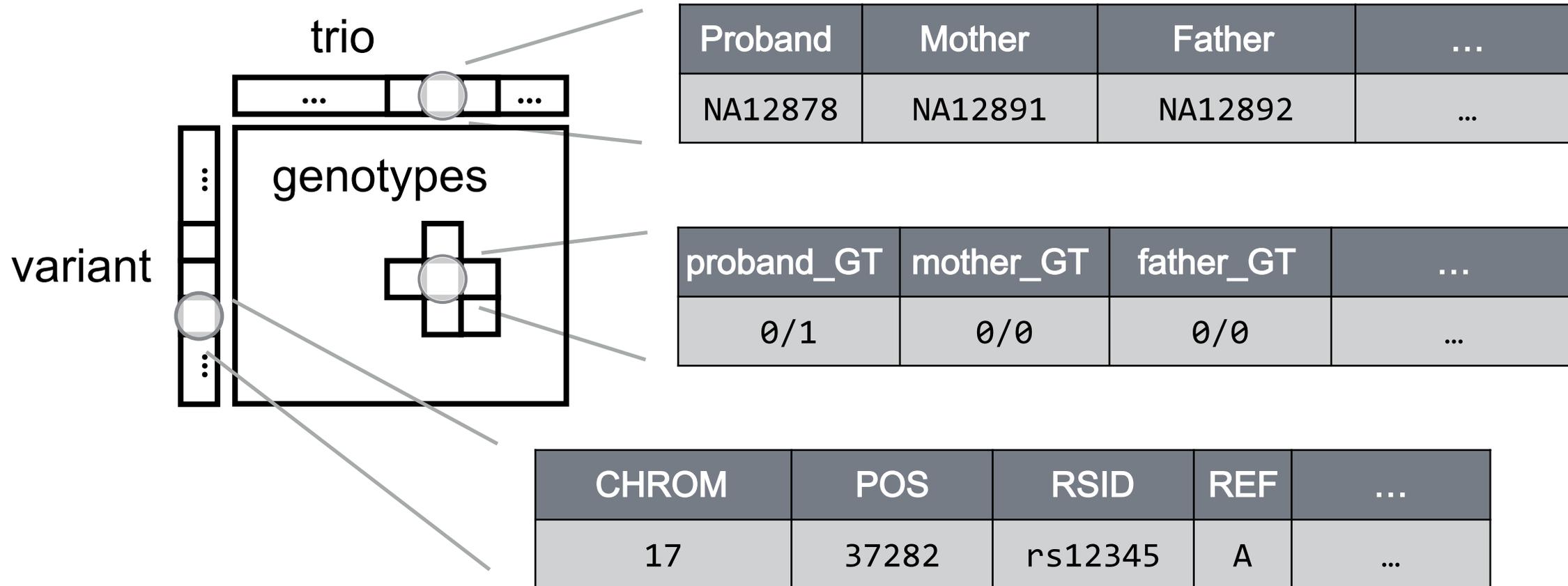
Variant Call Format (VCF)



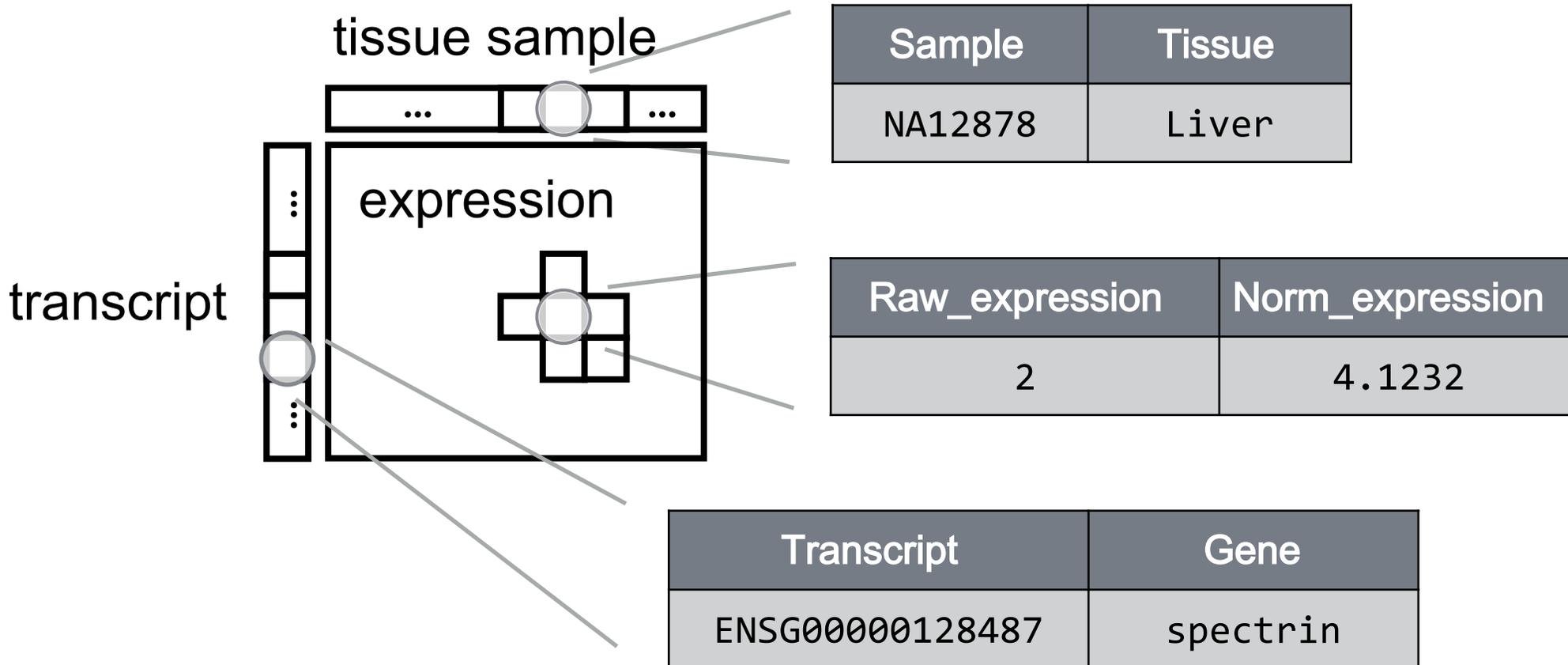
Rare variant aggregation



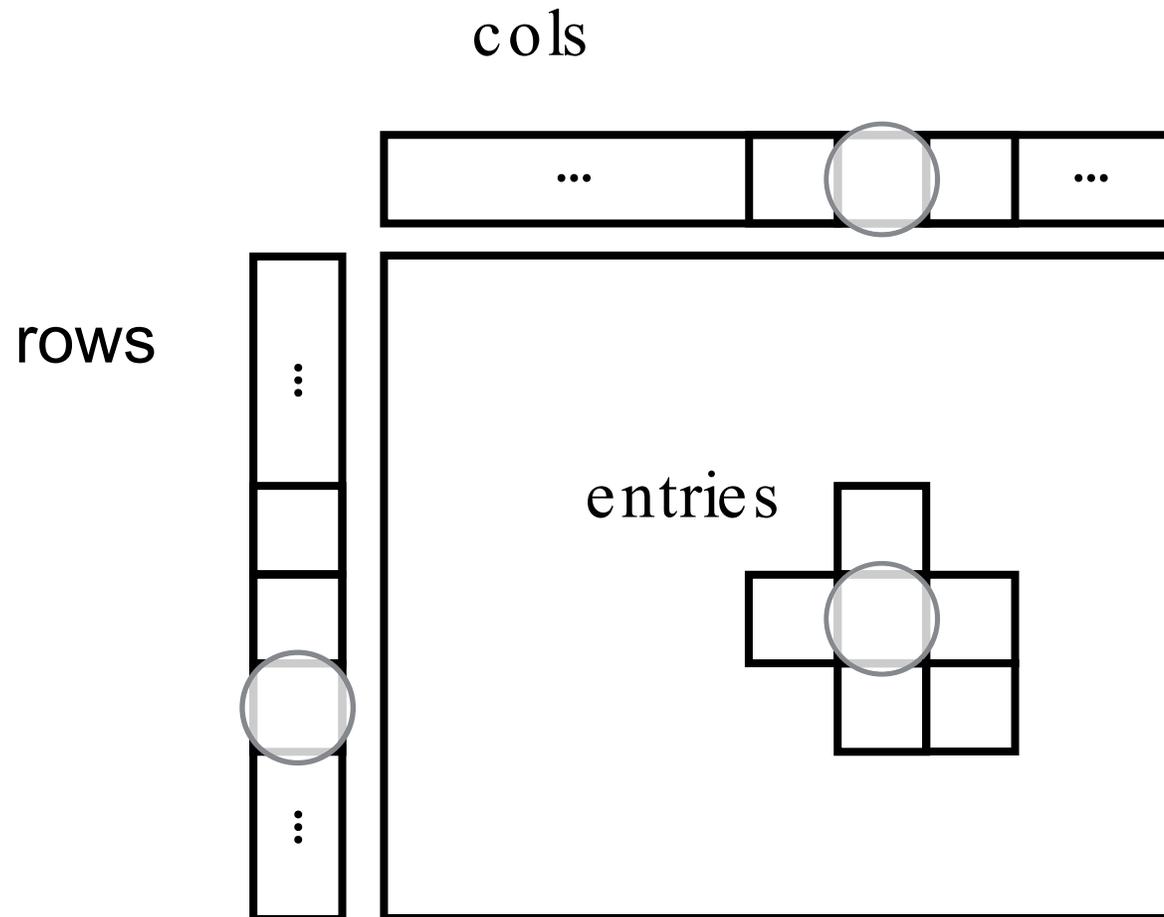
Trio data



Transcript expression



MatrixTable



Global fields:
None

Column fields:
's': str

Row fields:
'locus': locus<GRCh37>
'alleles': array<str>
'rsid': str
'qual': float64
'filters': set<str>
'info': struct {
 NEGATIVE_TRAIN_SITE: bool,
 AC: array<int32>,
 ...
 DS: bool
}

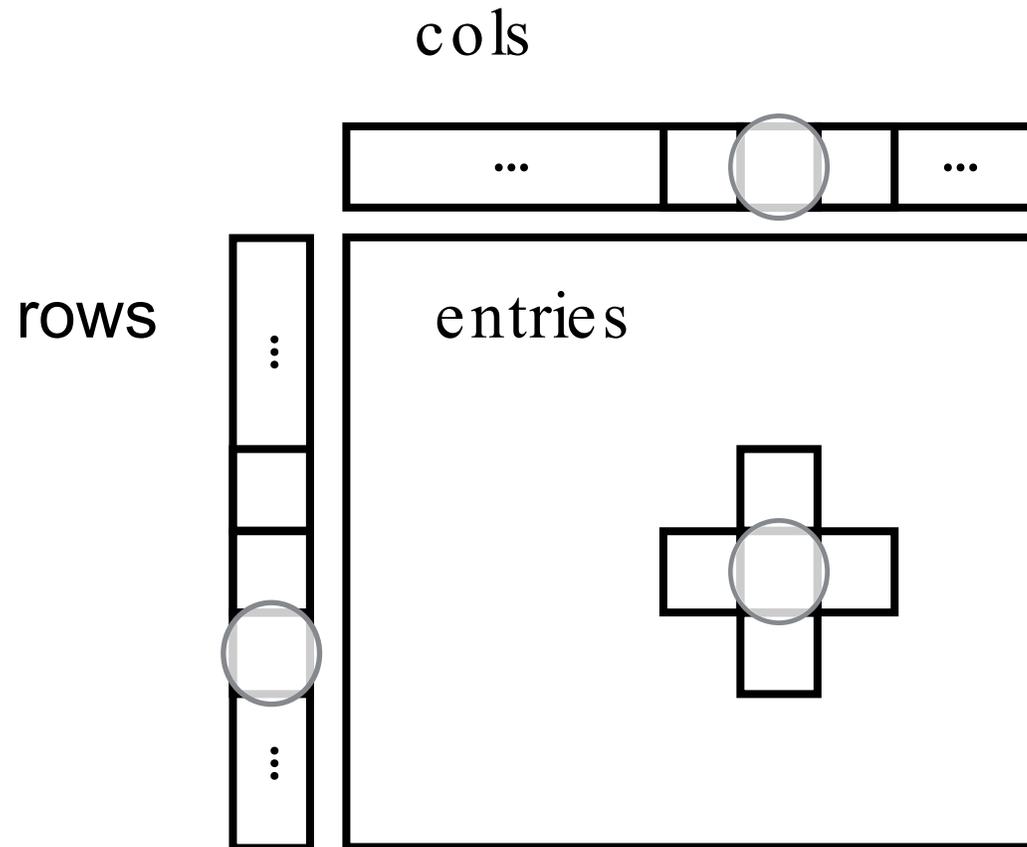
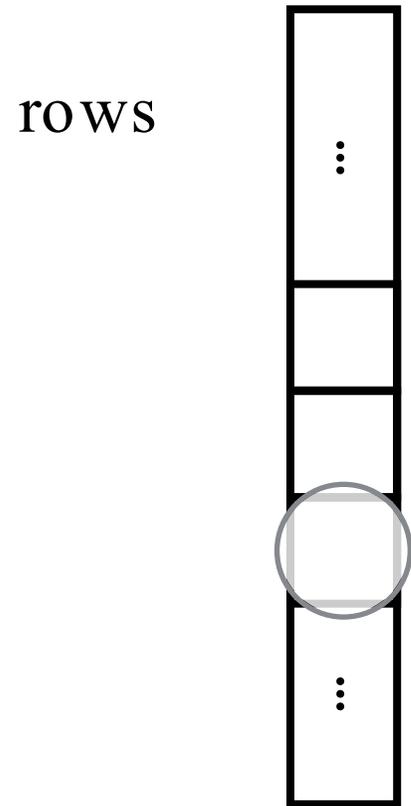
Entry fields:
'GT': call
'AD': array<int32>
'DP': int32
'GQ': int32
'PL': array<int32>

Column key:
's': str

Row key:
'locus': locus<GRCh37>
'alleles': array<str>

Table

MatrixTable



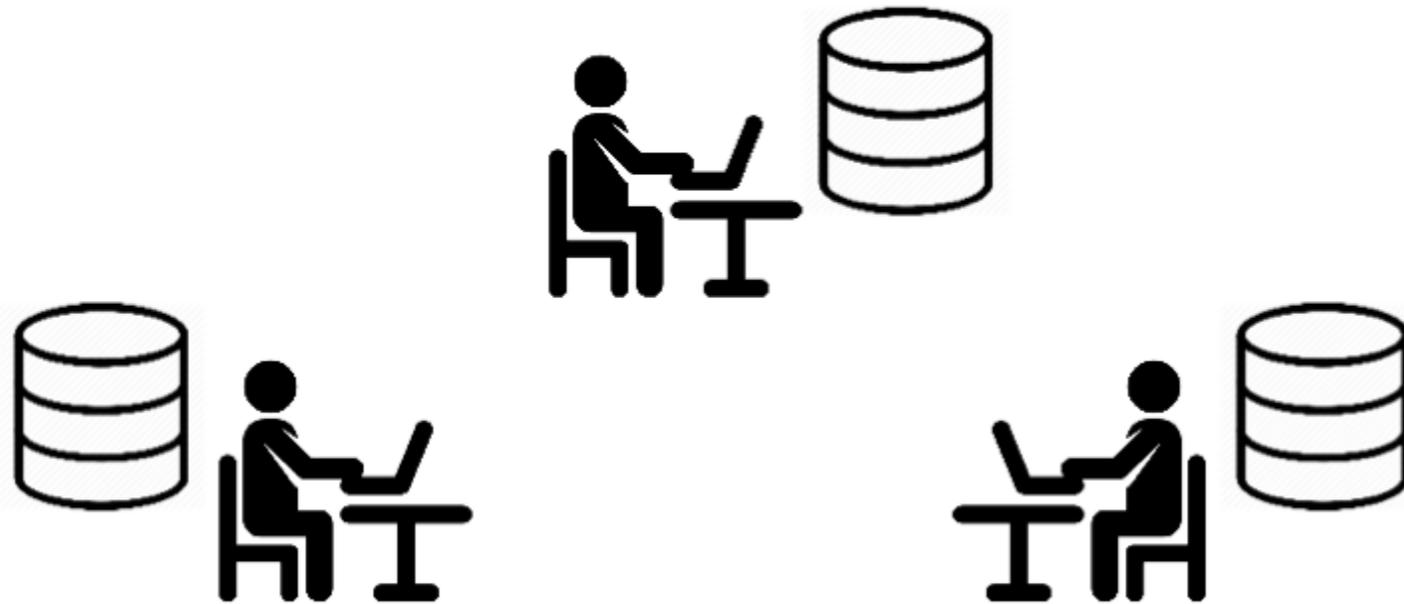
Mastering Hail takes practice

- Hail is harder to learn than command-line tools
 - It's not about memorizing command-line calls!
 - Foundational data science skills are necessary
- Prior experience with a data frame library* will help
 - * `R`, `dplyr`, `pandas`, etc
- Hail is about giving you the tools you need to indulge scientific curiosity on biological data, and that's not always easy.

Large-scale datasets

- UK Biobank 500K => 5M?
 - ... and many other biobanks
- gnomAD: 20K => 150K WGS
- TOPMed: >120K WGS
- All of Us: 1M
- Million Veterans Project: 1M

From Bringing Data to Researchers



To Bringing Researchers to Data



Computational Landscape

- Laptop/Desktop
- Server
- High Performance Computing (HPC) cluster
- Cloud

Computational Landscape

- Laptop/Desktop
development, small data (10s of genomes, 100s of exomes)
- Server
medium data (1Ks of genomes, 10Ks of exomes)
- High Performance Computing (HPC) cluster
large (1M genomes, 10M exomes)
- Cloud
large (1M genomes, 10M exomes)

Computational Landscape

- Laptop/Desktop
`pip install hail`
- Server, or a single node on High Performance Computing (HPC) cluster
`pip install hail`
- High Performance Computing (HPC) cluster
Institutional Spark cluster
Hail *does not support* HPC schedulers like SLURM, UGER, and LSF
- Cloud
Google Cloud Platform (GCP):
`pip install hail`
`hailctl dataproc start CLUSTER`

Amazon Web Services (AWS): some support
 - <https://github.com/hms-dbmi/hail-on-AWS-spot-instances>
 - <https://discuss.hail.is/t/spin-up-aws-emr-clusters-with-hail/818>

Your next steps

```
pip install hail
```

[DOCS](#)[FORUM](#)[POWERED-SCIENCE](#)[BLOG](#)[WORKSHOP](#)[Hail Docs \(0.2\)](#)[Installation](#)[Hail on the Cloud](#)[Tutorials](#)[Reference \(Python API\)](#)[Overview](#)[How-To Guides](#)[Cheatsheets](#)[Docs](#) » [Hail 0.2](#)hail.is/docs/0.2/[View page source](#)

Hail 0.2

Hail is an open-source library for scalable data exploration and analysis, with a particular emphasis on genomics. See the [overview](#) for a high-level walkthrough of the library, the [GWAS tutorial](#) for a simple example of conducting a genome-wide association study, and the [installation page](#) to get started using Hail.

[HOME PAGE](#)[HAIL DOCUMENTATION](#)[HAIL FORUM](#)[HAIL POWERED-SCIENCE](#)[HAIL BLOG](#)[HAIL WORKSHOPS](#)blog.hail.is/[GENOMICS](#)

Hail: An Introduction to an Efficient Genomic Analysis Tool

Hail is an open-source Python library for genomic data manipulation and analysis. Five years in the making, we want to (re)introduce our actively developed tool to you, our users!

discuss.hail.is[Sign Up](#)[Log In](#)[About](#)[FAQ](#)[Terms of Service](#)[Privacy](#)

About Hail Discussion

Discussion forum for Hail, an open-source, scalable framework for exploring and analyzing genomic data (<https://hail.is>)

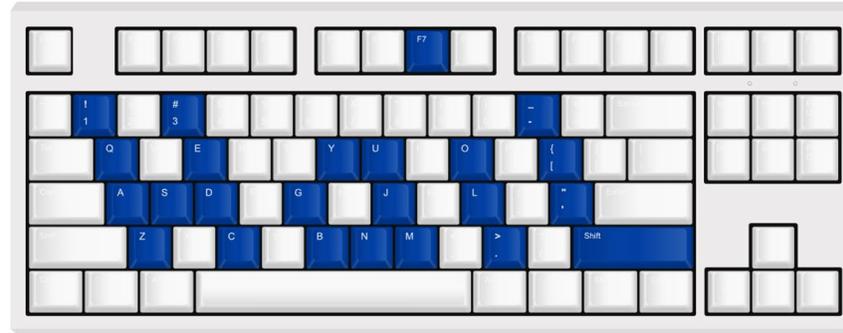


Learning Outcomes

- Hail is a useful software tool for analyzing genomic data
 - Hail is especially useful for large datasets
 - Hail team comprises of fabulous individuals
 - Hail can be used for almost every genomic and especially sequencing based questions that I have

Learning Objectives for practical session

- To be able to use basic Hail functions
- To apply basic GWAS analysis techniques using Hail on their own datasets
- To describe the use of PCA in Hail to decipher ancestries
- To obtain resources to further explore the extent of Hail capabilities



Sequencing and Introduction to Hail: Analysis of sequencing data using Hail

2021 Virtual Workshop on Statistical Genetics Methods for Human Complex Traits

June 16th, 2021 (practical)

Hosted by the Institute for Behavioral Genetics, University of Colorado, Boulder

Kumar Veerapen, PhD

Hail Support and Community Outreach Manager

Tim Poterba, Carolin Diaz, Dan Howrigan, John Compitello



<https://hail.is>
@mkveerapen / @hailgenetics
veerapen@broadinstitute.org

#scalableGenomics

#hailGenetics #ATGUstrong



Sequencing and Introduction to Hail: Unlocking the power of the cloud with Hail

2021 Virtual Workshop on Statistical Genetics Methods for Human Complex Traits

June 16th, 2021 (practical)

Hosted by the Institute for Behavioral Genetics, University of Colorado, Boulder

Kumar Veerapen, PhD

Hail Support and Community Outreach Manager

Tim Poterba, Carolin Diaz, Dan Howrigan, John Compitello



<https://hail.is>
@mkveerapen / @hailgenetics
veerapen@broadinstitute.org

#scalableGenomics

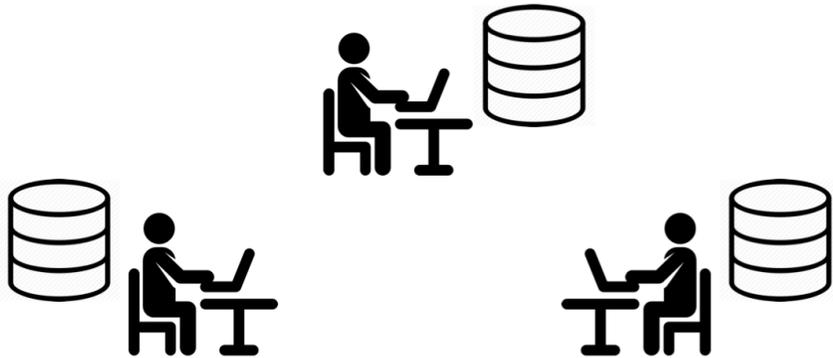
#hailGenetics #ATGUstrong

Learning Objectives

- How can you leverage Hail effectively on the cloud?
 - What are public compute clouds and how do they work?
 - How can you run Hail pipelines on the cloud?
 - What are best practices for cost management?

The cloud is the future of research computing

From Bringing Data to Researchers



To Bringing Researchers to Data



The cloud: pros and cons

Advantages:

- Many researchers can work on the same data at no additional cost
- No need to share resources with colleagues -- rent your own.
- High computer utilization means good cost efficiency
 - Pay for lots of CPUs when you need them, and pay nothing when you don't
- Great security and fault-tolerance for data
- Democratization: don't need access to institutional HPC cluster to participate in research (though do still need funding)

Disadvantages:

- Every operation has a cost, so an understanding of cost model is important
- Difficult to know how many resources to provision
 - Too small a cluster, you waste your time. Too large a cluster, you waste money.
- Cost overruns do happen
 - However, cloud providers will often refund accidental spend

Cloud Computing Platforms



best Hail
support
right now

Cloud computing products from Google

Google Storage (sometimes referred to as "Google Buckets")

- Store data, Python notebooks, anything you want.
- ~\$25 per TB per month.

Google Compute Engine (GCE)

- Rent a virtual machine, use it however you want.
- ~\$0.05 per CPU per hour for standard VMs
- ~\$0.01 per CPU per hour for preemptible VMs

Google Dataproc

- Rent a cluster running Apache Spark, which is Hail's distributed computing engine.
- GCE price, plus \$0.01 per CPU per hour.

hailctl, the manager for Hail on the cloud

hailctl = “hail control”

- **hailctl dataproc** is the Hail cloud manager for Google.
- **hailctl emr** (Amazon) and **hailctl azure** (Microsoft) planned.

Common cluster operations:

```
hailctl dataproc start MYCLUSTER --max-age 4h
```

```
hailctl dataproc connect MYCLUSTER notebook
```

```
hailctl dataproc submit MYCLUSTER script.py
```

```
hailctl dataproc stop MYCLUSTER
```

gsutil, file manager for Google Storage

gsutil = Google Storage Utilities

- Amazon and Microsoft clouds have their own analogs.

Create a new bucket (root directory)

```
gsutil mb gs://mybucket
```

List files in a bucket:

```
gsutil ls gs://mybucket  
gsutil ls gs://mybucket/subfolder
```

Copy data to/from the cloud

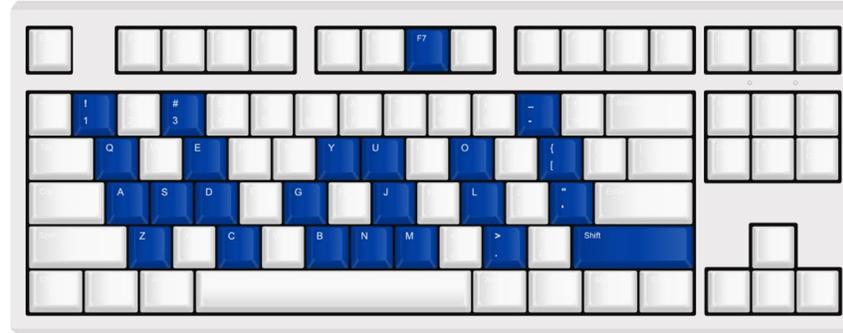
```
gsutil cp gs://mybucket/file /Users/me/data/file  
gsutil cp /Users/me/data/file gs://mybucket/file
```

Cost management best practices

- **Develop small, run big**
 - Iterate on pipelines using a piece of the full dataset (make chr22 your bestie)
 - Run pipelines on large clusters when ready
- **Manage risk**
 - Set billing limits and alerts (you'll get an email if you start to overspend)
 - Always use `--max-age` or `--max-idle` flags on cluster creation
 - Use Buckets with retention policies (data deleted after X days) when possible
- **Plan Ahead**
 - Calculate costs ahead of time where possible
 - <https://cloud.google.com/products/calculator/>

Learning Outcomes

- **You** can leverage Hail effectively on the cloud.
 - There are multiple public compute clouds, but Google Cloud has the most mature infrastructure for working with Hail.
 - Tools like `hailctl` and `gsutil` can help you easily get started running Hail pipelines on Google Dataproc.
 - In order to be a responsible cloud user, you should develop your scripts on small data, plan ahead when running large pipelines, and manage risk by setting up alerts and lifetime limits for expensive resources.



Sequencing and Introduction to Hail: Unlocking the power of the cloud with Hail

2021 Virtual Workshop on Statistical Genetics Methods for Human Complex Traits

June 16th, 2021 (practical)

Hosted by the Institute for Behavioral Genetics, University of Colorado, Boulder

Kumar Veerapen, PhD

Hail Support and Community Outreach Manager

Tim Poterba, Carolin Diaz, Dan Howrigan, John Compitello



<https://hail.is>
@mkveerapen / @hailgenetics
veerapen@broadinstitute.org

#scalableGenomics

#hailGenetics #ATGUstrong