The challenges of analyzing hundreds of thousands genomes

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

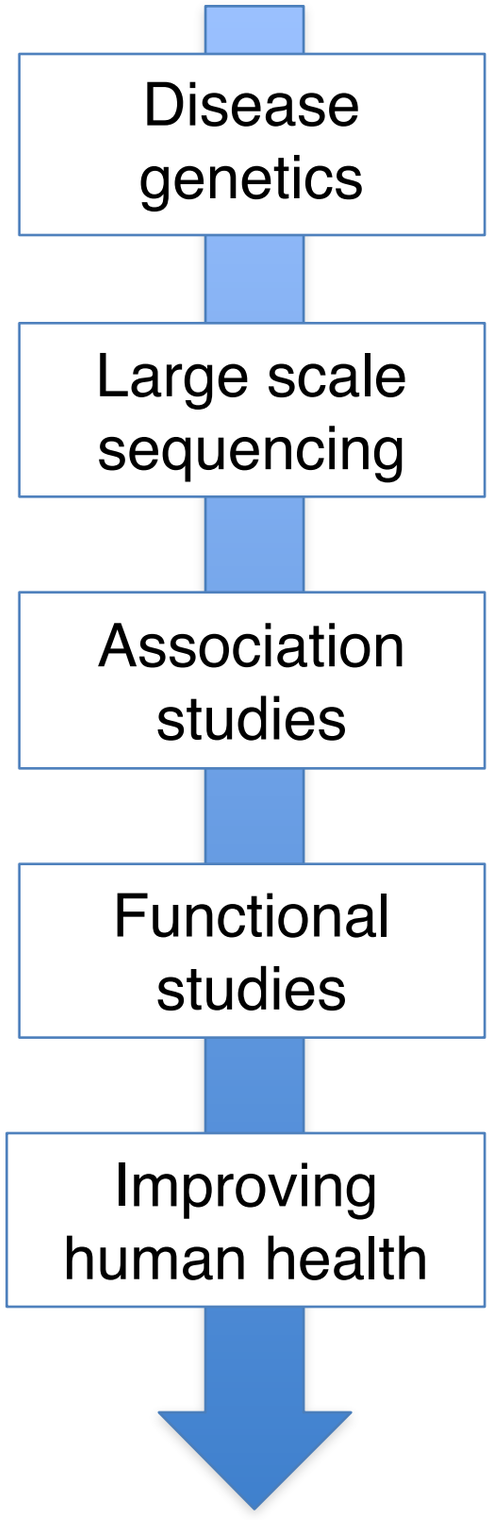
What is the BROAD INSTITUTE ?

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.

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

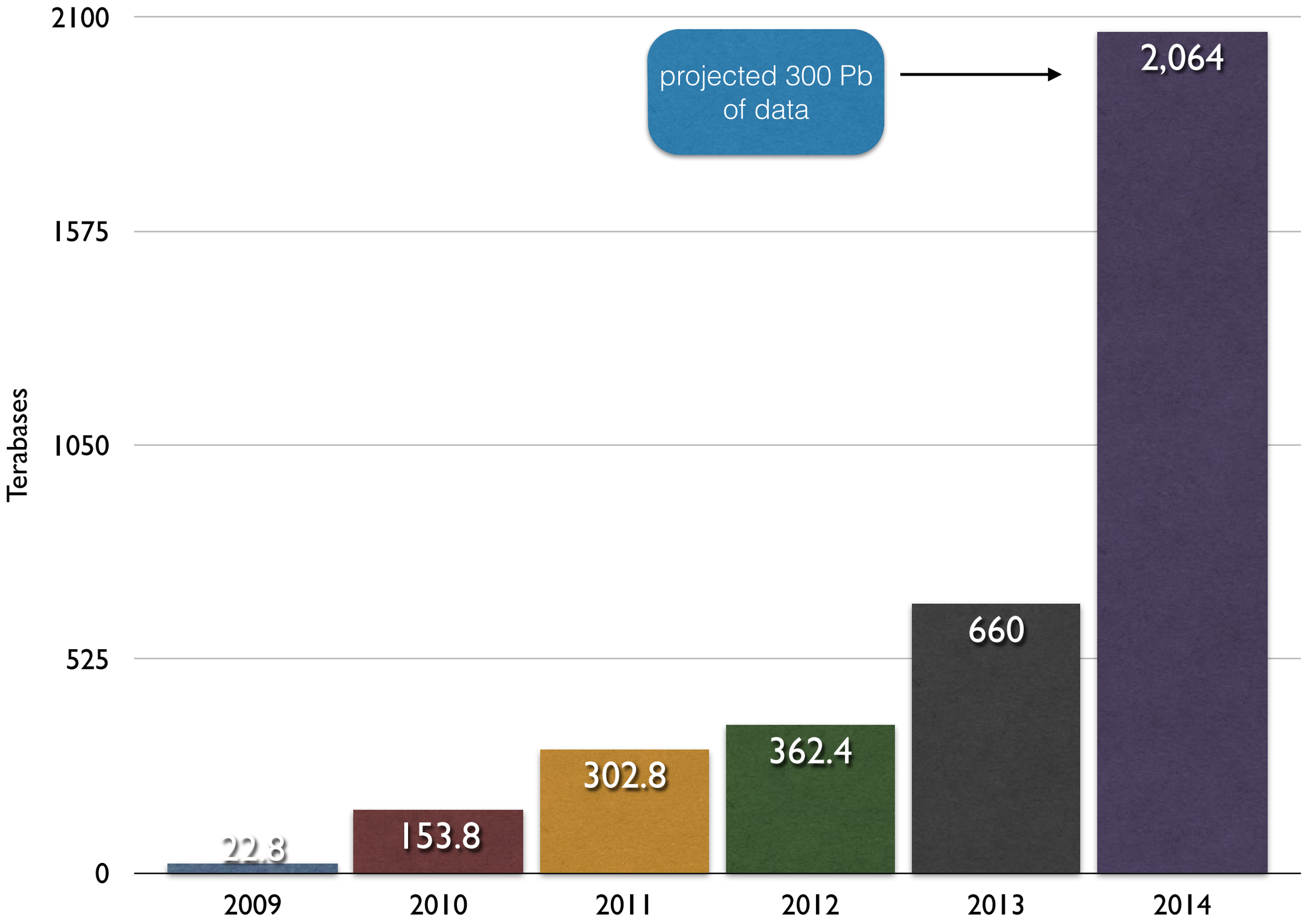
8,189
RNA

9,788
16S

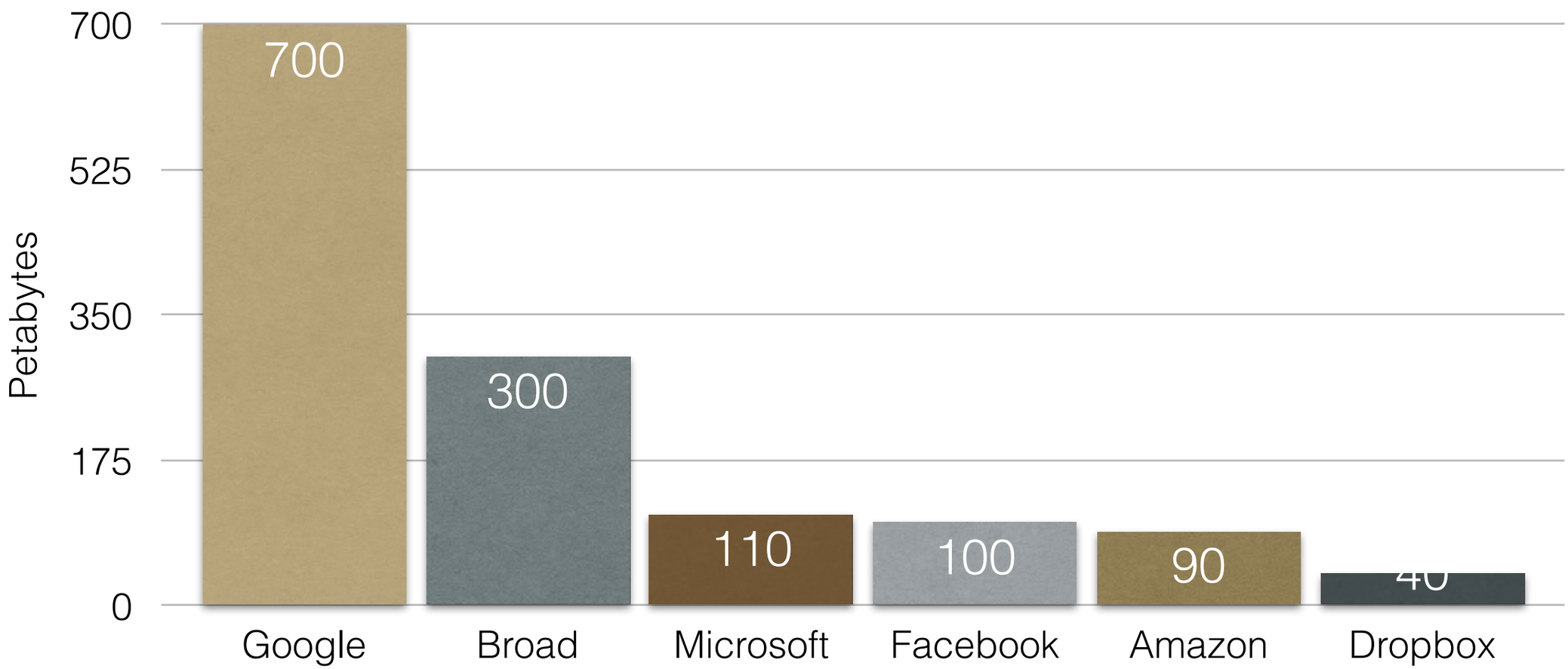
47,764
arrays

228
cell lines

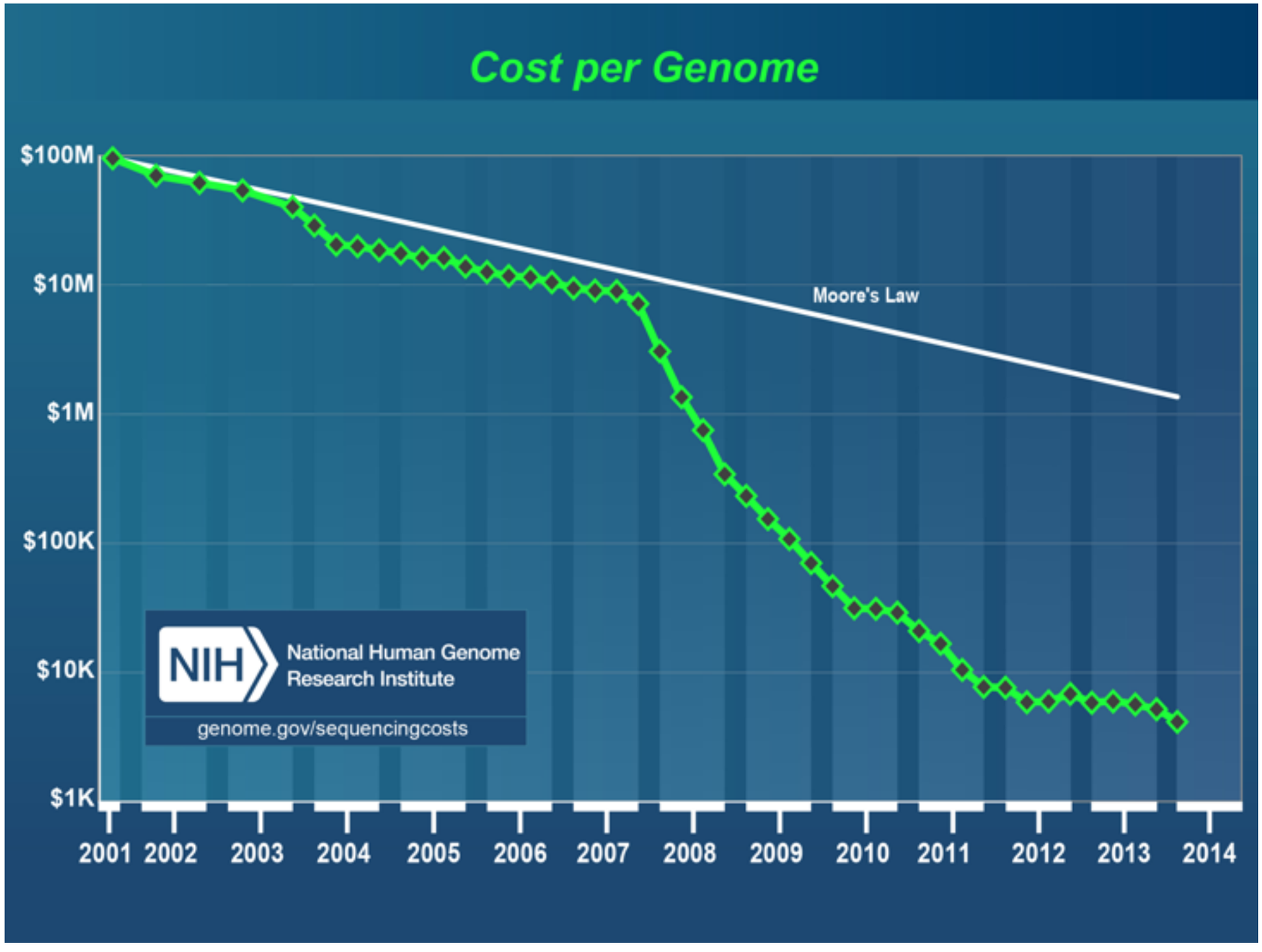
Terabases of Data Produced by Year



We produce as much data as the big cloud providers

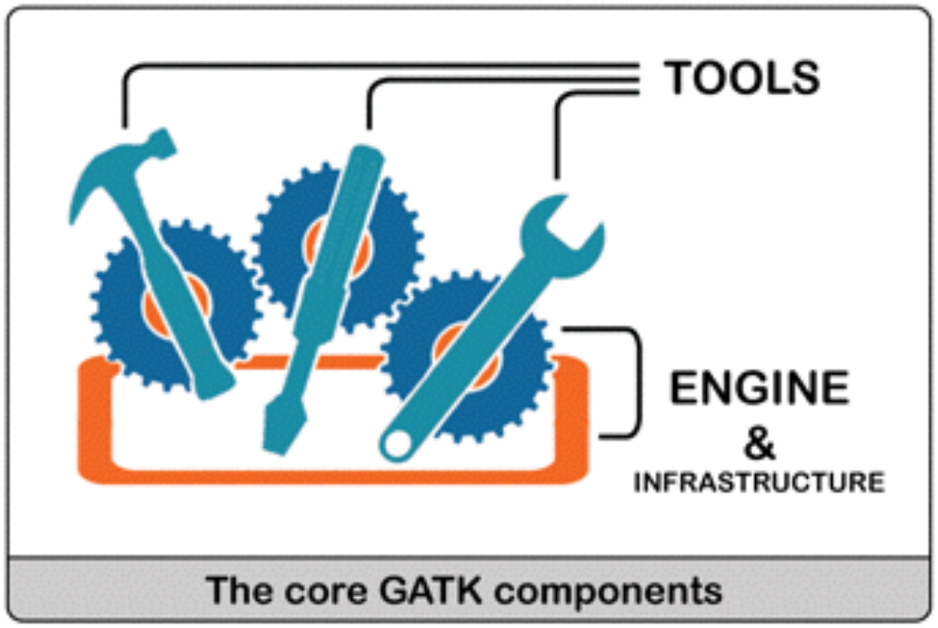


and these numbers will continue to grow faster than Moore's law



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

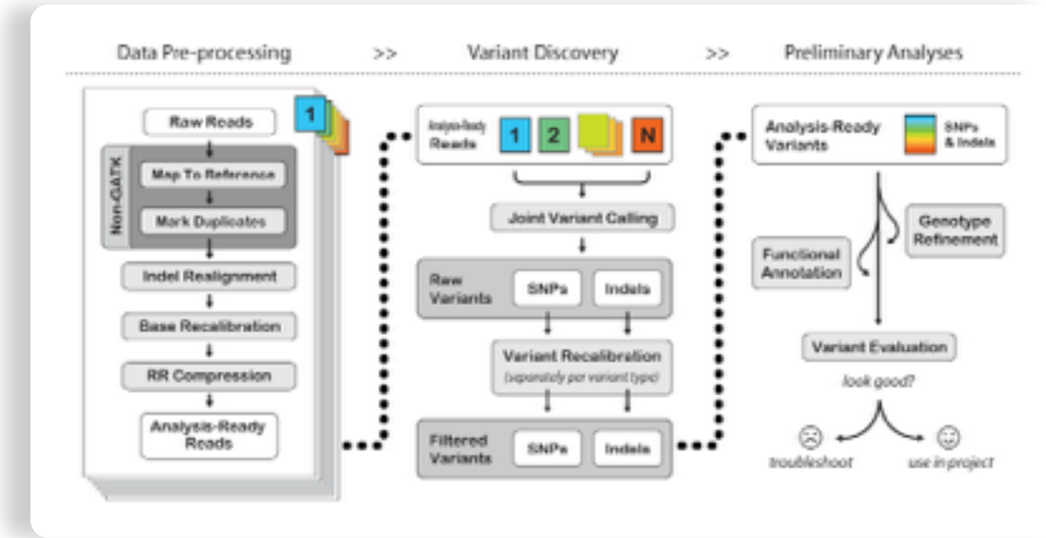
Toolkit & framework packages



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

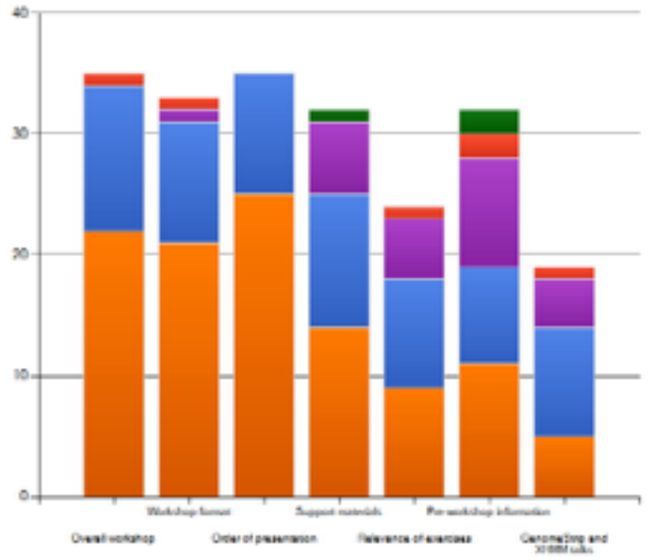
Workshop series educates local and worldwide audiences

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

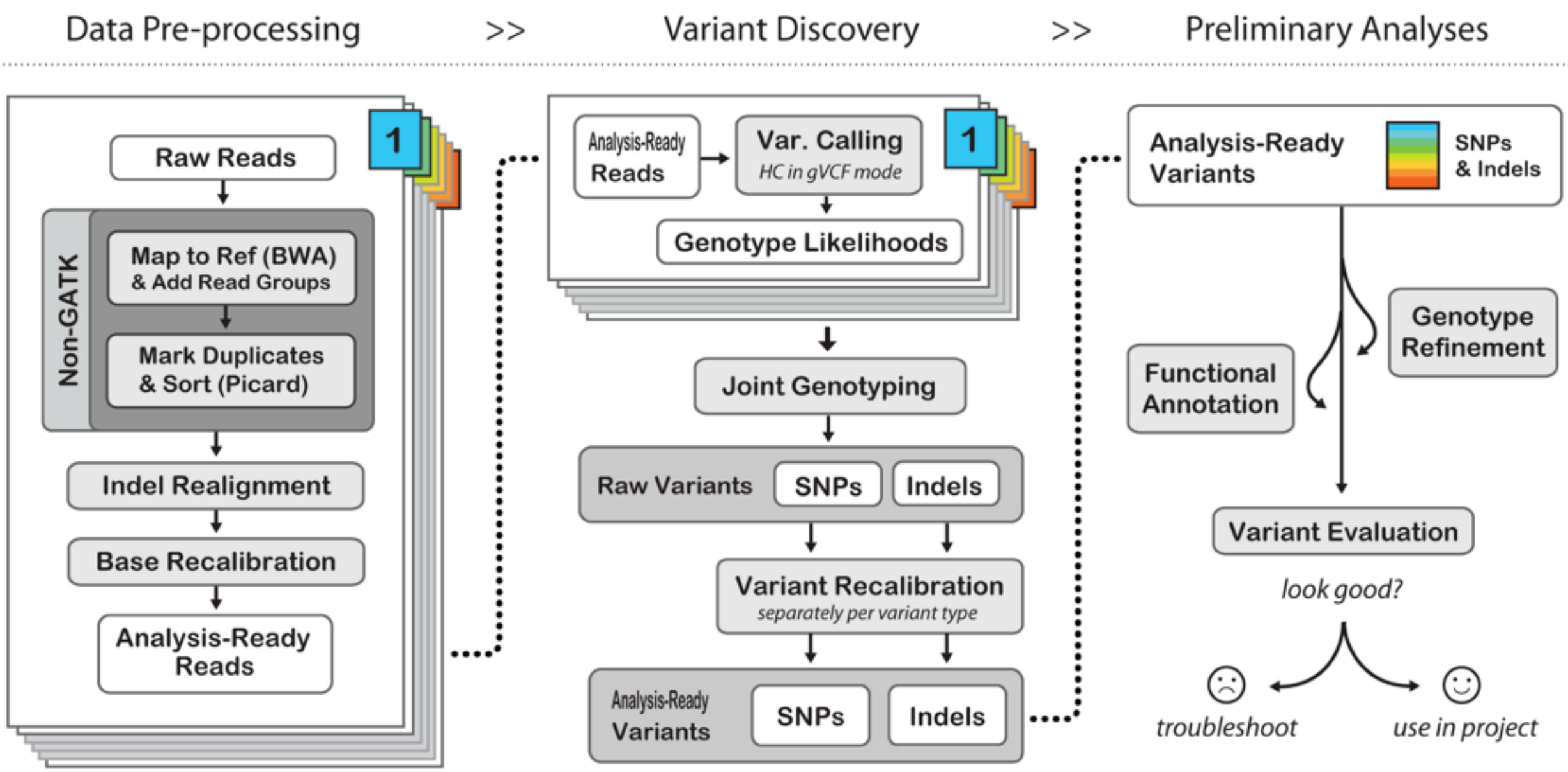
BroadE: Overview of GATK & best practices

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 ...

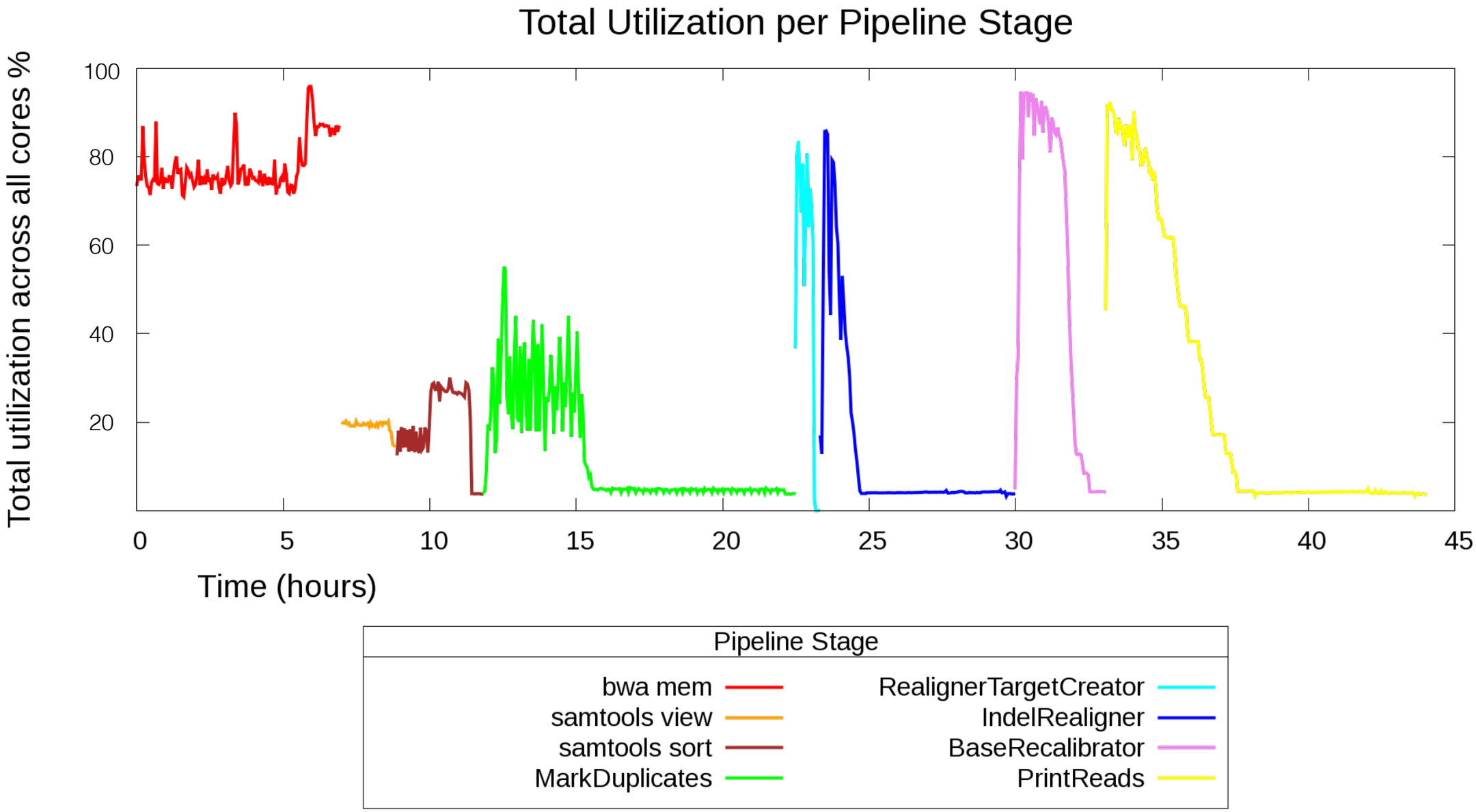
NEW HD

We have defined the best practices for sequencing data processing



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

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**

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

Rare Variant Association Study (RVAS)

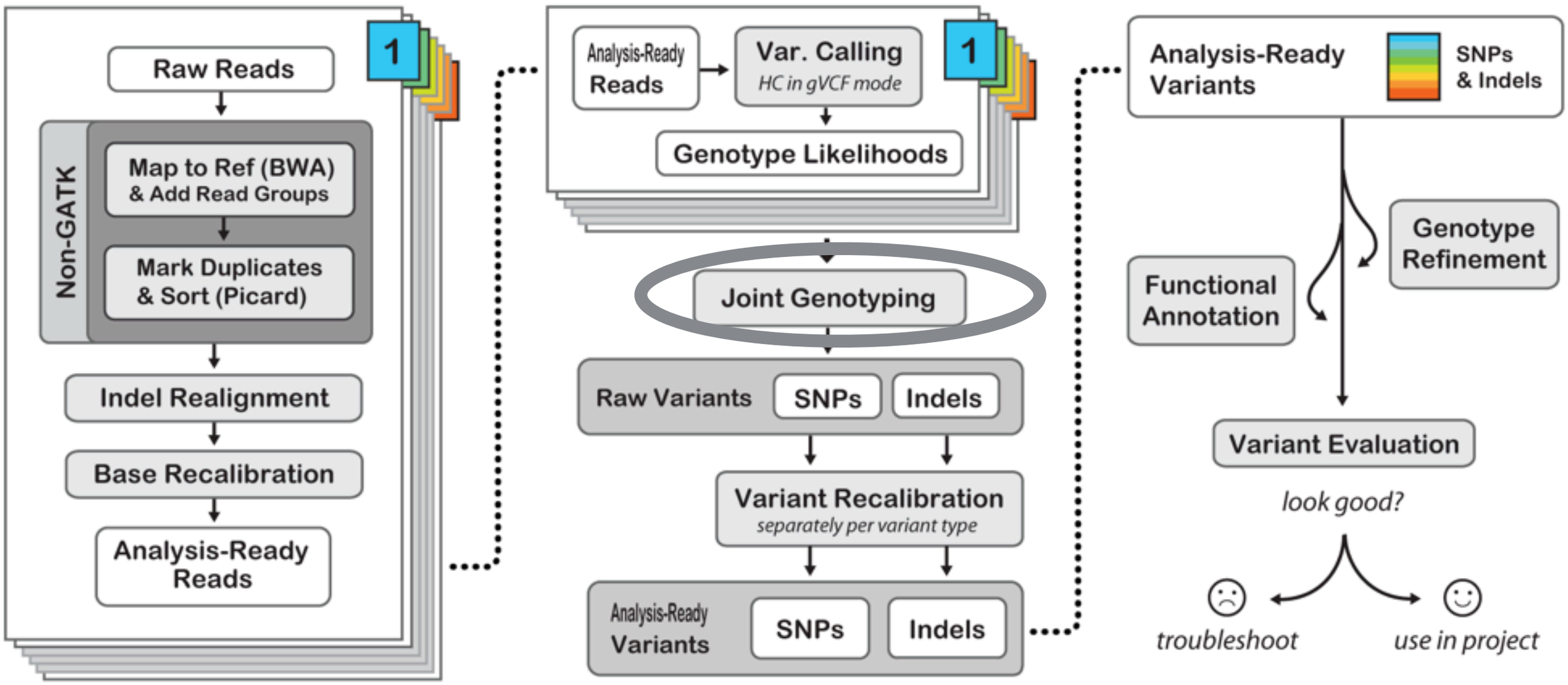
vs

Common Variant Association Study (CVAS)

vs

Technical challenge
all samples must be *jointly* called

Joint genotyping is an important step in Variant Discovery



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

The ideal database for RVAS and CVAS studies would be a complete matrix

All case and control samples

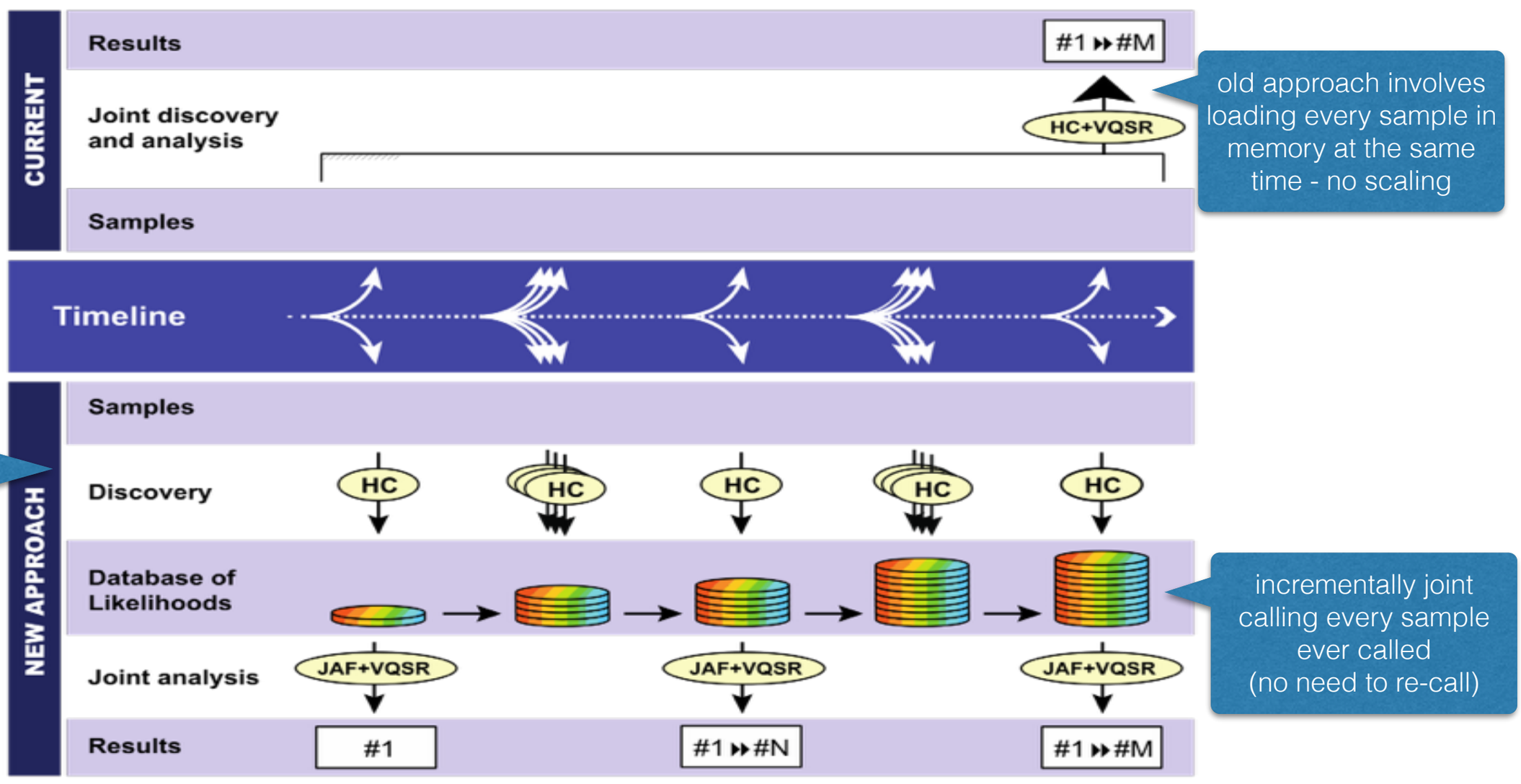
~3M variants

	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

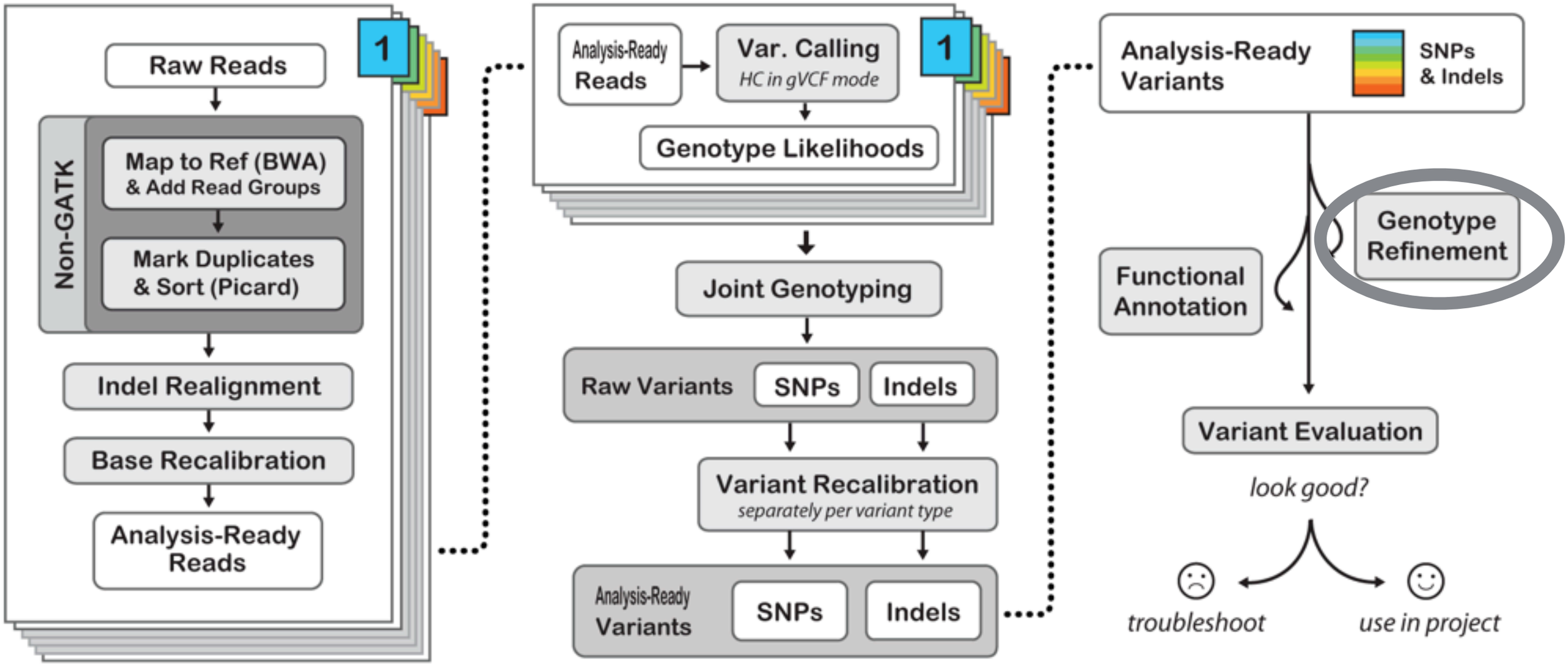
Likelihoods:
A/B/C phred-scaled probability of hom (A), het (B), hom-alt (C) genotypes given NGS data

The reference model enables incremental calling



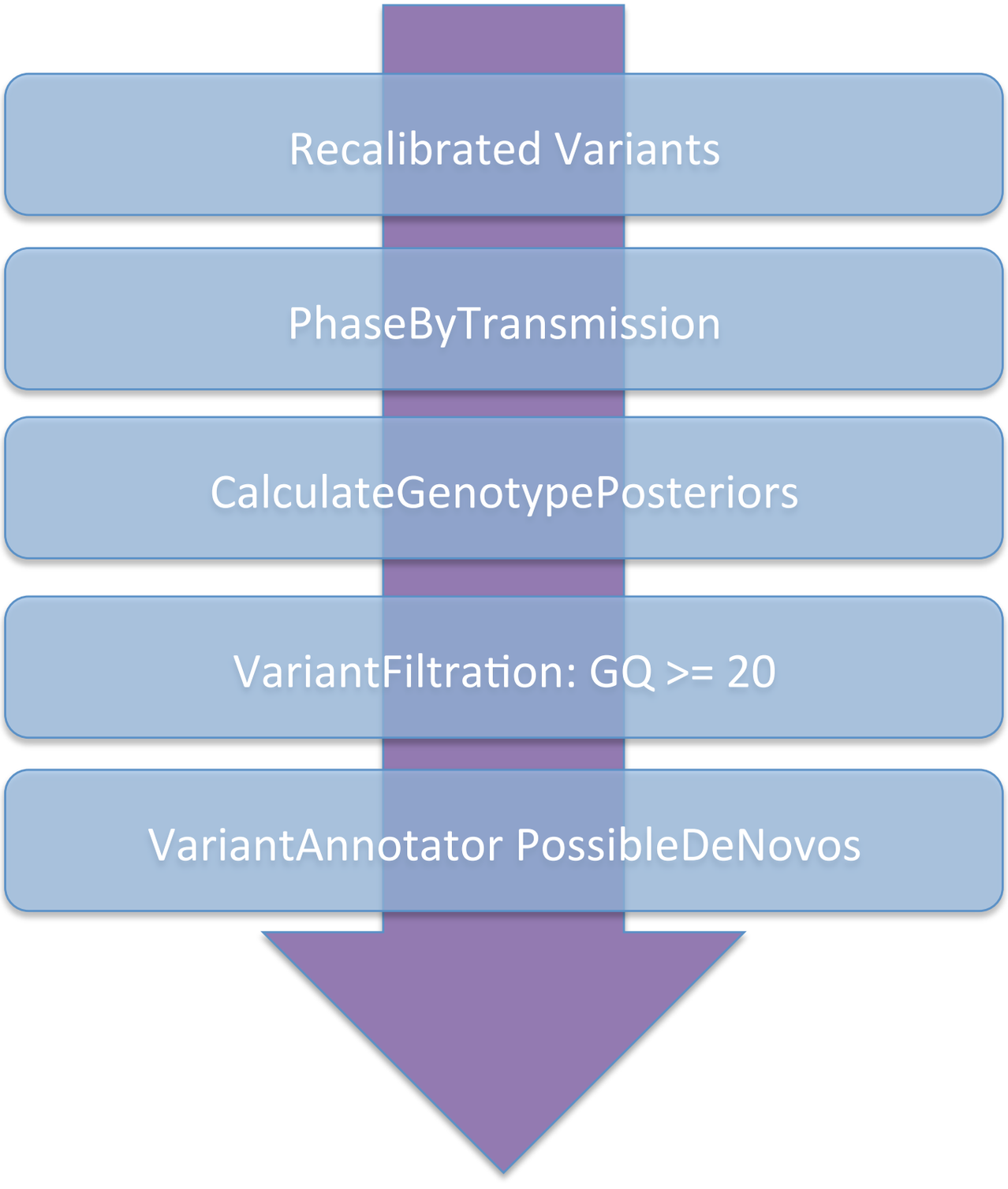
by separating discovery from joint analysis, we can now jointly call any arbitrary number of samples

Joint genotyping is an important step in Variant Discovery



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

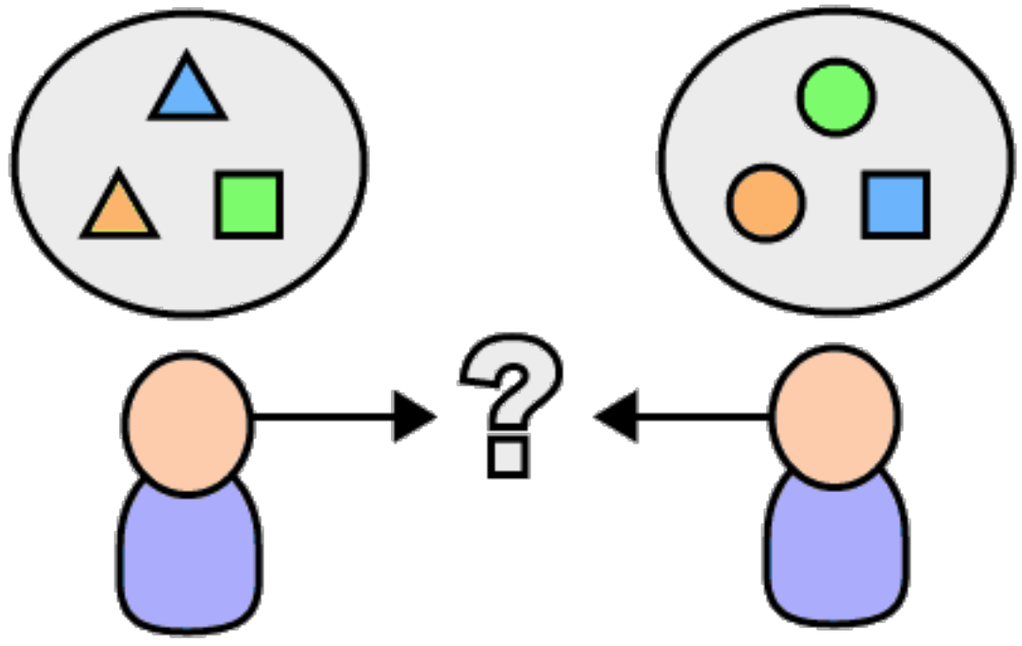
New Genotype Refinement Pipeline



- 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

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).



To fully understand **one** genome we need **hundreds 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.

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

Rare Variant Association Study (RVAS)

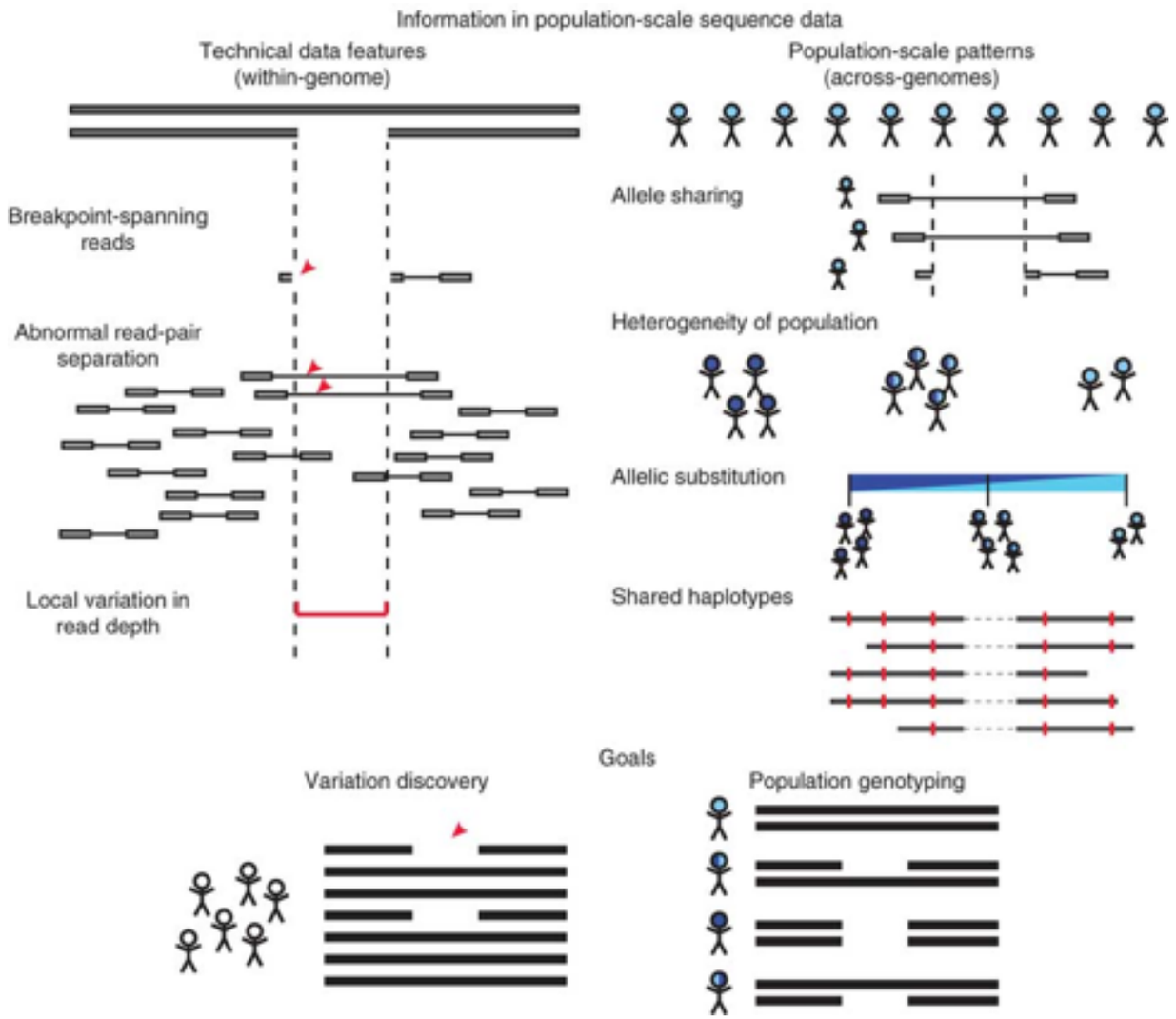
vs

Common Variant Association Study (CVAS)

vs

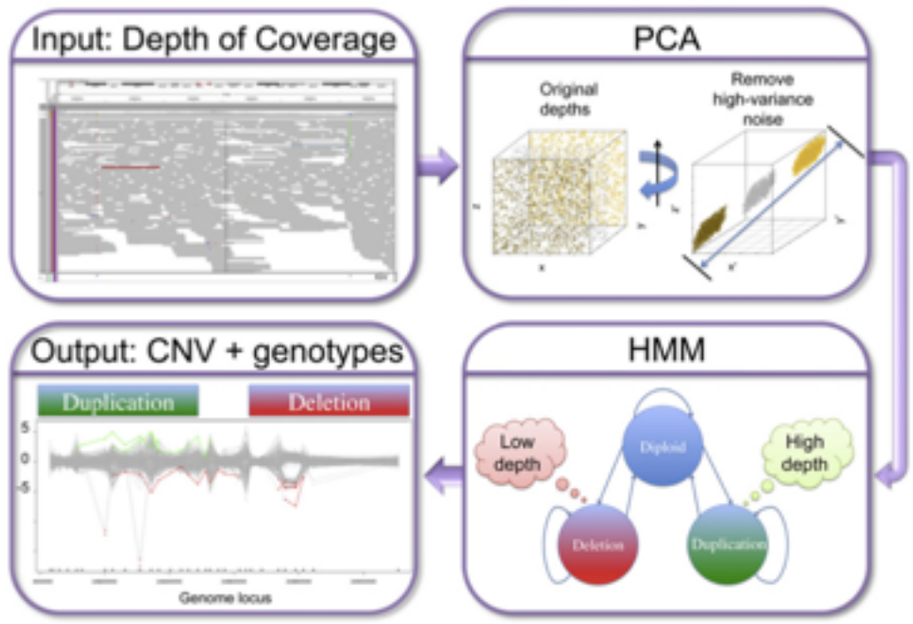
Technical challenges
RNA-seq, Structural Variation and Cancer

Structural variation is an important missing piece in the standard analysis



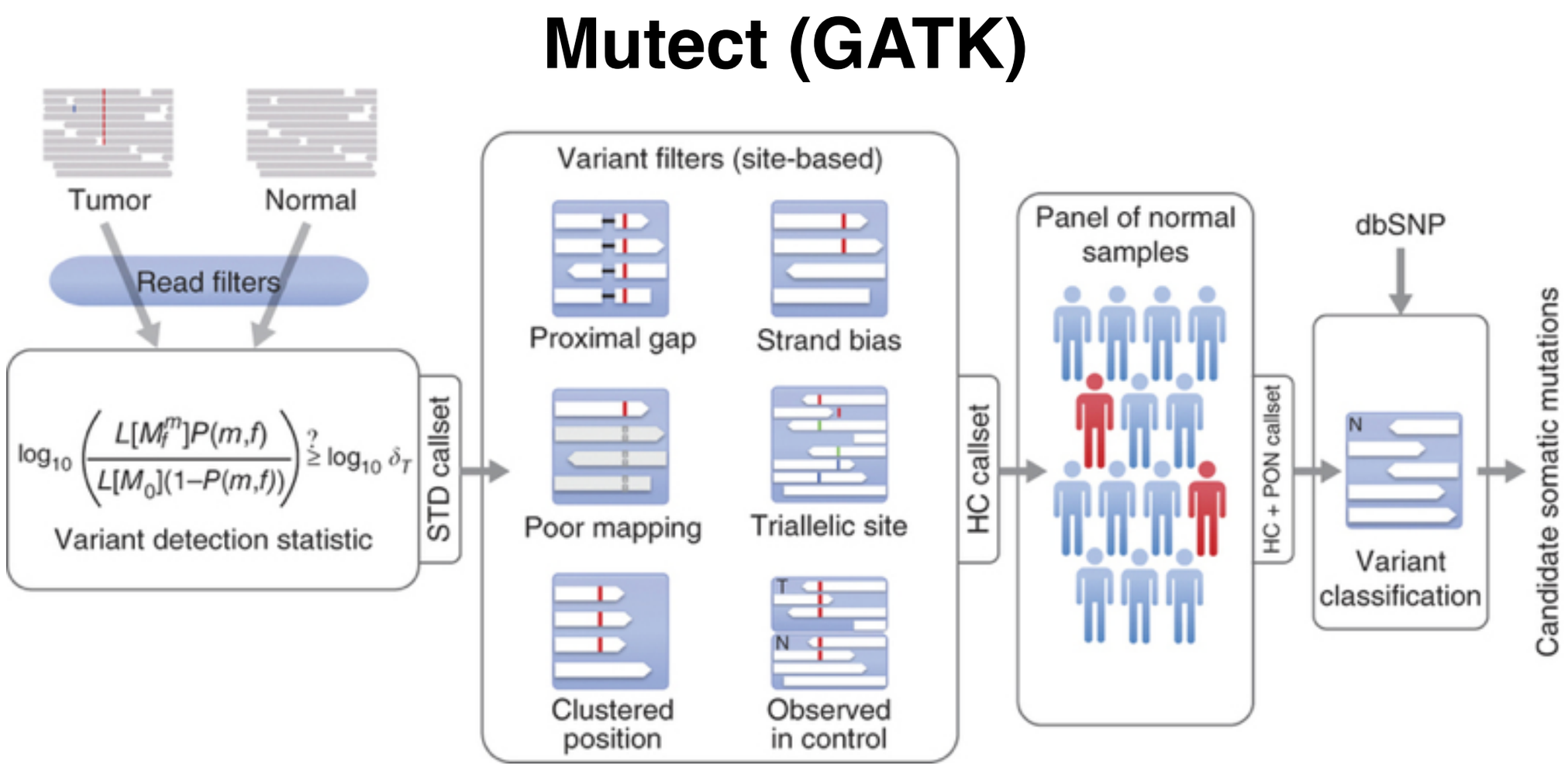
Handsaker, R et al. *Nature Genetics* (2011)

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



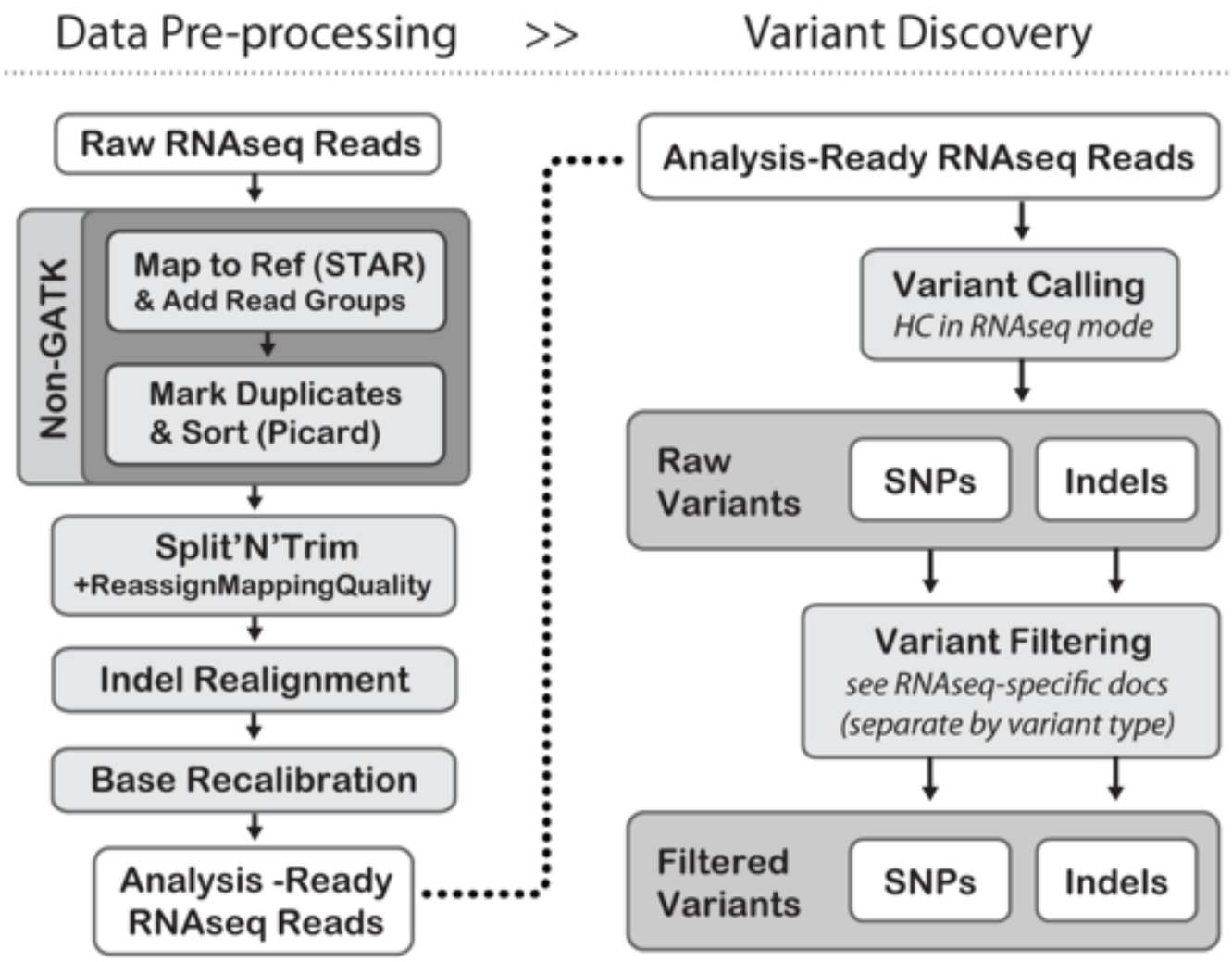
Cibulskis, K et al. *Nature Biotechnology* (2013)

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.

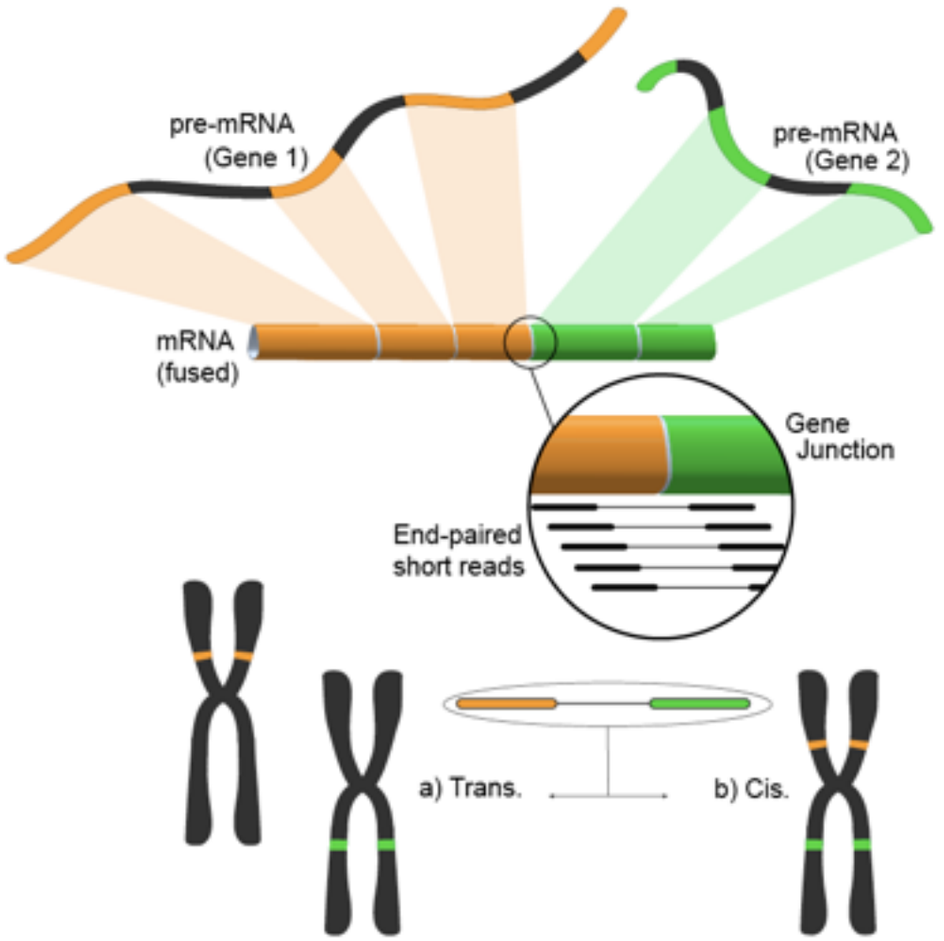
Milestone 2:
Build new tools to address the needs



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