Future directions in sequencing data analysis

Mauricio Carneiro

carneiro@broadinstitute.org

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

This is the work of many…

team












colleagues



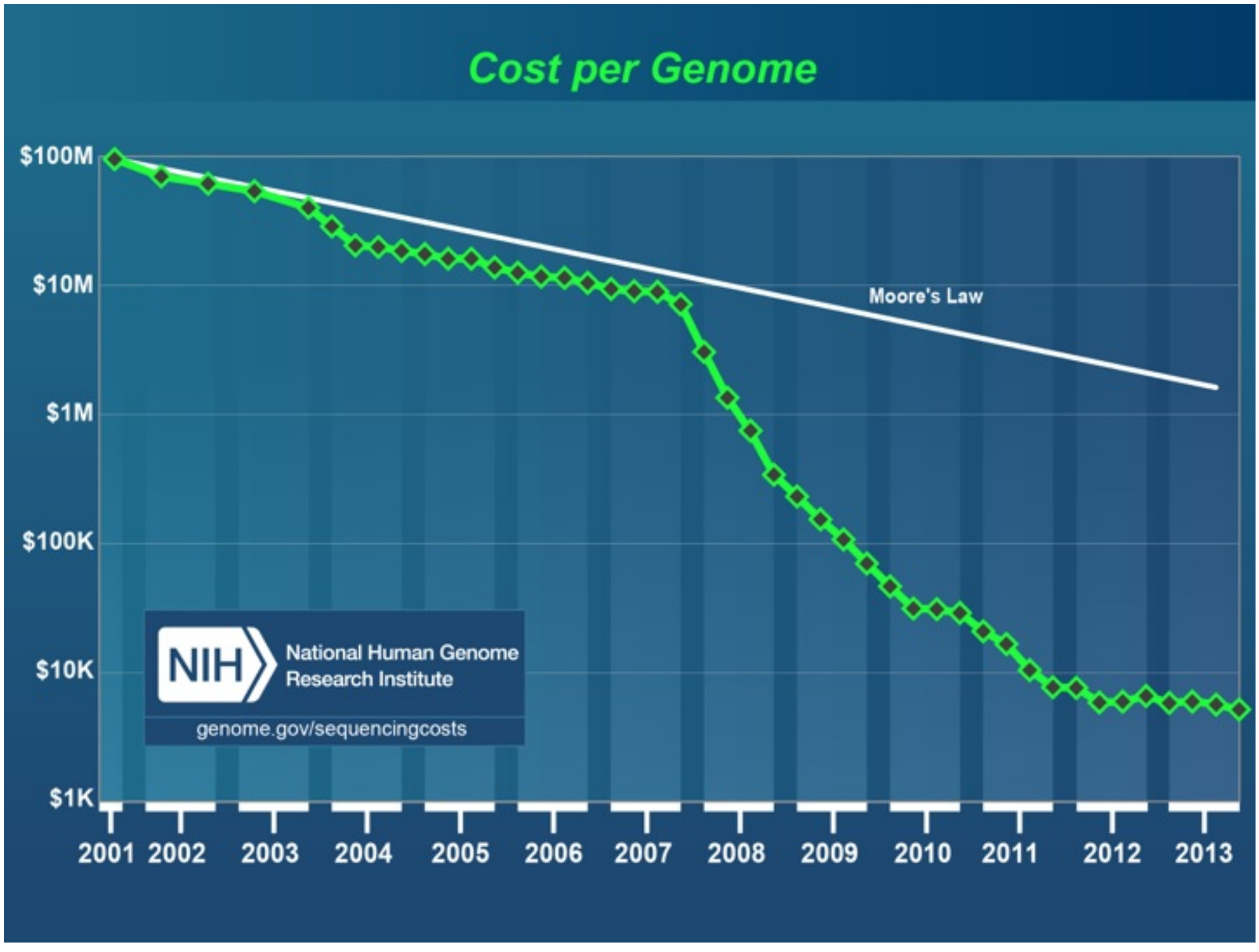





collaborators



vastly outpacing moor's law



Genomics Platform in 2013

50
HiSeqs

10
MiSeqs

2
NextSeqs

14
HiSeq X

6.5
Pb of data

427
projects

180
people

2.1
Tb/day

Genomics Platform in 2013

44,130
exomes

2,484
exome express

2,247
genomes

2,247
assemblies

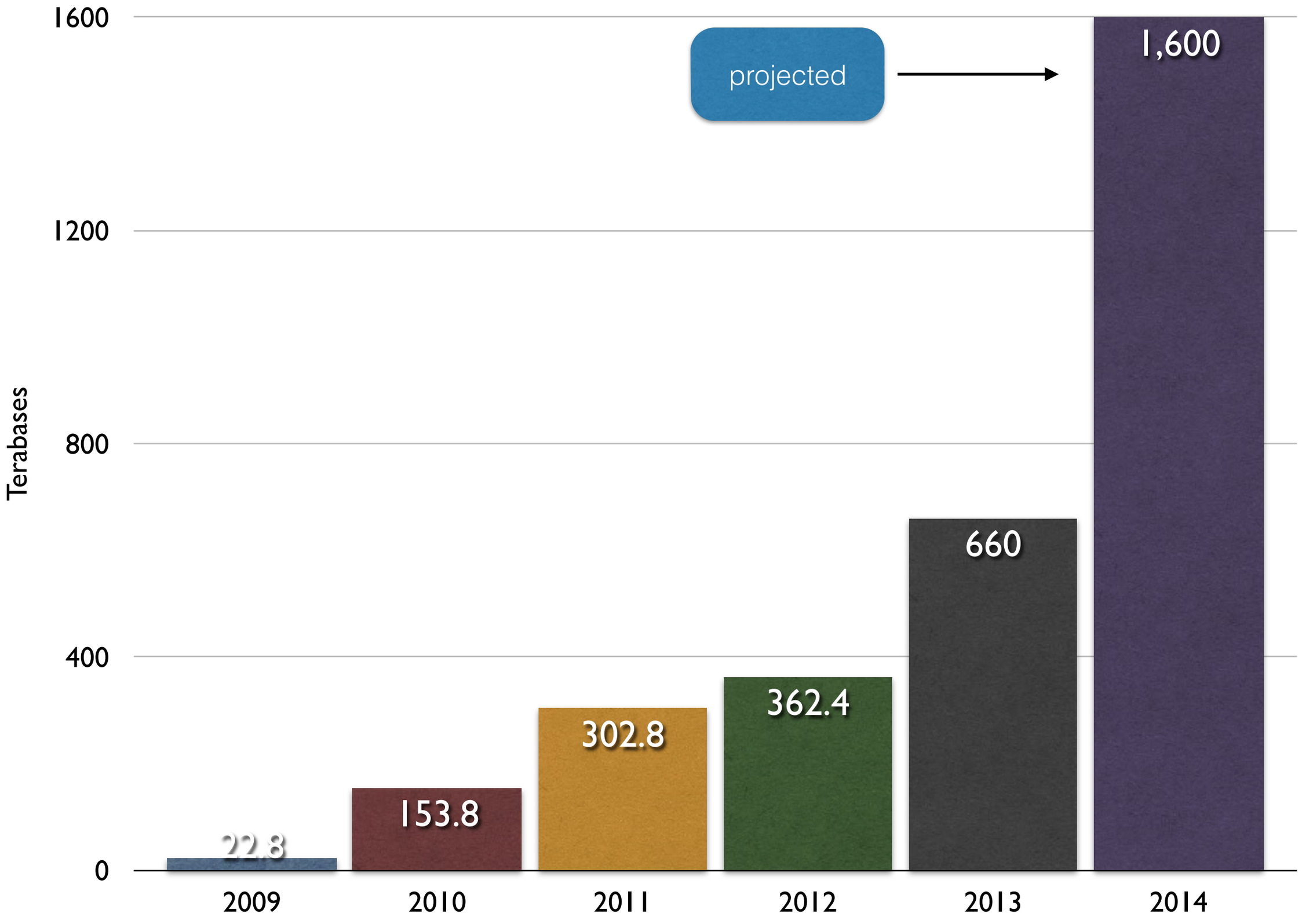
8,189
RNA

9,788
16S

47,764
arrays

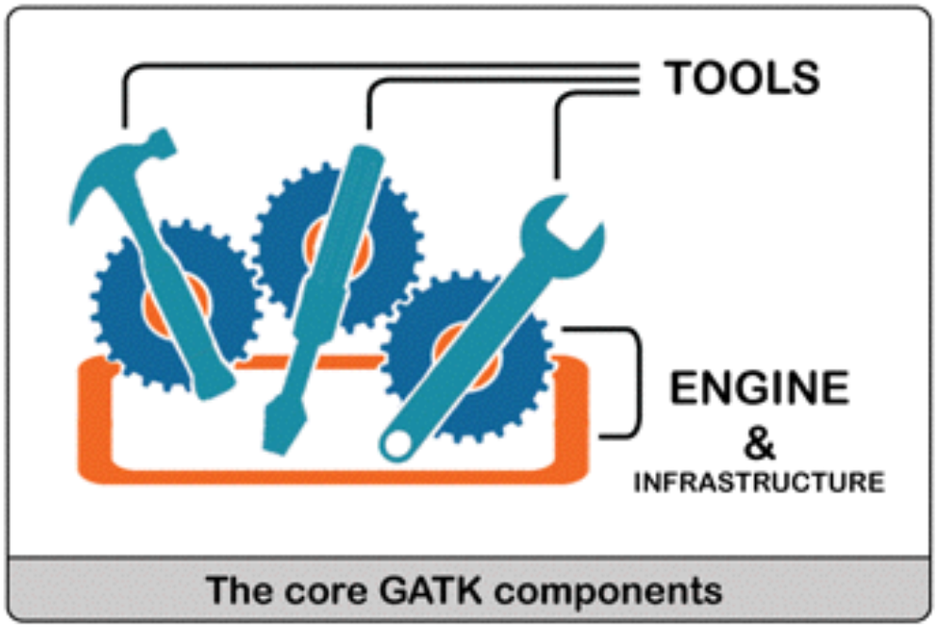
228
cell lines

Terabases of Data Produced by Year



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

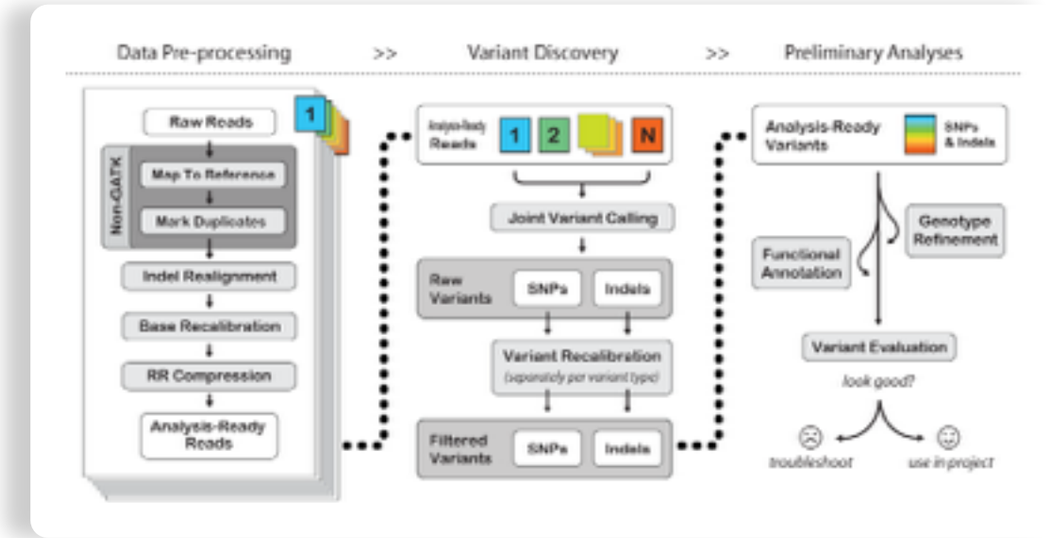
Toolkit & framework packages



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

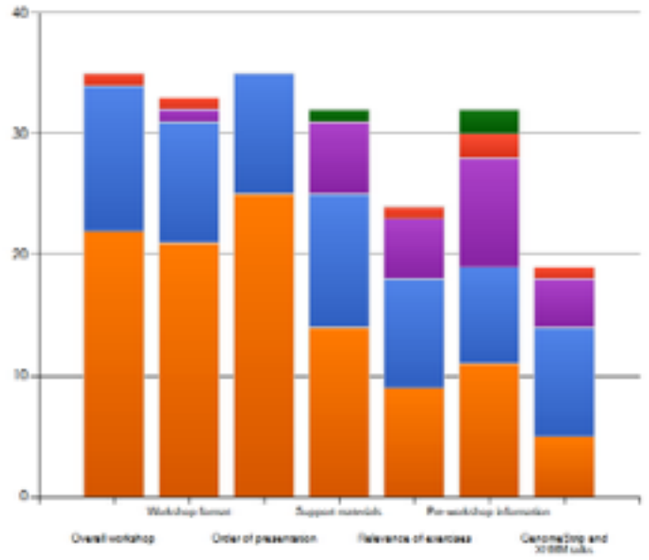
Workshop series educates local and worldwide audiences

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

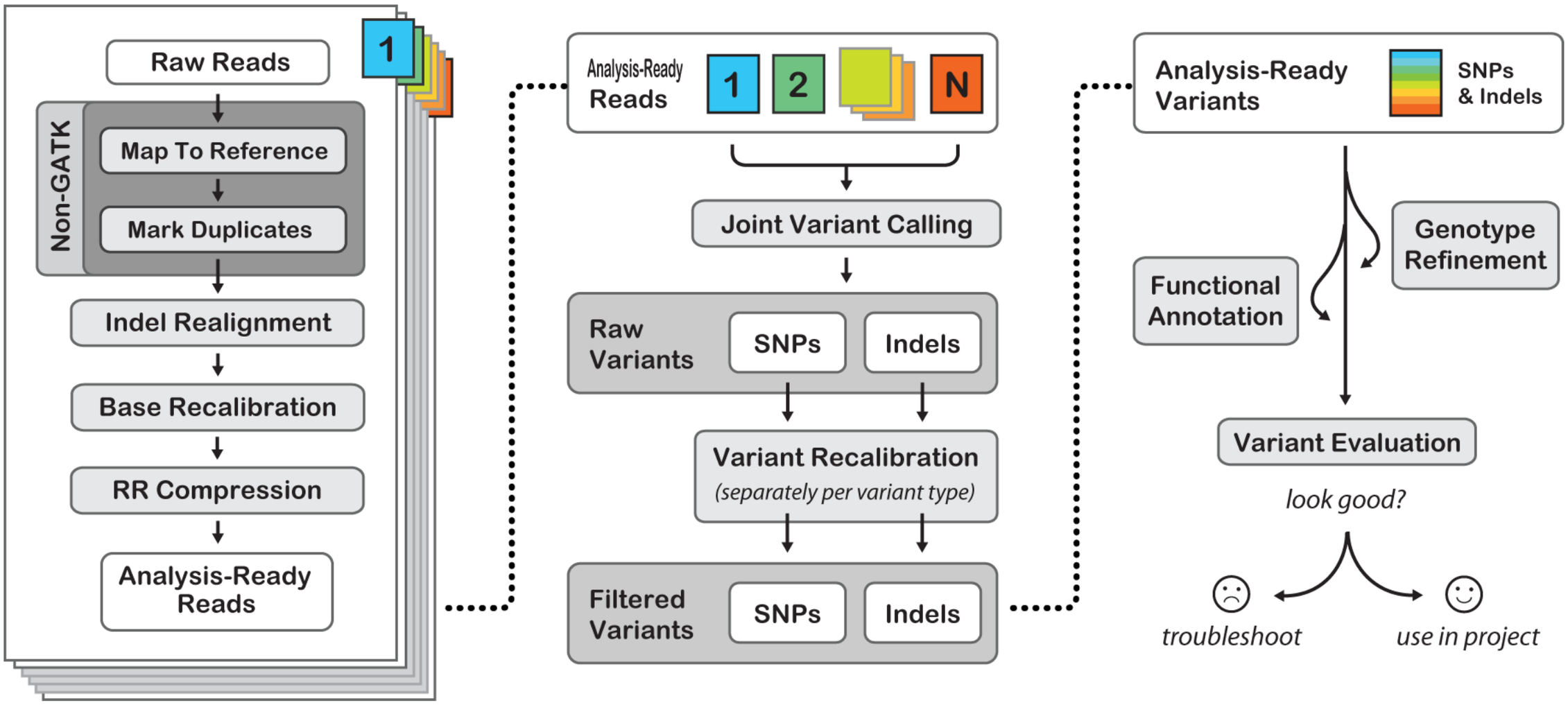
BroadE: Overview of GATK & best practices

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)

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

Rare Variant Association Study (RVAS)

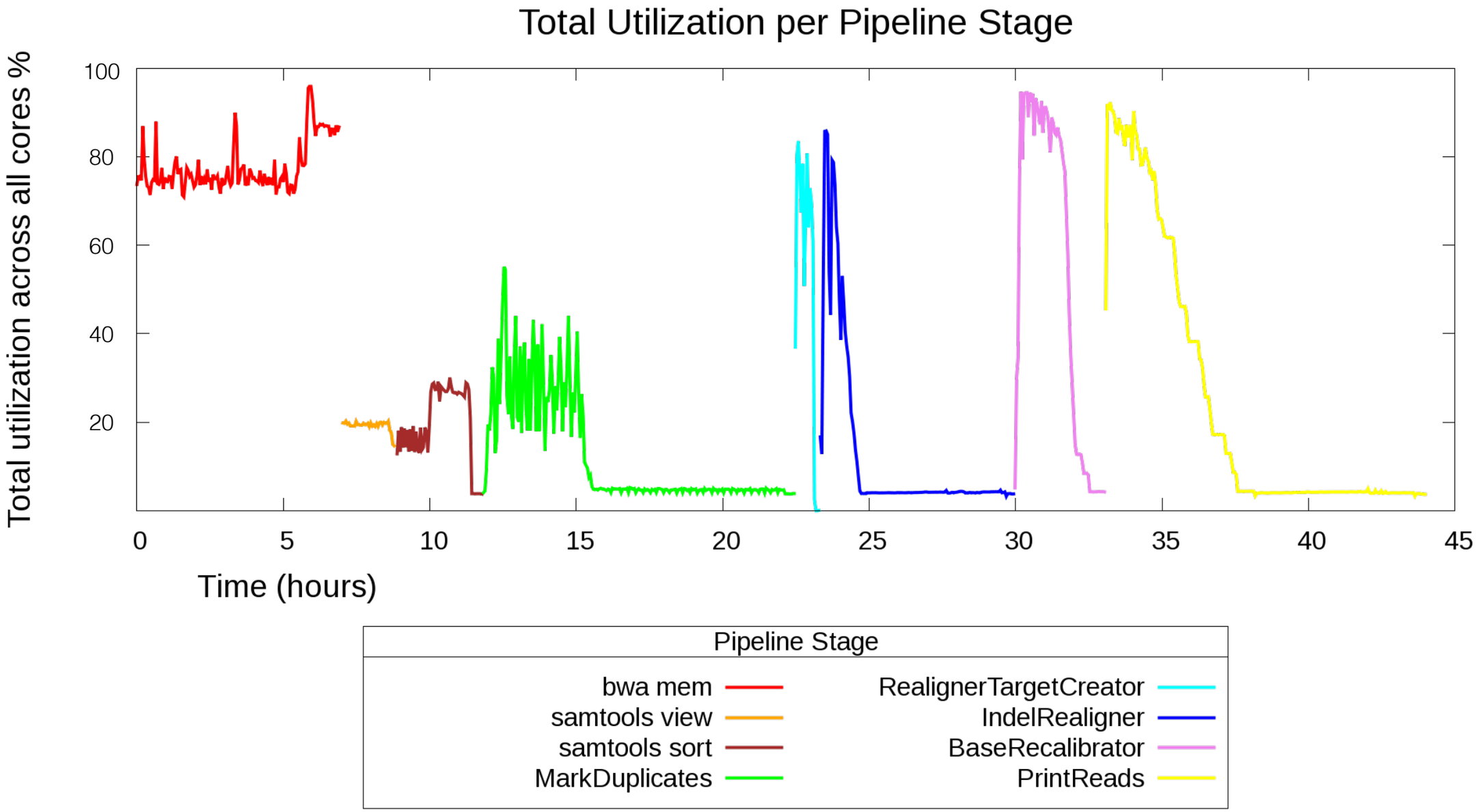
vs

Common Variant Association Study (CVAS)

vs

Technical challenge
all samples need to be consistently processed

Processing times on a whole genome sample



Future challenges to scale up the pre-processing pipeline

- Eliminate disk read/writing in between pipeline steps
- Redesign algorithms with performance in mind (e.g. Duplicate Marking)
- Make use of heterogeneous compute alternatives (e.g. SSE/AVX, GPU, …)
- Reduce time spent doing unnecessary calculations (e.g. Base Recalibration on *good* data)
- Eliminate Indel Realignment
- New methods (entirely) that can replace long running processes (e.g. different statistical approach for Base Recalibration?)

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**

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

Rare Variant Association Study (RVAS)

vs

Common Variant Association Study (CVAS)

vs

Technical challenge
all samples must be *jointly* called

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

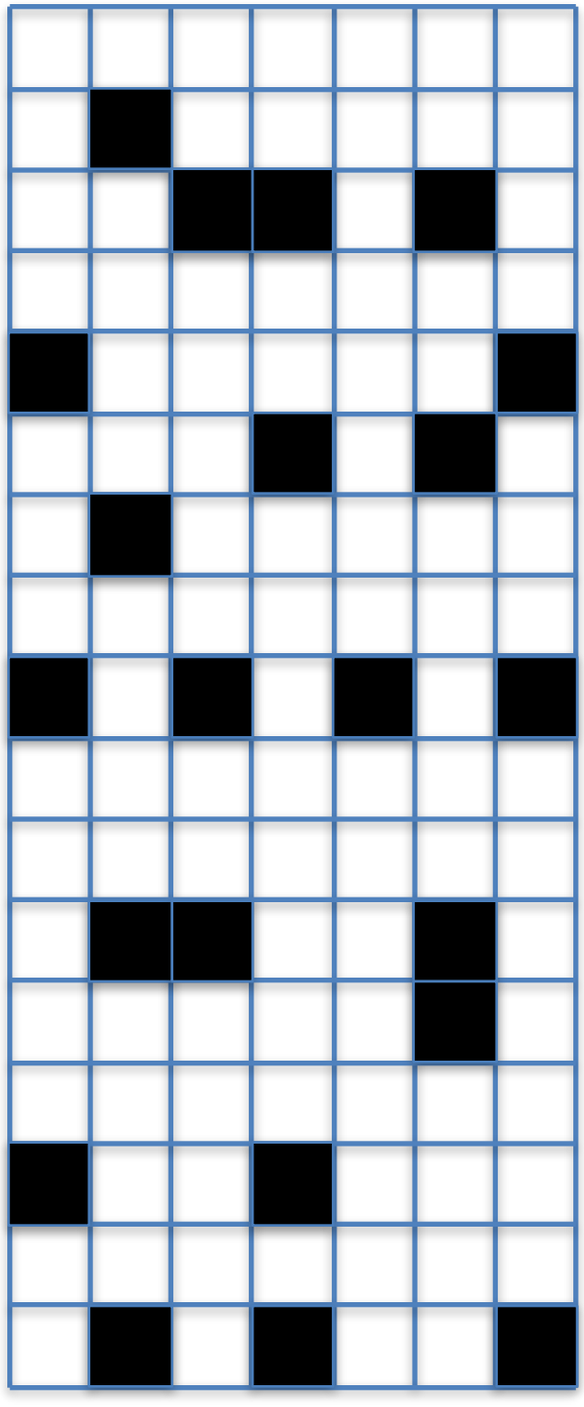
Likelihoods:
A/B/C phred-scaled probability of hom (A), het (B), hom-alt (C) genotypes given NGS data

Joint Calling

The importance of Joint calling

Individuals

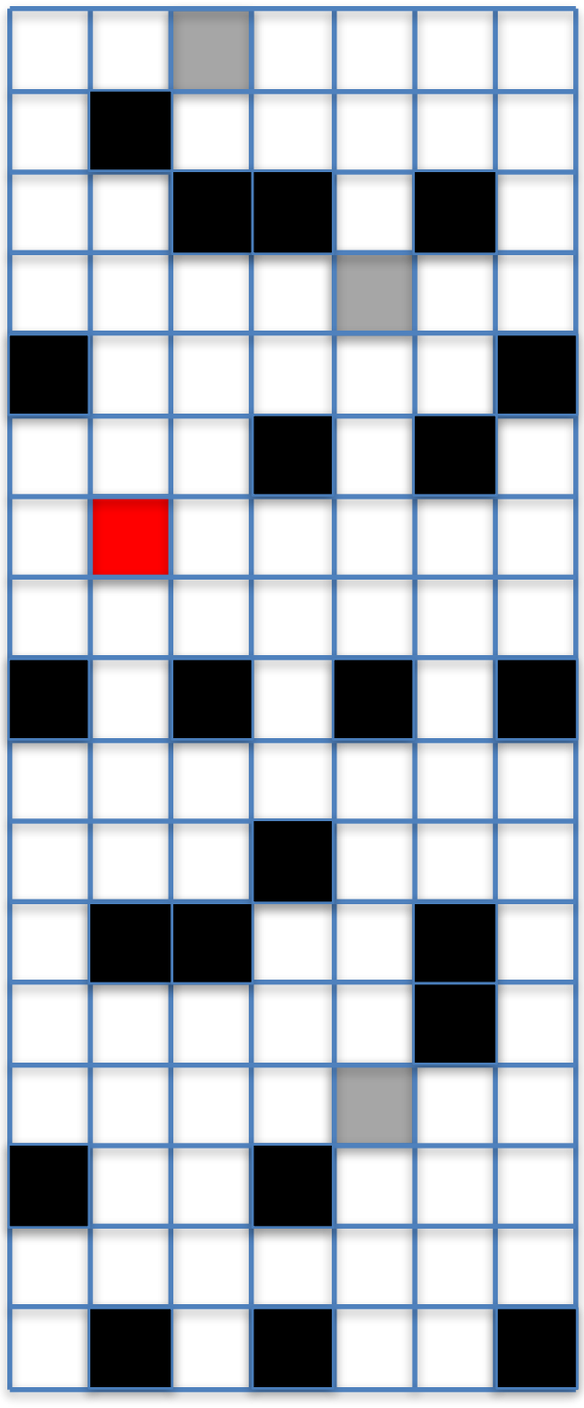
Variant positions



The importance of Joint calling

Individuals

Variant positions



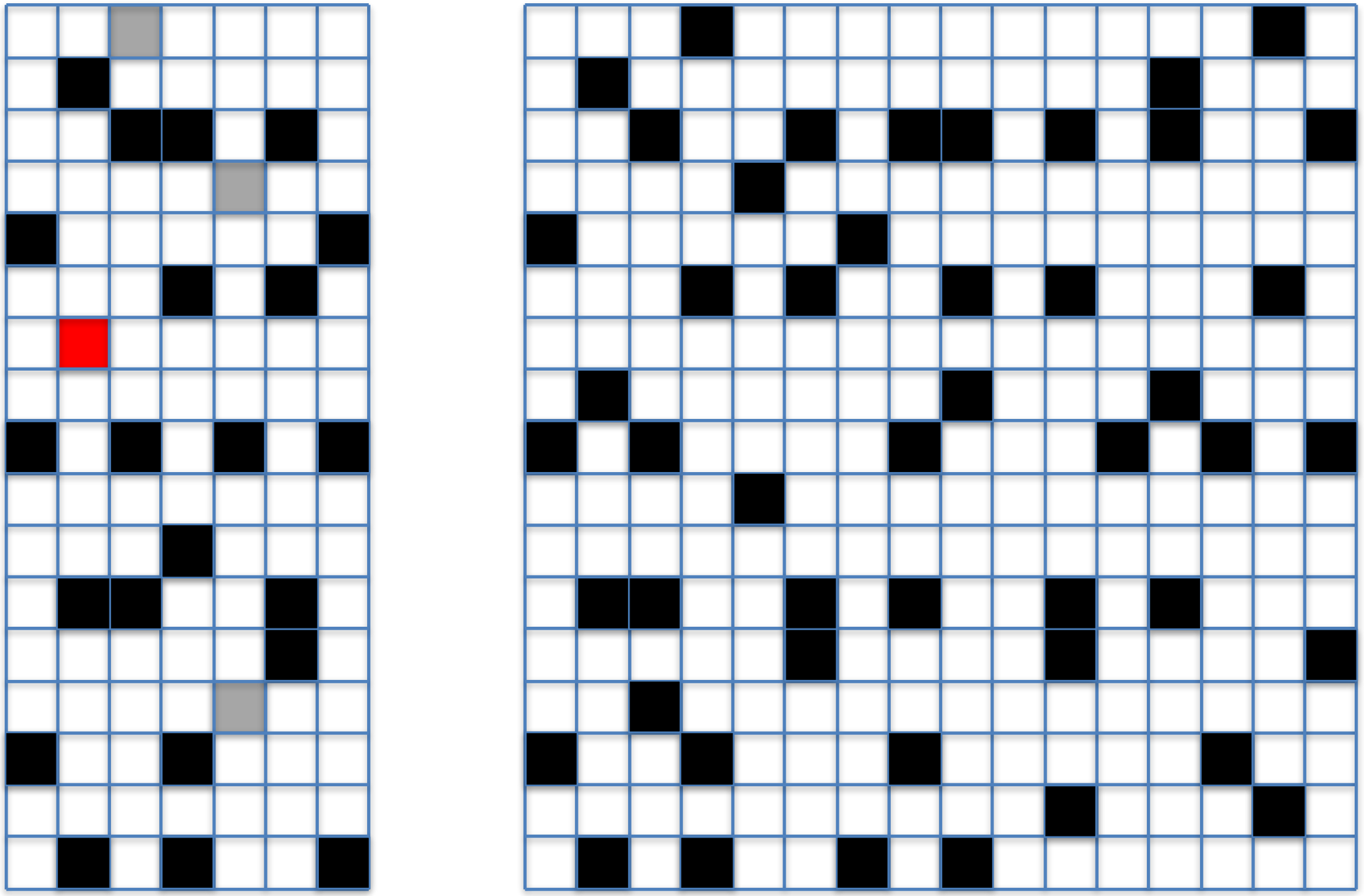
false negatives

false positives

The importance of Joint calling

Individuals

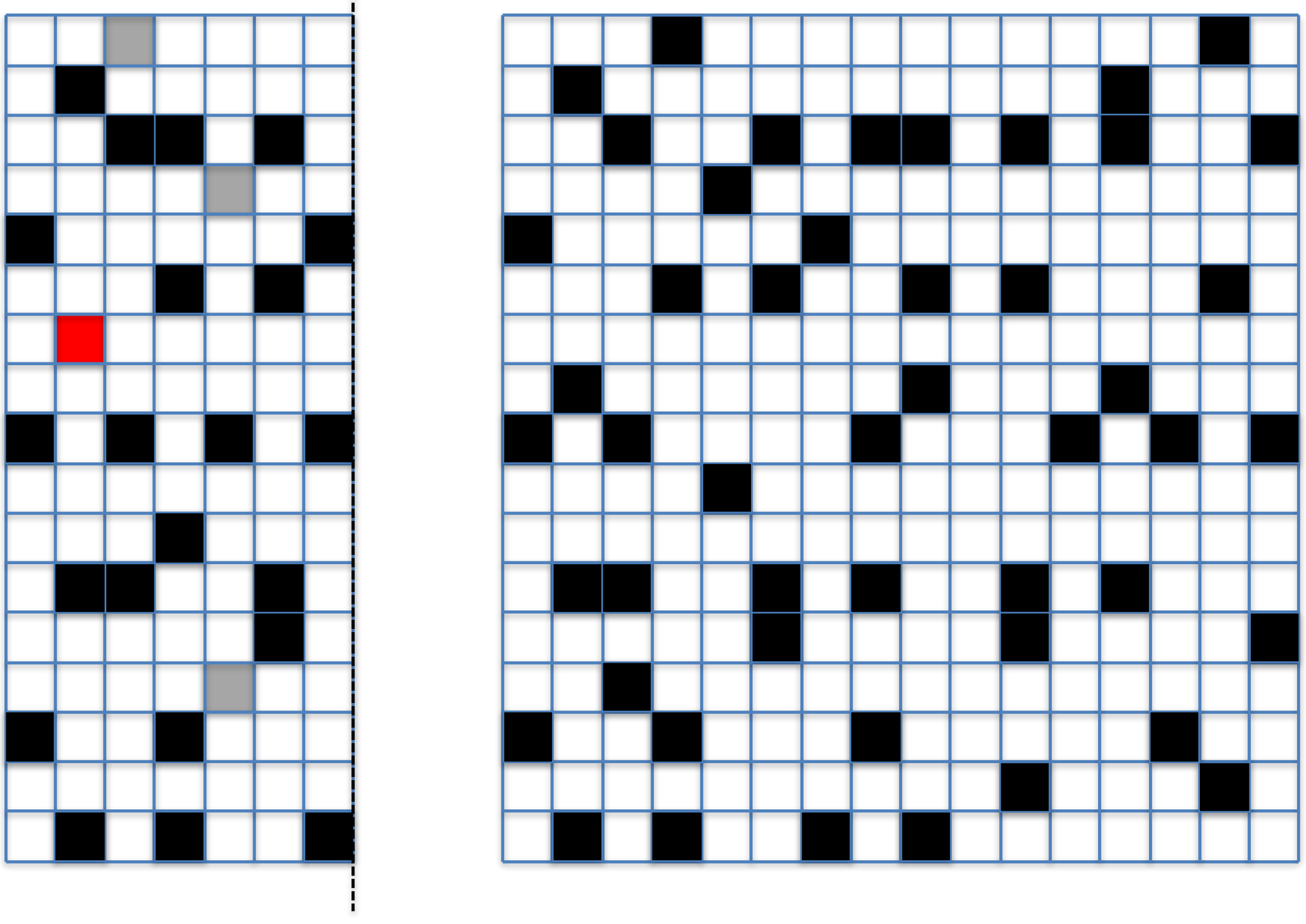
Variant positions



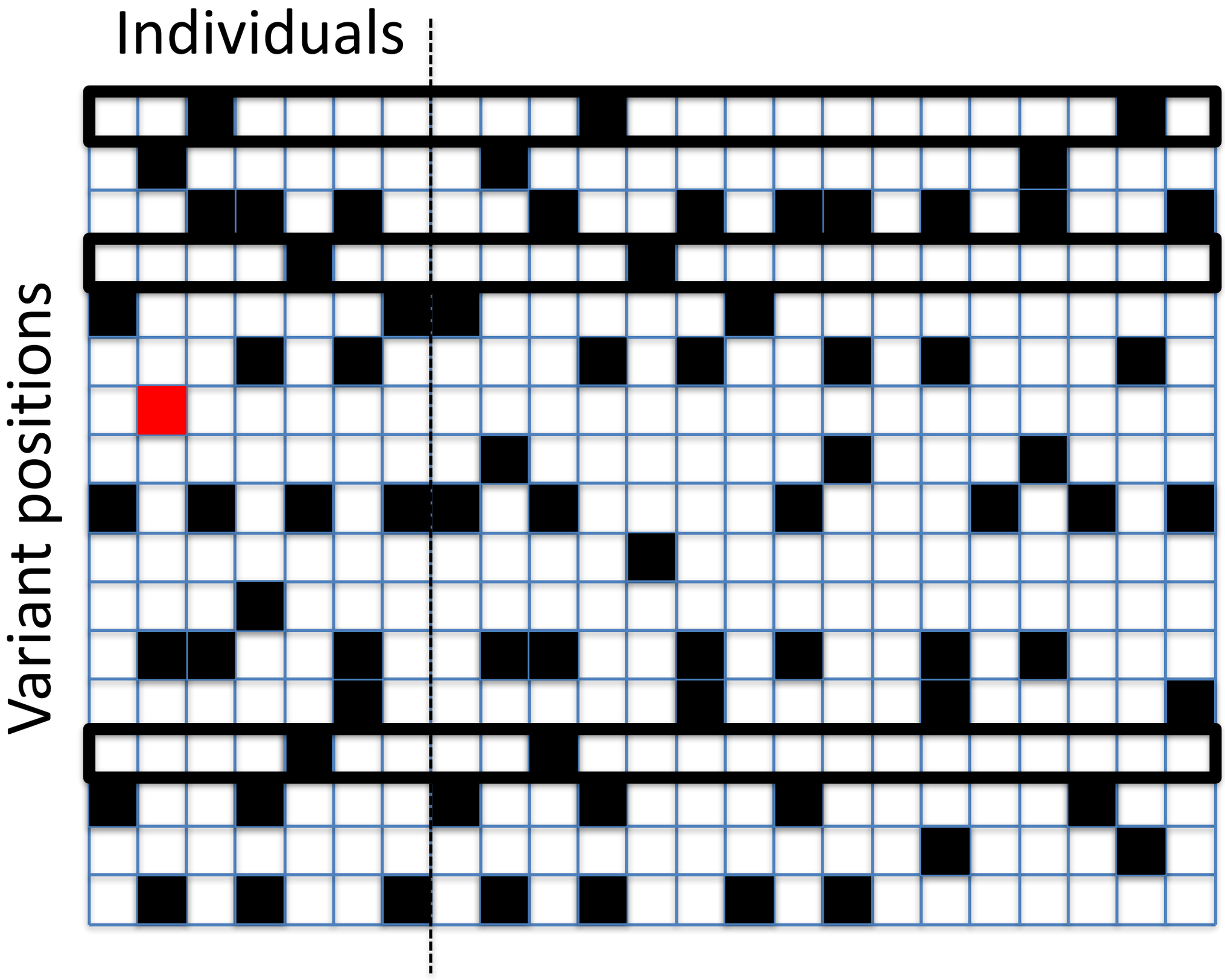
The importance of Joint calling

Individuals

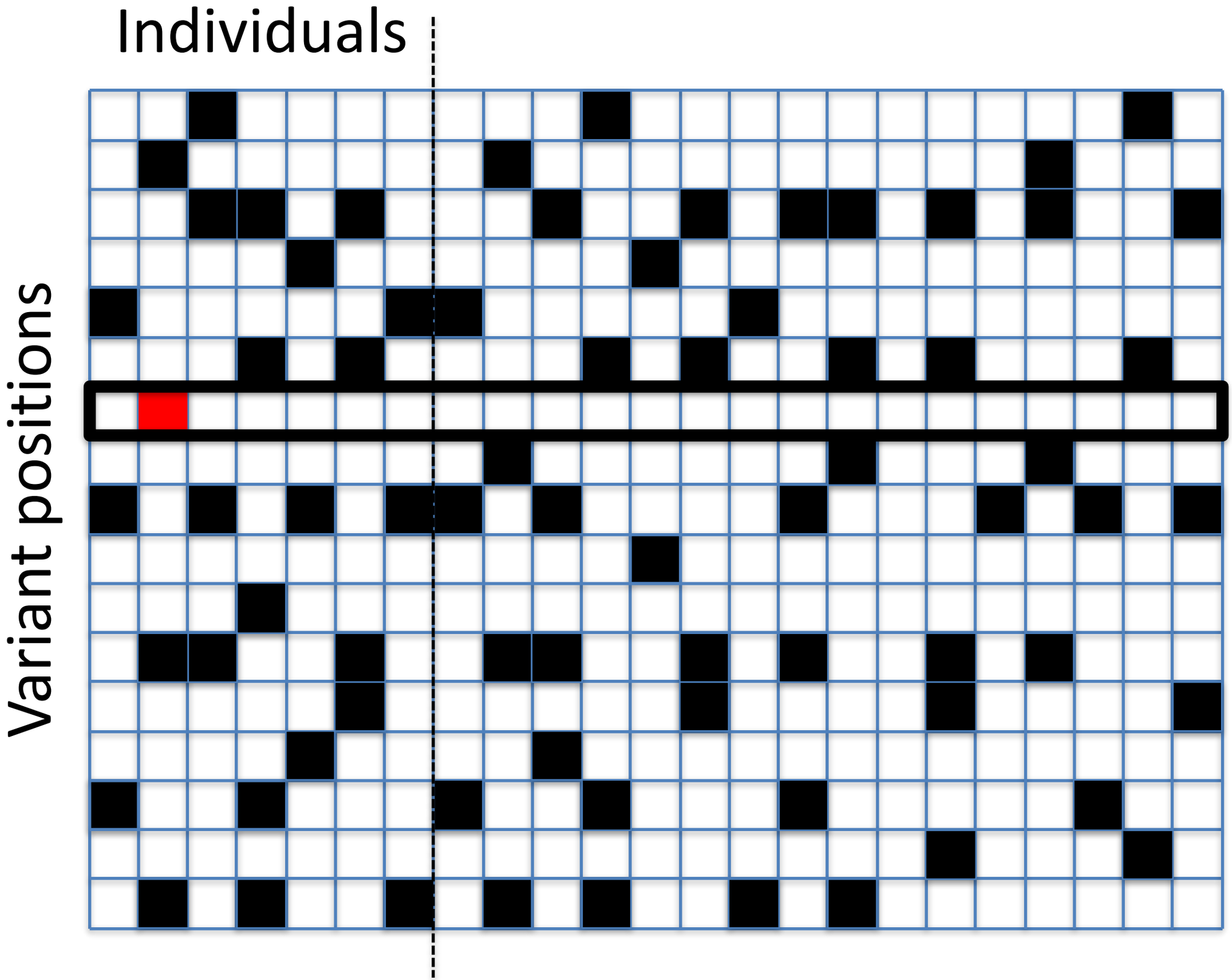
Variant positions



The importance of Joint calling



The importance of Joint calling

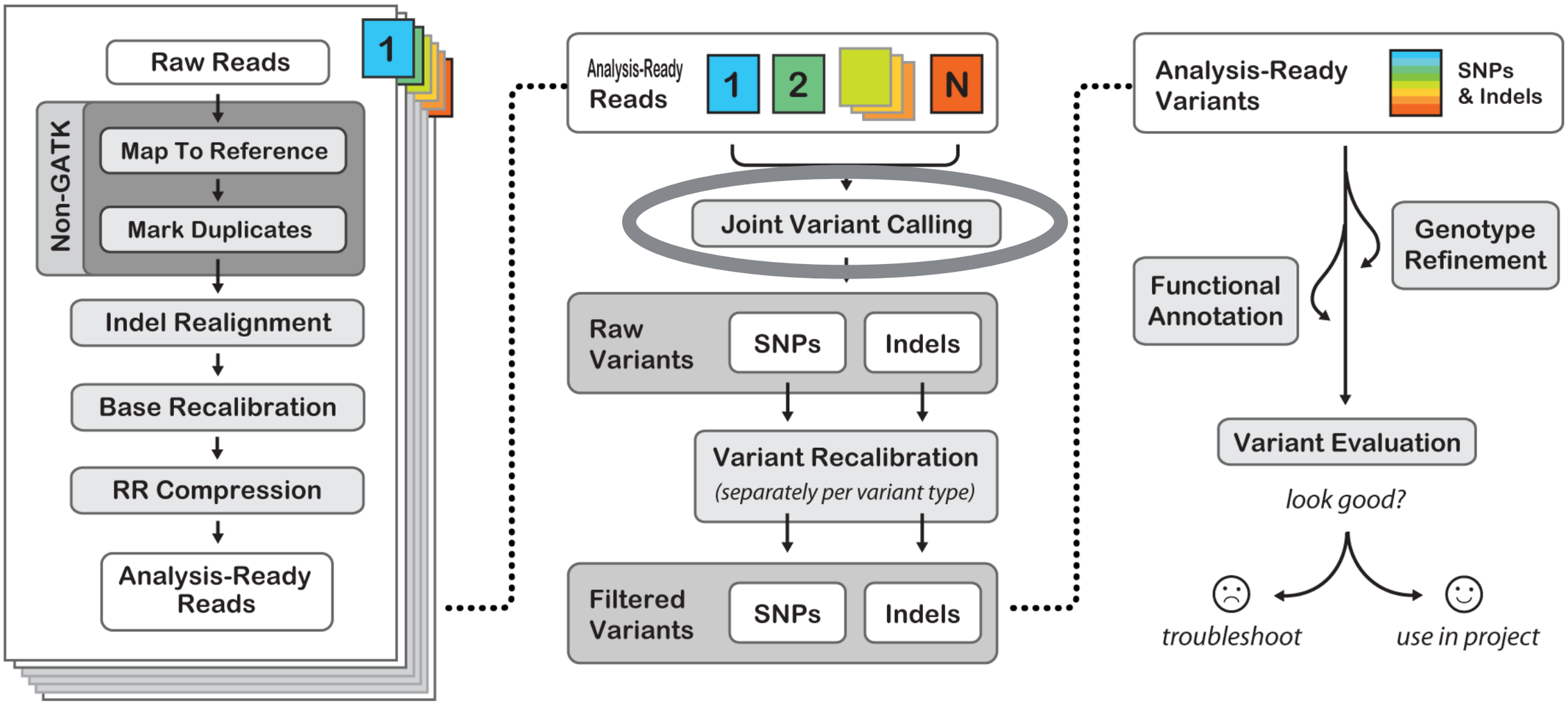


The importance of Joint calling

Individuals

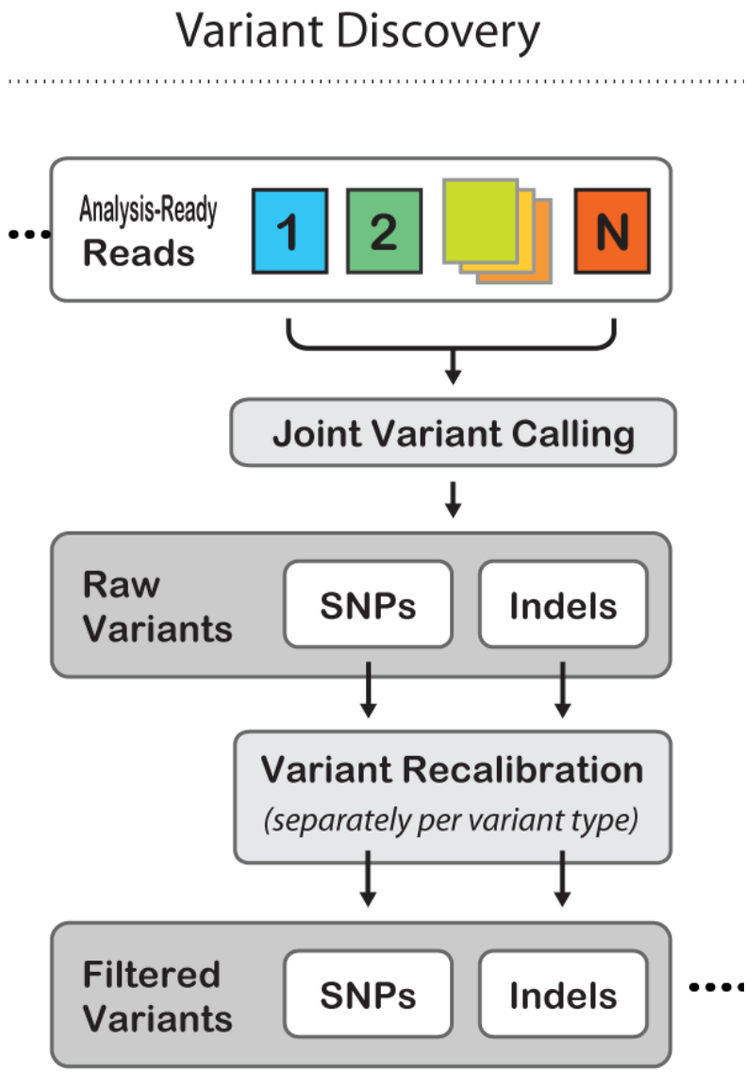
Variant positions

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



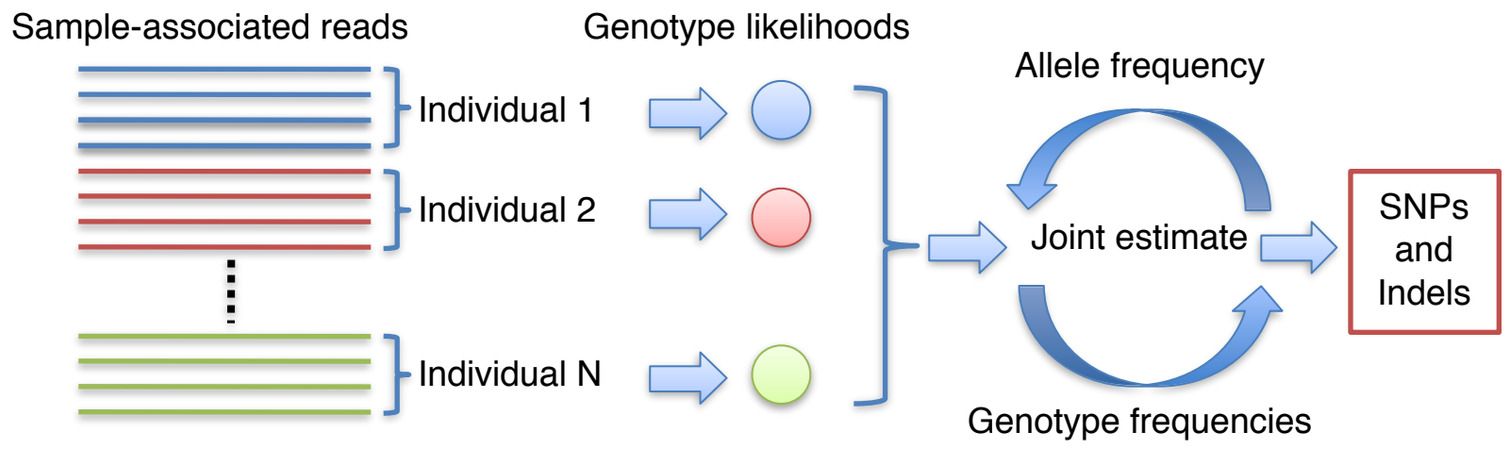
prior Likelihood

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

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)

Calling 57,000 samples jointly is not easy

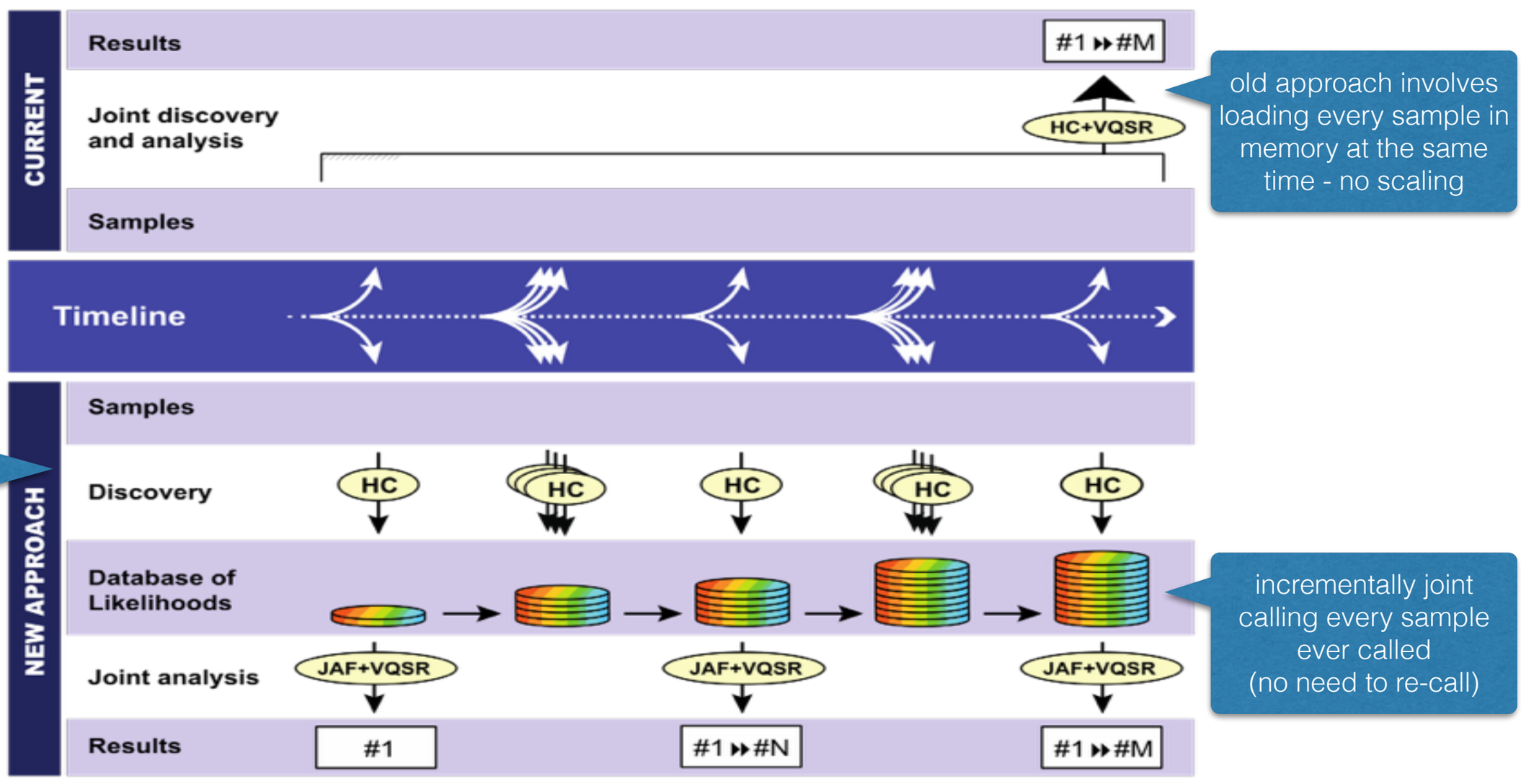
- Raw BAM files are **~1 Pb,** Reduced BAM files are **~100 Tb**
- **2,508,000 CPU hours** for data processing and **84,000** for variant calling with the *Unified Genotyper*.
- Some regions required **64 Gb RAM** (most required under 16 Gb)
- Final compressed VCFs are **2.35 Tb**

MacArthur, D. *in prep. (2014)*

Heterogeneous compute speeds up variant calling significantly

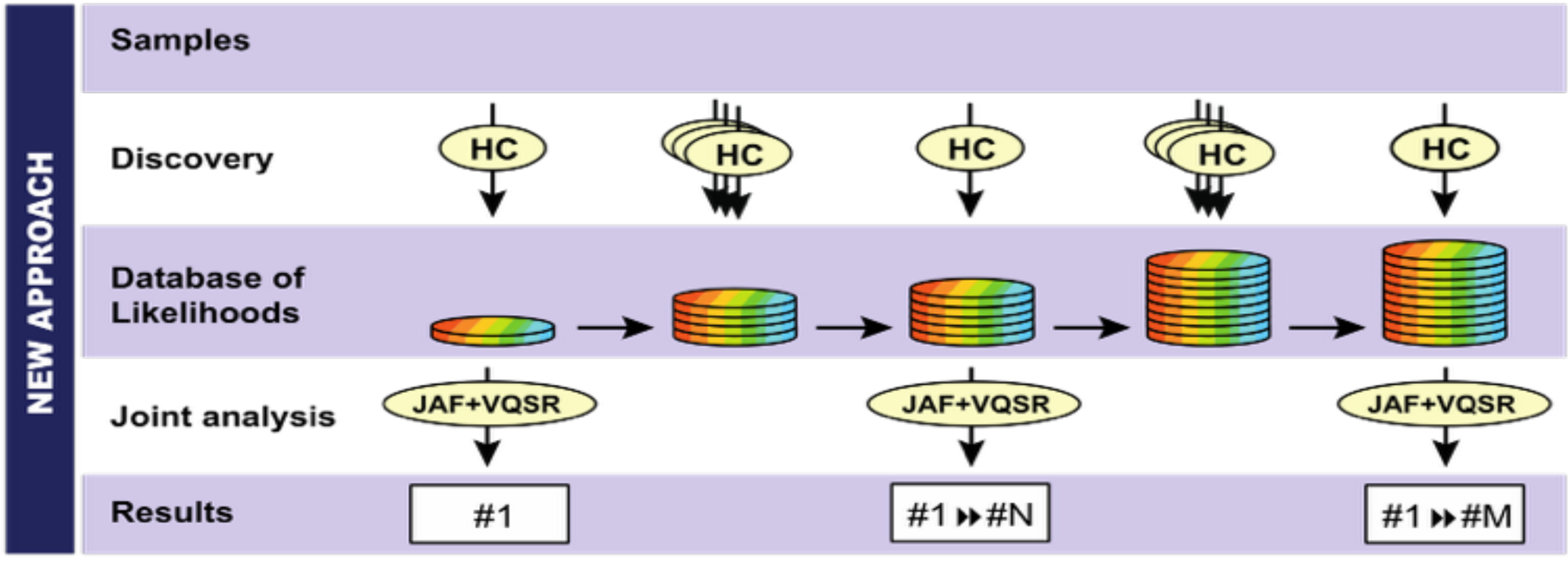
Technology	Hardware	Runtime	Improvement
GPU	NVidia Tesla K40	70	154x
GPU	NVidia GeForce GTX Titan	80	135x
GPU	NVidia GeForce GTX 480	190	56x
GPU	NVidia GeForce GTX 680	274	40x
GPU	NVidia GeForce GTX 670	288	38x
AVX	Intel Xeon 1-core	309	35x
FPGA	Convey Computers HC2	834	13x
-	C++ (baseline)	1,267	9x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

The reference model enables incremental calling



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

But the joint analysis needs better infrastructure to become a trivial step



currently joint analysis has to be done hierarchically and takes approximately the same amount of time as the discovery step

Future challenges to scale up the variant calling pipeline

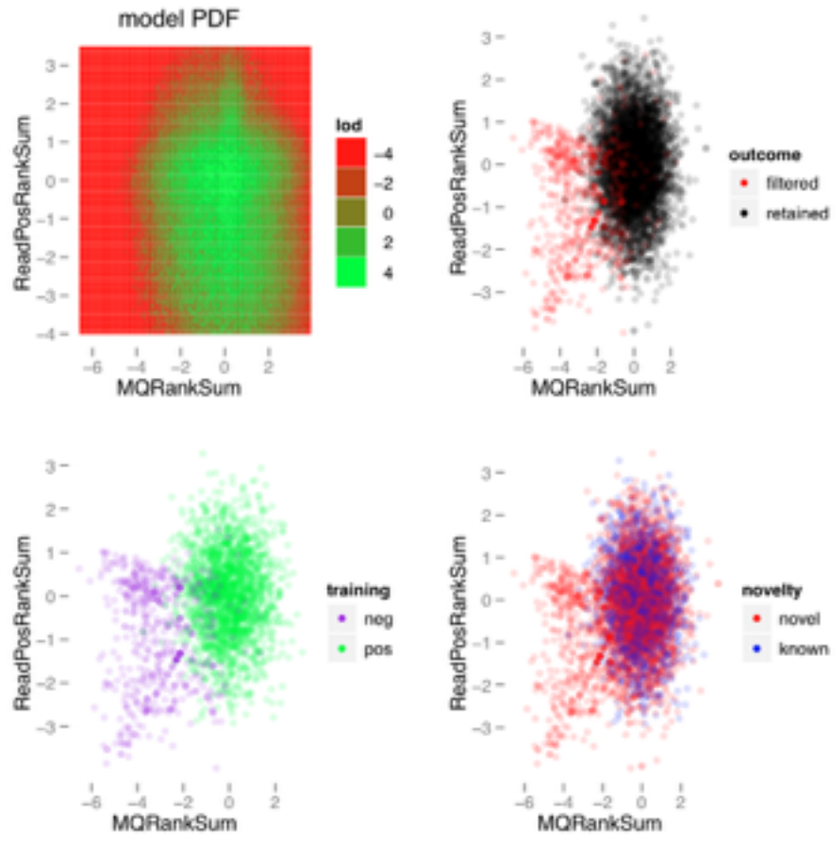
- Improve runtime of the Haplotype Caller
  (e.g. faster assembly, better prediction of haplotypes)

- Expand on our experience with heterogeneous compute.
  (e.g. specialize sections of the code for GPU/AVX)

- Improve the performance of the joint analysis by building dedicated high performance tools.

- Improve statistical filtering of indels using non gaussian based methods.
  (e.g. random forests or other machine learning techniques)

- Revisit all statistical annotations (they behave differently at this scale) identify differences and implement new alternatives.



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

Rare Variant Association Study (RVAS)

vs

Common Variant Association Study (CVAS)

vs

Technical challenge
data integration

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.

Post-processing analysis pipeline standardization and scaling

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

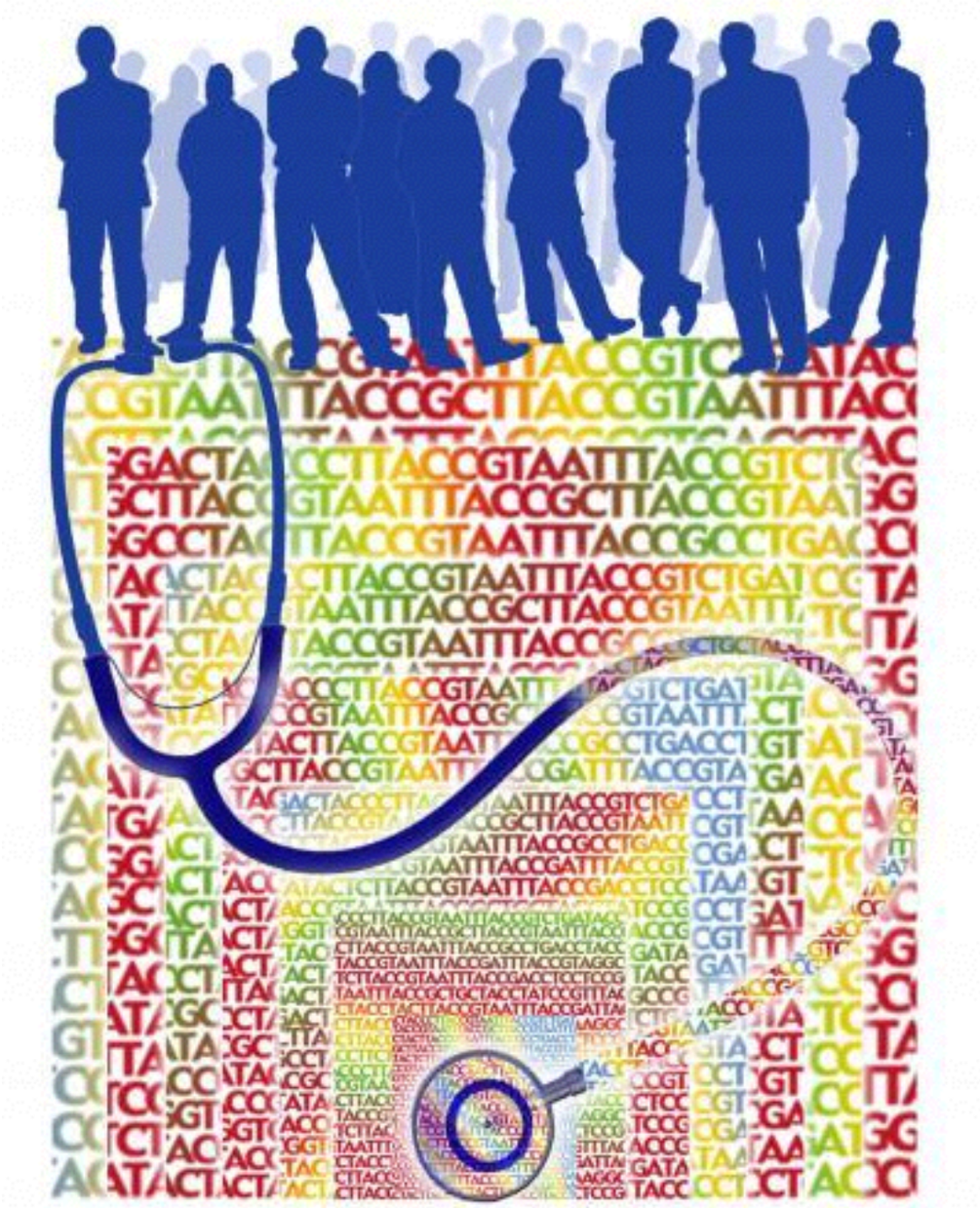
Future challenges to enable large scale post-processing analysis

- Build a generic *variant database* that would enable large scale data analysis of many thousands of patients and controls

- Establish best practices and revisit the available tools to create a uniform framework where developers can create reusable and maintainable tools.

- Standardize workflows for common analyses such as RVAS and CVAS.

- Establish best practices for use of such pipelines in the clinic.



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

Rare Variant Association Study (RVAS)

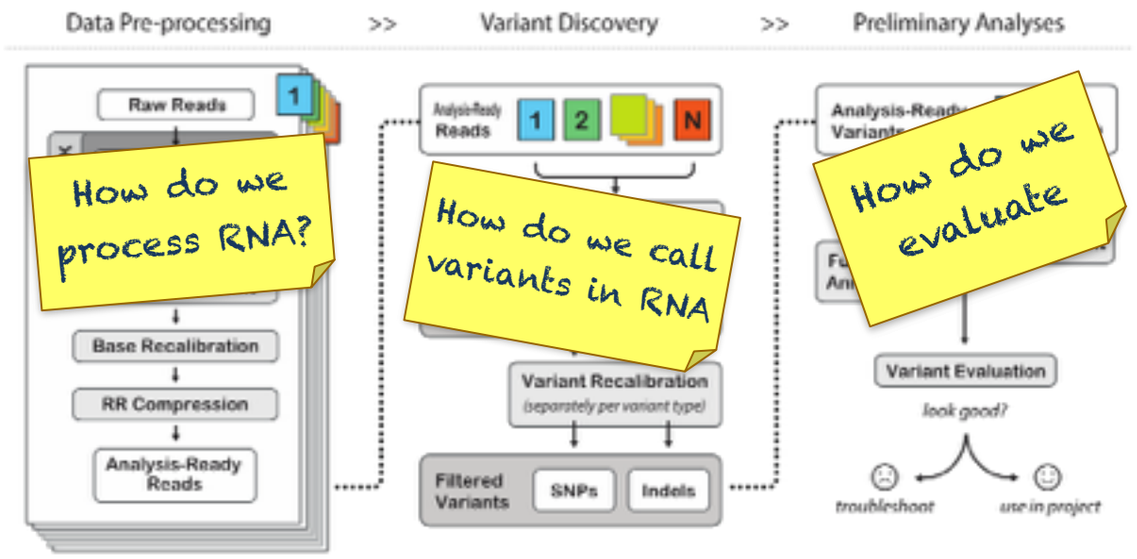
vs

Common Variant Association Study (CVAS)

vs

Technical challenges
RNA-seq, Structural Variation and Cancer

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

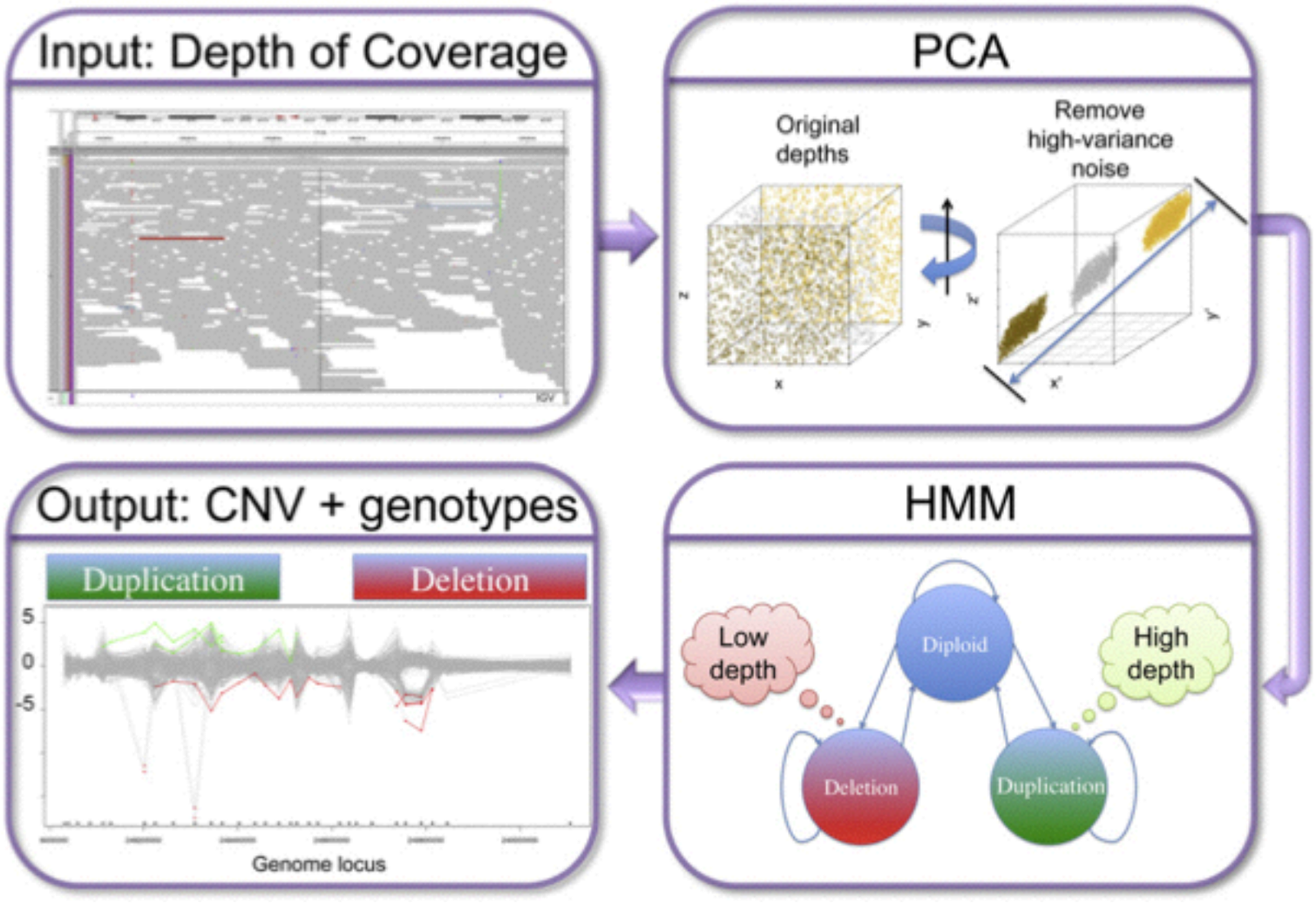


Milestone 1:
provide a best practice processing and analysis pipeline for RNA-seq with existing tools and minor modifications to the GATK. **released in GATK 3.0**

Milestone 2:
update GATK tools to make best use of RNA data including contrastive variant calling with DNA. Improve accuracy and overall performance.

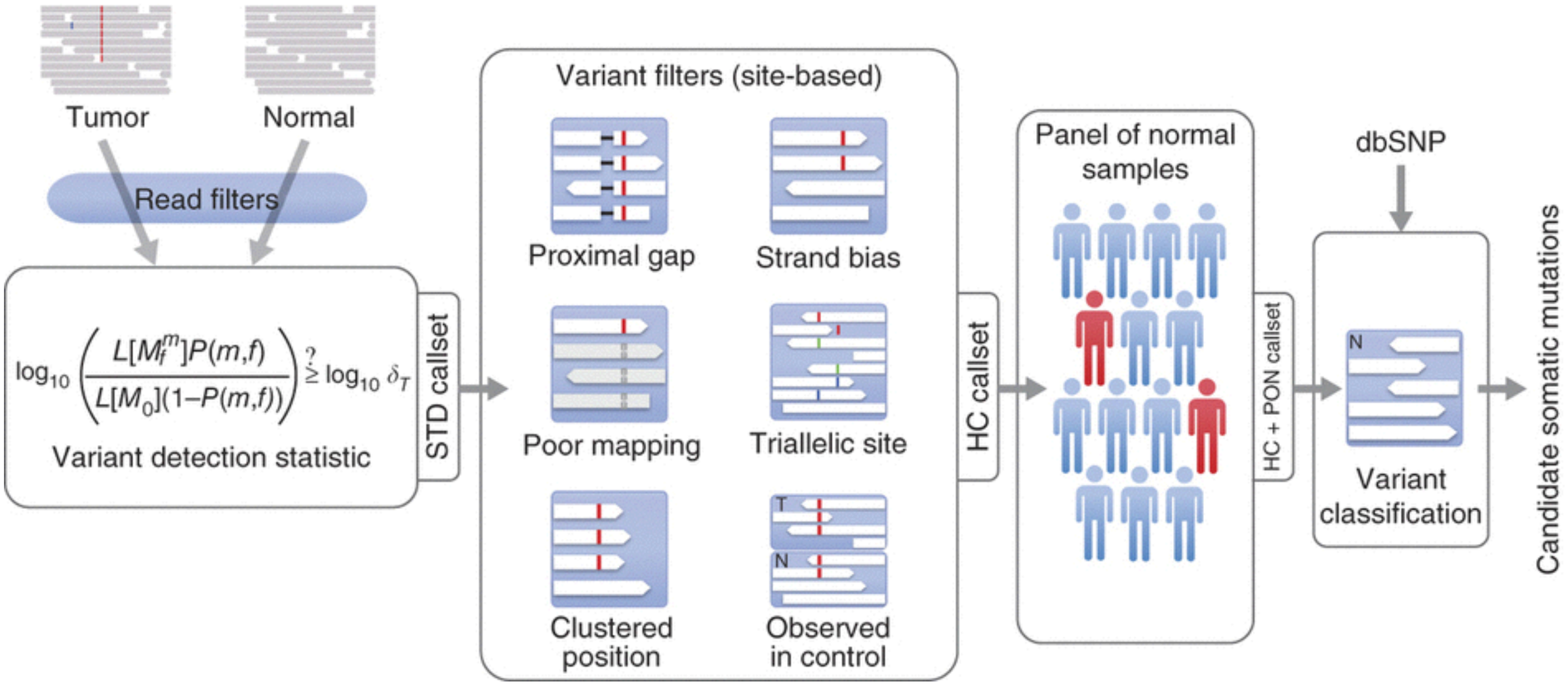
Milestone 3:
Build new tools to address the needs of the RNA-seq world. (e.g. Heterogeneous expression, RNA editing, expression levels, alternate splicing…)

Structural variation is an important missing piece in the analysis pipeline



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

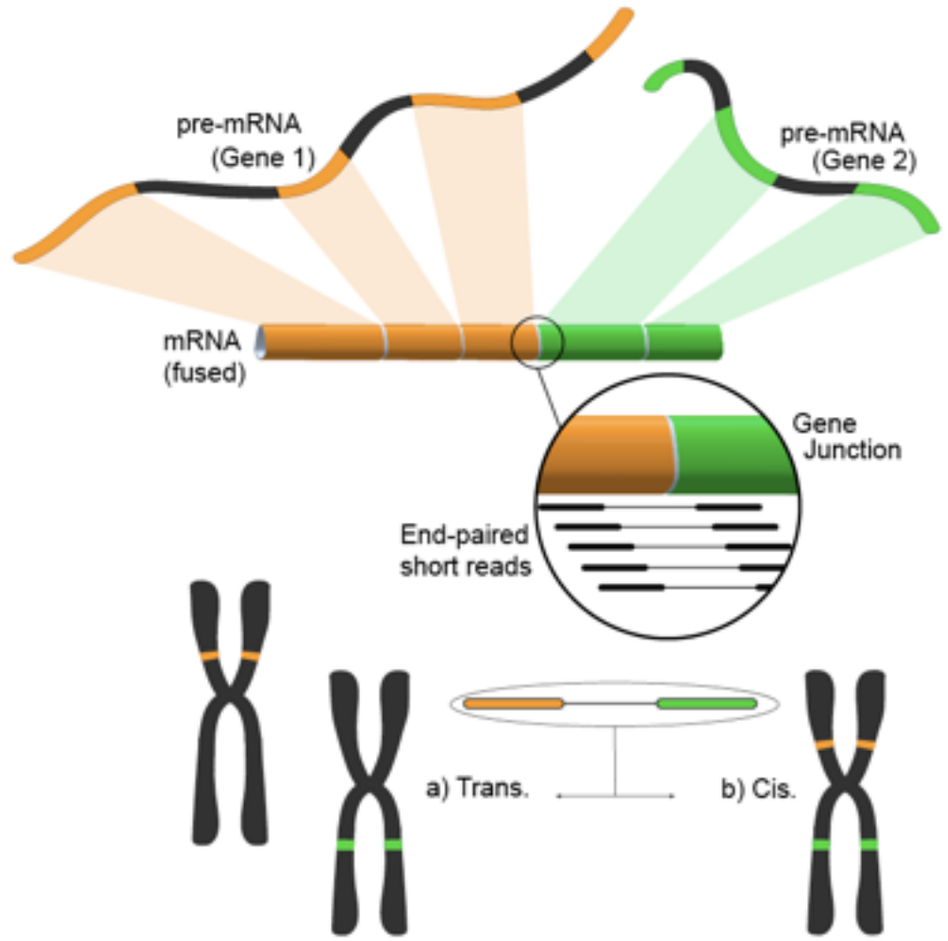
Cancer needs the same rigorous standardization from study design to data processing and analysis



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

Future challenges to scale up the variant calling pipeline

- RNA-seq has great potential to further complement discoveries from DNA-seq.
- The world of clinical applications has yet to explore the plethora of information provided by the RNA — we need tools!
- Structural variations are known to play an important role. Tools should be integrated into the standard research pipeline for CVAS and RVAS.
- Cancer still has many open questions from the processing side to the specialized downstream analysis.



There is a lot to do… but it's definitely worth it.

- samples must be consistently pre-processed 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!)

- Data needs to be integrated to vastly accelerate discovery. New technology needs to be built in order to enable the creation of methods and tools that leverage the aggregated data.

- RNA-seq and other analyses need to be integrated and standardized for scientists and clinicians to understand the whole picture.

Rare Variant Association Study (RVAS)

Common Variant Association Study (CVAS)

“Thank you!”

–Mauricio Carneiro