# Touching the limits of large scale disease research

**Mauricio Carneiro, PhD**
Group Lead, Computational Technology Development
Broad Institute

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

Rare Variant Association Study (RVAS)

*vs*

Common Variant Association Study (CVAS)

*vs*

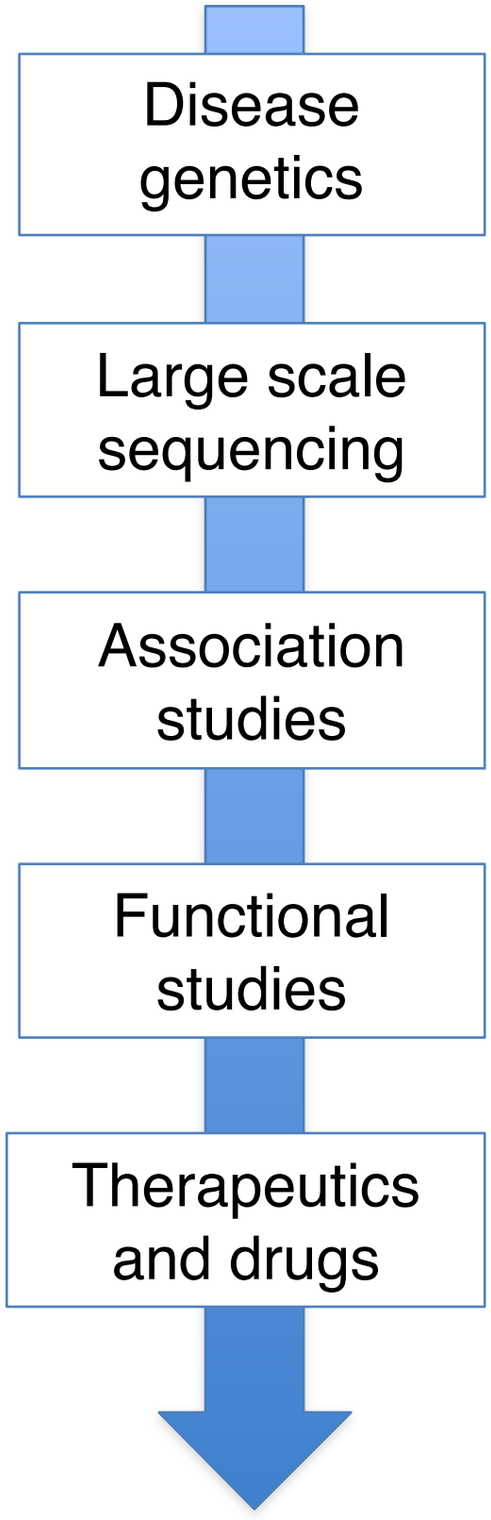
# What is the BROAD INSTITUTE ?

***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.

# Improving human health in 5 easy steps




Many simple and complex human diseases are heritable. Pick one.

Affected and unaffected individuals differ systematically in their genetic composition

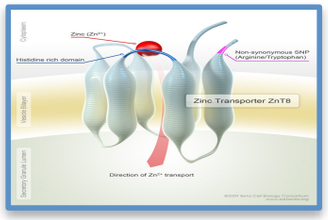
These systematic differences can be identified by comparing affected and unaffected individuals

These associated variants give insight into the biological mechanisms of disease

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

# The Importance of Scale…Early Success Stories (at 1,000s of exomes)

## Type 2 Diabetes

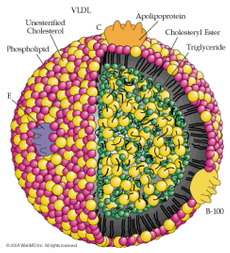



- 13,000 exomes
- SLC30A8
  (Beta-cell-specific $Zn^{++}$ transporter)
- 3-fold protection against T2D!
- ***1 LoF per 1500 people***

## Schizophrenia

- 5,000 exomes
- Pathways
  - Activity-regulated cytoskeletal (ARC) of post-synaptic density complex (PSD)
  - Voltage-gated $Ca^{++}$ Channel
- 13-21% risk in carriers
- ***Collection of rare disruptive mutations (~1/10,000 carrier frequency)***

## Coronary Heart Disease




- 3,700 exomes
- APOC3
- 2.5-fold protection from CHD
- ***4 rare disruptive mutations (~1 in 200 carrier frequency)***

## Early Heart Attack

- 5,000 exomes
- APOA5
- 22% risk in carriers
- ***0.5% Rare disruptive / deleterious alleles***

# 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

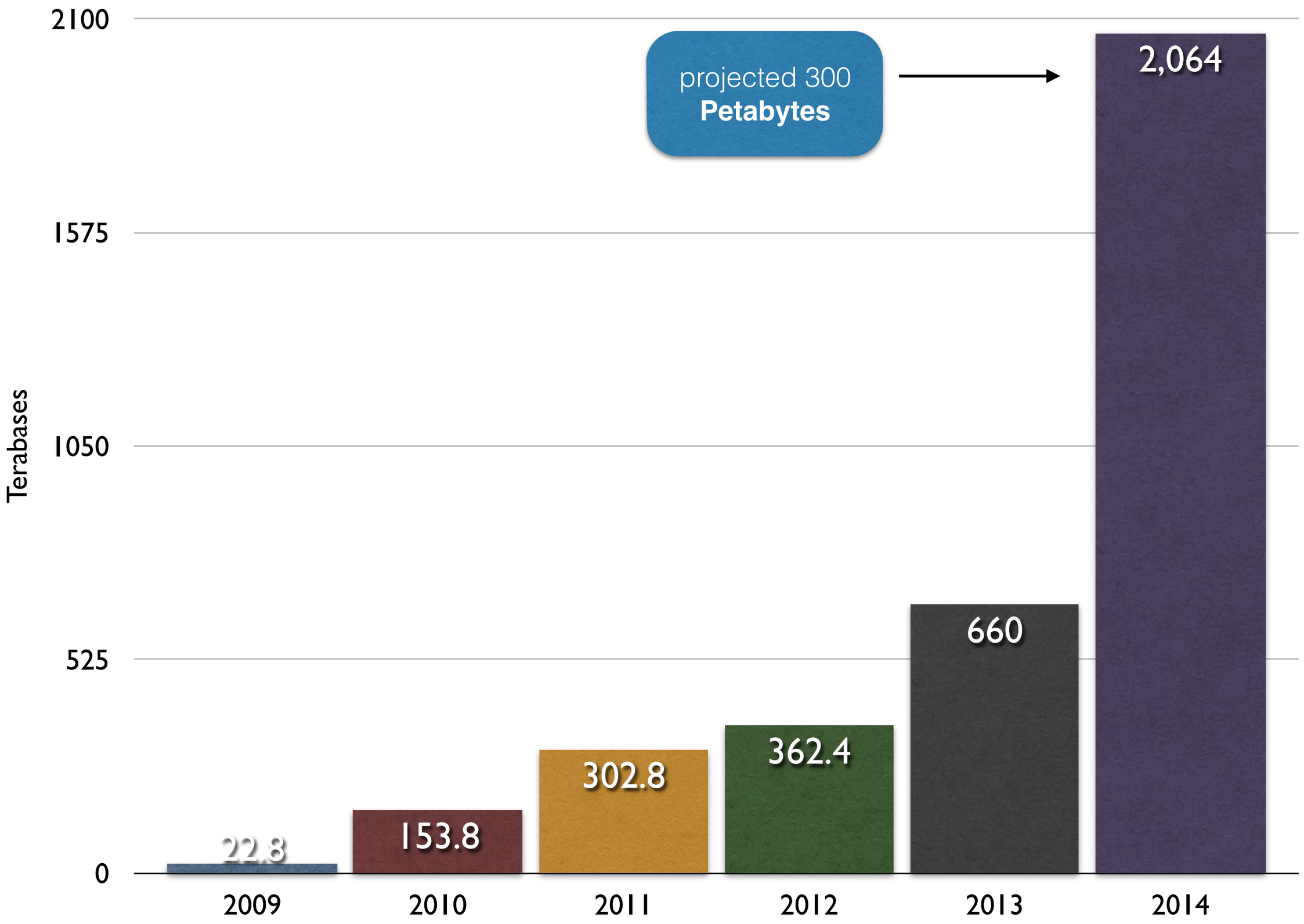
**8,189**
RNA

**9,788**
16S

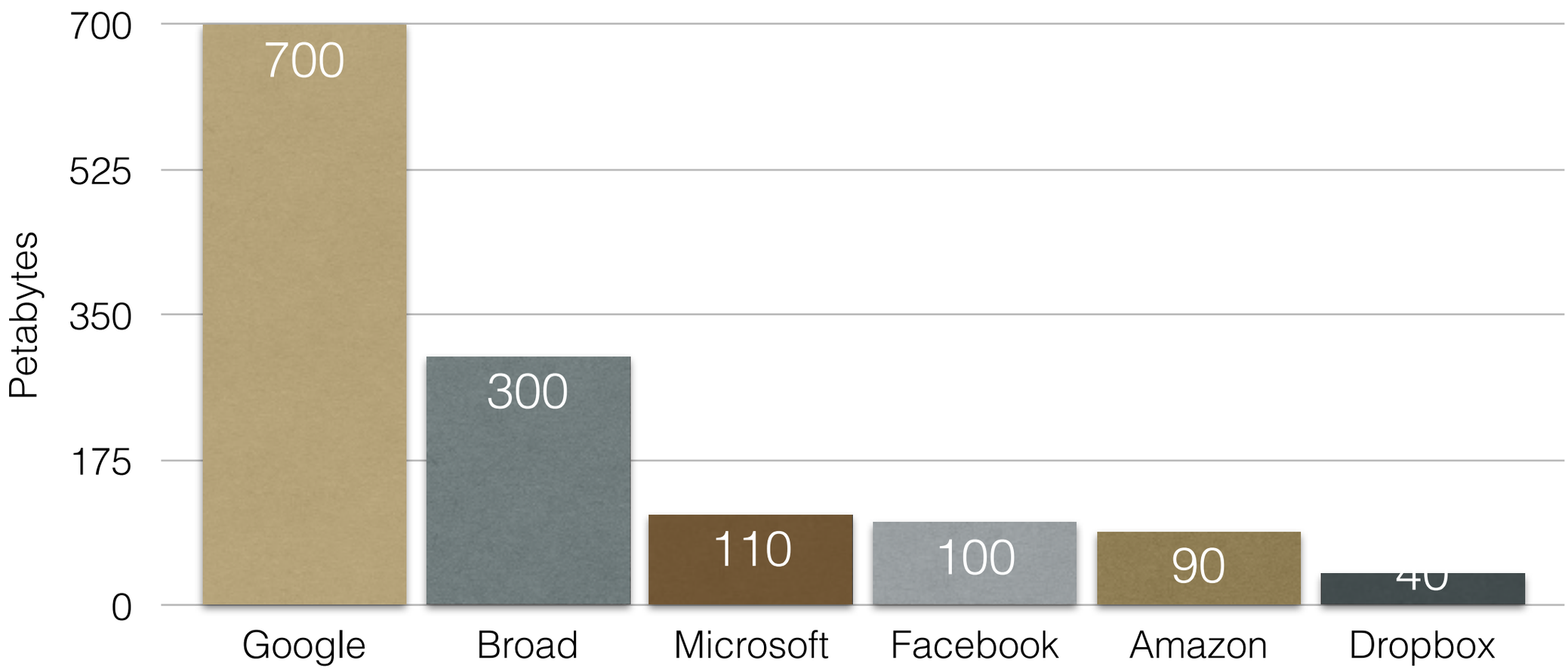
**47,764**
arrays

**228**
cell lines

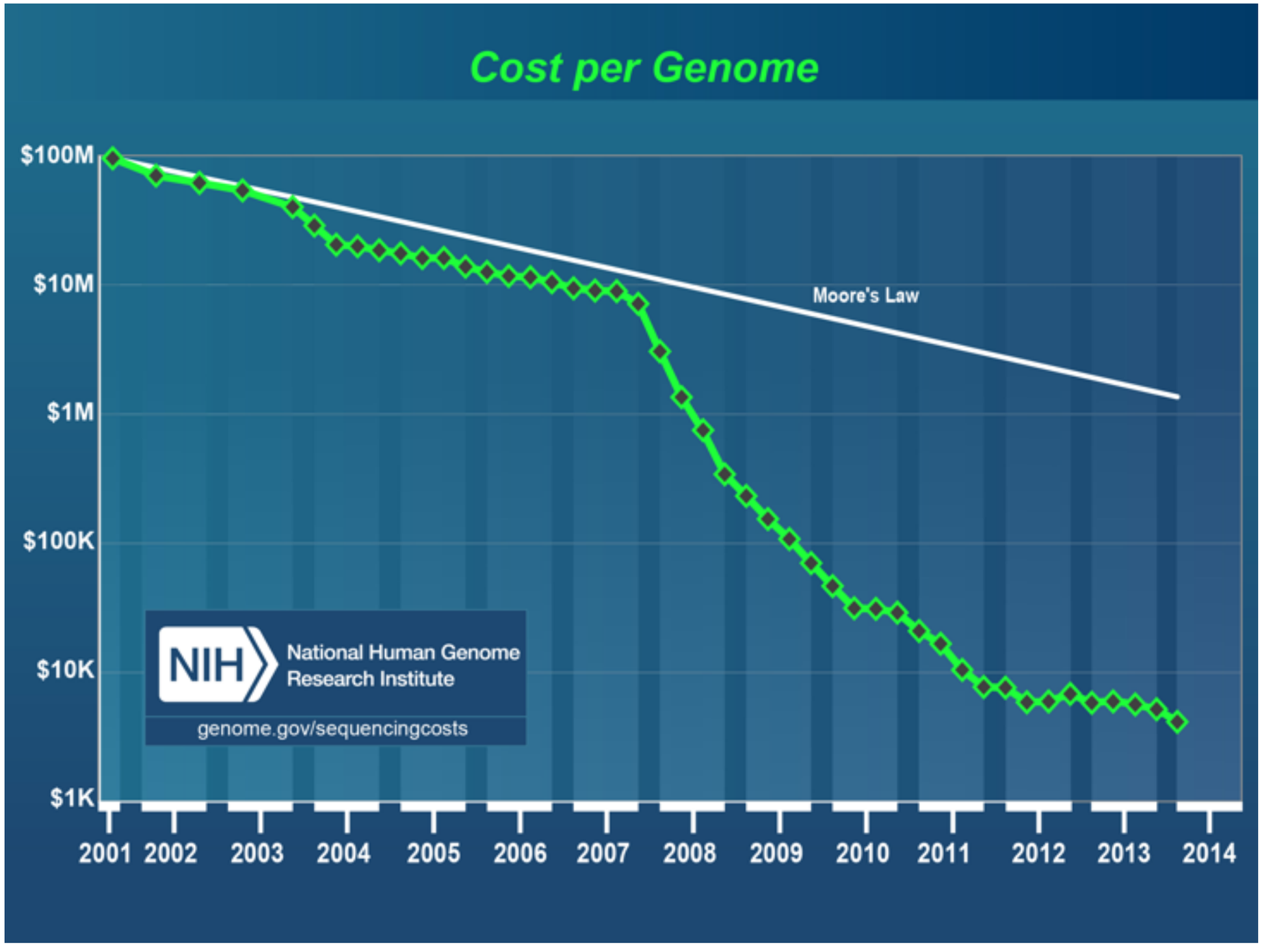
# Terabases of Data Produced by Year



# We produce as much data as the big cloud providers

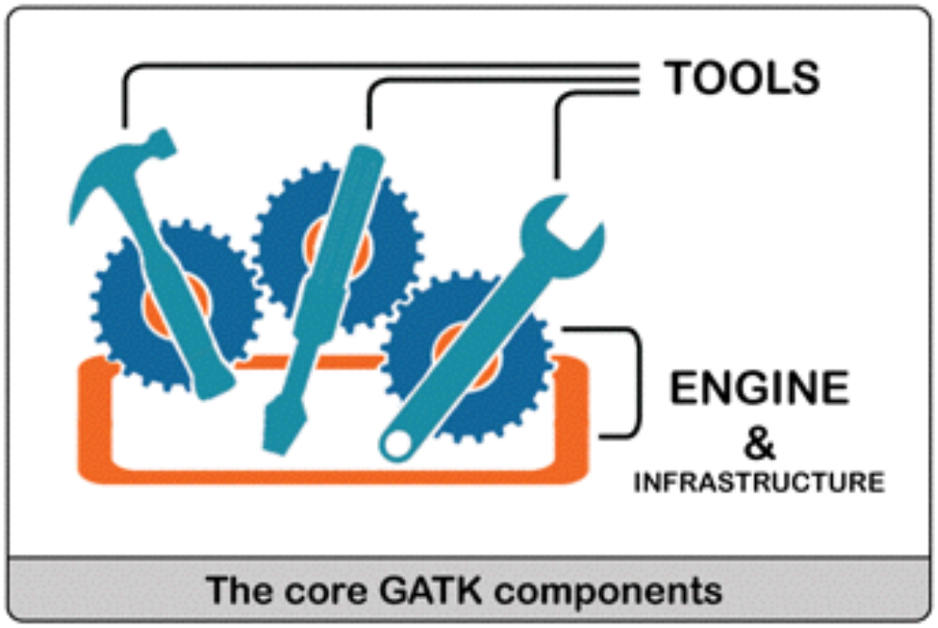


# and these numbers will continue to grow faster than Moore's law



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

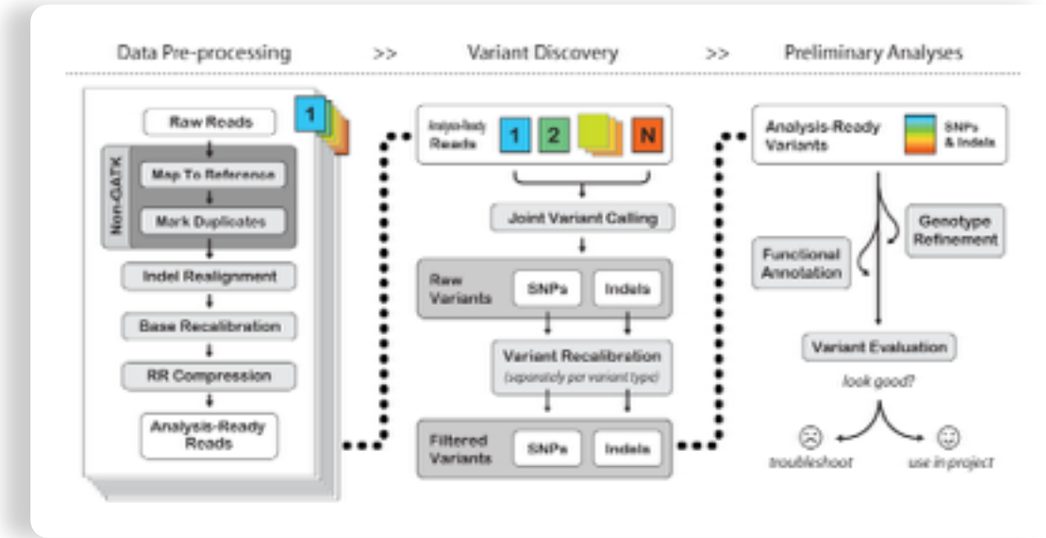
## Toolkit & framework packages



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

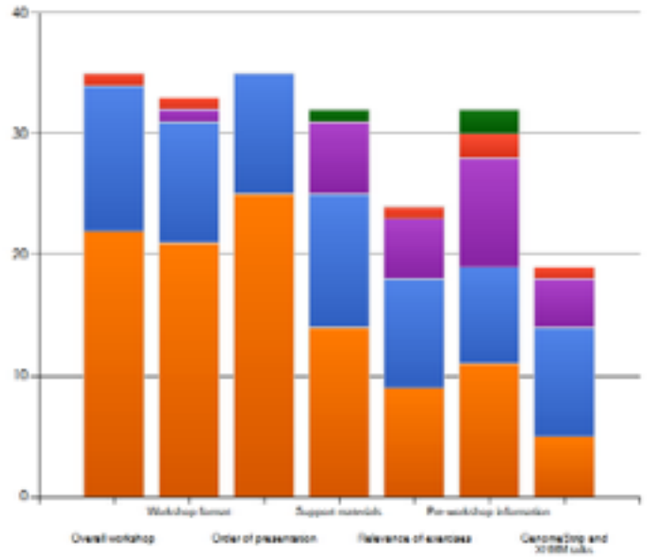
# Workshop series educates local and worldwide audiences

**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**



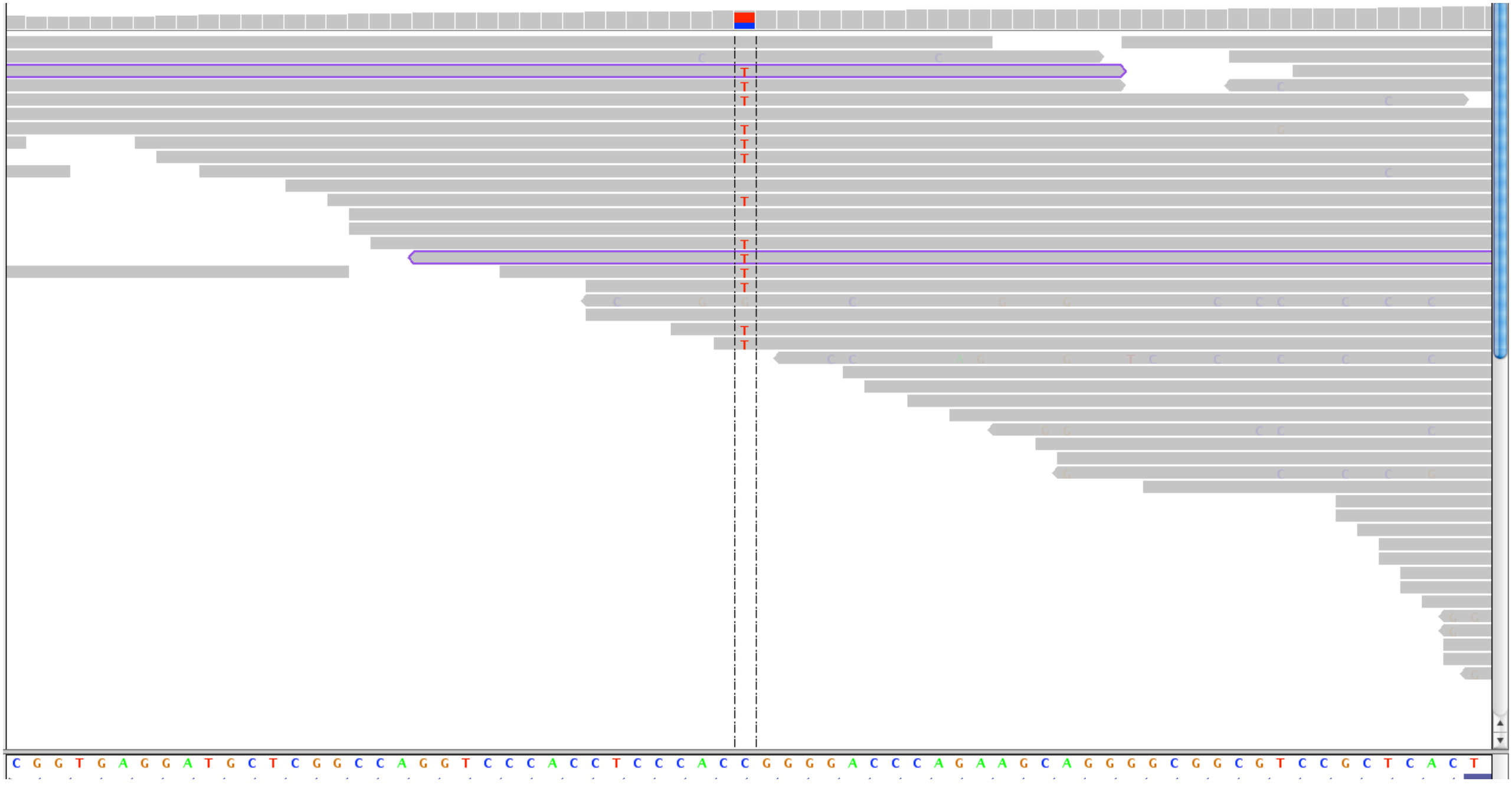
**BroadE: Overview of GATK & best practices**

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

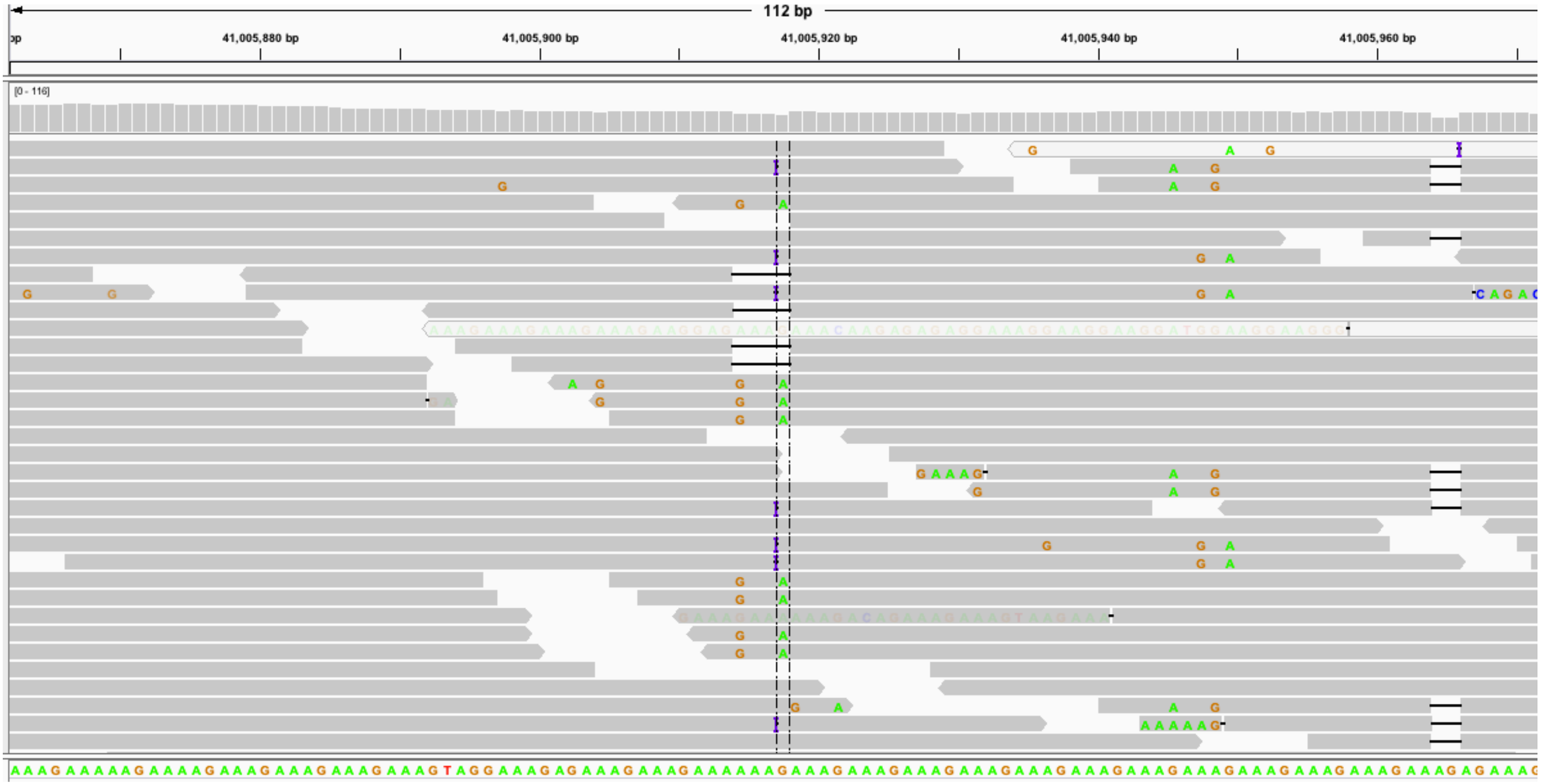
Copyright Broad Institute, 2013. All rights reserved. The presentations below were filmed during the 2013 **GATK** Workshop, part of ...

NEW HD

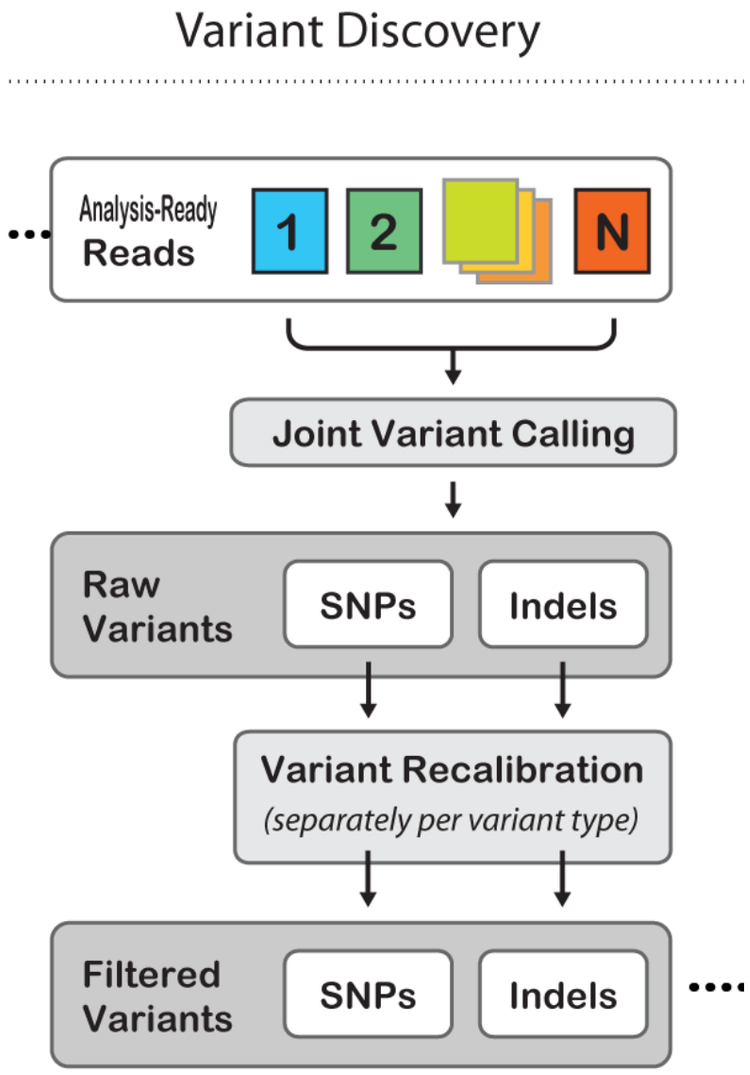
# Identifying mutations in a genome is a simple “find the differences” problem



# Unfortunately, real data does not look that simple



# Variant calling is a large-scale bayesian modeling problem



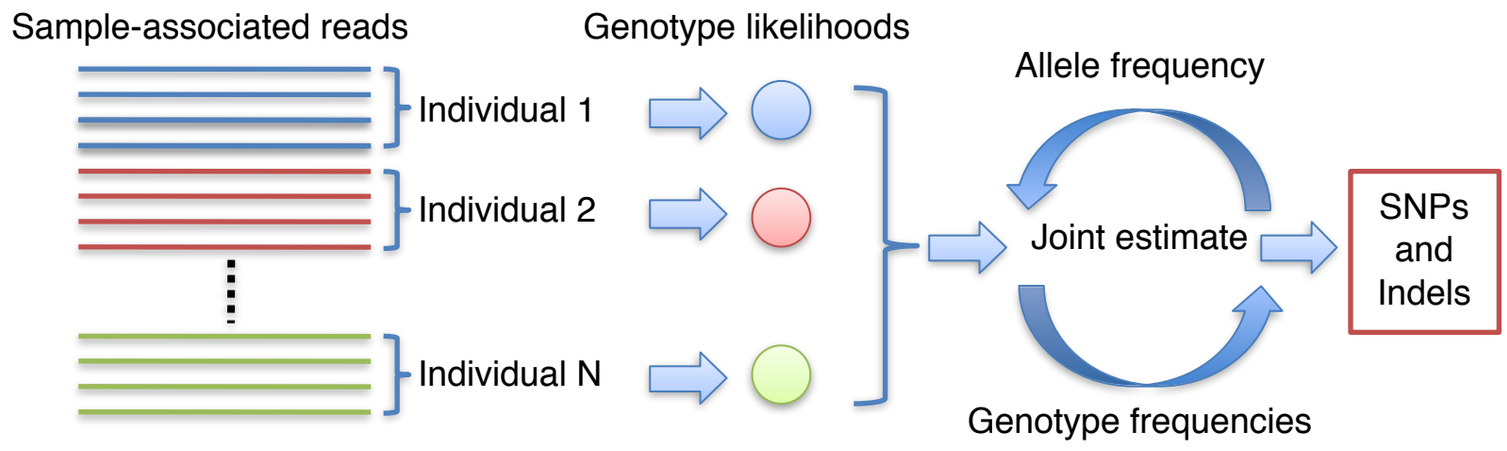

prior Likelihood

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

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 ideal database for RVAS and CVAS studies would be a complete 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

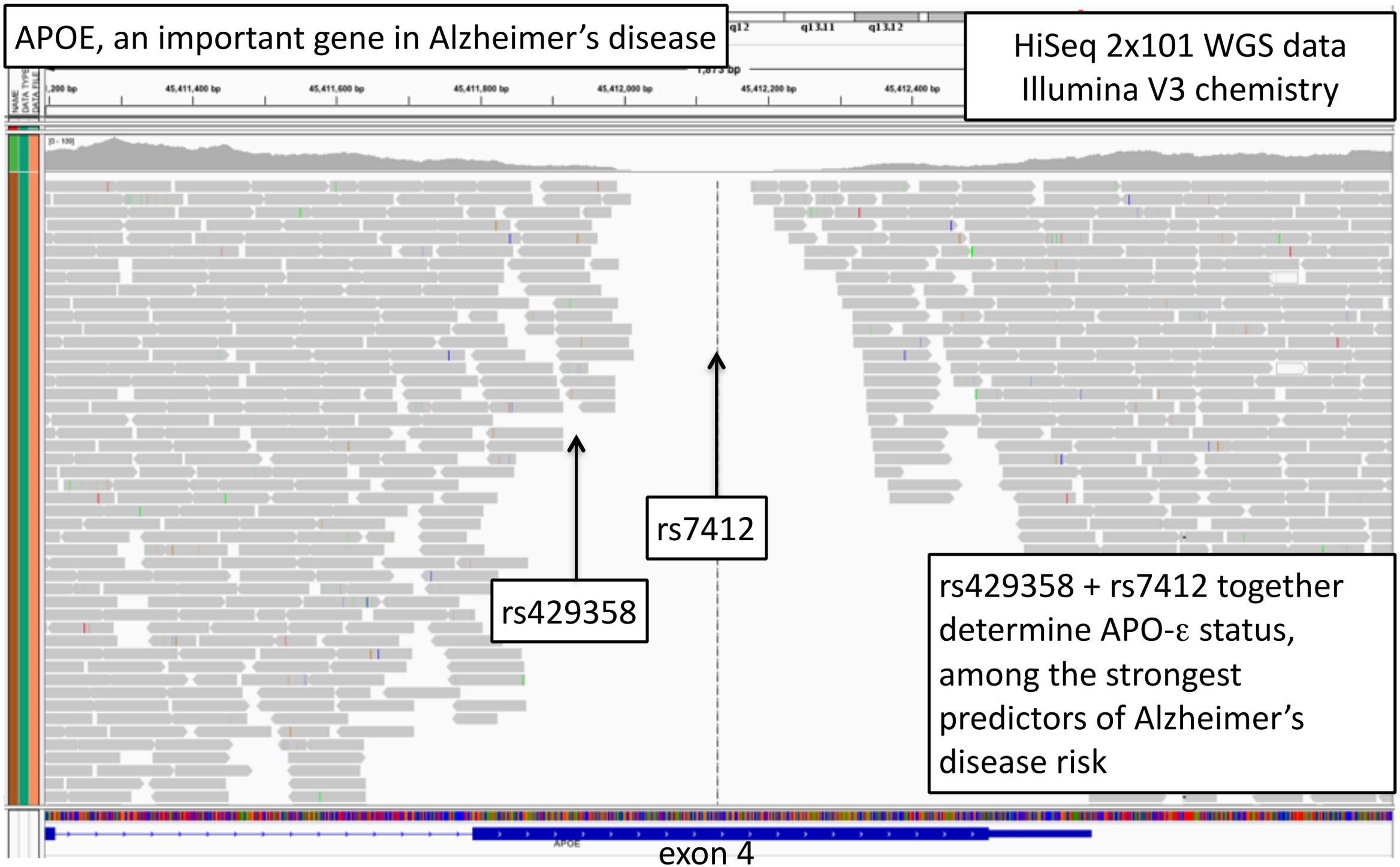
# What is limiting our ability simply use NGS to find disease-causing variants?

- **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

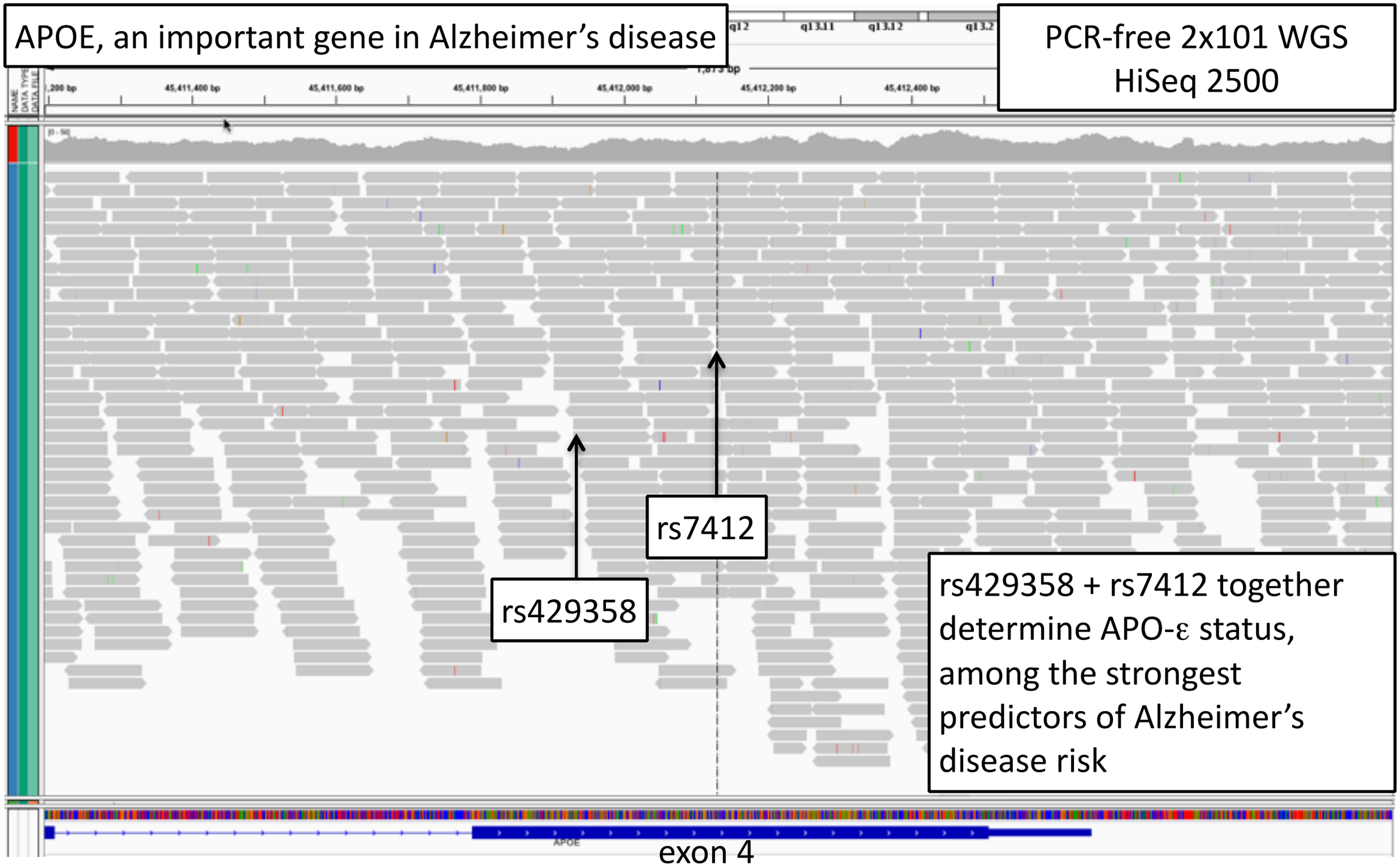
- **Analytic limitations**
  - Misinterpret our data right before our eyes, calling the wrong variant near the right place (esp. indels)

- **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

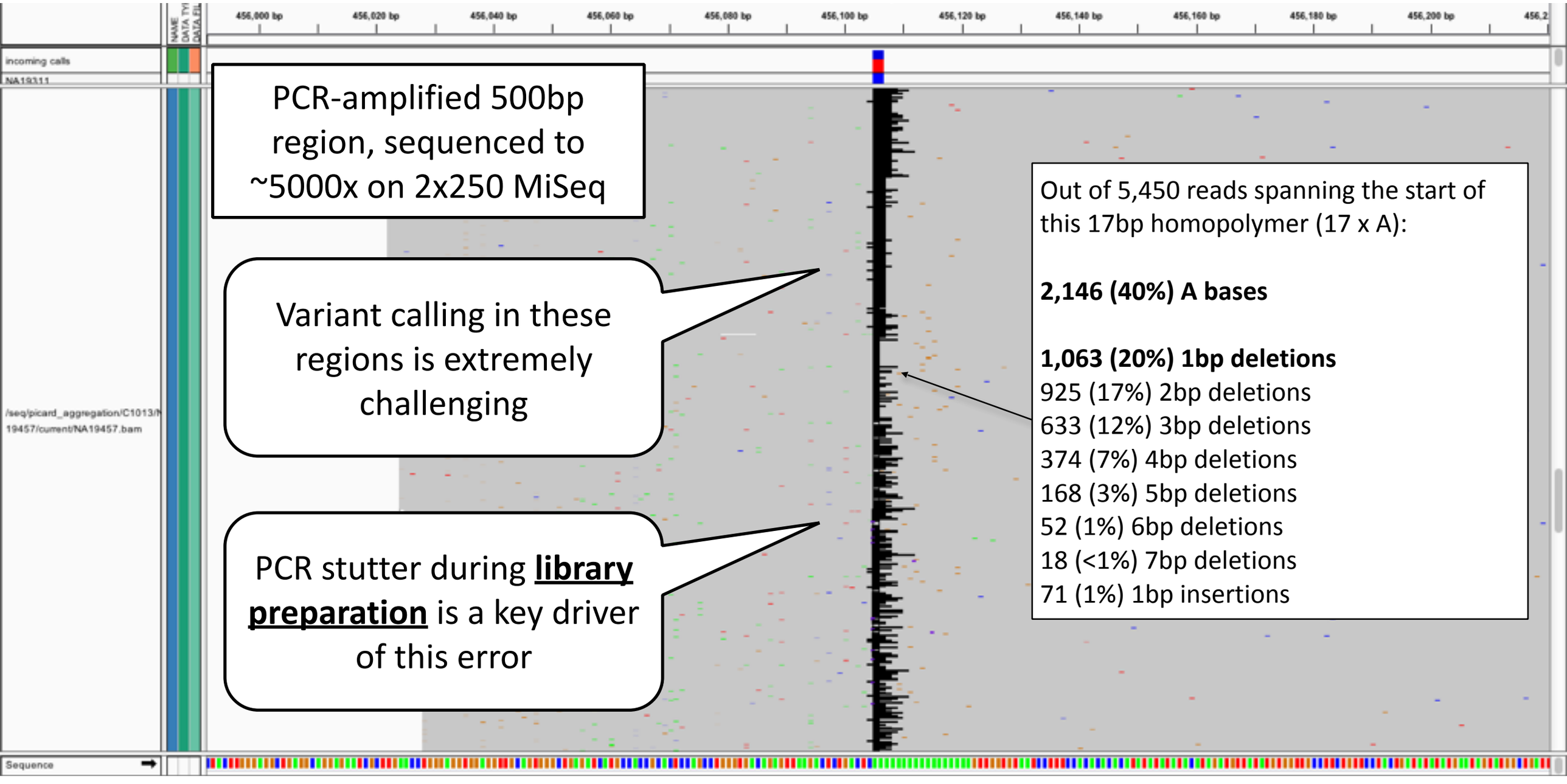


# PCR-free libraries improve sequencing experiments



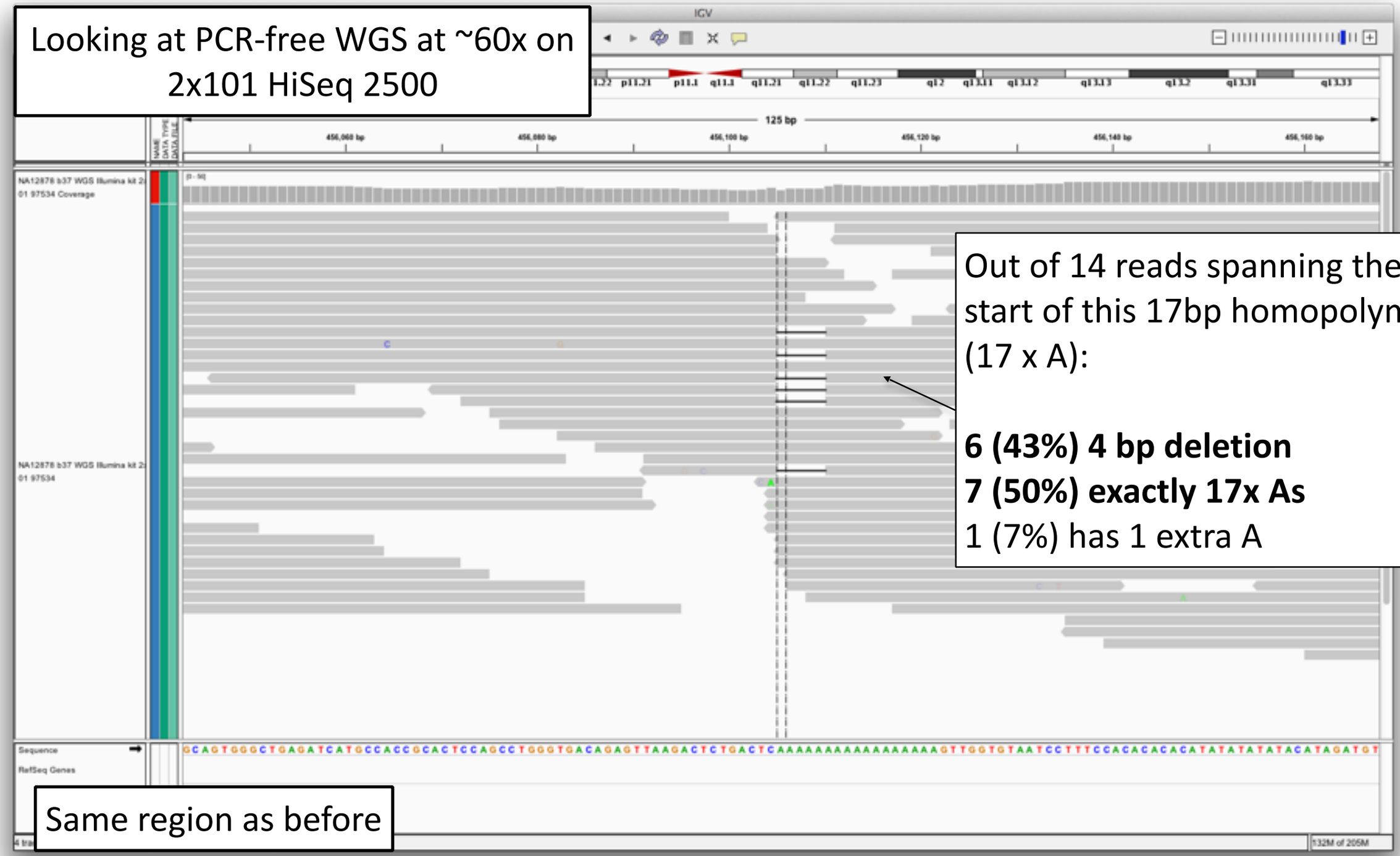
# complex error process caused by PCR inundates us with false positives

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

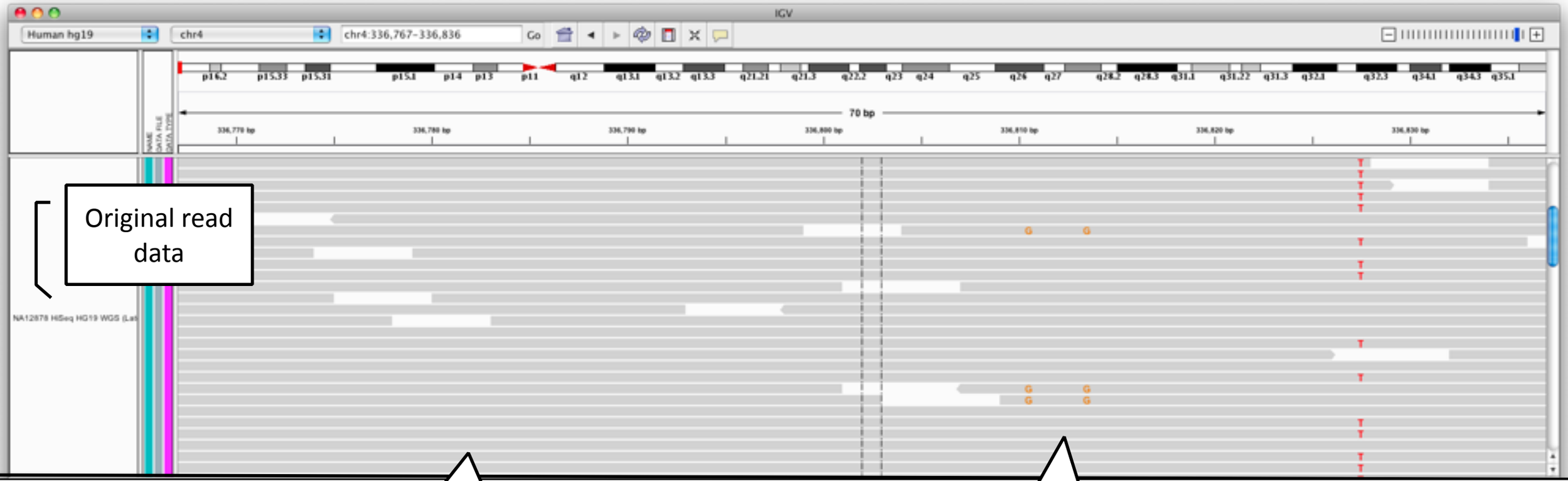



Out of 14 reads spanning the start of this 17bp homopolymer (17 x A):

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

Same region as before

# Regions of the genome with no apparent variation (or plain confusing)



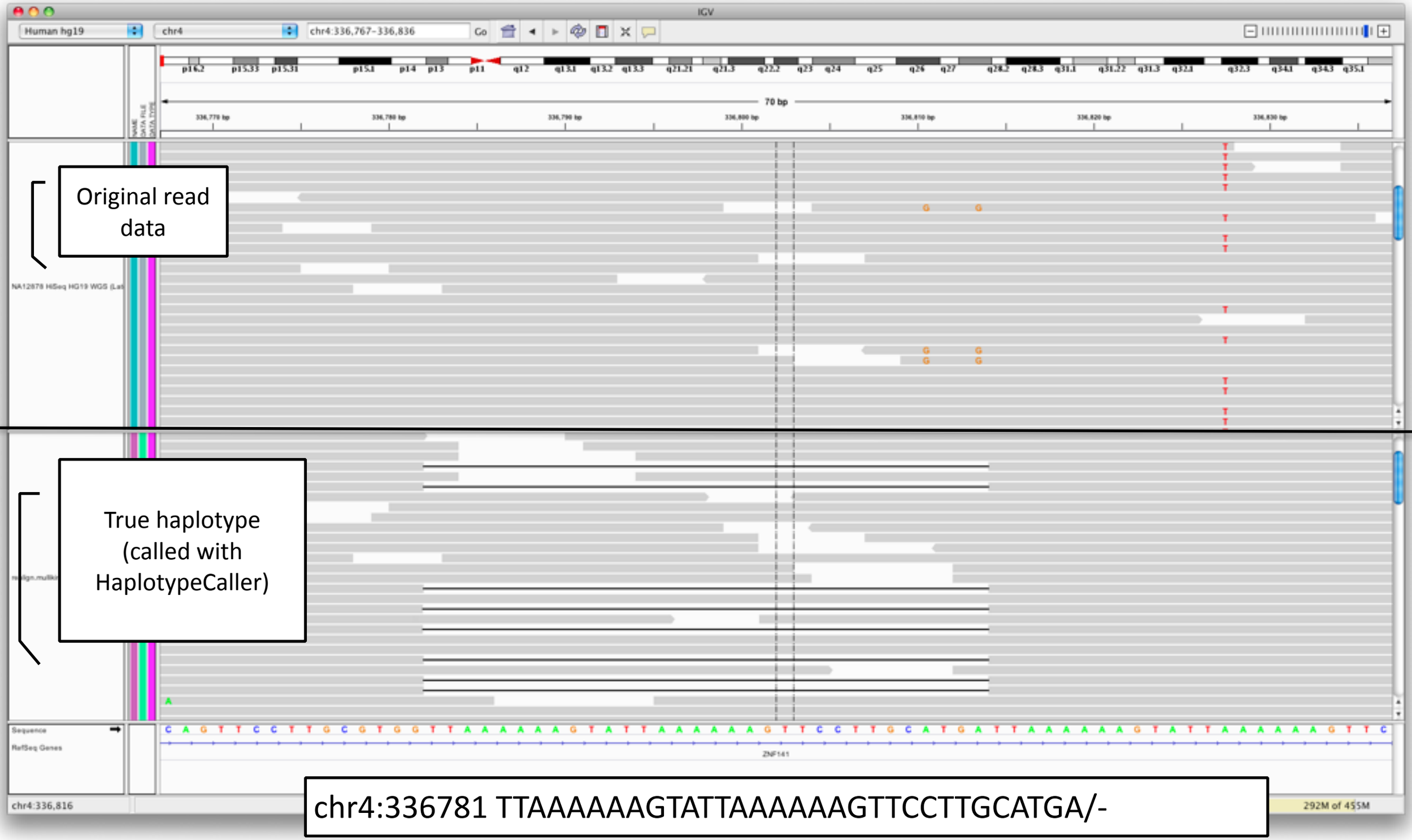

Looks confidently reference.

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

# Is correctly detangled by the GATK's Haplotype Caller using *local denovo* assembly




chr4:336781 TTAAAAAAGTATTAAAAAAGTTCCTTGCATGA/-

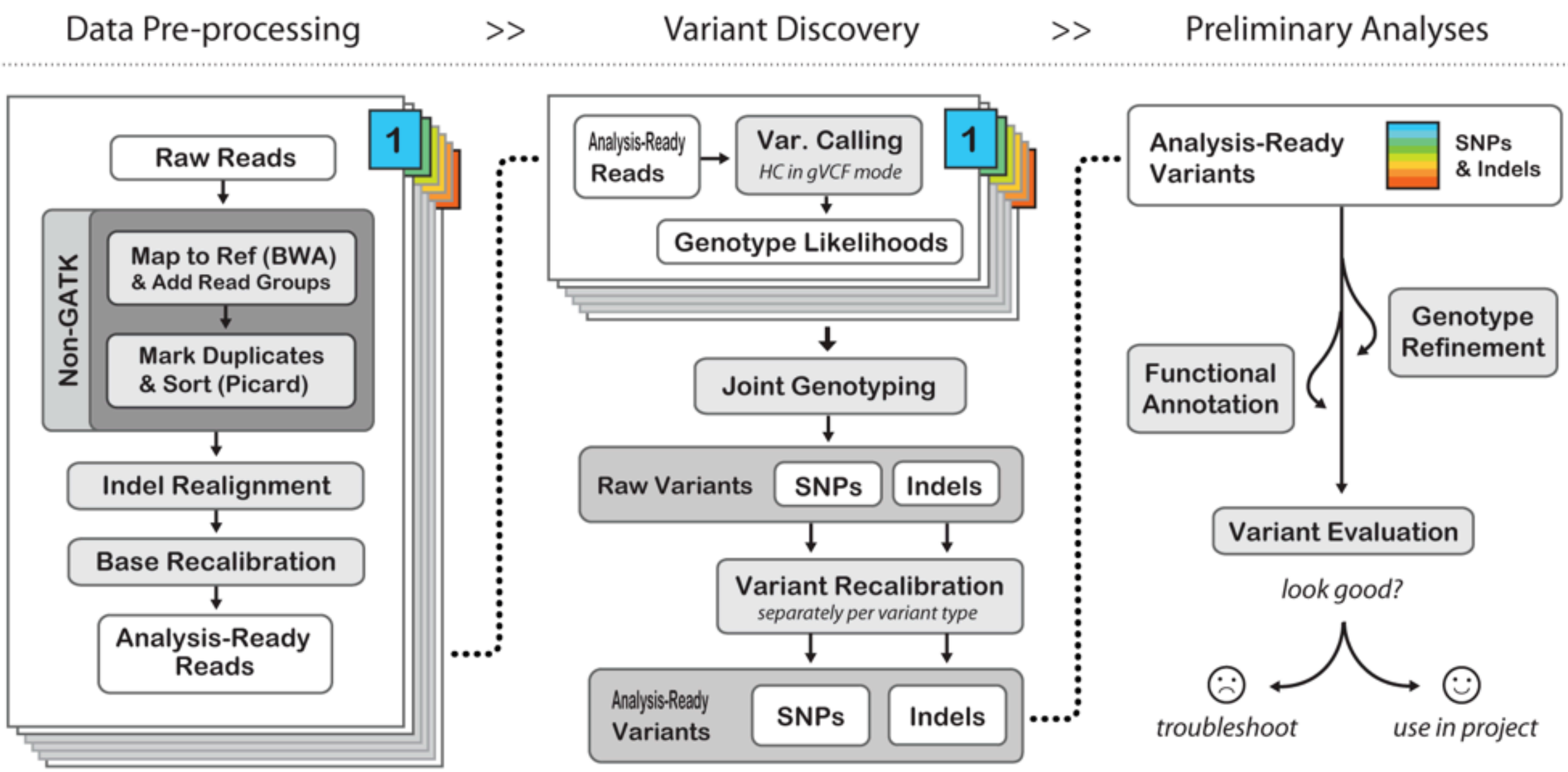
# …but today's biggest limitation is **sample size**

- 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)

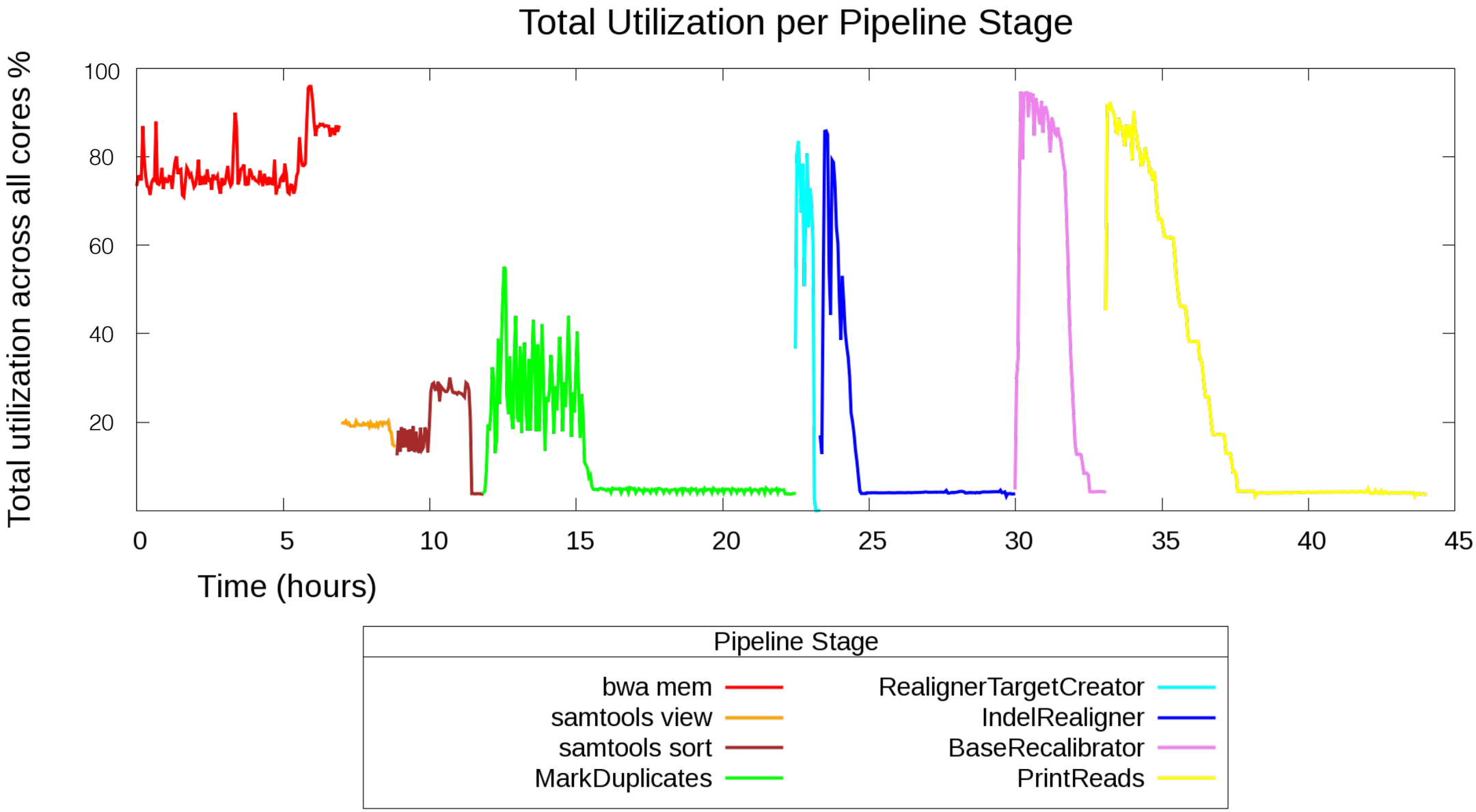
| Comparison | Just my 500 samples | Analyze with 500 samples | Analyze with 1K samples | Analyze with 10K samples | Analyze with 100K samples |
|---|---|---|---|---|---|
| ***Affected*** | ***10/1000*** | ***10/1000*** | ***10/1000*** | ***10/1000*** | ***10/1000*** |
| Unaffected | None | 0/1000 |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

# We have defined the best practices for sequencing data processing




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

# Processing is a big cost on whole genome sequencing



# Challenges to scale up the processing pipeline

- Eliminate disk read/writing in between pipeline steps
- Reduce time spent doing unnecessary calculations
  (e.g. Base Recalibration on *good* data)
- High performance native I/O libraries
  (*Gamgee*: https://github.com/broadinstitute/gamgee)
- Redesign algorithms with performance in mind

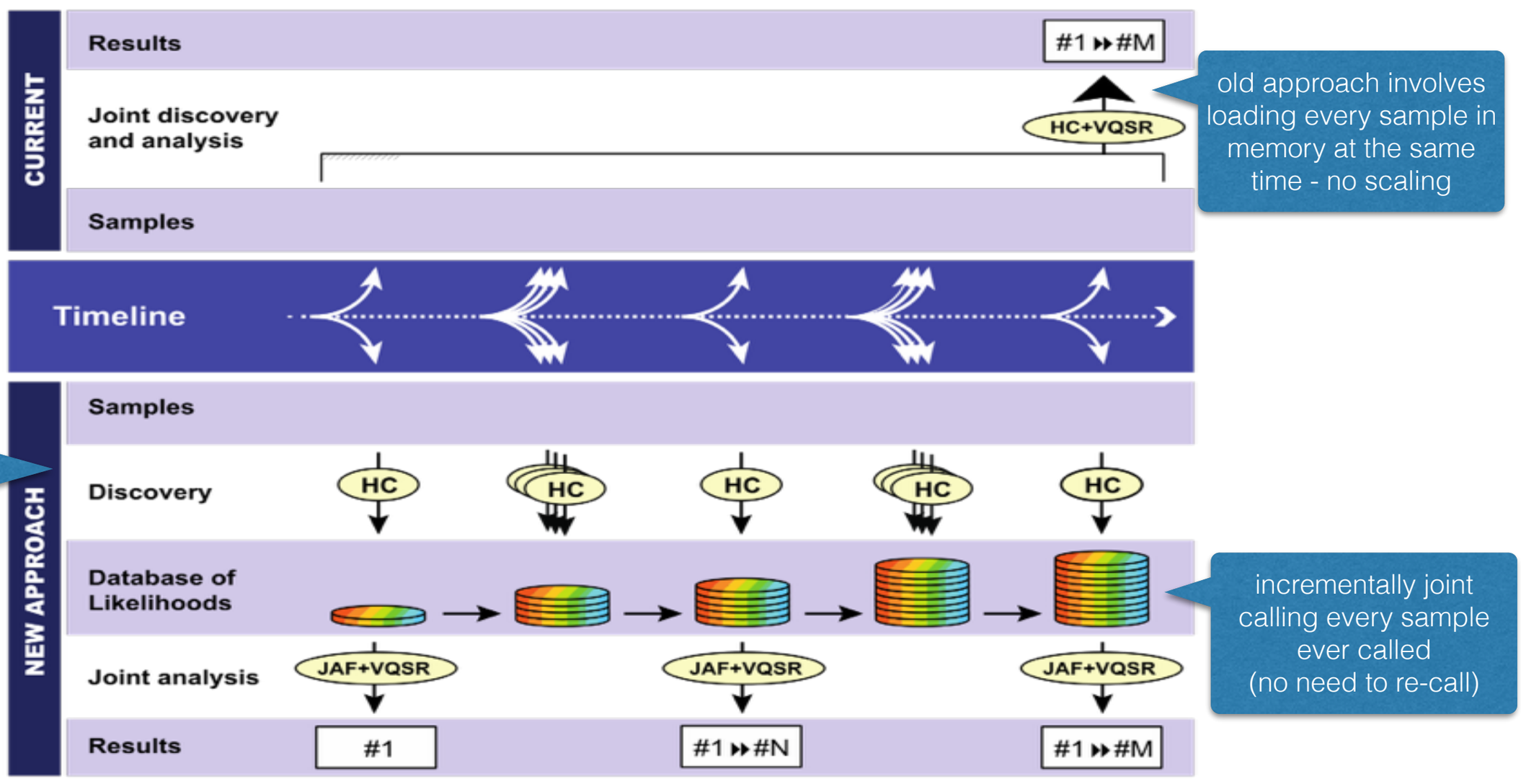
| step | threads | time |
|---|---|---|
| BWA | 24 | 7 |
| samtools view | 1 | 2 |
| sort + index | 1 | 3 |
| MarkDuplicates | 1 | 11 |
| RealignTargets | 24 | 1 |
| IndelRealigner | 24 | 6.5 |
| BaseRecalibrator | 24 | 1.3 |
| PrintReads + index | 24 | 12.3 |
| **Total** | | **44** |

# Heterogeneous compute speeds up variant calling significantly

| Technology | Hardware | Runtime | Improvement |
|---|---|---|---|
| - | Java (gatk 2.8) | 10,800 | - |
| - | C++ (baseline) | 1,267 | 9x |
| FPGA | Convey Computers HC2 | 834 | 13x |
| AVX | Intel Xeon 1-core | 309 | 35x |
| GPU | NVidia GeForce GTX 670 | 288 | 38x |
| GPU | NVidia GeForce GTX 680 | 274 | 40x |
| GPU | NVidia GeForce GTX 480 | 190 | 56x |
| GPU | NVidia GeForce GTX Titan | 80 | 135x |
| GPU | NVidia Tesla K40 | 70 | 154x |
| AVX | Intel Xeon 24-core | 15 | 720x |

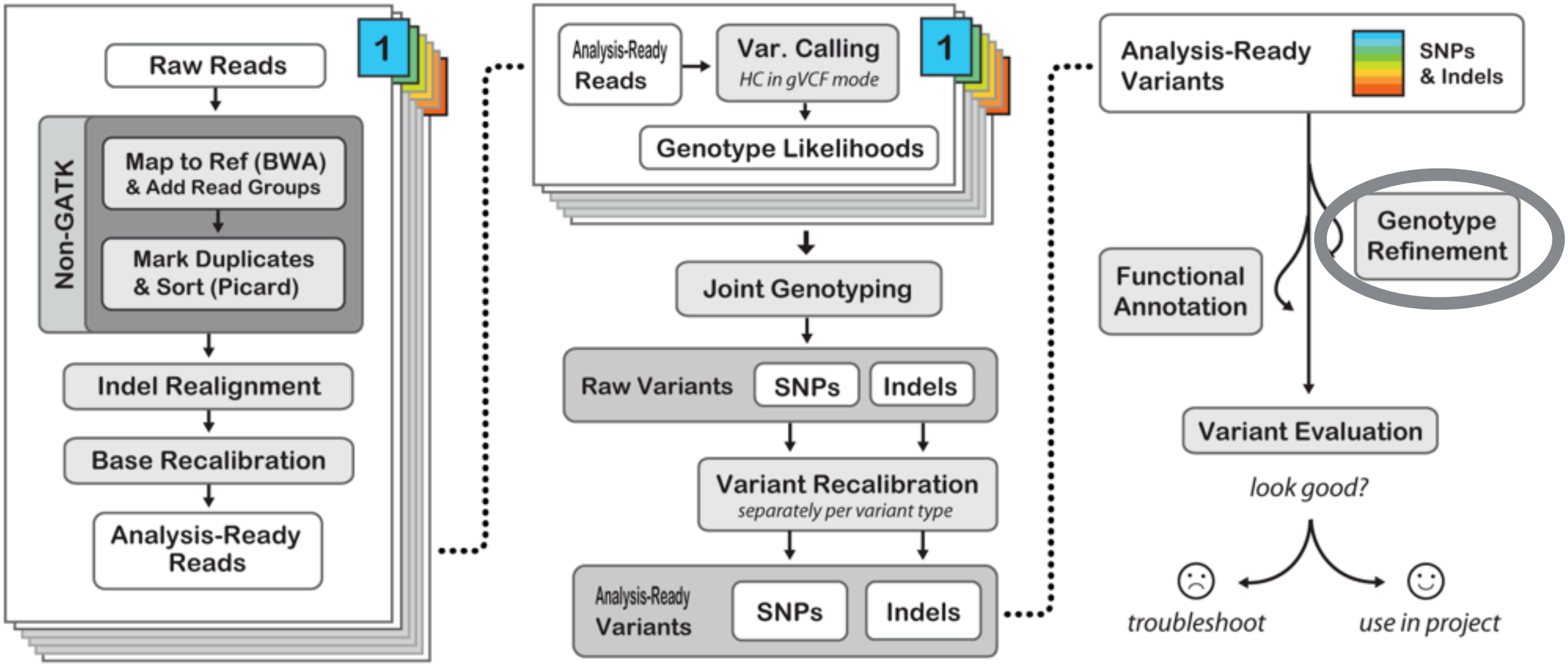
all times in seconds

# The reference model enables incremental calling



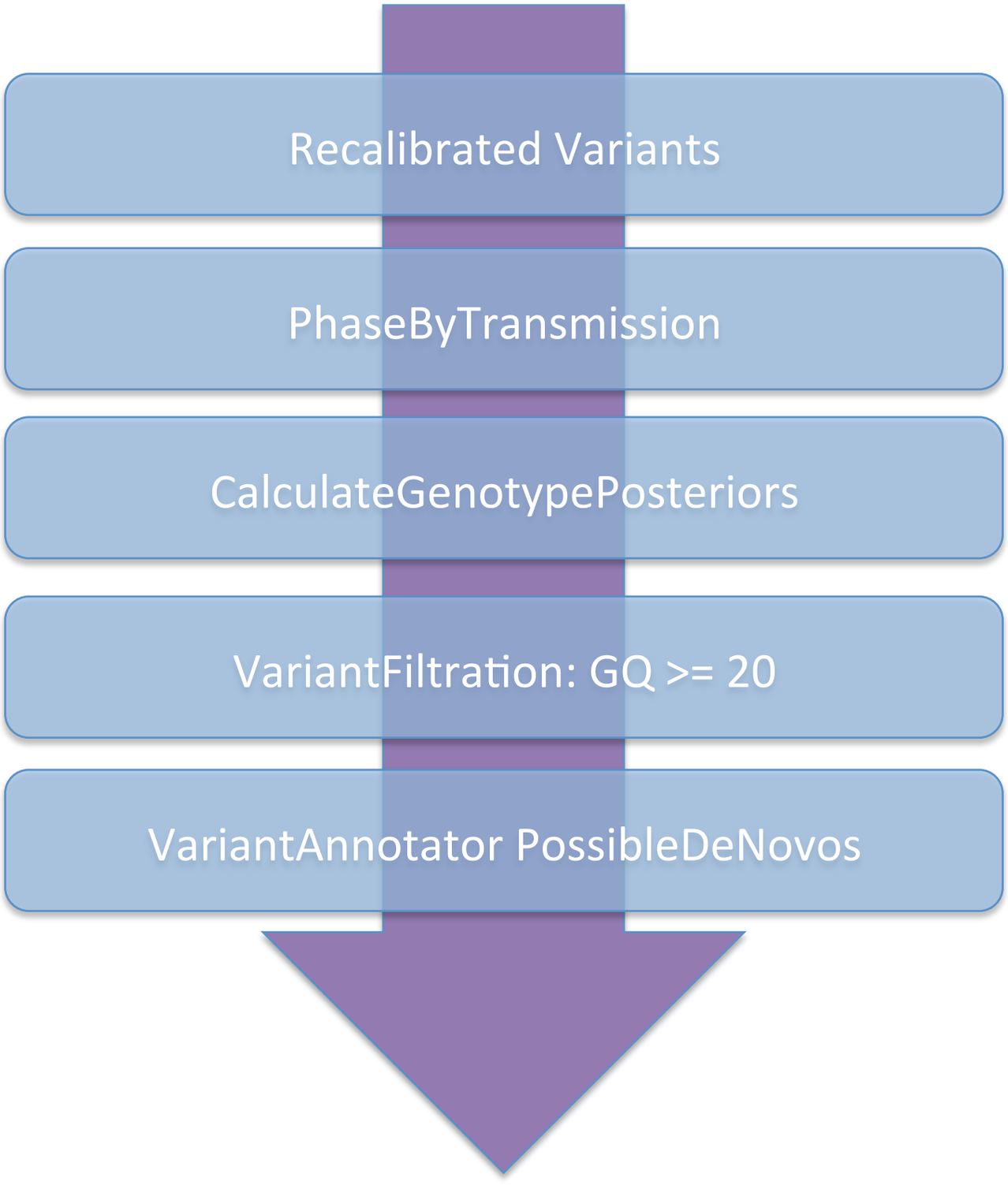

**by separating discovery from joint analysis, we can now jointly call any arbitrary number of samples**

# Genotyping is only the first step to establish accurate likelihoods




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

# New Genotype Refinement Pipeline




- Use VQSR-passing sites
- Likelihood adjustments based on pedigree
- Likelihood adjustments based on population
- Use confident genotypes
- Tag significantly likely de novo mutations in offspring

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

Rare Variant Association Study (RVAS)

*vs*

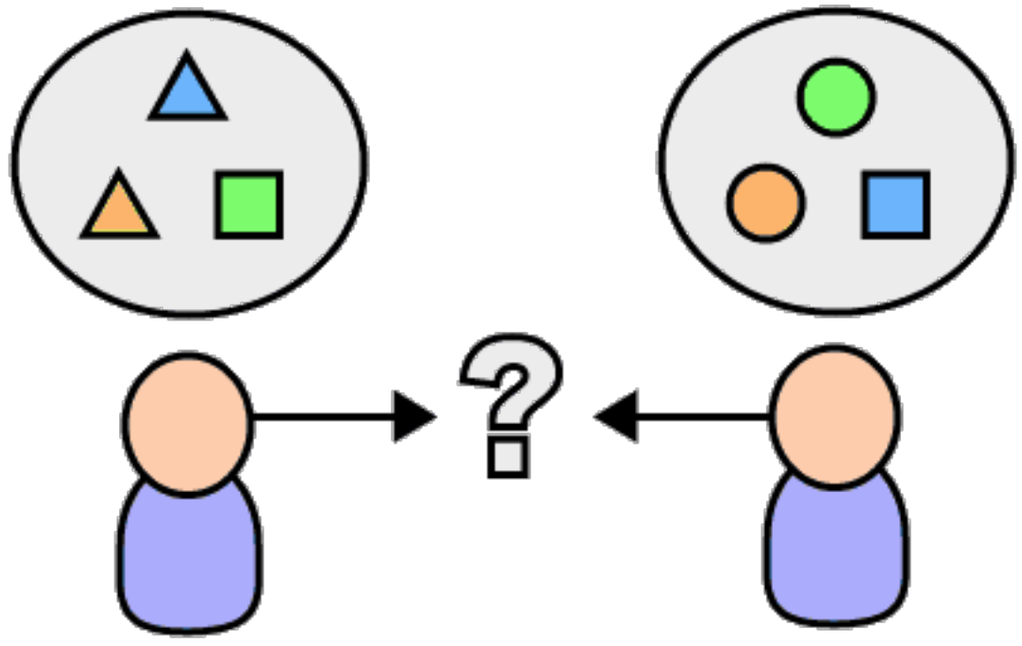
Common Variant Association Study (CVAS)

*vs*

**Associations, RNA, SV, cancer**

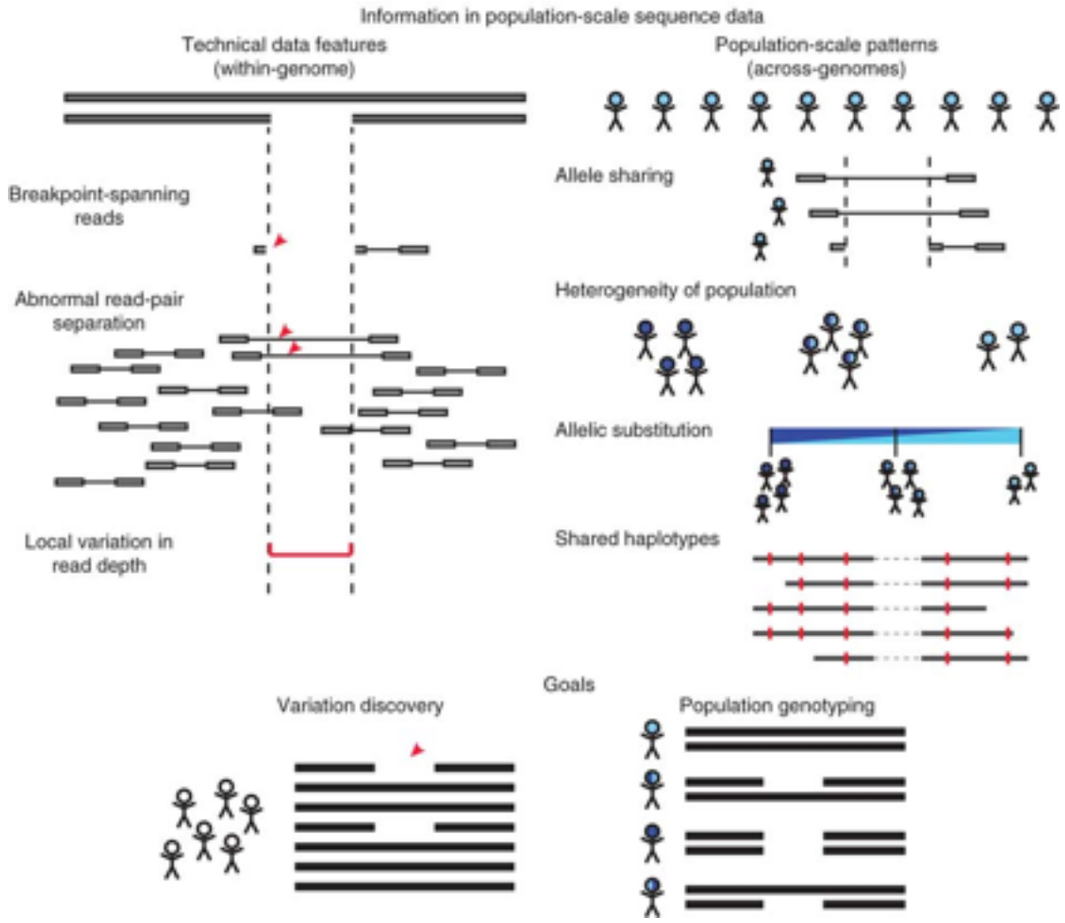
# Post-calling pipeline standardization and scaling is the next big challenge

- 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).

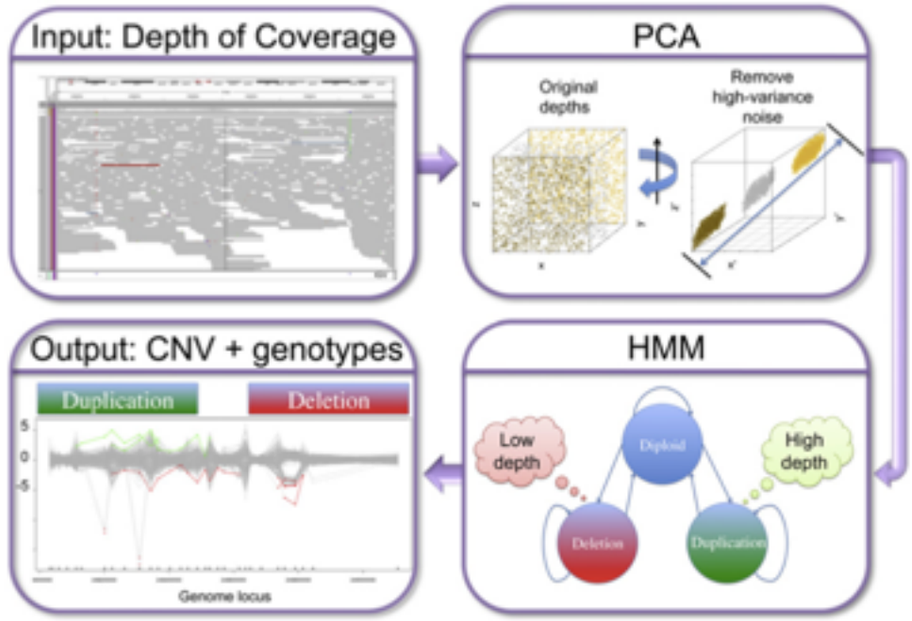
# Structural variation is an important missing piece in the standard analysis

**Genome Strip (GATK)**




Handsaker. R et al. *Nature Genetics*

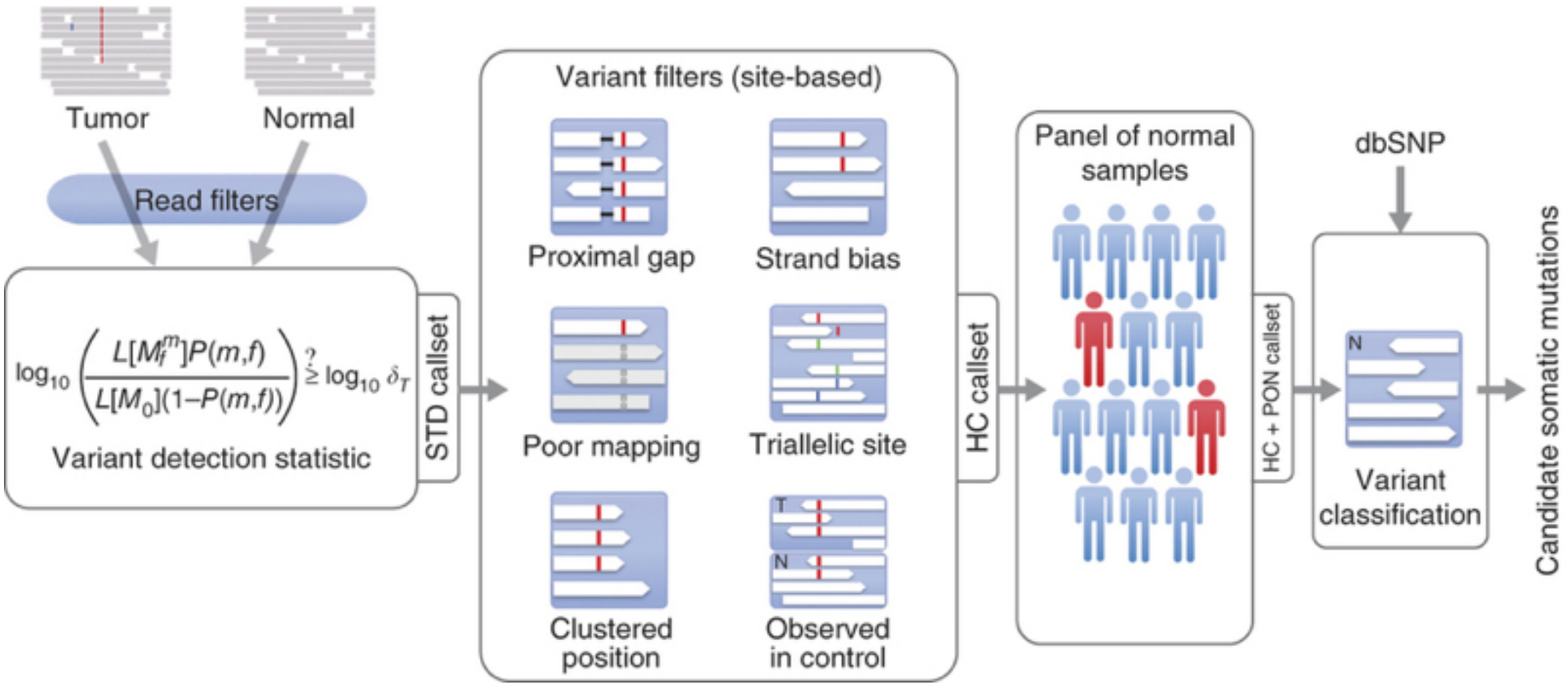
**XHMM (GATK)**




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

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

# Cancer tools also need the same rigorous standardization and scalability

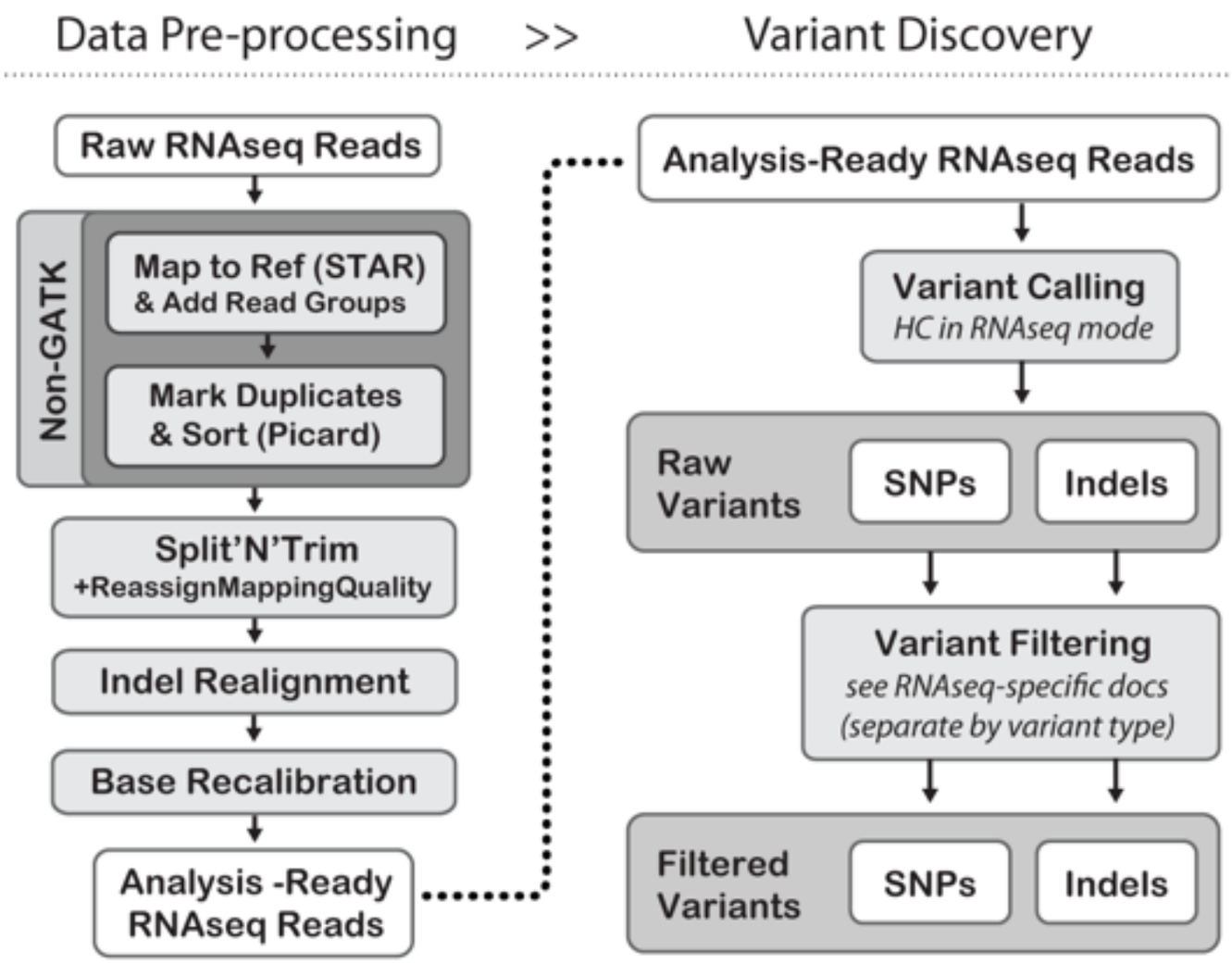



Cibulskis, K et al. *Nature Biotechnology* (2013)

# 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!**

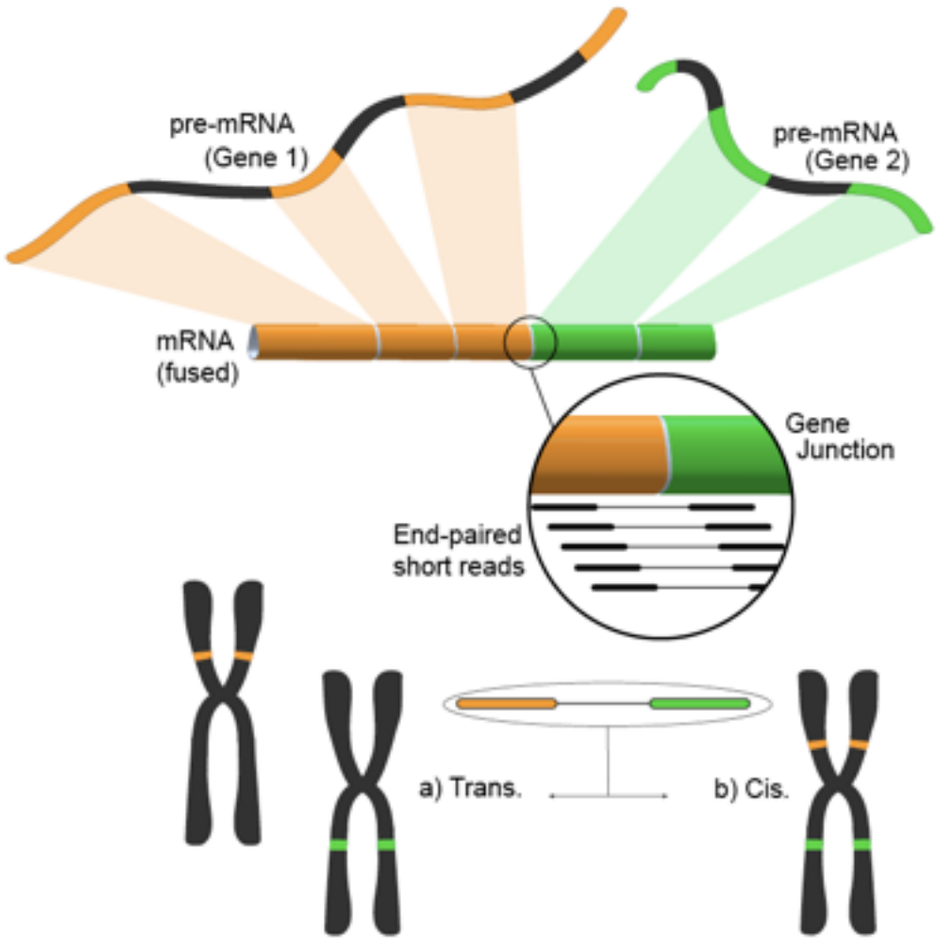
- Improve alignments to comply with our standards of quality
- Specific tools for **allele specific expression** and **RNA editing**
- Design and develop methods for combined analysis between RNA and DNA.




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

# Bonus slides?

**Accelerating tools and the future of the GATK**

# The GATK java codebase has severe limitations

- More than 70% of the instructions in the current GATK pipeline are memory access — the processor is just waiting.

- Excessive use of strings, maps and sets to handle basic data structures that are frequently used in the codebase.

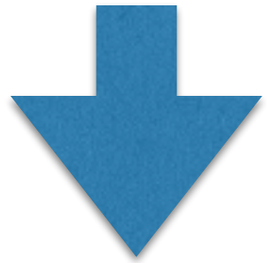
- Java makes it extremely difficult to explore memory contiguity in its data structures.

- Java floating point model is incompatible with modern x86 hardware.

- Java does not offer access to the hardware for optimizations even when desired. As a result, we are forced to underutilize modern hardware.

# A typical GATK-Java Data Structure: A Map-of-Maps-of-Maps

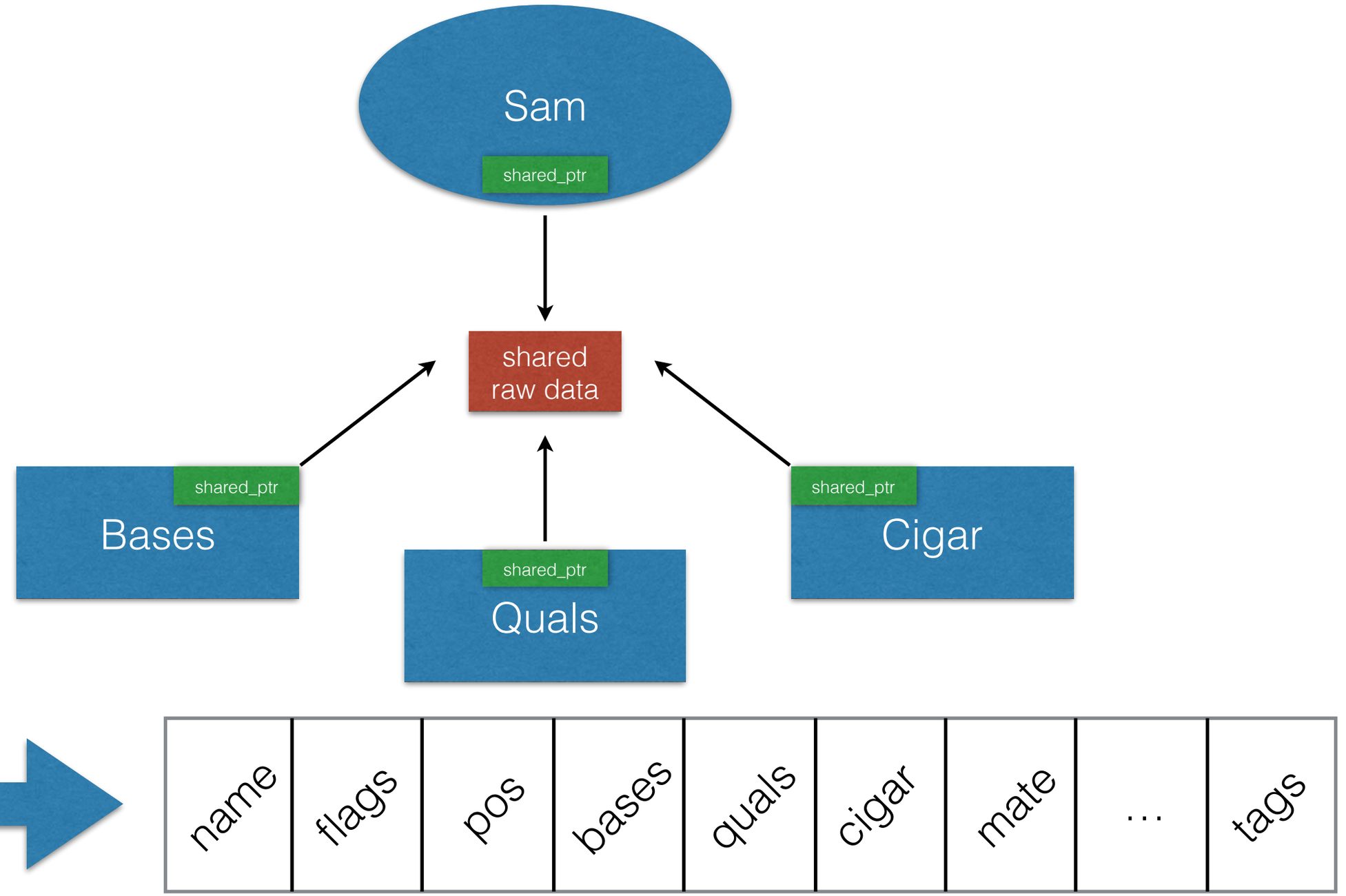
```
Map<String,PerReadAlleleLikelihoodMap> map;
```



```
public class PerReadAlleleLikelihoodMap {
    protected Map<GATKSAMRecord,
                  Map<Allele, Double>> likelihoodReadMap
                = new LinkedHashMap<>();
...
```

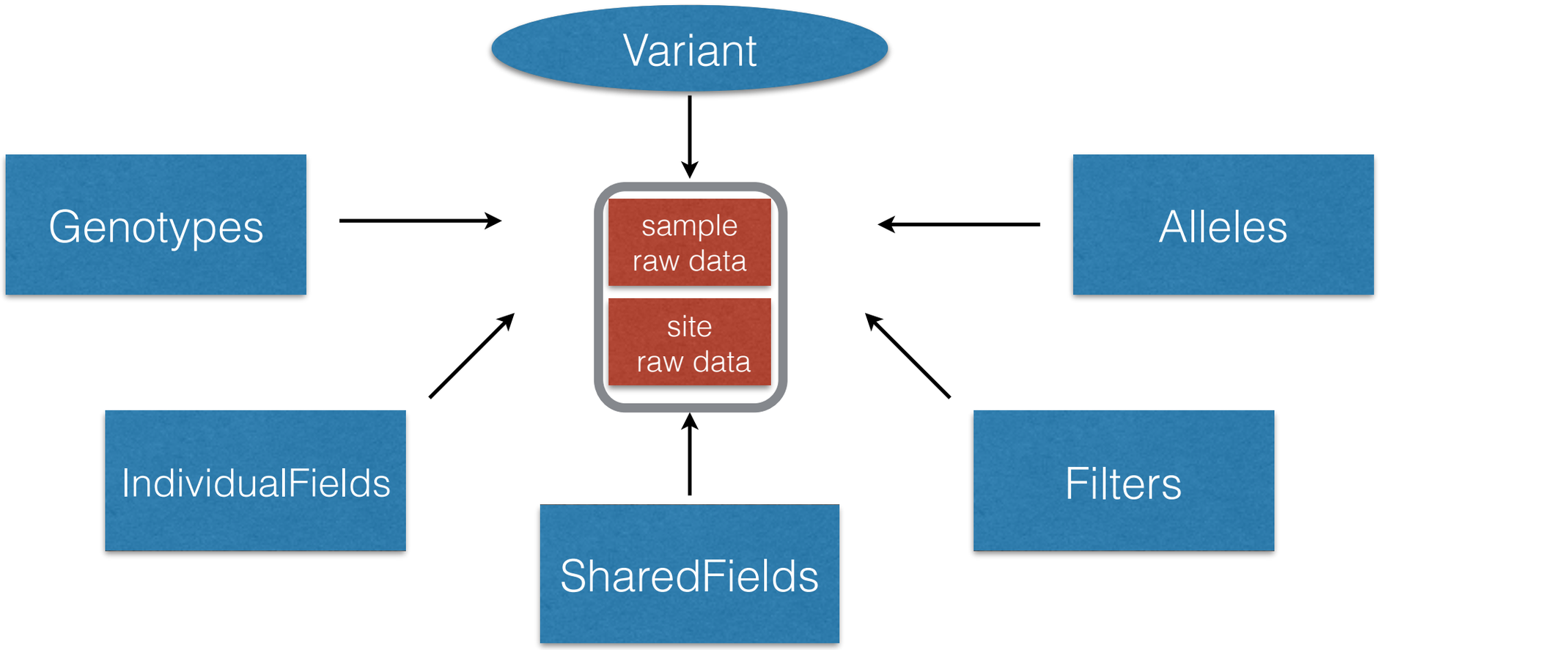
No data locality – most lookups will consist of a series of cache misses

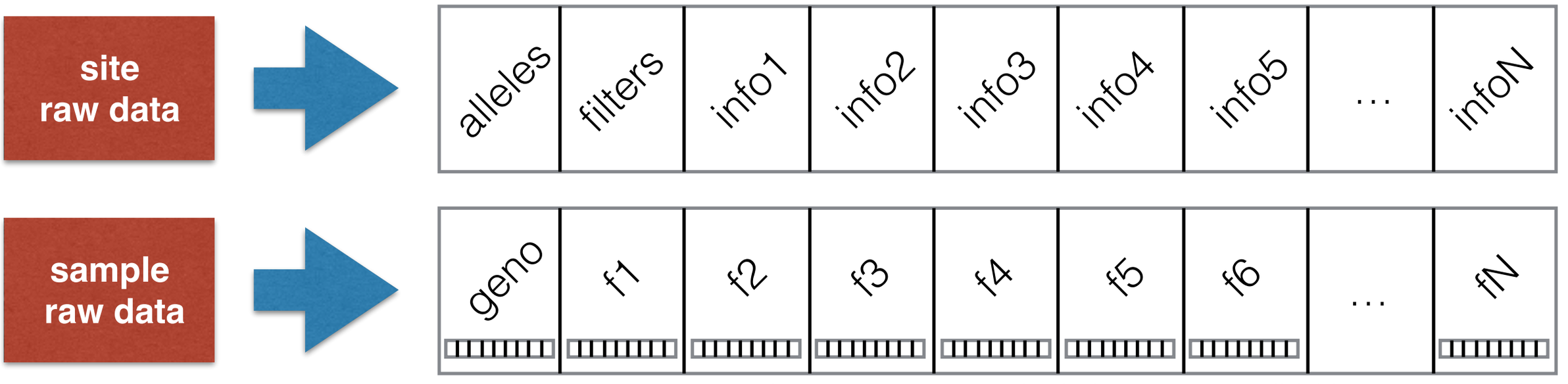
# Gamgee memory model




*in-memory representation is the same as on-disk binary representation*

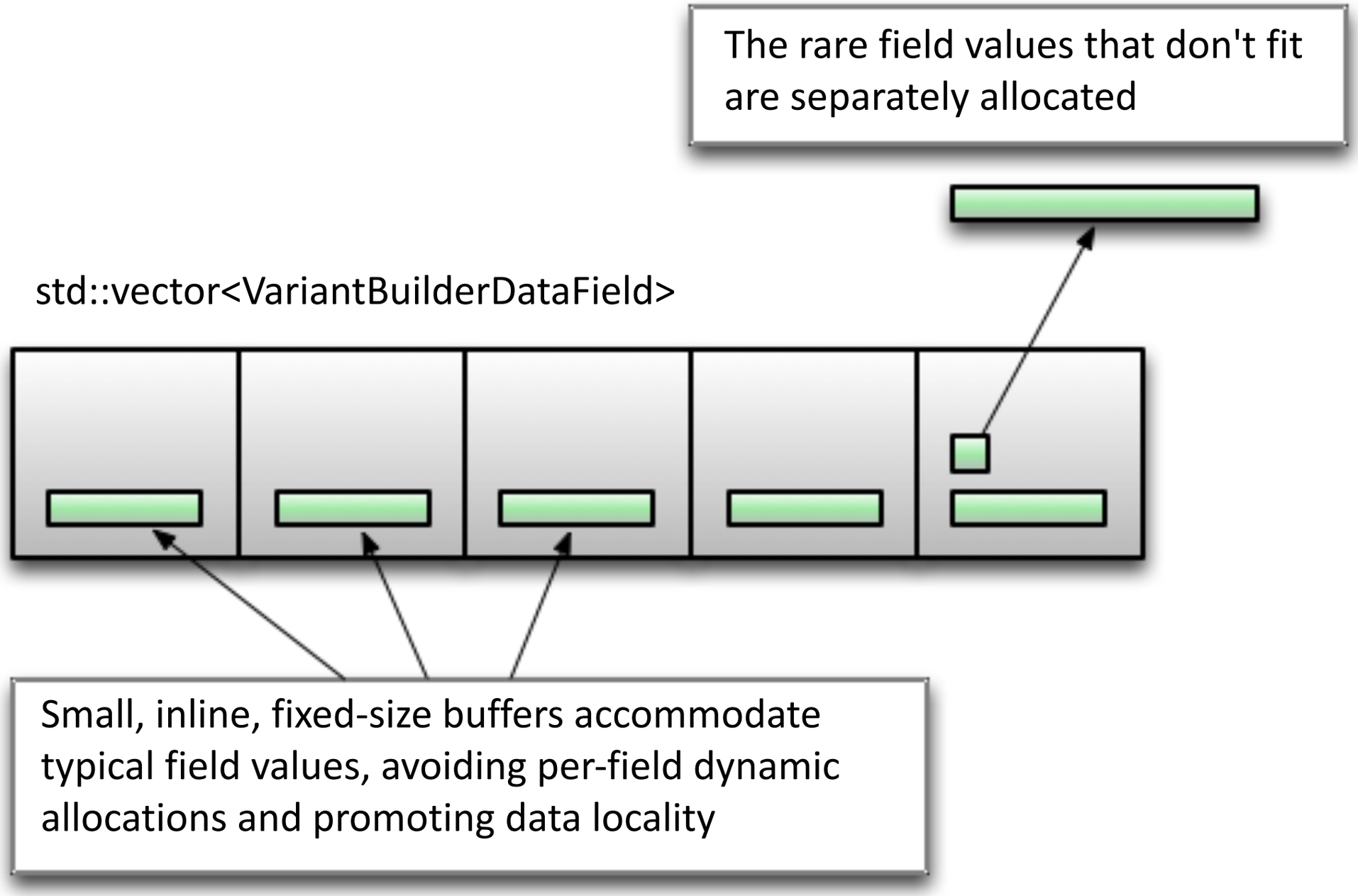
# Gamgee memory model



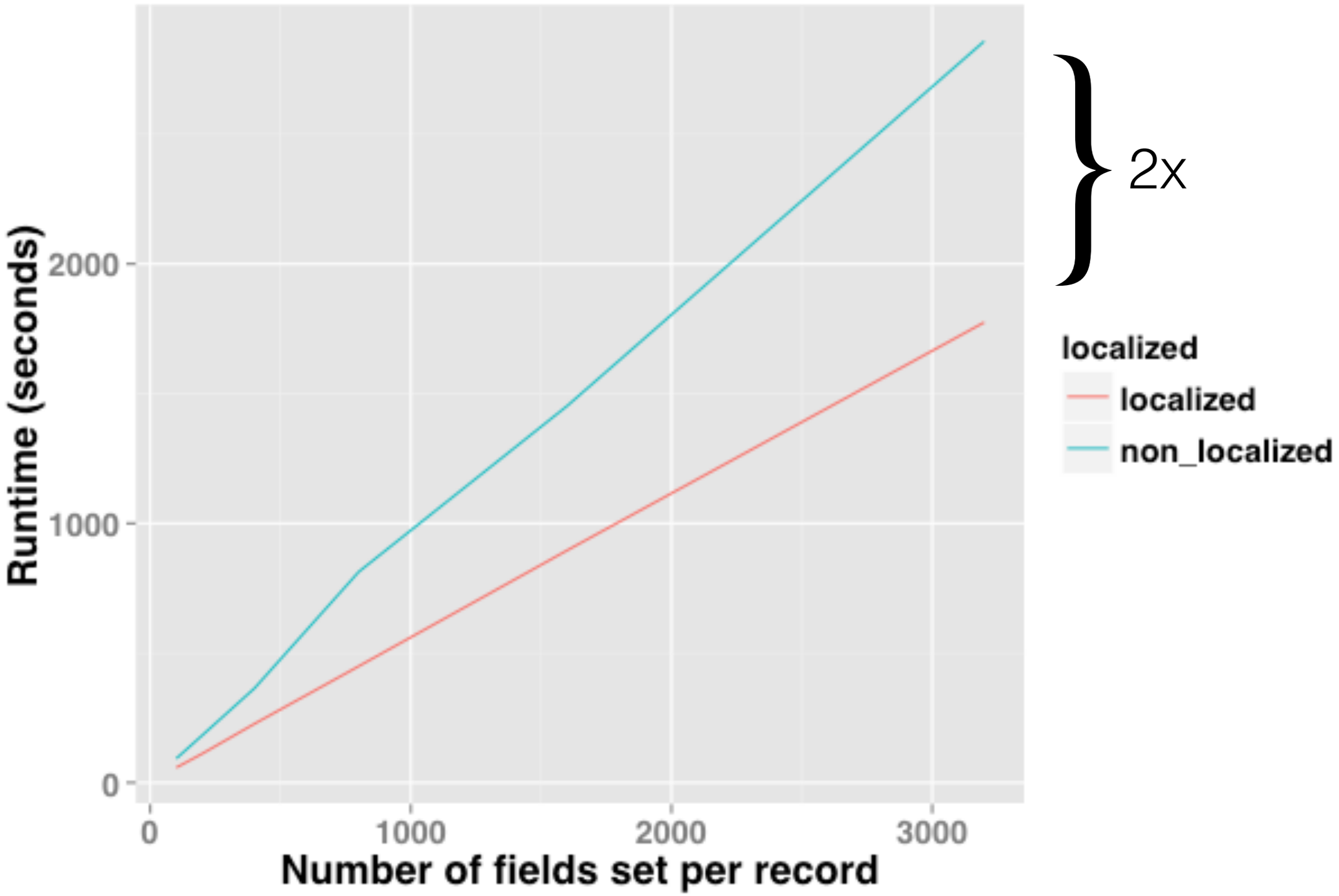

*in-memory representation is the same as on-disk binary representation*

# Preserving data locality

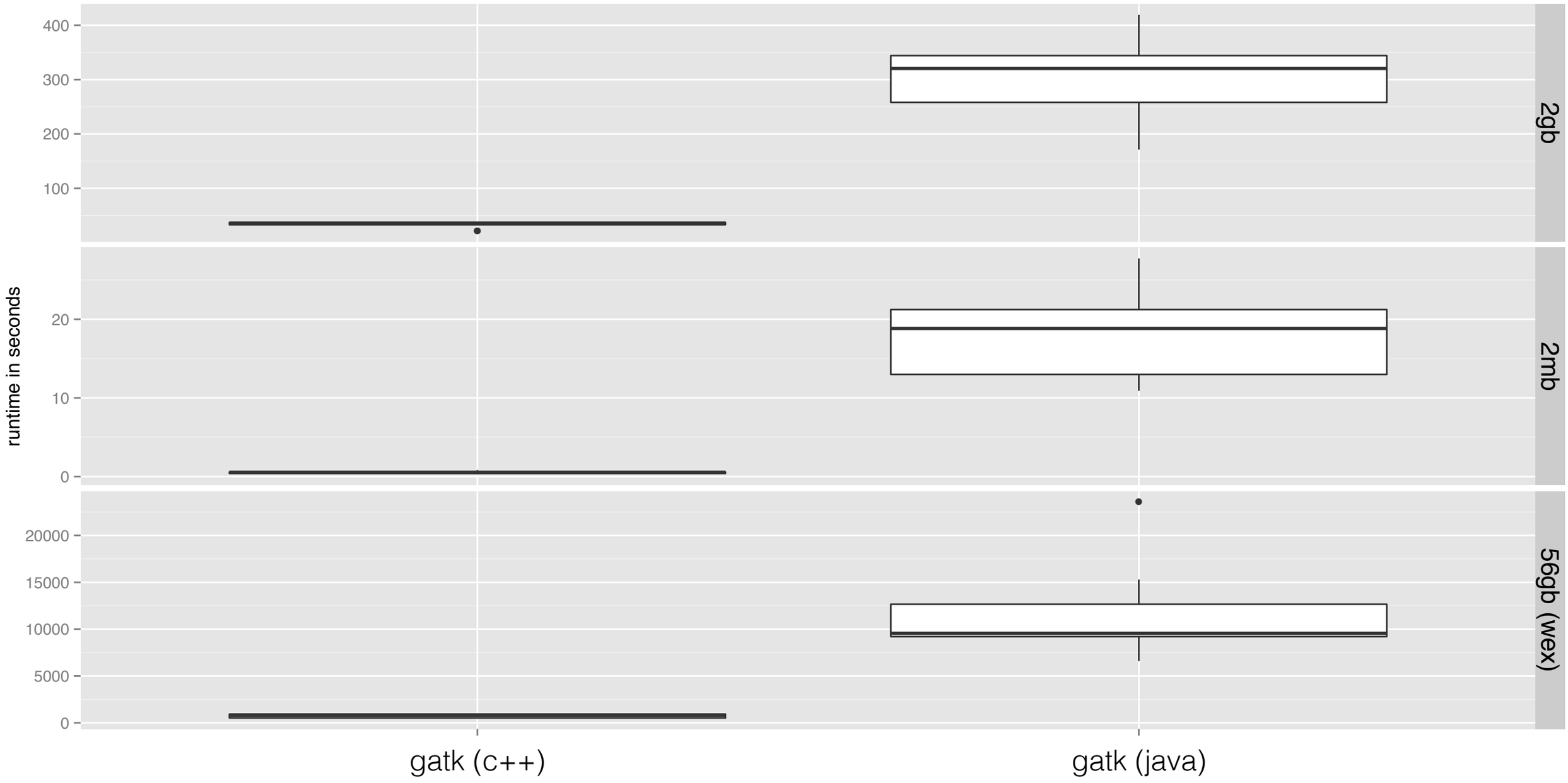



- Same idea as Short String Optimization (SSO) in std::string
- Almost impossible to achieve in Java

# Time to create 3,000,000 variant records in VariantBuilder, with and without data locality optimizations



# Reading BAM files is 17x faster in gamgee



# Reading variant files is much faster in gamgee

| 2GB (1KG) | GATK C++ | GATK Java |
|---|---|---|
| **Text Variant File (VCF)** | 32.71s | 137.57s |
| **Binary Variant File (BCF)** | 4.61s | 242.33s |

the new memory model makes the binary version of the file extremely fast to read and write

# MarkDuplicates is 5x faster

| | GATK C++ | new Picard (java) | old Picard (java) |
|---|---|---|---|
| **Exome** | 4m | 20m | 2h23m |
| **Genome** | 1h15m | 4h47m | 11h06m |

exact same implementation in Java after our C++ version was presented

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

Rare Variant Association Study (RVAS)

*vs*

Common Variant Association Study (CVAS)

*vs*

**The future of the GATK**

# The future of the GATK

Libraries
frameworks

MIT License

gamgee

+

GATK tool developer framework
**c++**

GATK tool developer framework
**java**

Toolkits

GATK License

**c++**

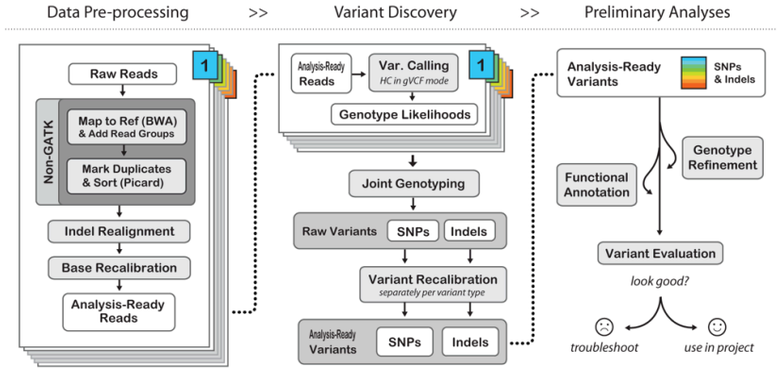
gatk

+

**java**

gatk

# Research tools need this scalability for the next wave of scientific advances




**Data Processing from DNA to Variants**

ready for ~1 million genomes

(will need more work to reach tens-hundreds of millions)

**Variant analysis and association studies**

fails today at just a few thousand genomes