Disease research in *big data* scale

Mauricio Carneiro

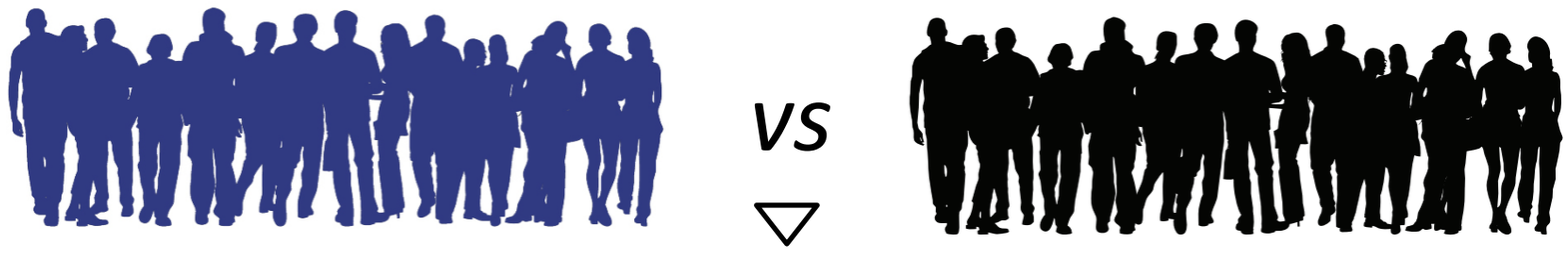
carneiro@broadinstitute.org

Group Lead, Computational Technology Development
Broad Institute of MIT and Harvard

To fully understand **one** genome we need **tens of thousands** of genomes

Rare Variant Association Study (RVAS)

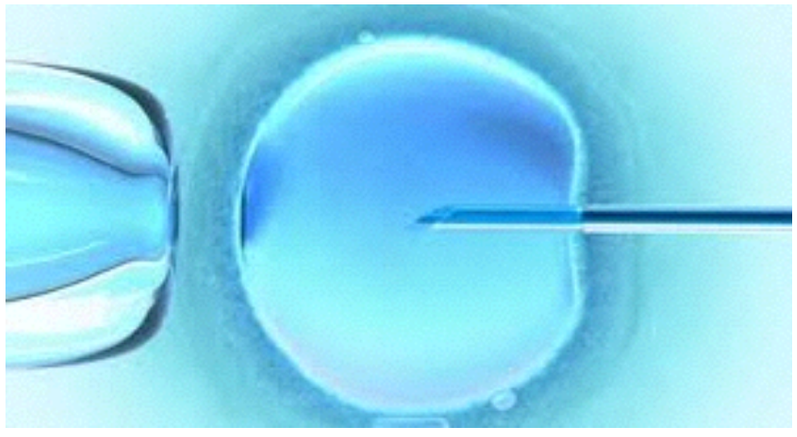
Common Variant Association Study (CVAS)



Personalized genomics for rare disease patients is already a success

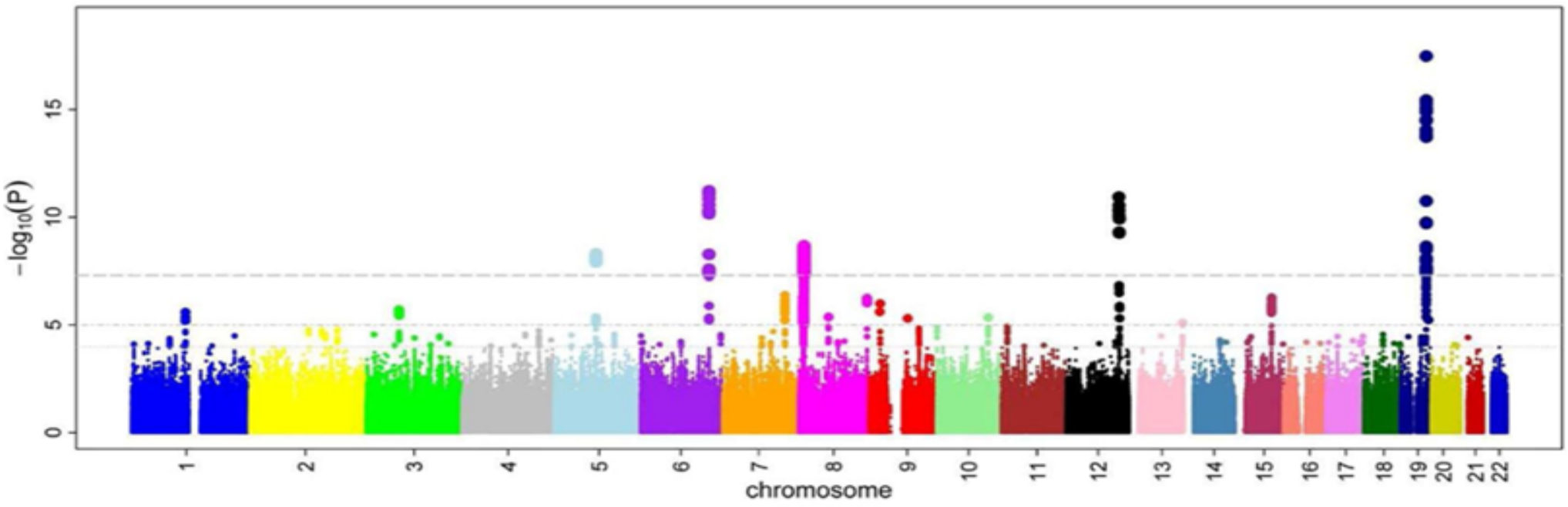
- patients with rare conditions can use DNA sequencing technology **today** to make sure that their children don't carry the rare disease.

- patient gets sequenced and an RVAS study identifies causal mutation
- couple undergoes *in vitro* fertilization and sequences the embryos
- embryos without the causal mutation are selected and a healthy baby is born.
- today we are limited on very rare mutations or diseases that research has provided data

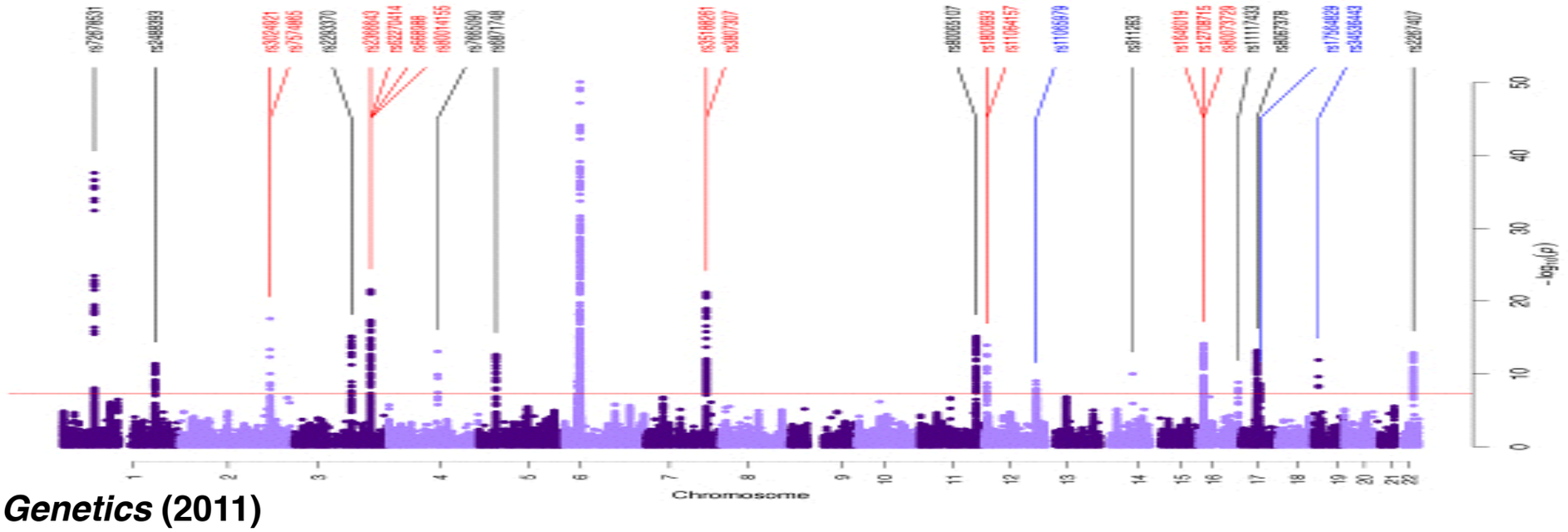


We need to understand the diseases in order to expand personalized genomics

1,000 samples



13,000 samples



Handsaker, R et al. *Nature Genetics* (2011)

Why personalized genomics for cancer patients?

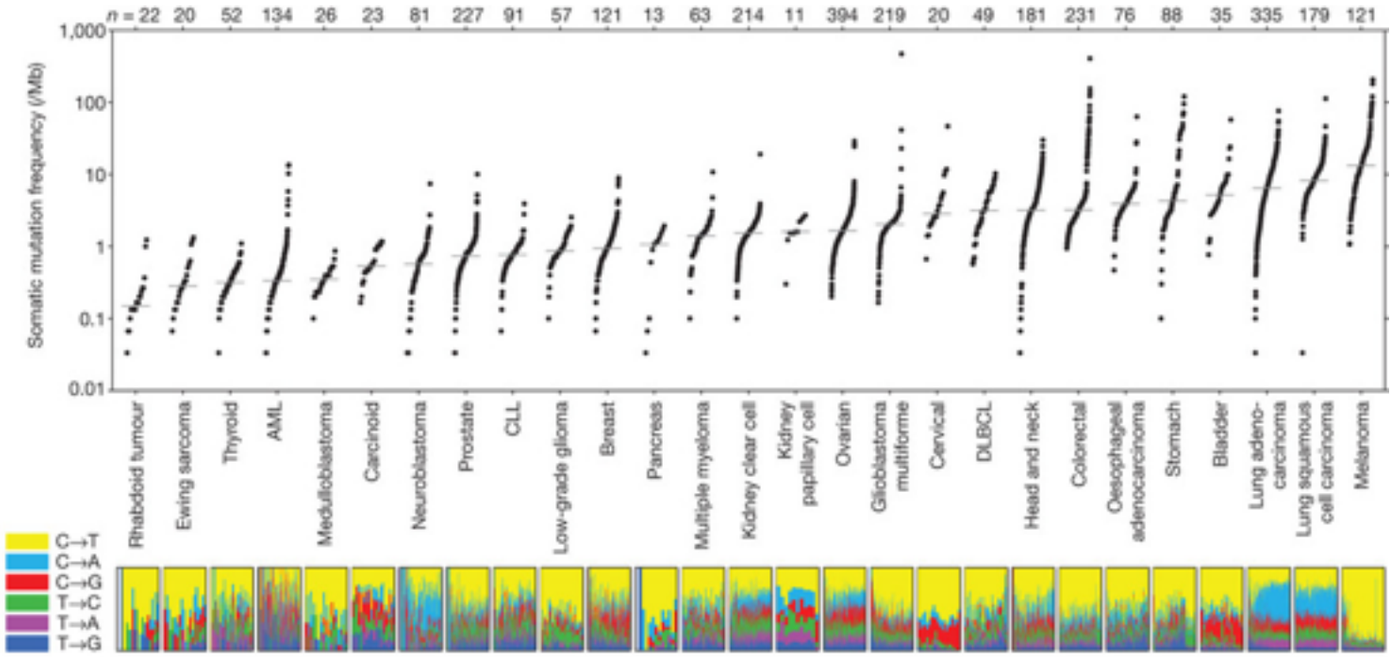
- patients rely on doctor's intuition to prescribe the right tests to identify the condition

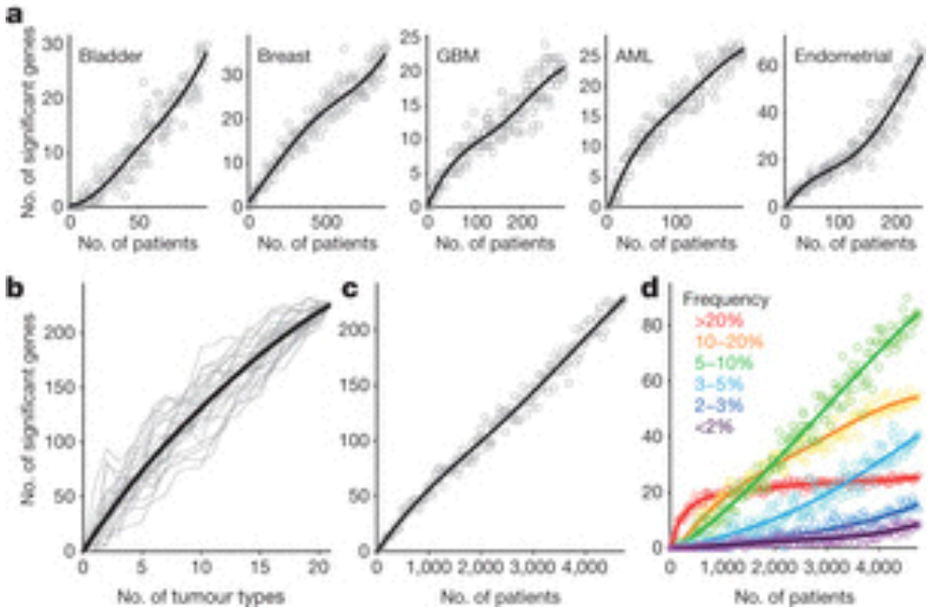
- tests are not always clear and diagnose typically is not possible until cancer is advanced

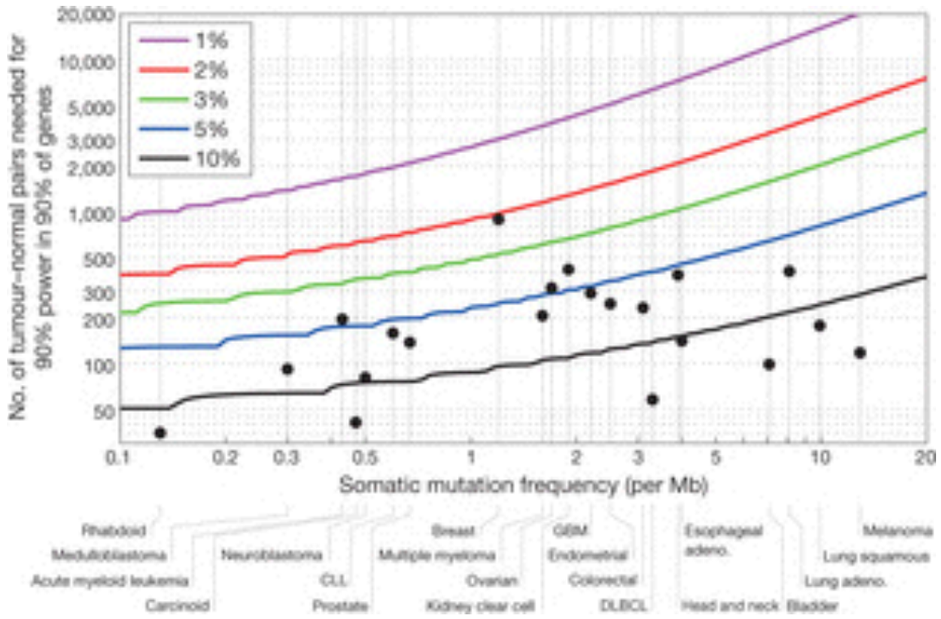
- treatments are disruptive and imprecise

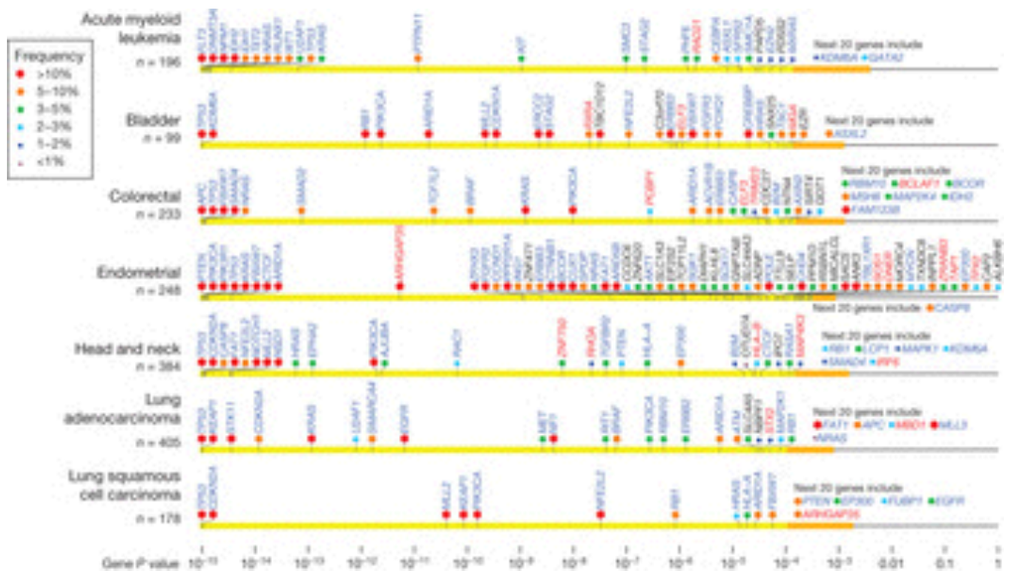
- personalized genomics could serve as a better diagnostics and targeted treatment tool.

Cancer research also needs more samples and a population based analysis









Lawrence M. *Nature* 2013

Lawrence M, *Nature* 2014

To enable personalized genomics we need to make progress in disease research

- personalized medicine will play an important role in:
  - ⋆ rare and common disease diagnostics
  - ⋆ common disease treatment guidance

- in order to get there, both paths are depending on:
  - ⋆ better knowledge of disease markers
  - ⋆ functional pathways and implications of variants
  - ⋆ treatment options from the pharmaceutical companies

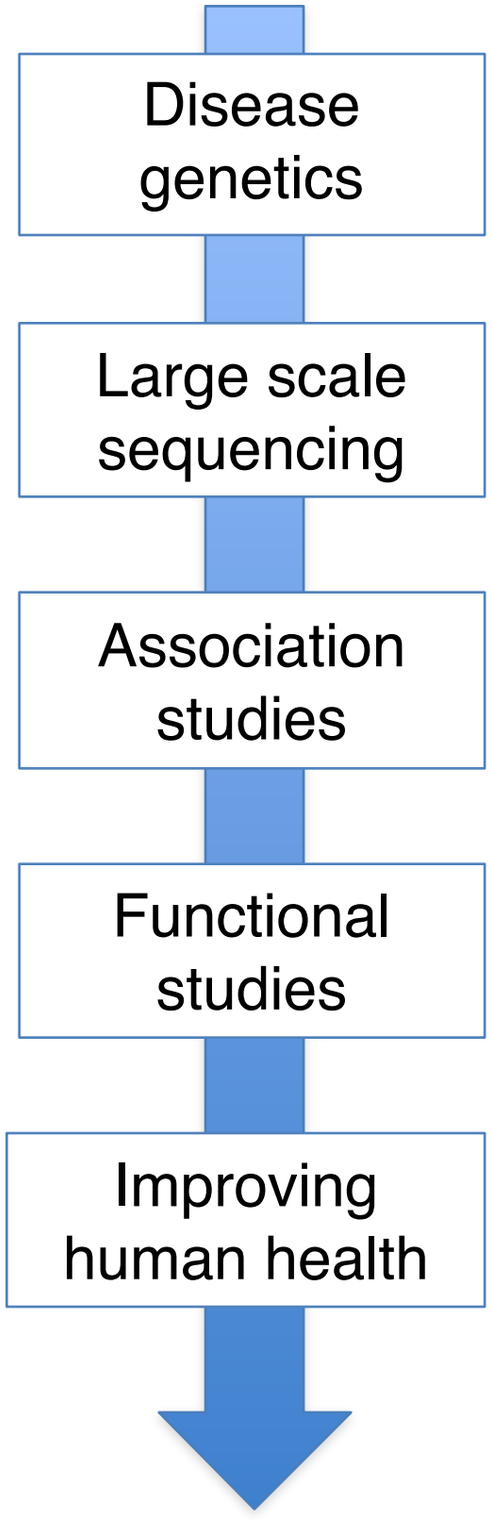
What is the BROAD INSTITUTE ?

The Broad Institute mission

This generation has a historic opportunity and responsibility to transform medicine by using systematic approaches in the biological sciences to dramatically accelerate the understanding and treatment of disease.

To fulfill this mission, we need new kinds of research institutions, with a deeply collaborative spirit across disciplines and organizations, and having the capacity to tackle ambitious challenges.

How is the Broad achieving these goals?



Many simple and complex human diseases are heritable

Affected and unaffected individuals differ systematically in their genetic composition

These systematic differences can be identified by comparing affected and unaffected individuals

These associated variants give insight into the biological mechanisms of disease

These insights can be used to intervene in the disease process itself

Broad Institute in 2013

50
HiSeqs

10
MiSeqs

2
NextSeqs

14
HiSeq X

6.5
Pb of data

427
projects

180
people

2.1
Tb/day

* we also own 1 *Pacbio RS* and 4 *Ion Torrent* for experimental use

Broad Institute in 2013

44,130
exomes

2,484
exome express

2,247
genomes

2,247
assemblies

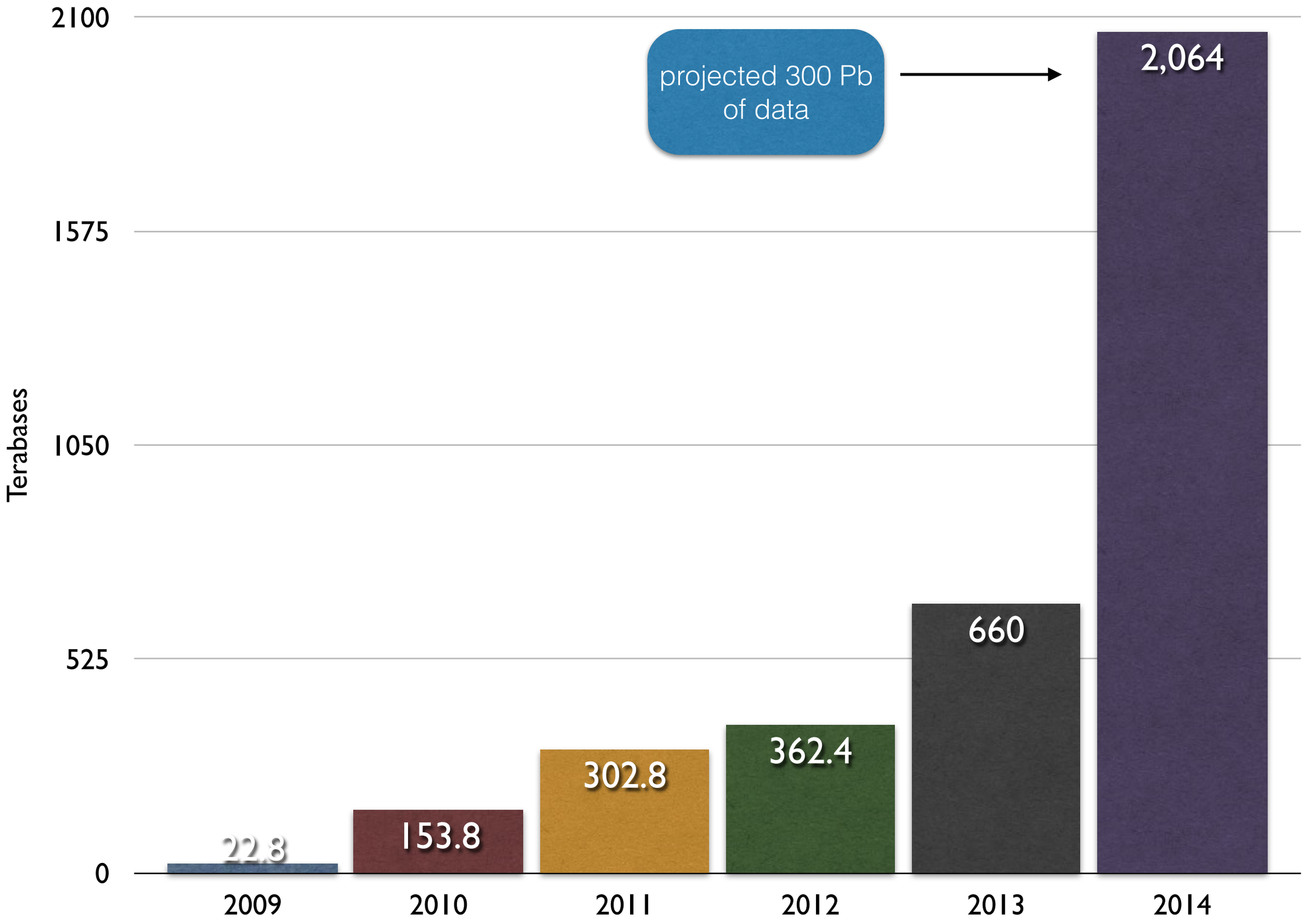
8,189
RNA

9,788
16S

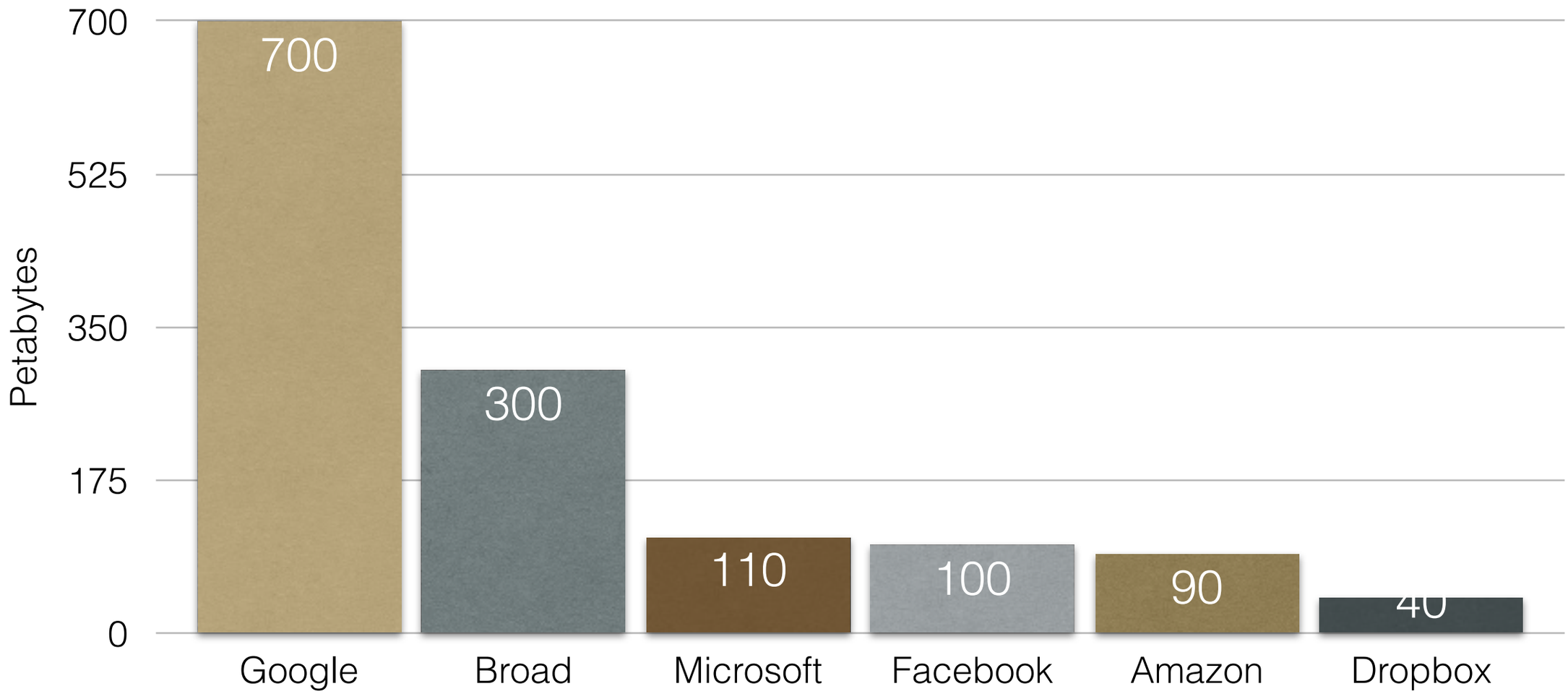
47,764
arrays

228
cell lines

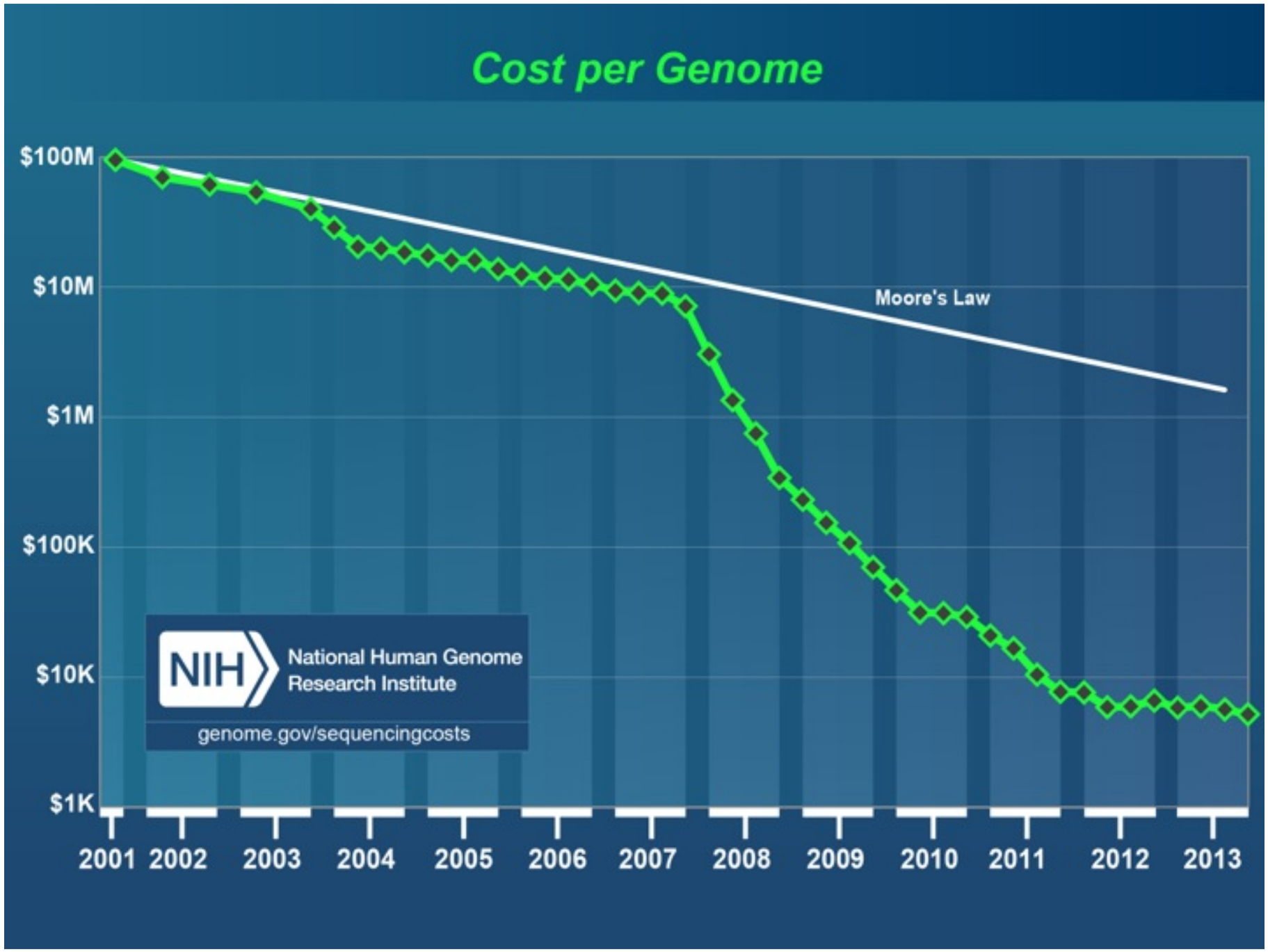
Terabases of Data Produced by Year



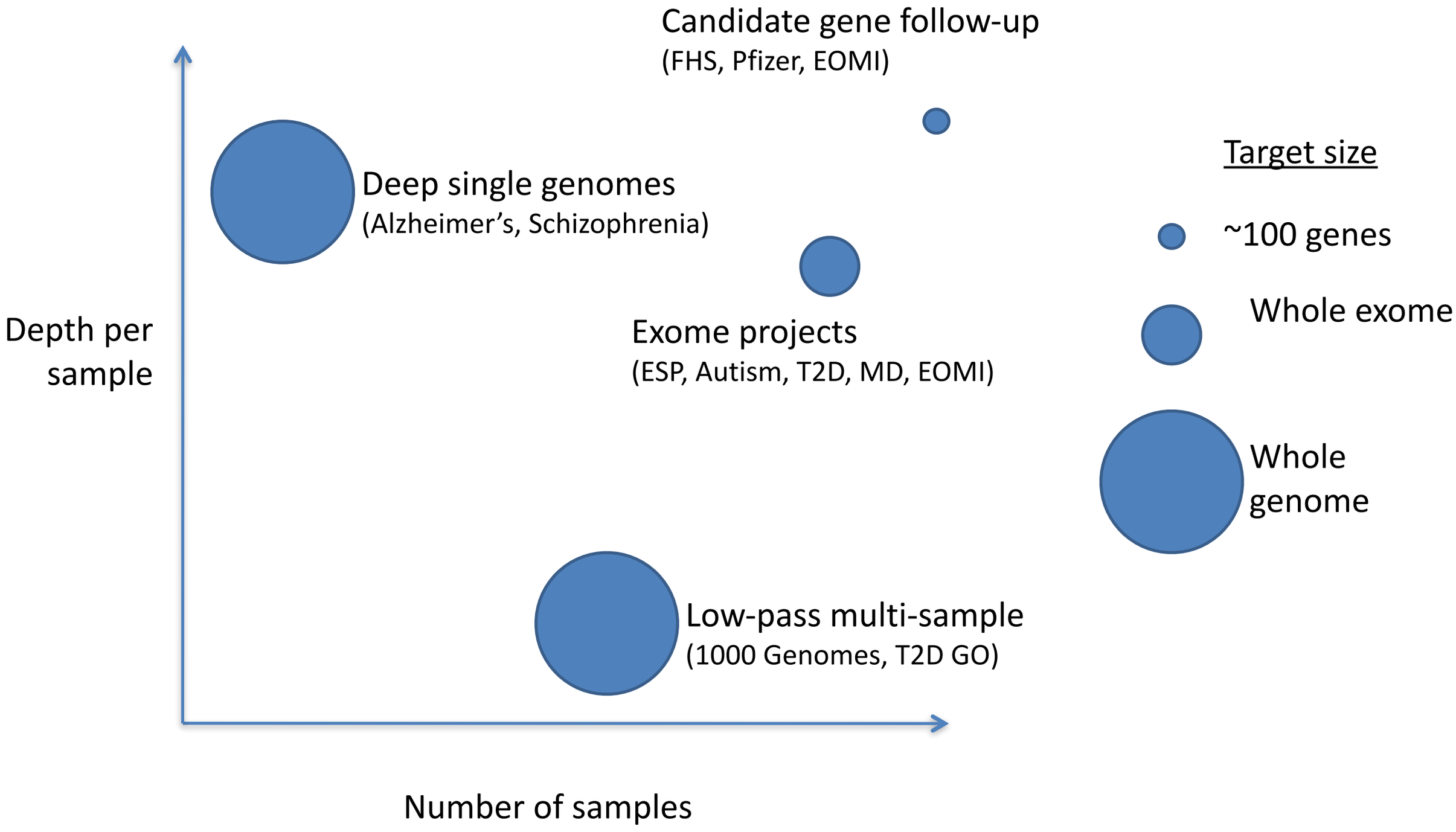
We produce as much data as the big cloud providers



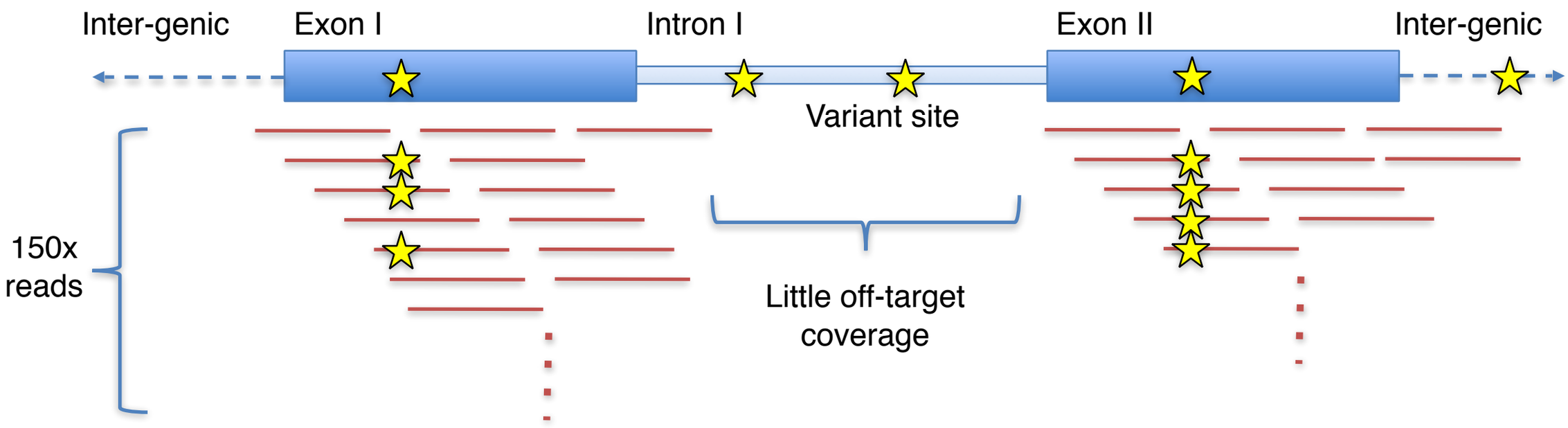
and these numbers will continue to grow faster than Moore’s law



A variety of experimental designs, mostly selected to maximize disease sample size



Exome capture sequencing design: >100K samples generated so far



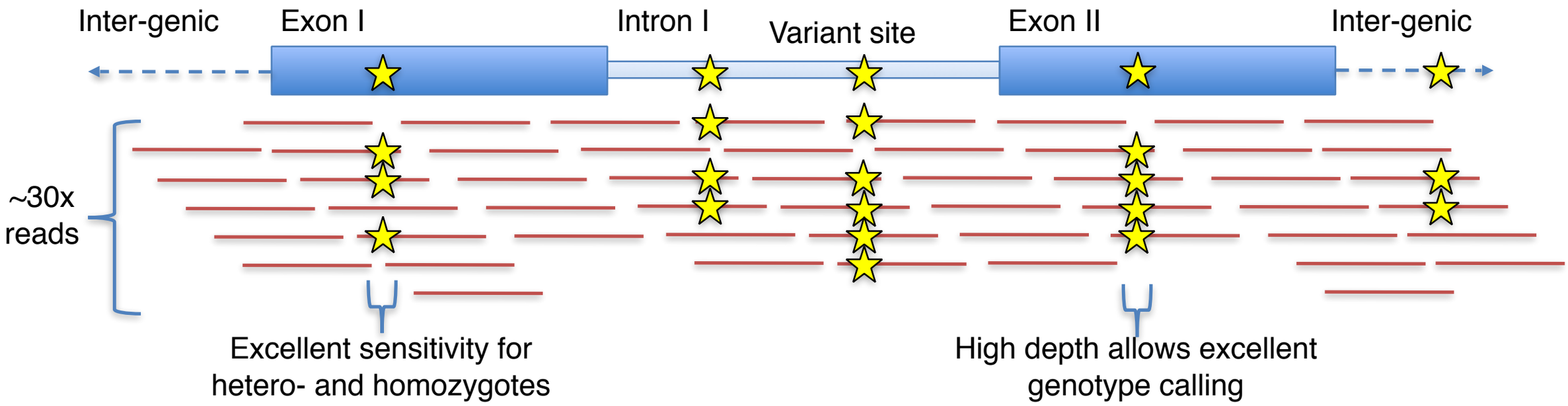
Data requirements per sample

Targeted bases	***~32Mb***
Coverage	***>80% 20x***
# sequenced bases	***5 Gb***
# lanes of HiSeq	***~0.33***

Variant detection among multiple samples

Variants found per sample	***~20K***
Percent of variation in genome	***0.5%***
Pr{singleton discovery}	***~95%***
Pr{common allele discovery}	***~95%***

High-pass sequencing design:
~10K generated, mostly in cancer



Data requirements per sample

Targeted bases	***~3 Gb***
Coverage	***Avg. 30x***
# sequenced bases	***100 Gb***
# lanes of HiSeq	***~8 lanes***

Variant detection among multiple samples

Variants found per sample	***~3-5M***
Percent of variation in genome	***>99%***
Pr{singleton discovery}	***>99%***
Pr{common allele discovery}	***>99%***

All of this NGS data processing requires a lot of CPU power and high-performance storage

$$\text{storage} \approx 2\frac{\text{bytes}}{\text{bp}} \times \text{targeted area}$$

Per sample data processing

The read data for each sample can be aligned, locally realigned, and recalibrated independently, for only **a few CPU hours per exome**.

Multi-sample variant calling

Joint calling takes approximately **1.5 cpu/days per exome**

Example Storage requires

	Target	Storage
Per sample		
Single WEx	32 Mb	25 Gb
Deep WGS	2.85 Gb	250 Gb
Complete Project		
700 EOMI exomes	32 Mb	20 Tb
Deep Trio WGS	2.85 Gb	750 Tb

Resources required for 57,000 samples project

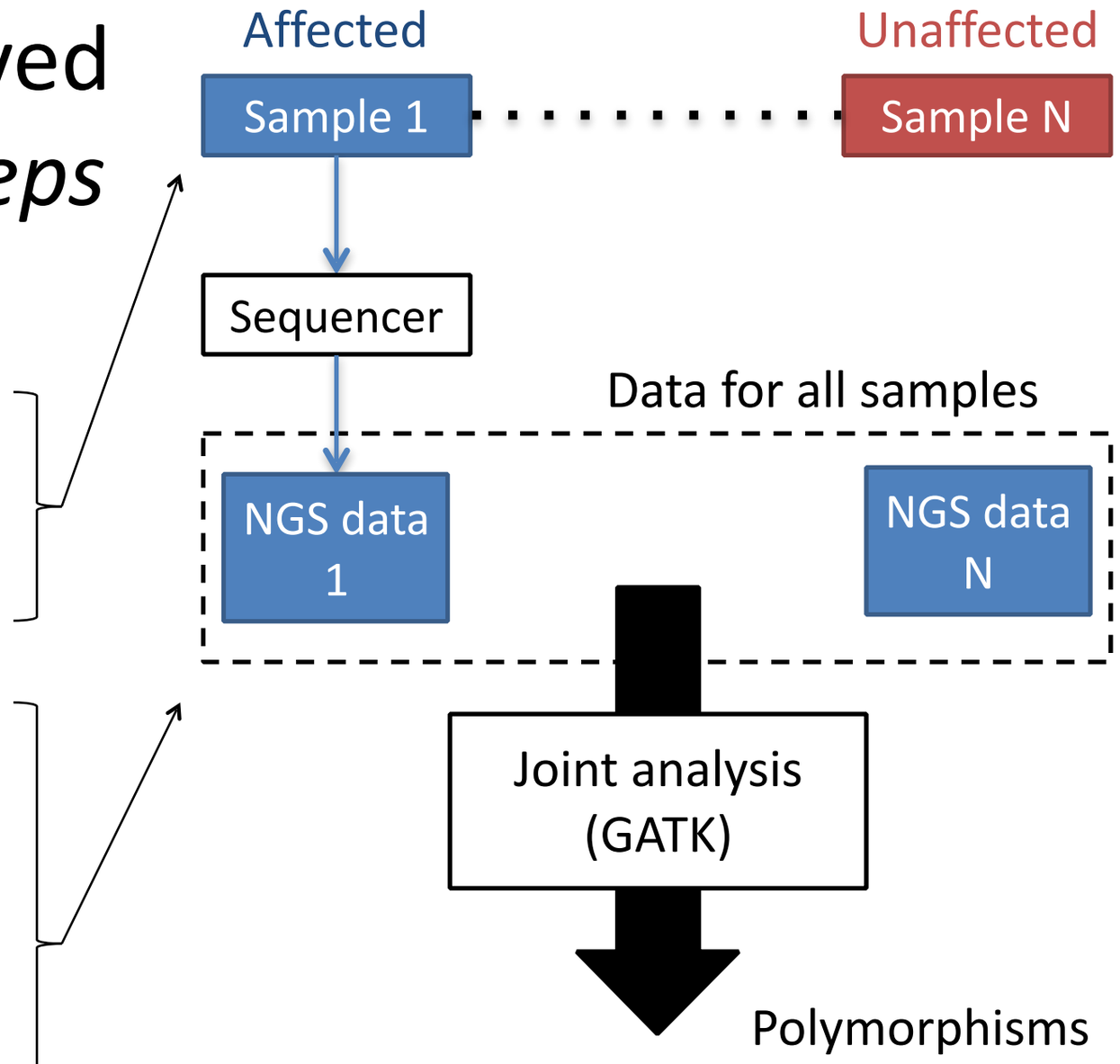
Processing	Storage	Result VCF
2,508,000 CPU/hours	1 Petabyte	2 Terabytes

How to discover loci involved in disease in *three easy steps*

1. Get thousands of affected and tens of thousands of unaffected samples

2. Sequence samples with next-generation sequencer
   - Discover polymorphisms (SNPs, indels, etc) across samples
   - Determine genotype of every sample at each variant site

3. Look for systematic differences in genotypes for affected samples vs. unaffected across all sites



	SNP	Indel
Affected	***10/1000***	***10/1000***
Unaffected	***25/2000***	***1/2000***
Association strength (P)	***0.72***	***0.0001***

Every RVAS and CVAS studies start with a complete variant matrix

All case and control samples

~3M variants

	Site	*Variant*	*Sample 1*	*Sample 2*	*…*	*Sample N*
SNP	***1:1000***	***A/C***	***0/0 0,10,100***	***0/1 20,0,200***	*…*	***0/0 0,100,255***
Indel	***1:1050***	***T/TC***	0/0 0,10,100	0/0 0,20,200	…	1/0 255,0,255
SNP	***1:1100***	***T/G***	***0/0 0,10,100***	***0/1 20,0,200***	…	***0/0 0,100,255***
	…	***…***	…	…	…	…
SNP	***X:1234***	***G/T***	***0/1 10,0,100***	***0/1 20,0,200***	…	***1/1 255,100,0***

Genotypes:
0/0 ref
0/1 het
1/1 hom-alt

Likelihoods:
A/B/C phred-scaled probability of hom (A), het (B), hom-alt (C) genotypes given NGS data

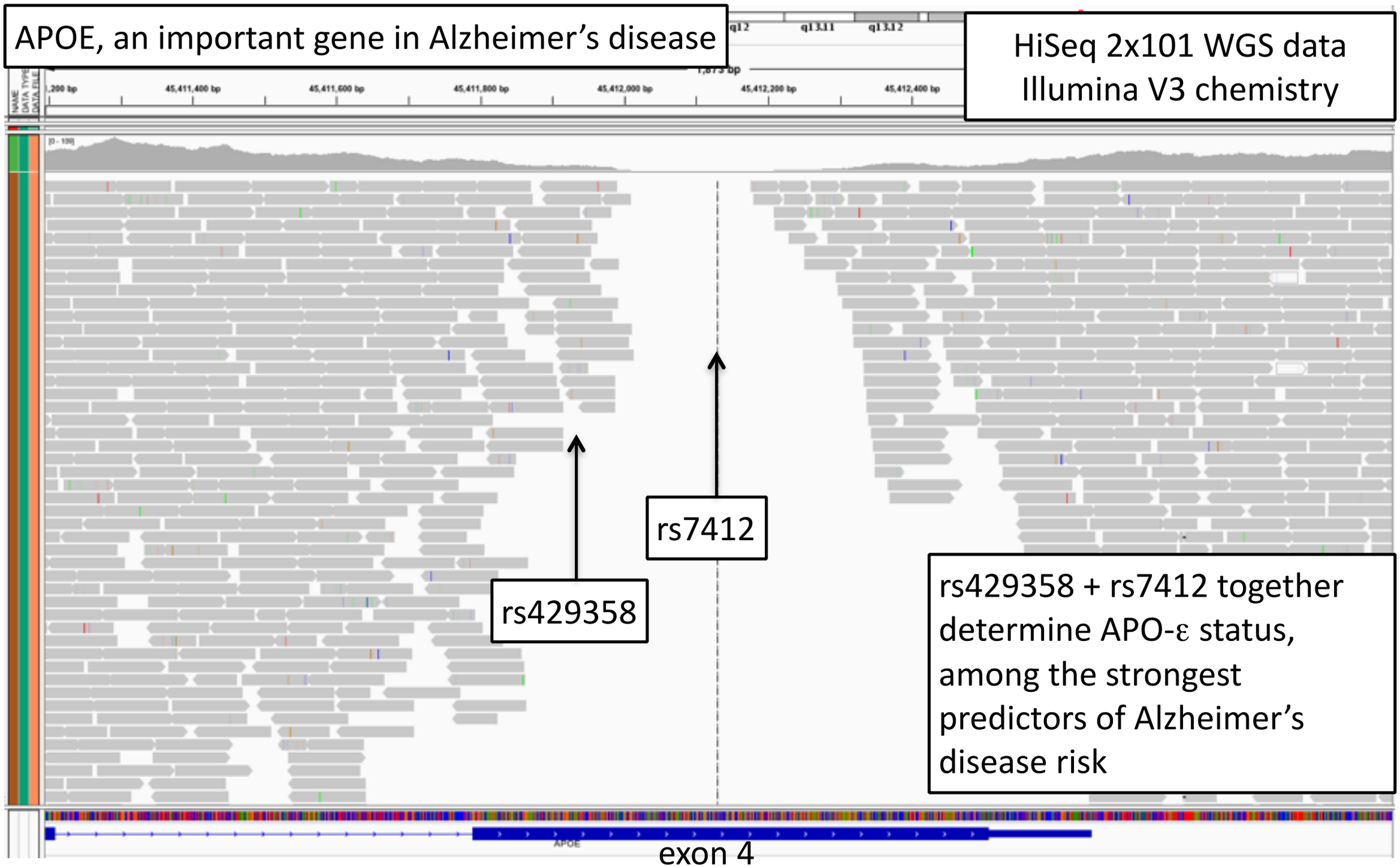
What is limiting our ability simply use NGS to find disease-causing variants?

- **Technical limitations**
  - Systematic lack of coverage in key regions (esp. genes!)
  - Complex error process
    - Mistake errors for true variants (false positives)
    - Overly conservative, making us miss real variants

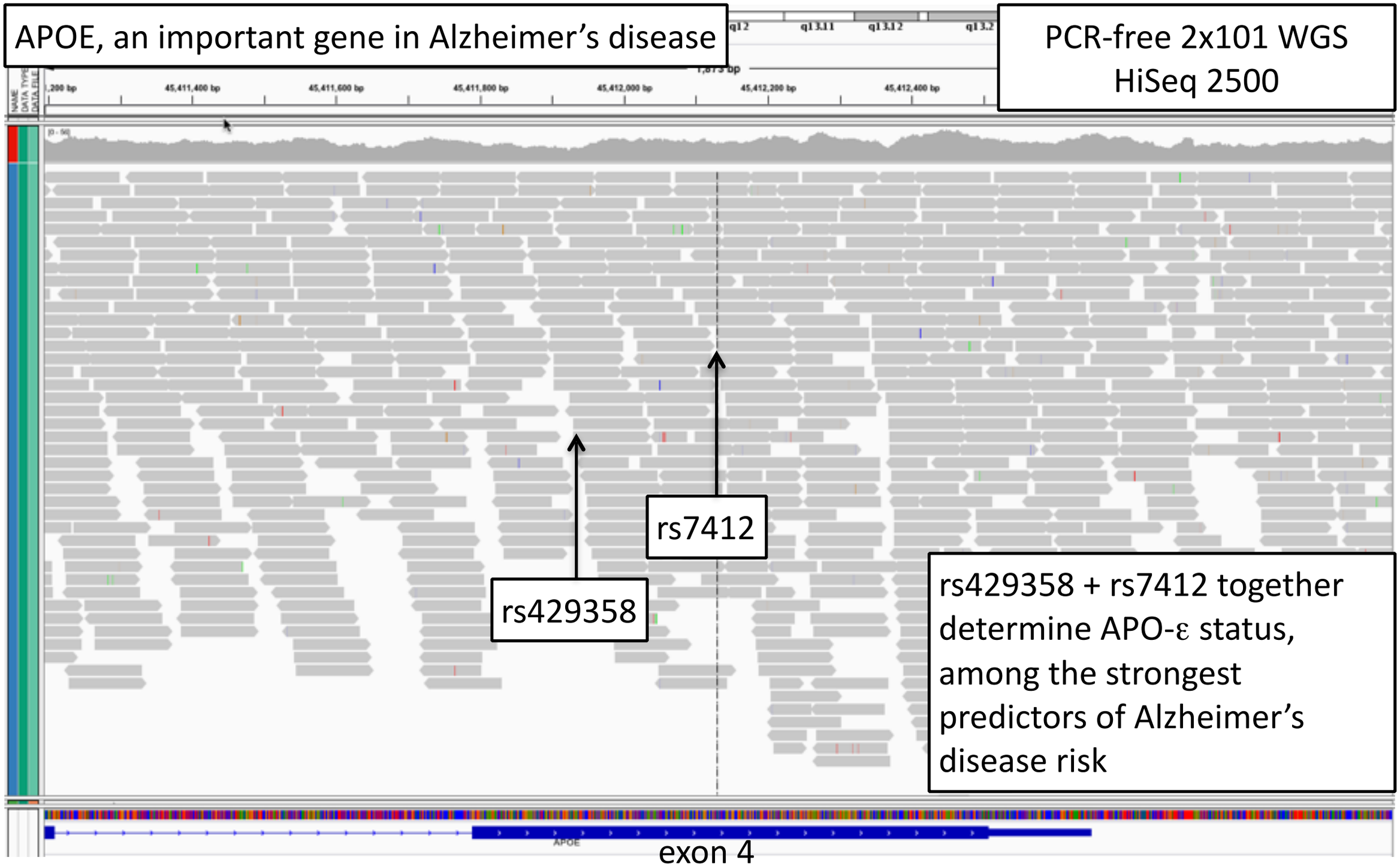
- **Analytic limitations**
  - Misinterpret our data right before our eyes, calling the wrong variant near the right place (esp. indels)

- **Don’t have power to associate variants with disease**
  - Need to analyze more samples, both with having lower cost sequencing and aggregating already existing data
  - Data needs to be consistently processed to be shared and analyzed effectively — *GATK Best Practices*

Technical problems such as poor coverage blinds us to many (important) genomic regions

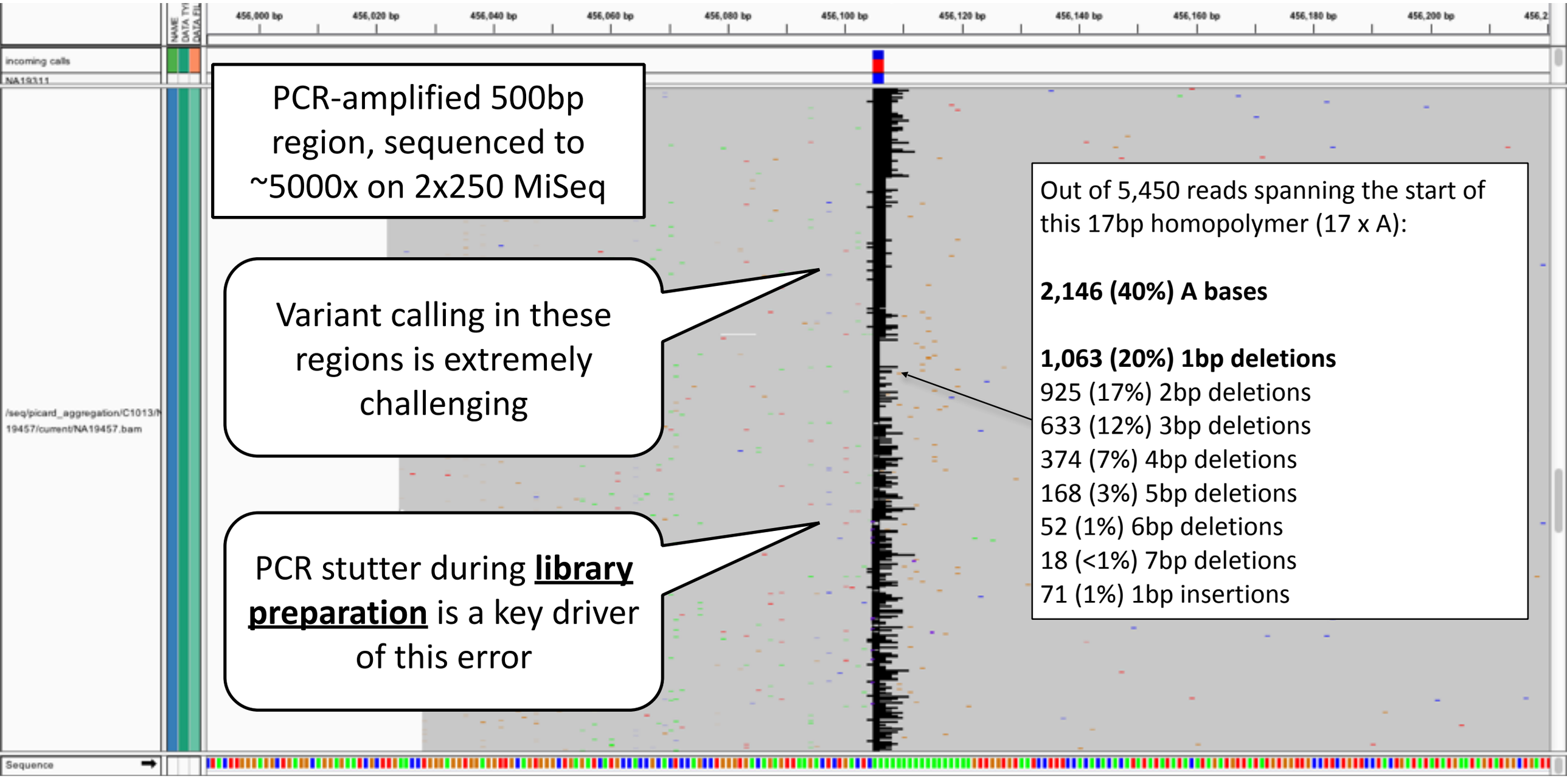


PCR-free libraries improve sequencing experiments



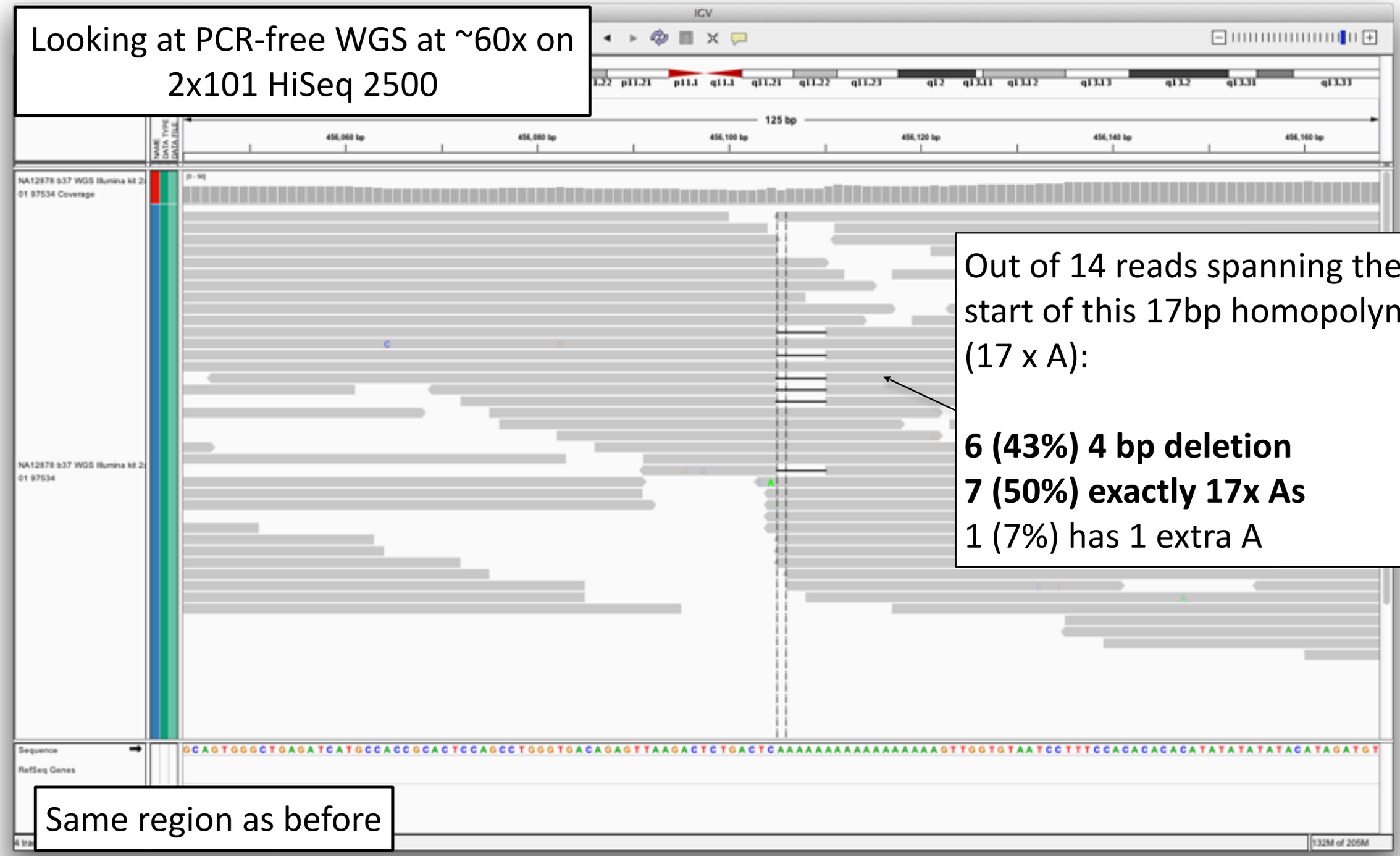
complex error process caused by PCR inundates us with false positives is also fixed in PCR-free libraries

Poorly-behaved region of the genome



Addressing systematic errors though better (PCR-free library) data

Looking at PCR-free WGS at ~60x on 2x101 HiSeq 2500

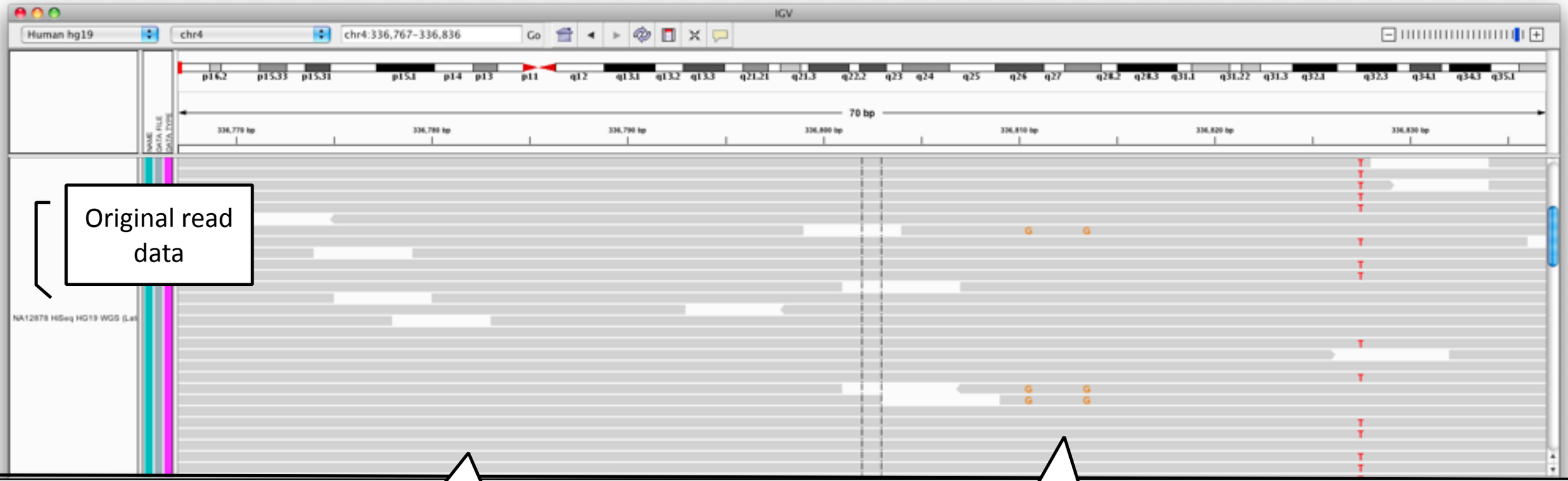


Out of 14 reads spanning the start of this 17bp homopolymer (17 x A):

6 (43%) 4 bp deletion
7 (50%) exactly 17x As
1 (7%) has 1 extra A

Same region as before

More advanced analysis tools will allow us to make more from today's data



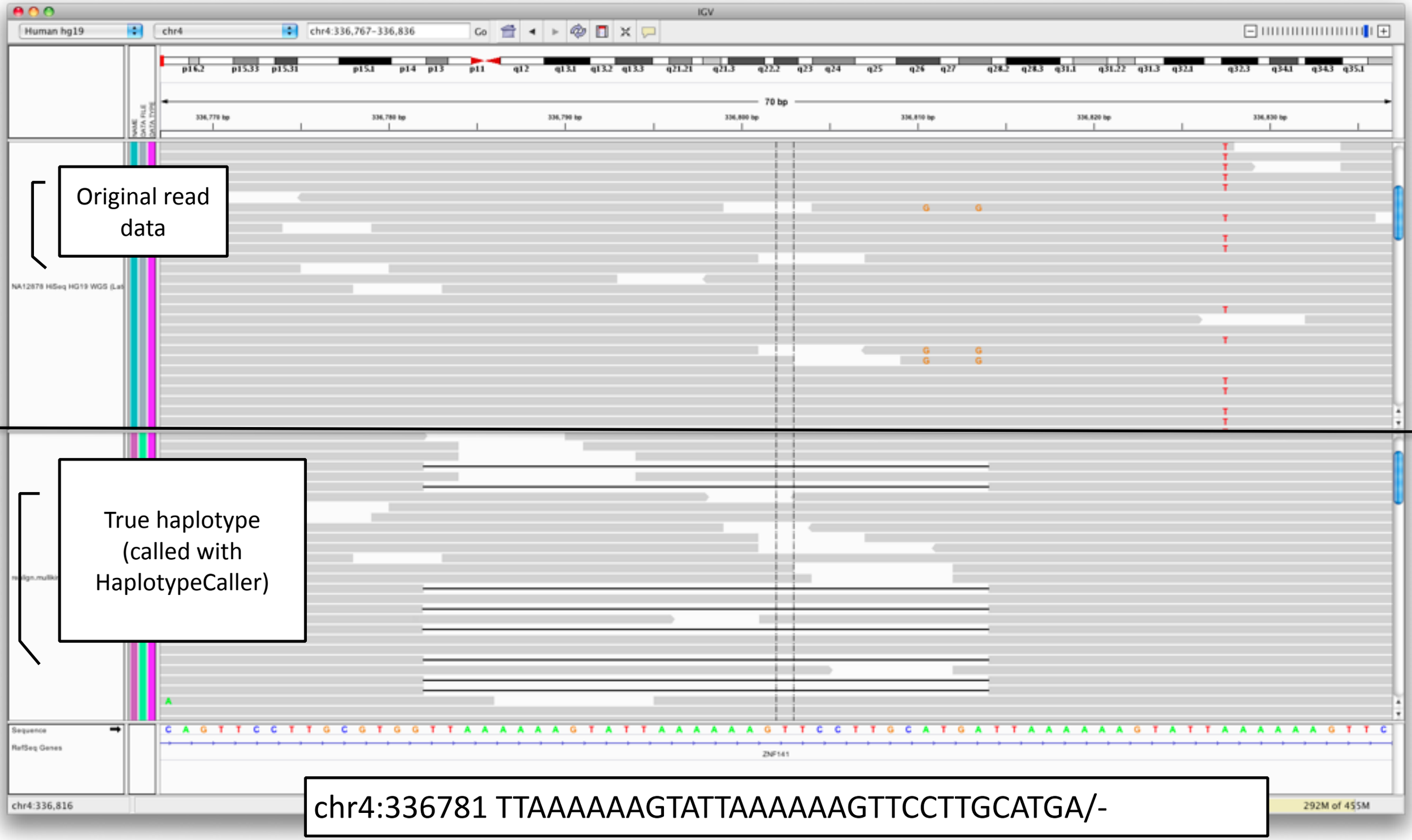
Looks confidently reference.

A gVCF file would undoubtedly say "confidently reference here"

Some indication that there's a multi-nucleotide substitution. Even on a different haplotype than the A/T SNP

Clean looking het A/T SNP

Such as the GATK's HaplotypeCaller



chr4:336781 TTAAAAAAGTATTAAAAAAGTTCCTTGCATGA/-

…but today's biggest limitation is **sample size**

- Suppose I sequence 500 people affected by Alzheimer's disease.
- I discover an loss-of-function indel in some interesting brain gene that is present in 10 samples
- What can I say about this variant?

Association of an indel with Alzheimer's disease risk (made up example)

Comparison	Just my 500 samples	Analyze with 500 samples	Analyze with 1K samples	Analyze with 10K samples	Analyze with 100K samples
Affected	***10/1000***	***10/1000***	***10/1000***	***10/1000***	***10/1000***
Unaffected	None	0/1000	1/2000	10/20000	100/200000
Association (P-value)	***None***	***10***	***10***	***10***	***10***
<u>What did I learn?</u>	I should have sequenced some controls	Not remotely significant	Still lost in the noise	Almost significant!	Important discovery!!

Massive data aggregation: The future of large-scale medical sequencing

Proposition

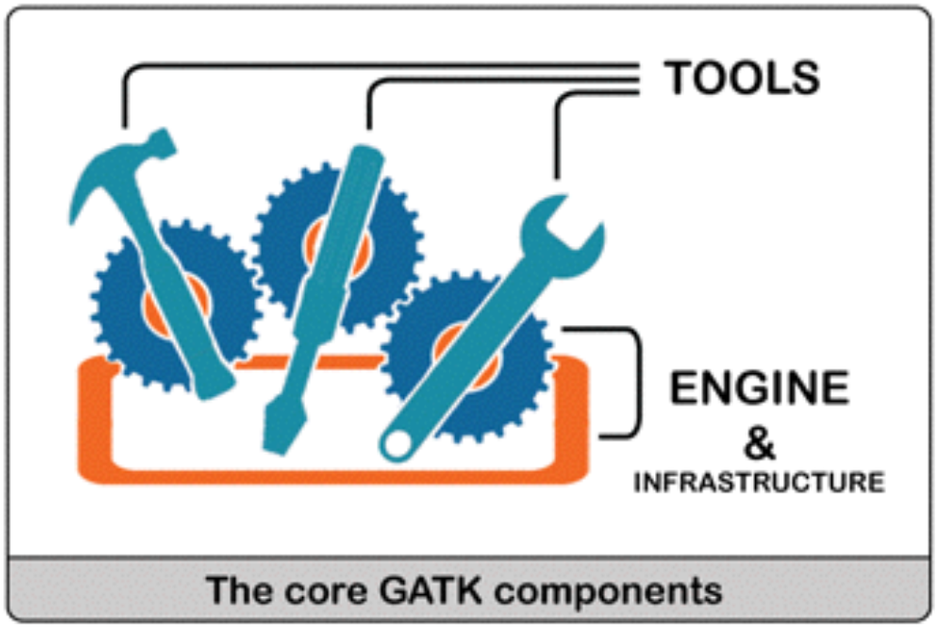
- cost of sequencing has fallen one-million-fold, enabling an explosion of information about the genetic basis of disease
- Learning from the world's combined genomic and clinical data will dramatically accelerate progress
- Aggregated sequence data will be needed to guide the interpretation of genome sequences in clinical practice

Technical challenge

- Aggregating hundreds of thousands of variants and patients requires a sophisticated database system with proper data protection, fast distributed access and a standardized API for tool development.
- For a truly global reach new protective legislation on consents, data access and sharing, IRB processes and patient education will be necessary.

GATK is both a toolkit and a programming framework, enabling NGS analysis by scientists worldwide

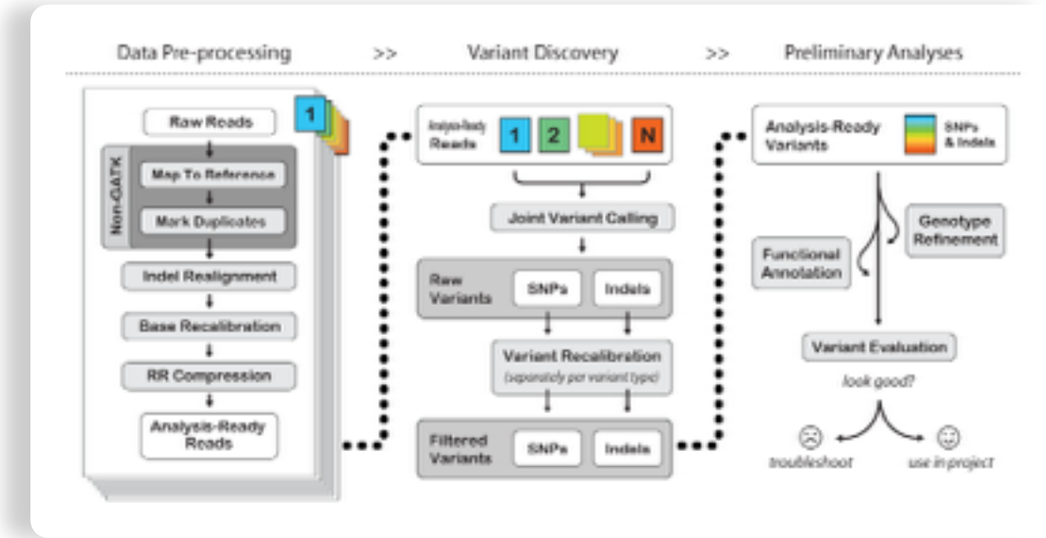
Toolkit & framework packages



The core GATK components

Toolkit

Best practices for variant discovery



Framework

MuTect, XHMM, GenomeSTRiP, ...

Tools developed on top of the GATK framework by other groups

Extensive online documentation & user support forum serving >10K users worldwide

http://www.broadinstitute.org/gatk

About
Overview of the GATK and the people behind it

Guide
Detailed documentation, guidelines and tutorials

Community
Forum for questions and announcements

Events
Materials from live and online events

iSeqTools

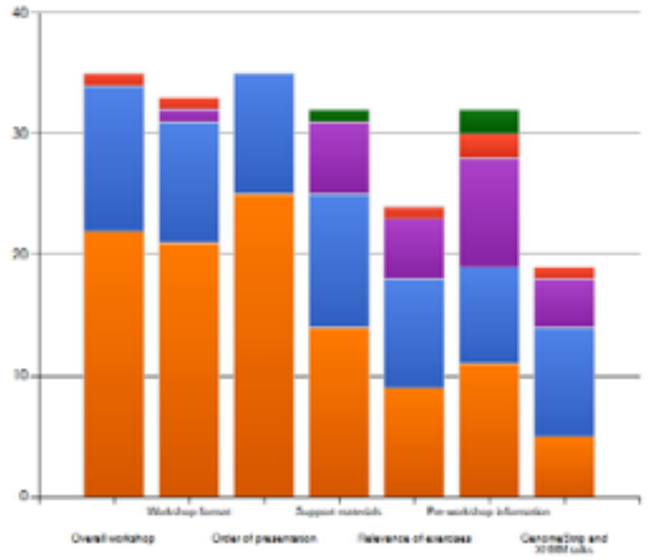
Workshop series educates local and worldwide audiences

Completed:
- Dec 4-5 2012, Boston
- July 9-10 2013, Boston
- July 22-23 2013, Israel
- Oct 21-22 2013, Boston

Planned:
- March 3-5 2014, Thailand
- Oct 18-29 2014, San Diego

Format
- Lecture series (general audience)
- Hands-on sessions (for beginners)

Portfolio of workshop modules
- GATK Best Practices for Variant Calling
- Building Analysis Pipelines with Queue
- Third-party Tools:
  - GenomeSTRiP
  - XHMM



- High levels of satisfaction reported by users in polls
- Detailed feedback helps improve further iterations

Tutorial materials, slide decks and videos all available online through the GATK website, YouTube and iTunesU

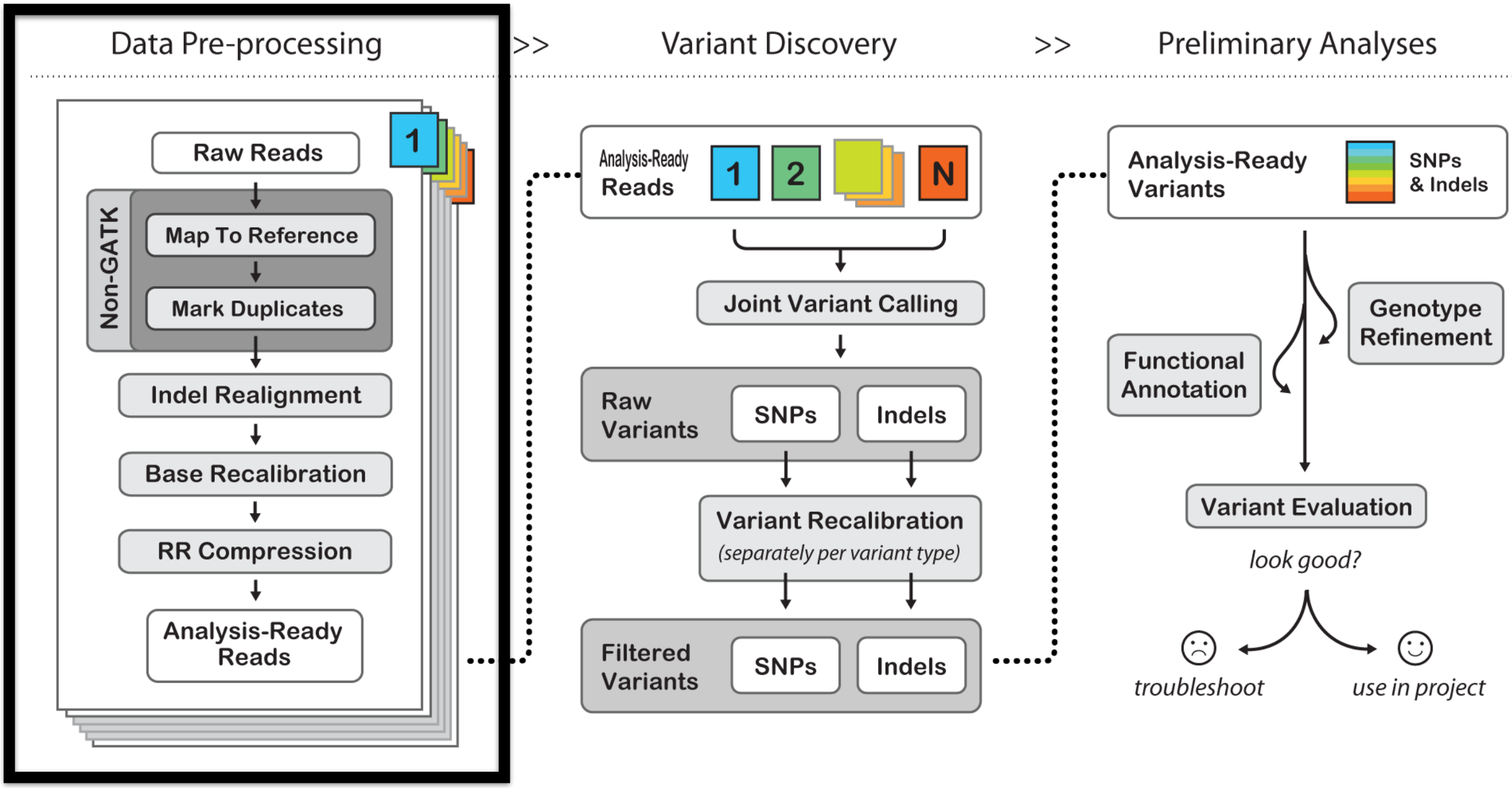
BroadE: Overview of GATK & best practices

by **broadinstitute** • 1 week ago • 1 view

Copyright Broad Institute, 2013. All rights reserved. The presentations below were filmed during the 2013 **GATK** Workshop, part of ...

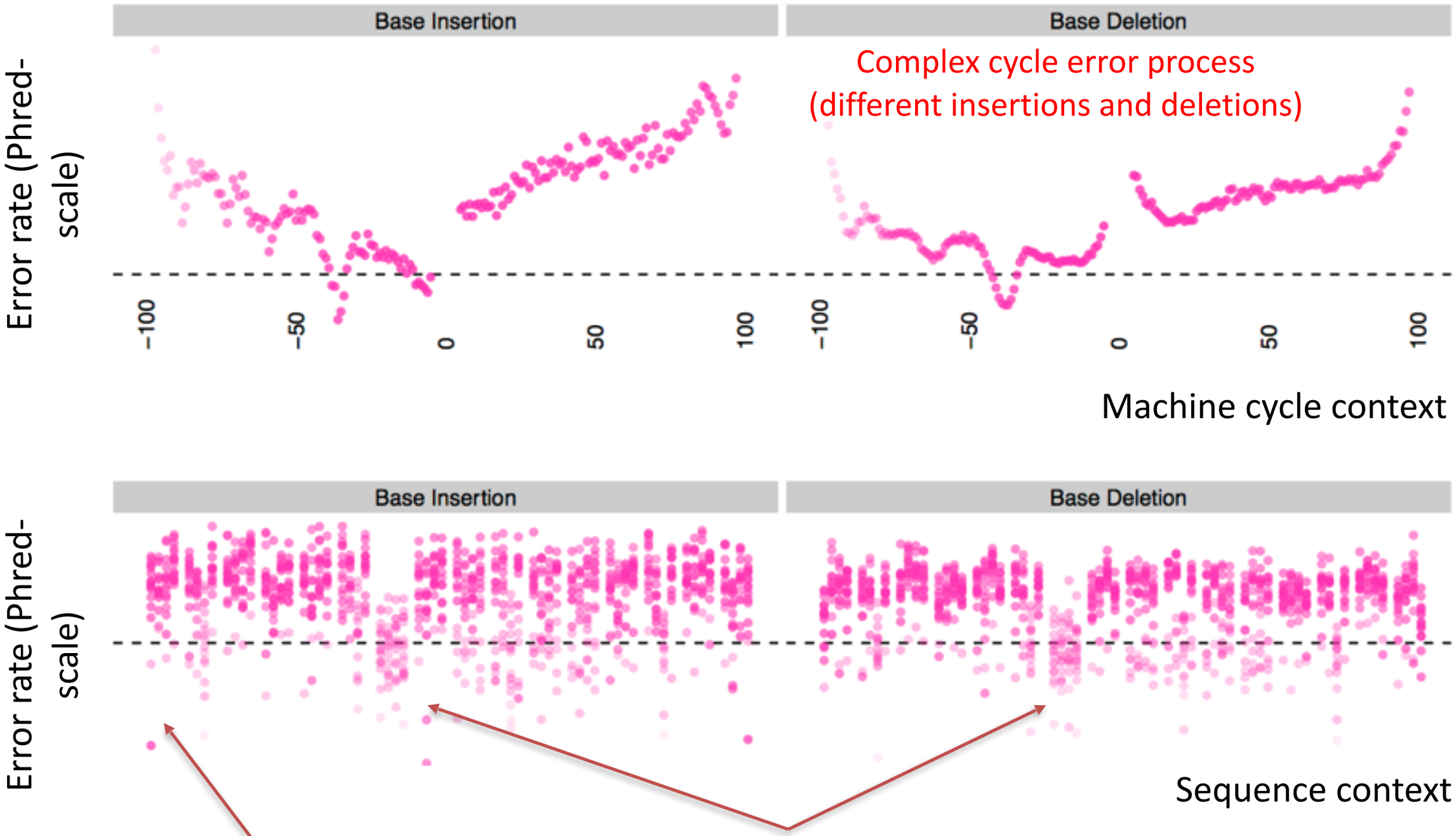
NEW HD

We have defined the best practices for sequencing data processing

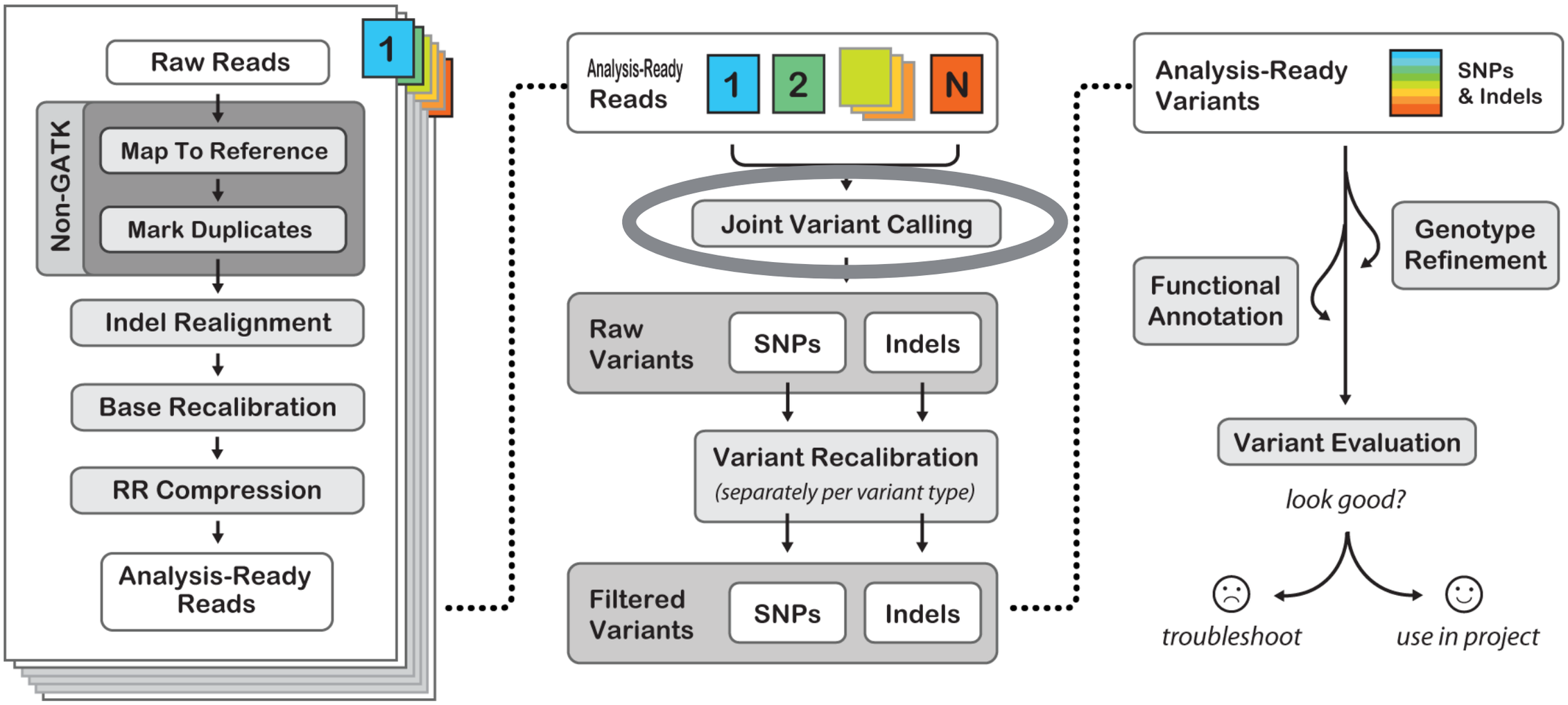


Auwera, GA et al. *Current Protocols in Bioinformatics* (2013)

We can model the error process of the machine in fine detail

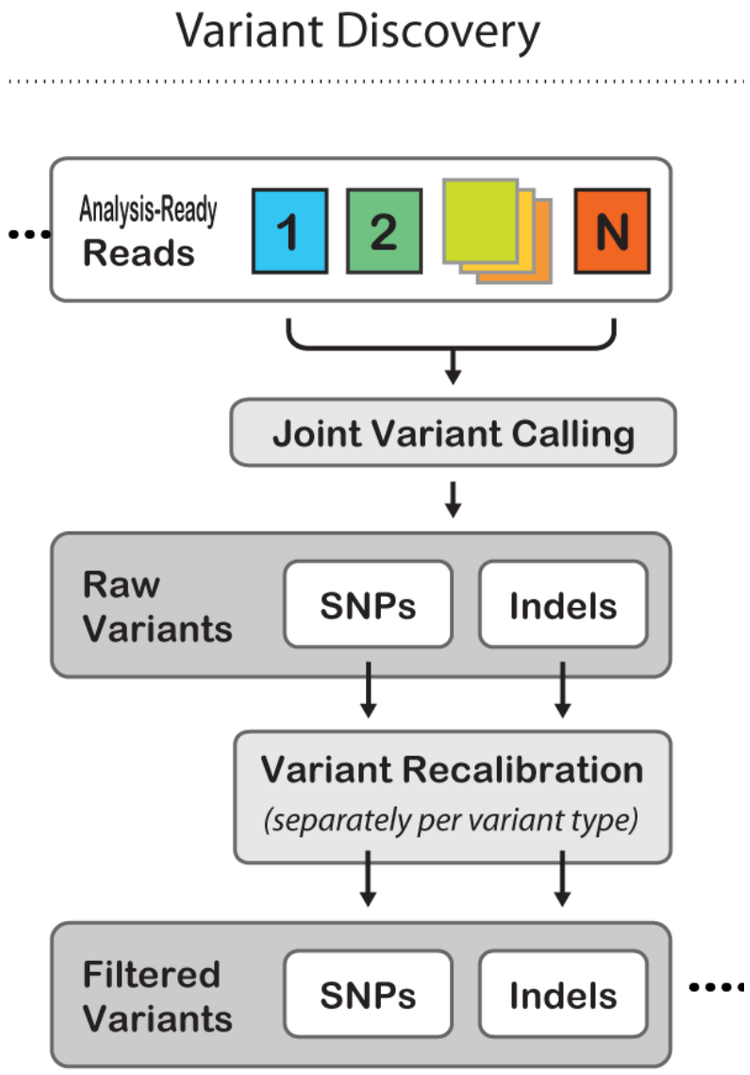


Joint calling is an important step in Variant Discovery



Auwera, GA et al. *Current Protocols in Bioinformatics* (2013)

Variant calling is a large-scale bayesian modeling problem



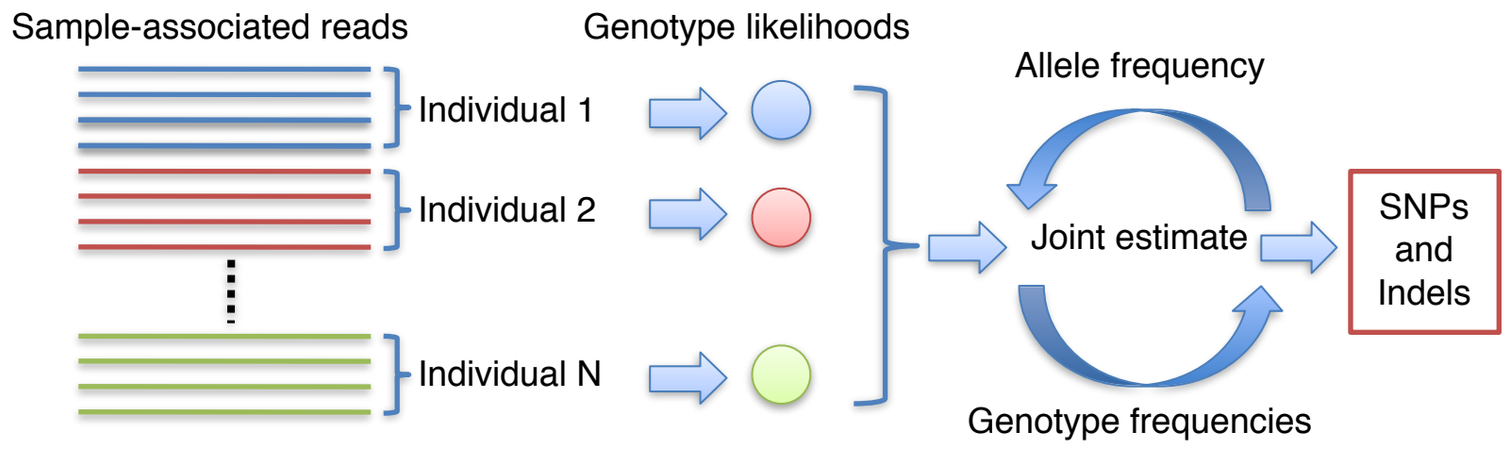
prior Likelihood

$$\Pr\{G|D\} = \frac{\Pr\{G\}\Pr\{D|G\}}{\Sigma_i \Pr\{G_i\}\Pr\{D|G_i\}}, \text{ [Bayes' rule]}$$

Diploid

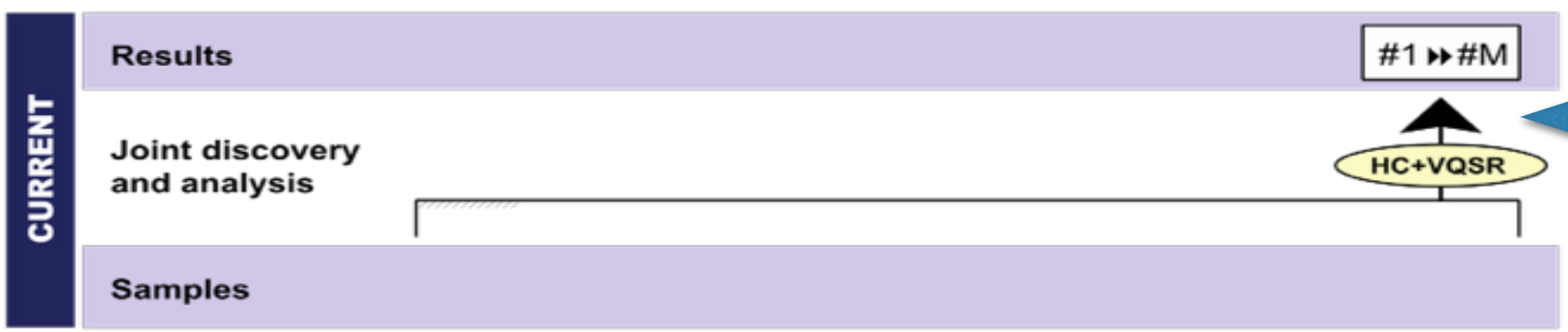
$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2}\right) \text{ where } G = H_1H_2$$

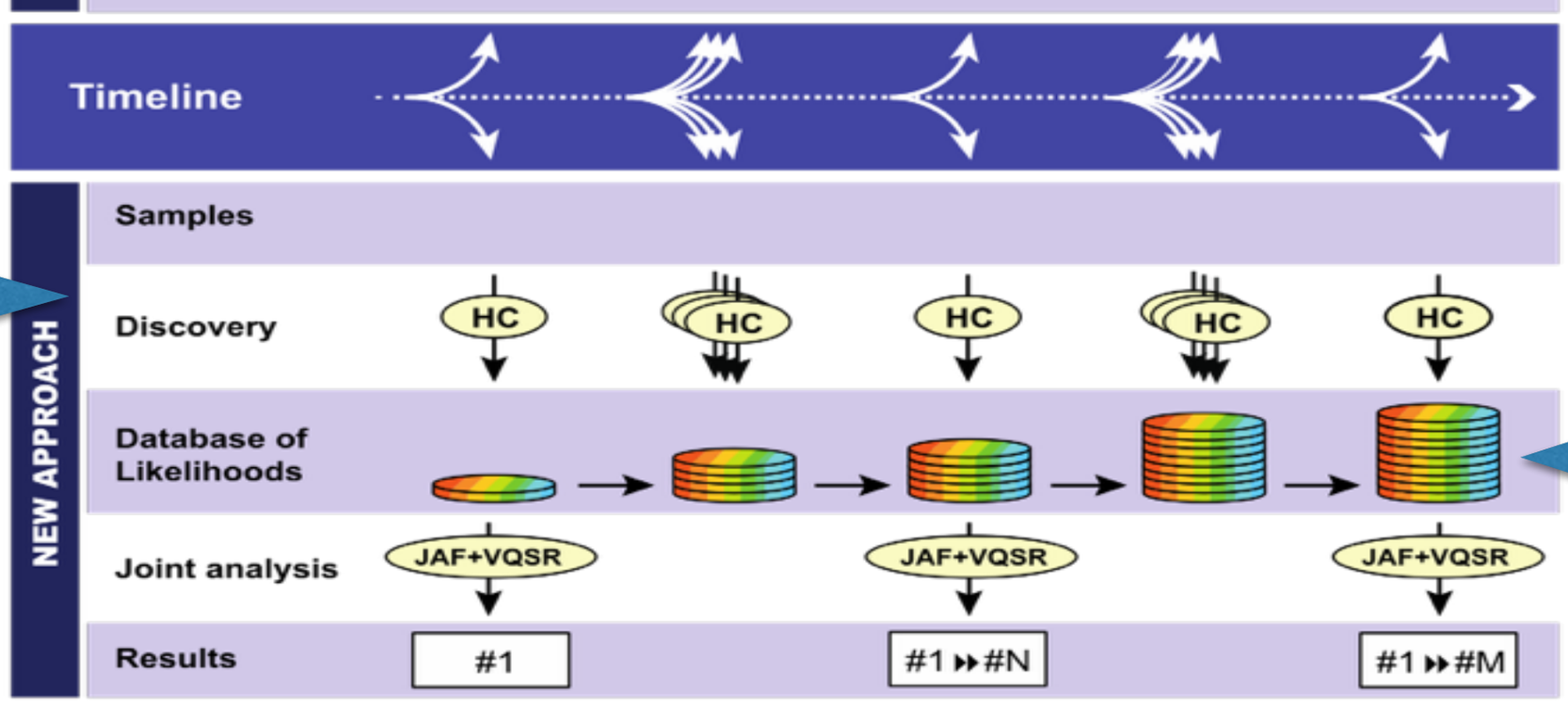
$\Pr\{D|H\}$ is the haploid likelihood function



DePristo et al. *Nature Genetics* (2011)

The *haplotype caller v3* enables incremental calling





by separating discovery from joint analysis, we can now jointly call any arbitrary number of samples

Variant Recalibration statistically separates good variants from bad ones

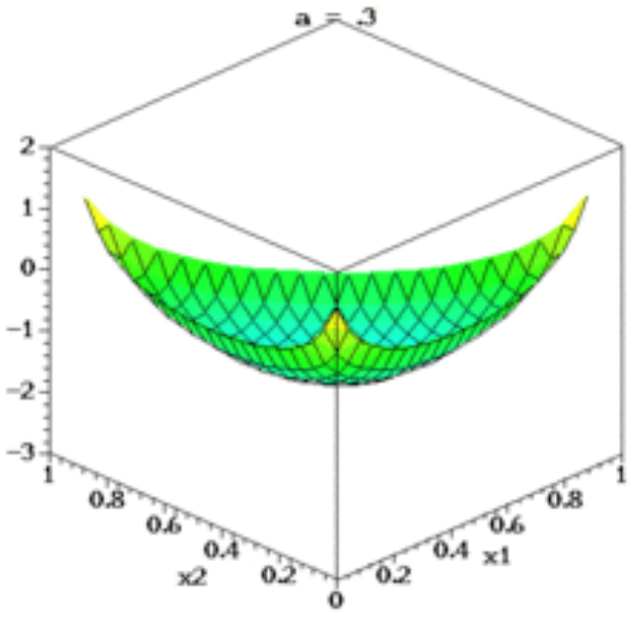
Gaussian Mixture Model trained on annotated variants, find MAP using VBEM:

$$p(\vec{c}) = \sum_{z} p(z)\, p(\vec{c} \mid z) = \sum_{k=1}^{K} \pi_k \, p(\pi_k) \, N(\vec{c} \mid \vec{\mu}_k, \Sigma_k) \, p(\vec{\mu}_k, \Sigma_k)$$

Dirichlet distribution

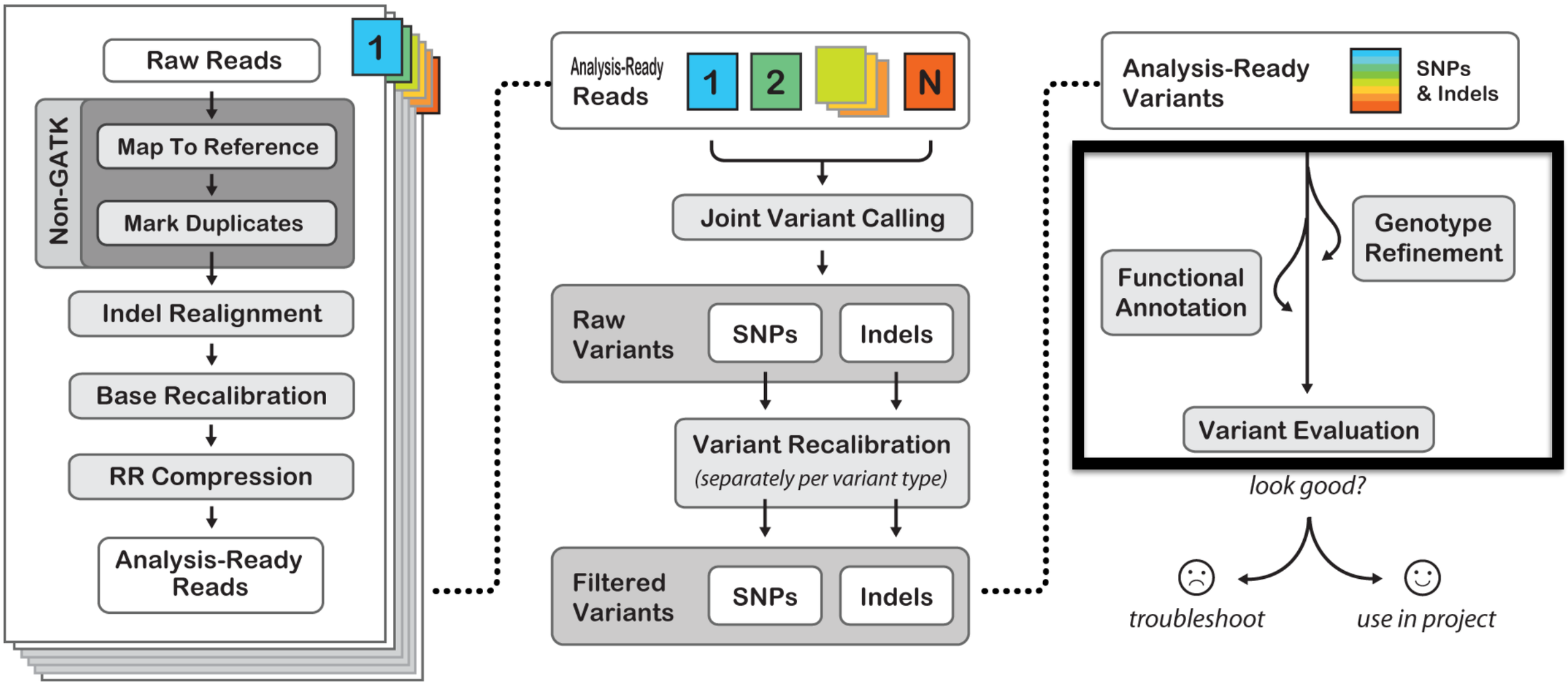
Normal – inverse Wishart distribution

Prior expectation is sparse set



Prior expectation is the empirical mean and empirical covariance of the data.
Bias away from singularities.

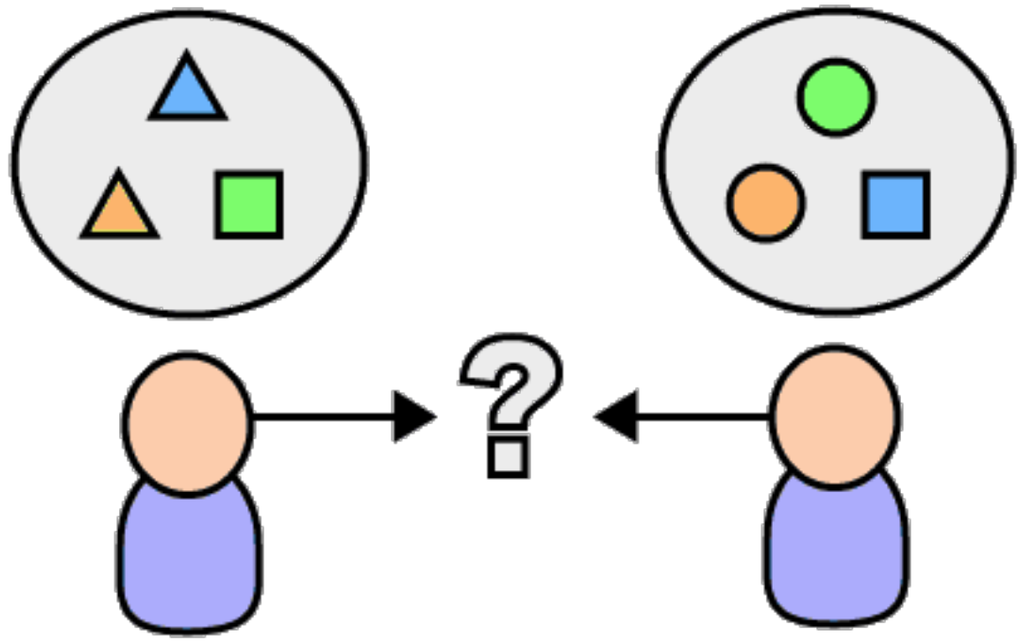
But the tools are missing from the preliminary analysis pane



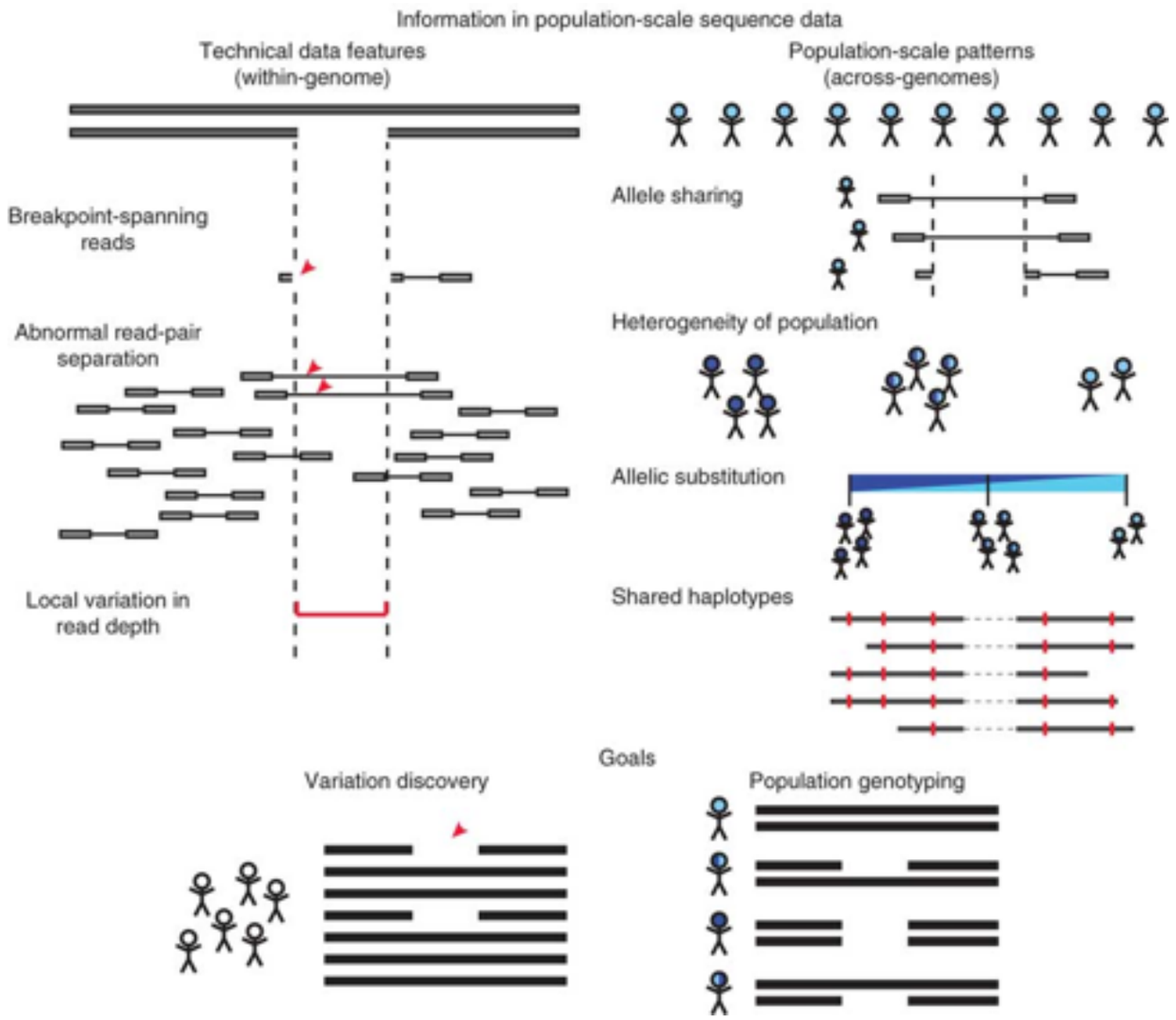
Auwera, GA et al. *Current Protocols in Bioinformatics* (2013)

Analysis pipeline standardization and scaling is the next big challenge

- What happens after variant calling is not standardized.
- Hundreds of completely unrelated tools are chained together with non-reusable scripts.
- Analyses are very often unrepeatable.
- Tools are not generalized and performance does not scale. (typically written in matlab, R, PERL and Python…)
- Most tools are written by one grad student/ postdoc and is no longer maintained.
- Complementary data types are not standardized (e.g. phenotypic data).

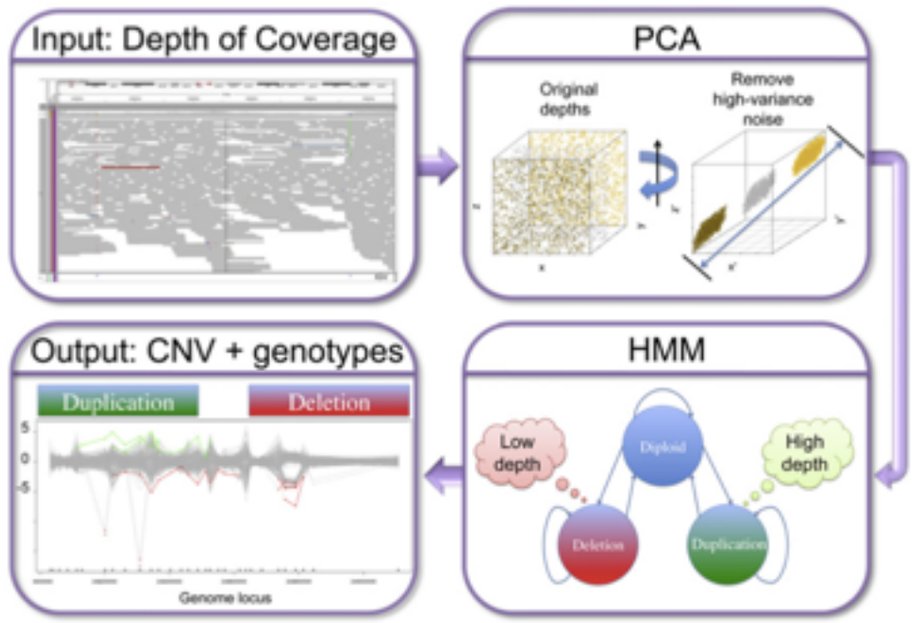


Structural variation is an important missing piece in the analysis pipeline



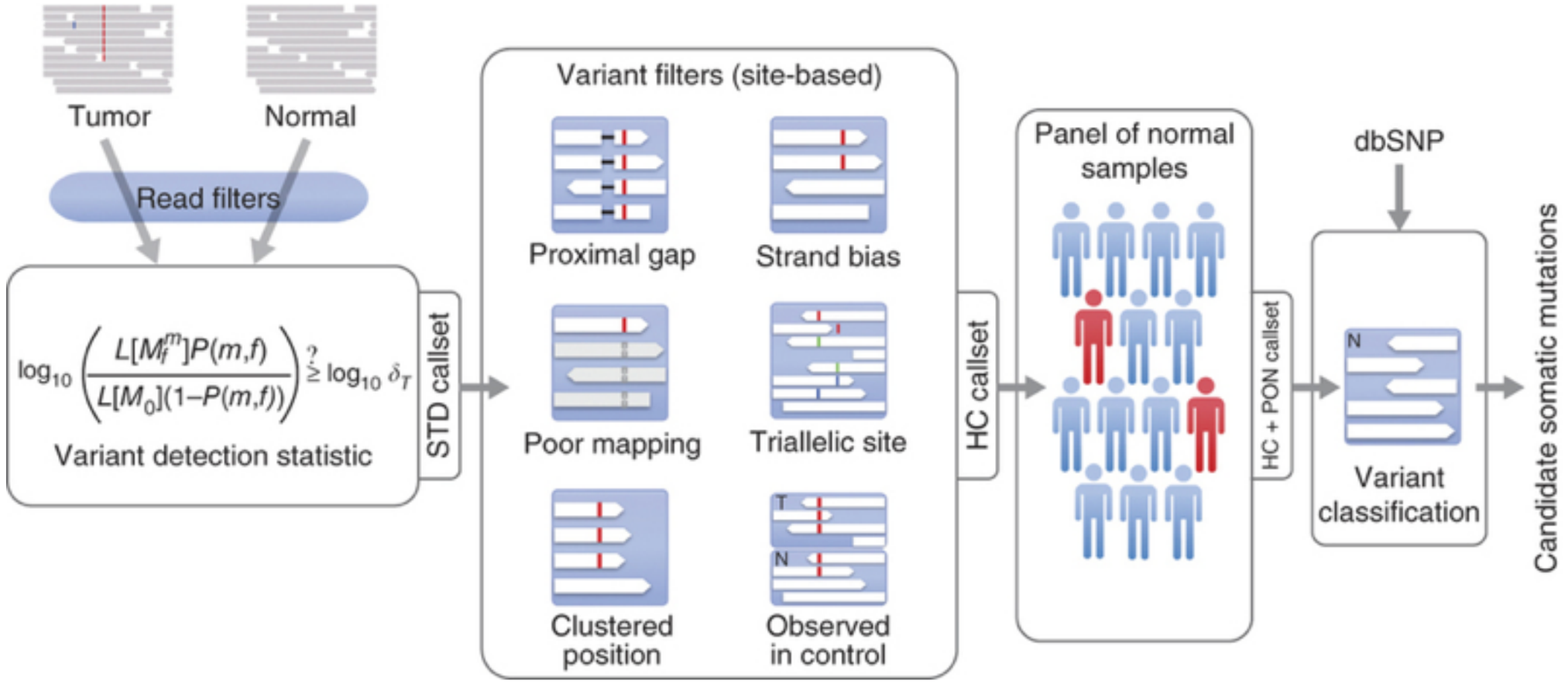
Handsaker, R et al. *Nature Genetics* (2011)

current implementations have confirmed the importance of structural variation calling for complex disease research but have not been *standardized* or *productionized.*



Fromer, M et al. *Nature* (2014)

Cancer tools also need the same rigorous standardization and scalability



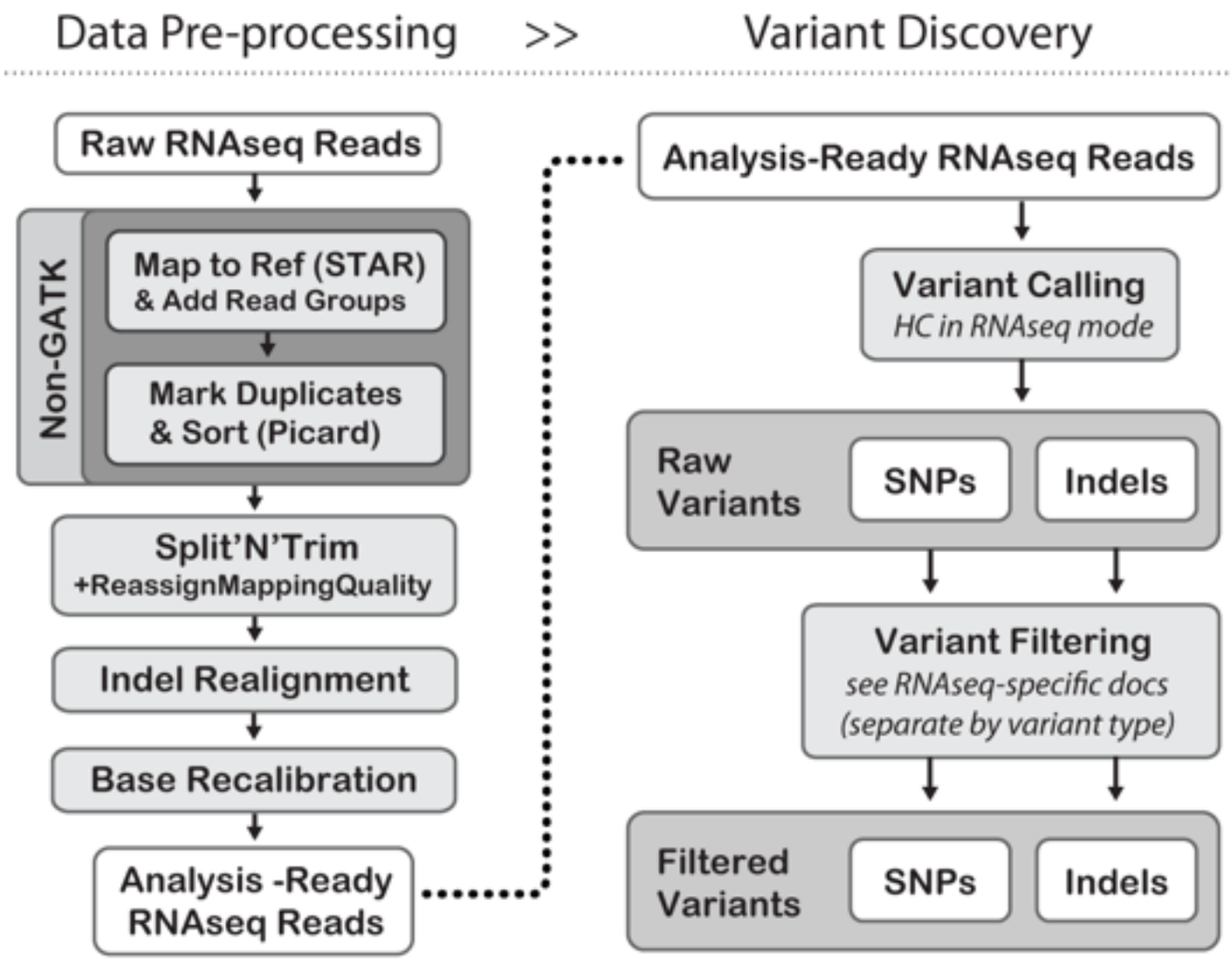
Cibulskis, K et al. *Nature Biotechnology* (2013)

DNA does not tell the whole story — there is RNA too!

We have standardized the RNA-seq pipeline, but now there is a lot of work to do!

Milestone 1:
update GATK tools to make best use of RNA data including contrastive variant calling with DNA. Improve accuracy and overall performance.

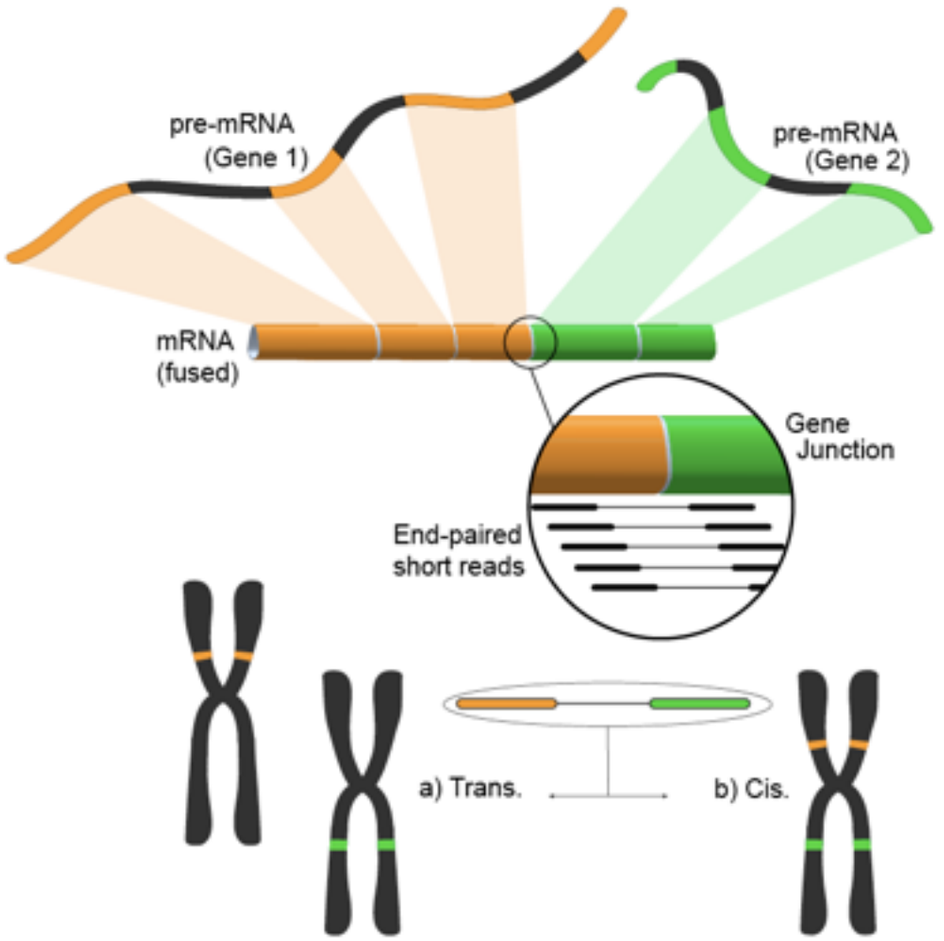
Milestone 2:
Build new tools to address the needs



Carneiro, M et al. *In preparation*

Personalized medicine depends immensely on disease research

- Samples must be consistently pre-processed worldwide and the processing pipelines need to scale in performance.
- Variants must be jointly called and currently available tools need to provide the necessary performance. (We solved the scaling problem!)
- Post variant calling analysis pipelines need to be rebuilt from scratch with performance and scalability in mind.
- We need to build new infrastructure to enable the aggregation of the massive wave of data that is coming our way
- RNA-seq and structural variation need to be integrated and standardized for scientists and clinicians to understand the whole picture.
- We need to start giving the same focus to functional analysis and therapeutics for all the associations identified.



This is the work of many…

the team

Eric Banks
Ryan Poplin
Khalid Shakir
David Roazen
Joel Thibault
Geraldine VanDerAuwera
Ami Levy-Moonshine
Valentin Rubio
Bertrand Haas
Laura Gauthier
Christopher Wheelan
Sheila Chandran

Broad colleagues

Heng Li
Daniel MacArthur
Timothy Fennel
Steven McCarrol
Mark Daly
Sheila Fisher
Stacey Gabriel
David Altshuler

collaborators

Menachem Fromer
Paolo Narvaez
Diego Nehab