Sequencing data processing in modern computers

Mauricio Carneiro

carneiro@broadinstitute.org

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

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

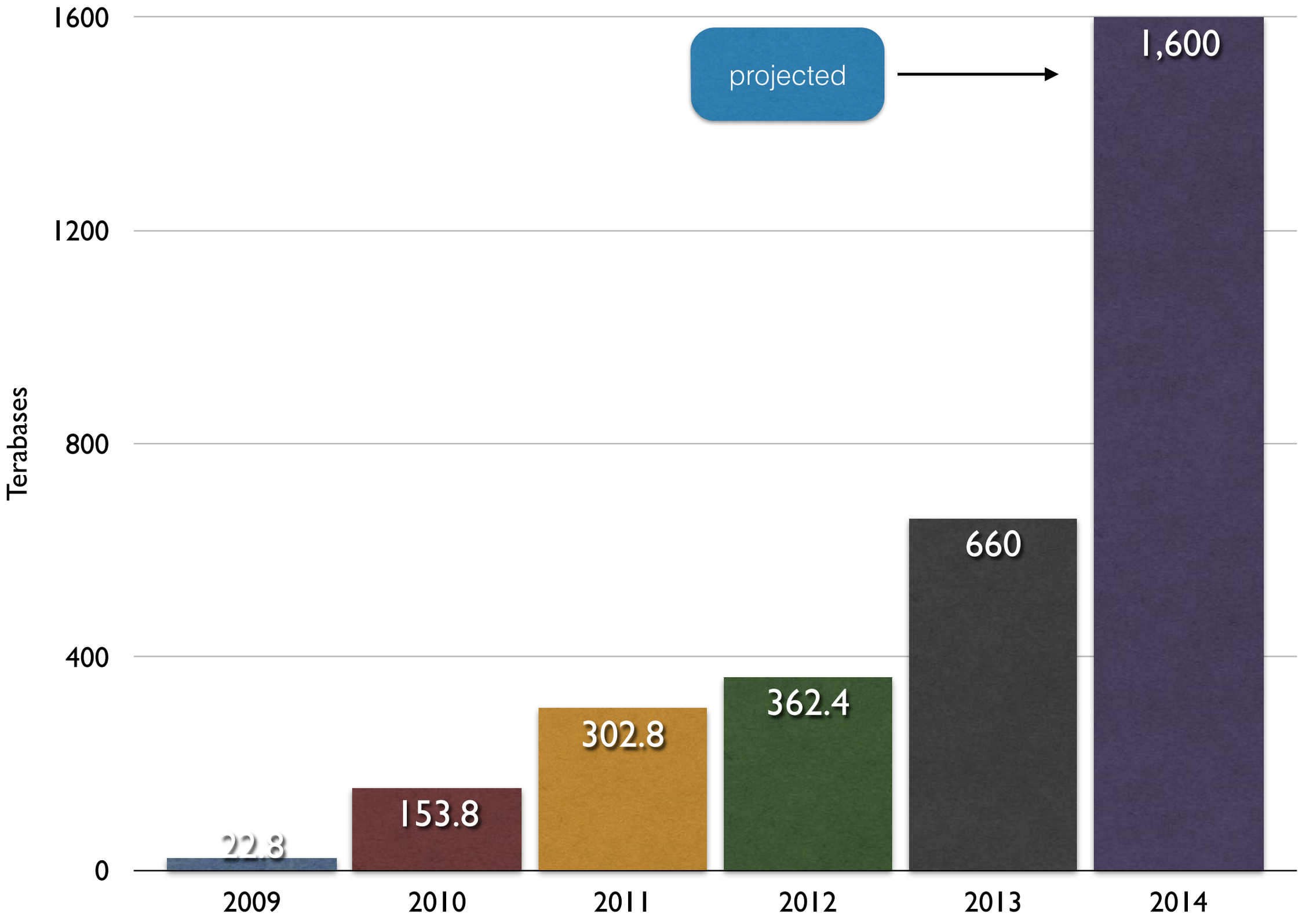
8,189
RNA

9,788
16S

47,764
arrays

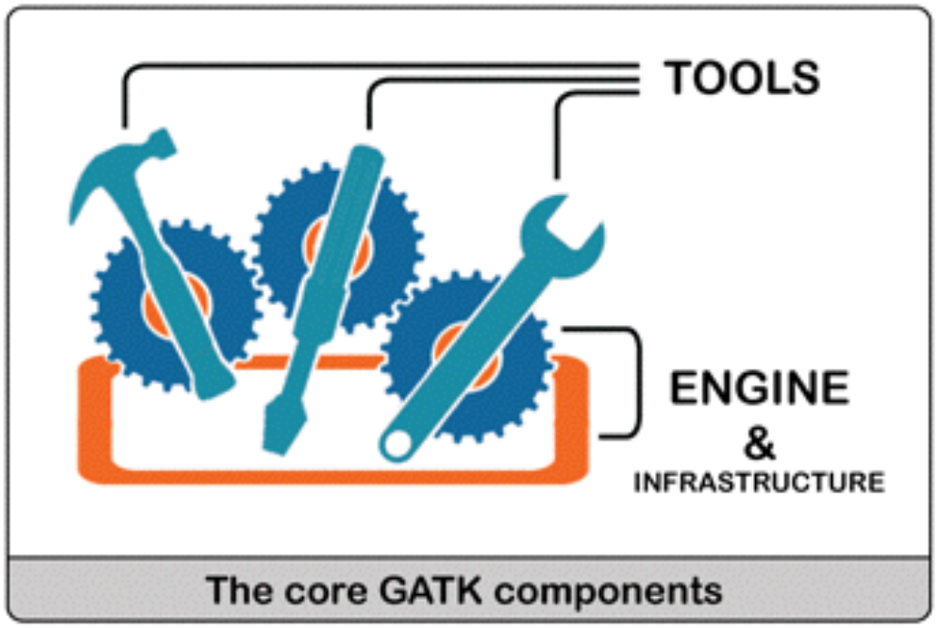
228
cell lines

Terabases of Data Produced by Year



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

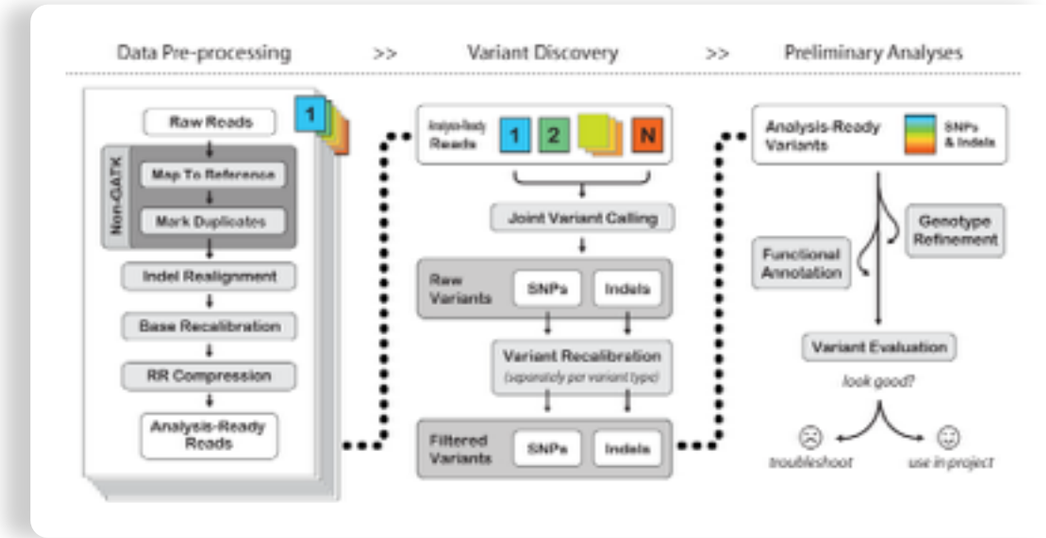
Toolkit & framework packages



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

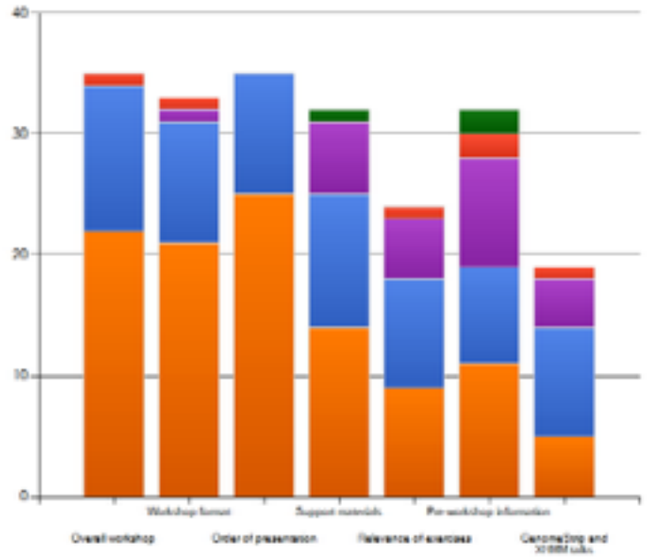
Workshop series educates local and worldwide audiences

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

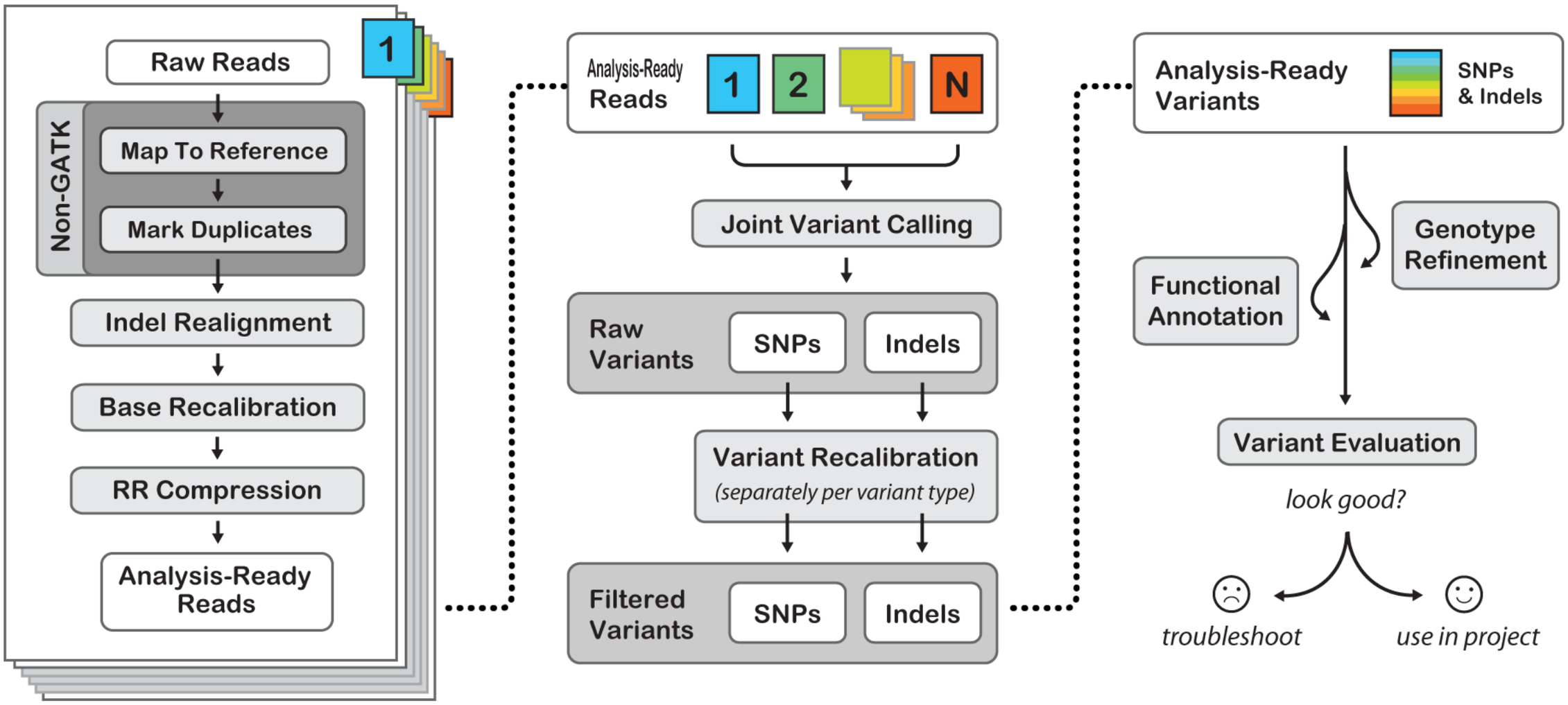
BroadE: Overview of GATK & best practices

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 must be *jointly* called

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

All case and control samples

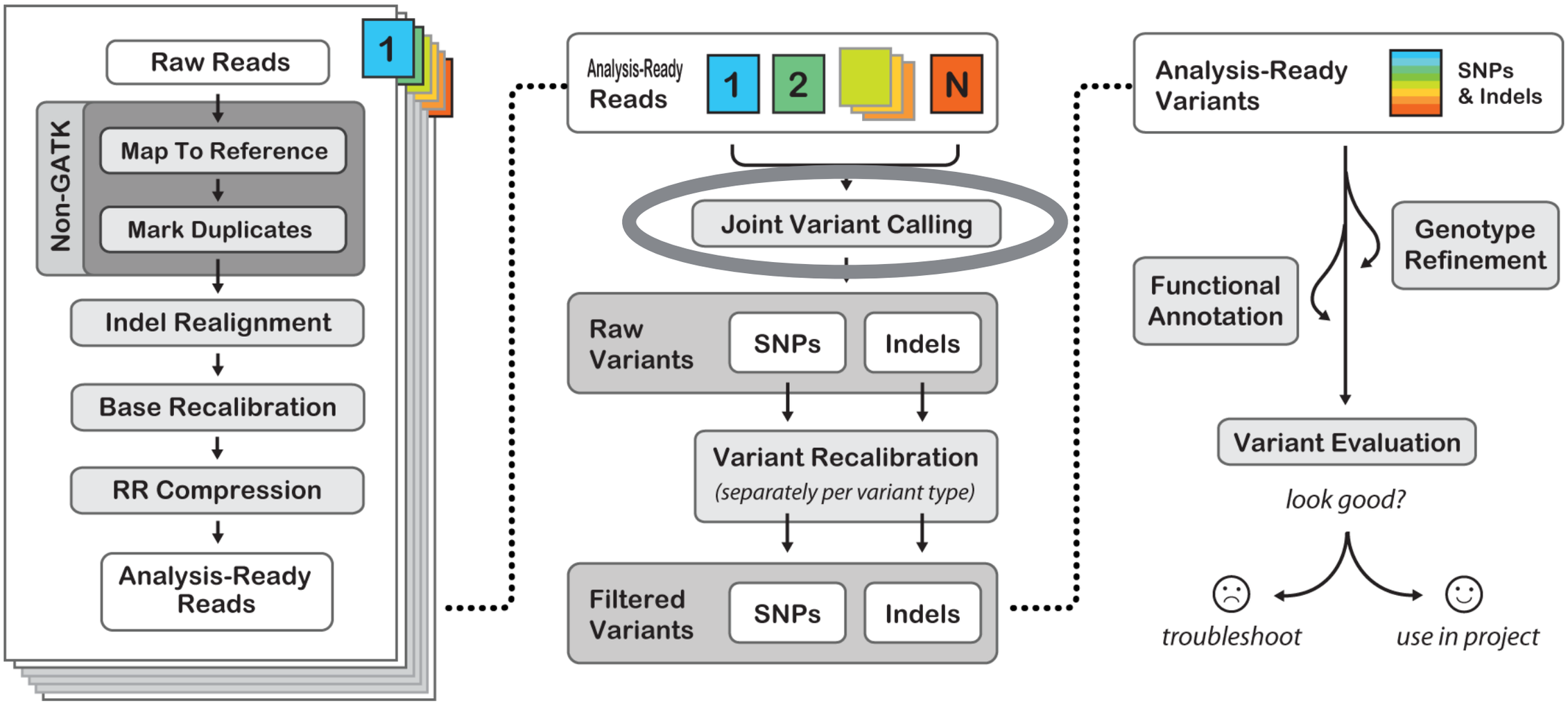
~3M variants

	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

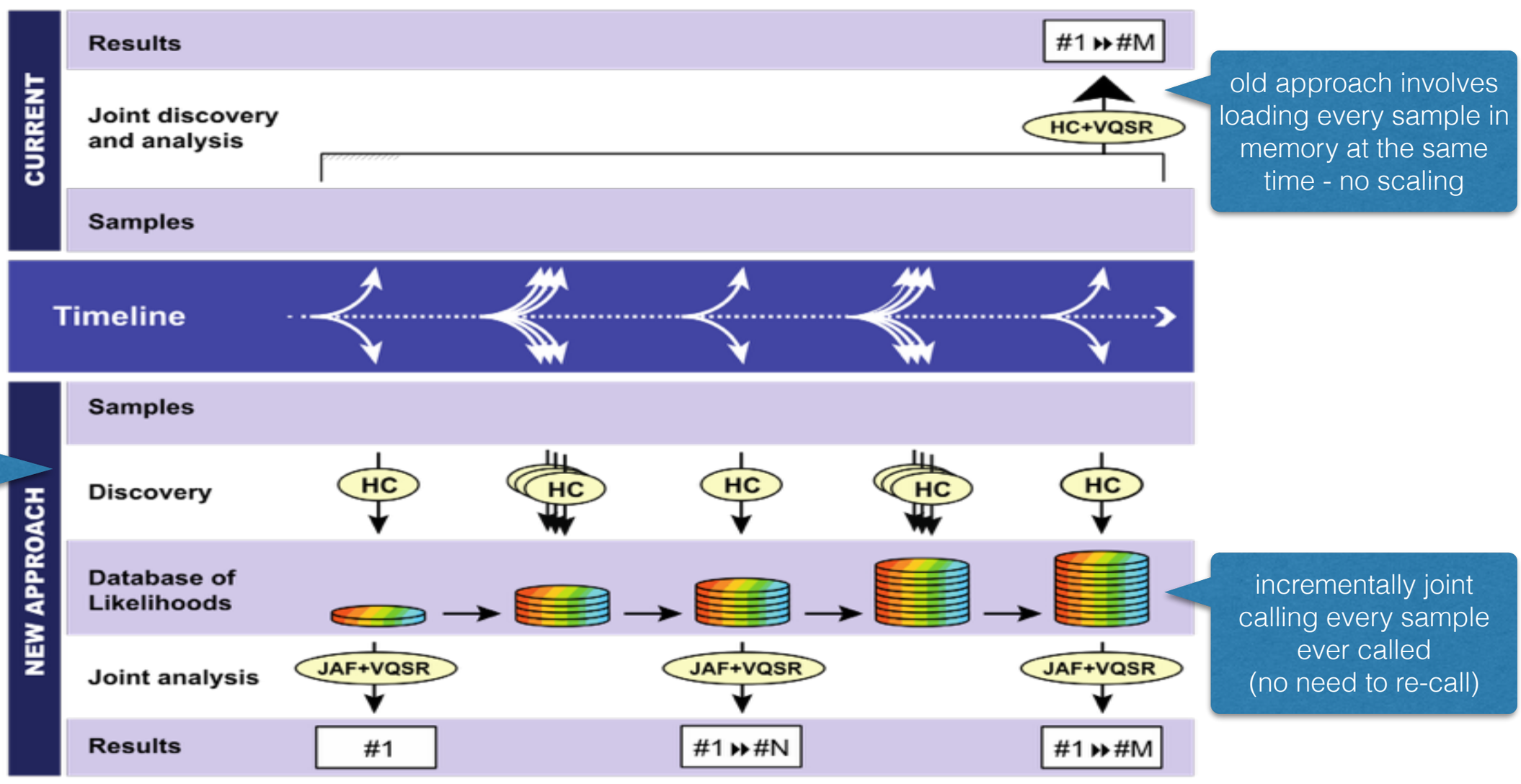
Likelihoods:
A/B/C phred-scaled probability of hom (A), het (B), hom-alt (C) genotypes given NGS data

Joint calling is an important step in Variant Discovery



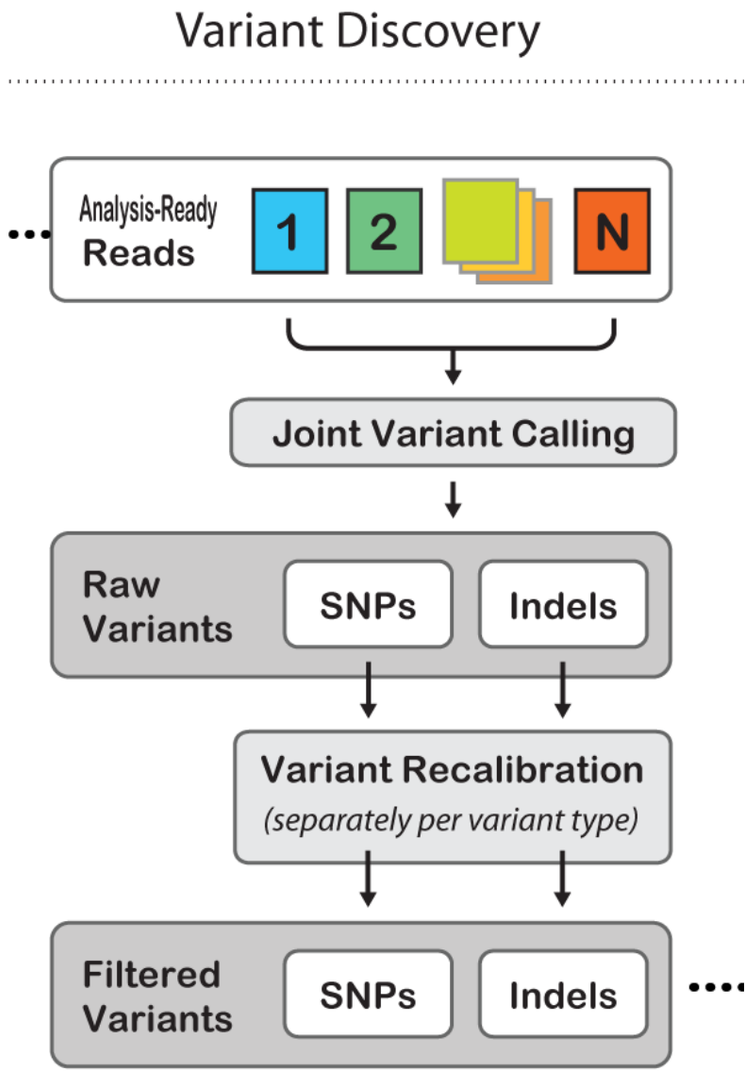
Auwera, GA et al. *Current Protocols in Bioinformatics* (2013)

The reference model enables incremental calling



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

Variant calling is a large-scale bayesian modeling problem



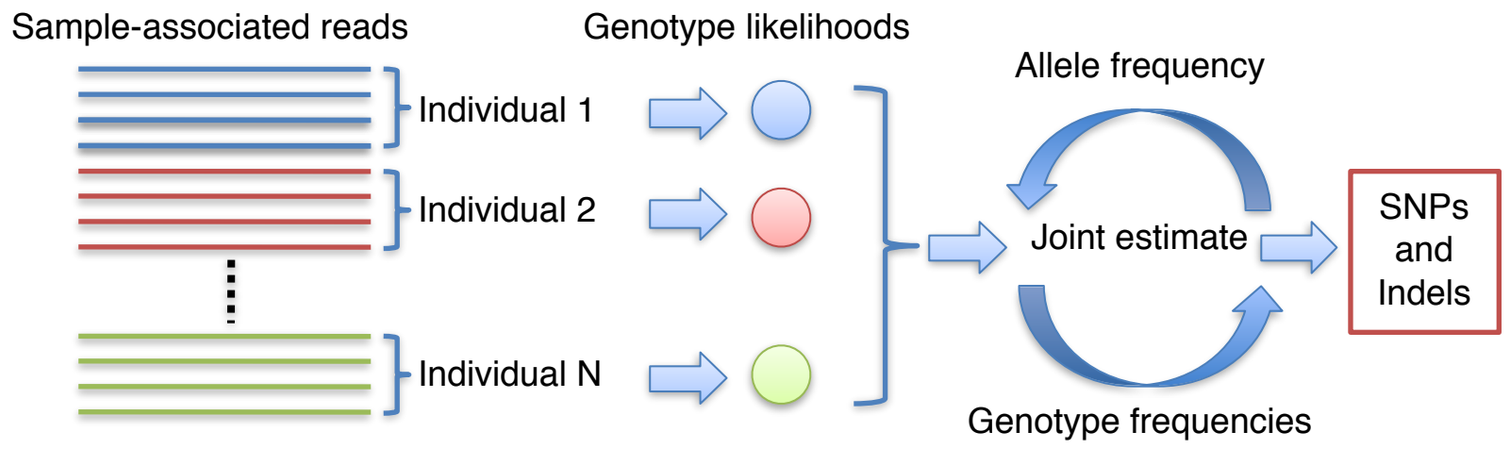
prior Likelihood

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

Diploid

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

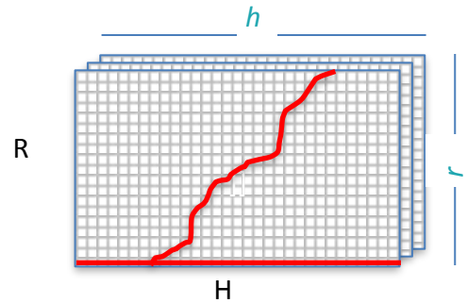
$\Pr\{D|H\}$ is the haploid likelihood function



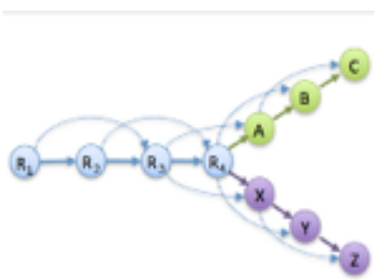
DePristo et al. *Nature Genetics* (2011)

Understanding the Haplotype Caller

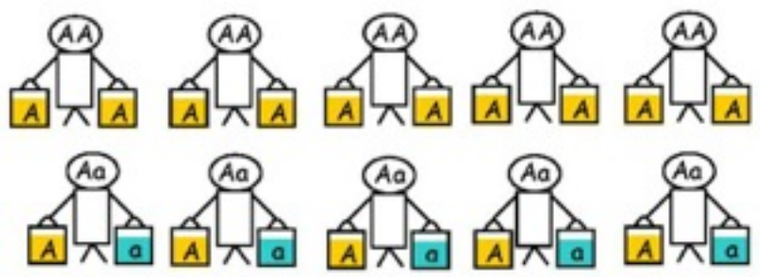
1. **Active region traversal**
identifies the regions that need to be reassembled



3. **Pair-Hmm evaluation** of all reads against all haplotypes
(scales exponentially)



2. **Local de-novo assembly**
builds the most likely haplotypes for evaluation



4. **Genotyping**
using the exact model

Pair-HMM is the biggest culprit for the low performance of the Haplotype Caller

Stage	Time	Runtime %
Assembly	2,598s	13%
Pair-HMM	14,225s	70%
Traversal + Genotyping	3,379s	17%

NA12878 80xWGS performance on a single core
chr20 time: 5.6h
whole genome: 7.6 days

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	-

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