GPU DNA Sequencing Base Quality Recalibration

Mauricio Carneiro
Nuno Subtil

carneiro@gmail.com

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

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

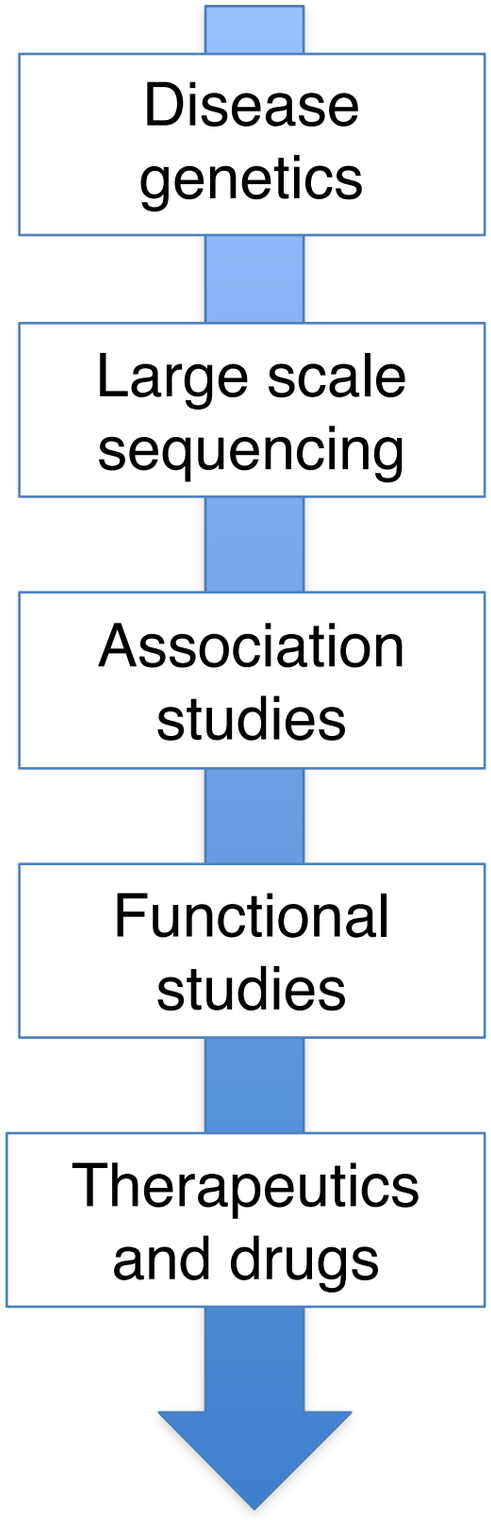
Rare Variant Association Study (RVAS)

vs

Common Variant Association Study (CVAS)

vs

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

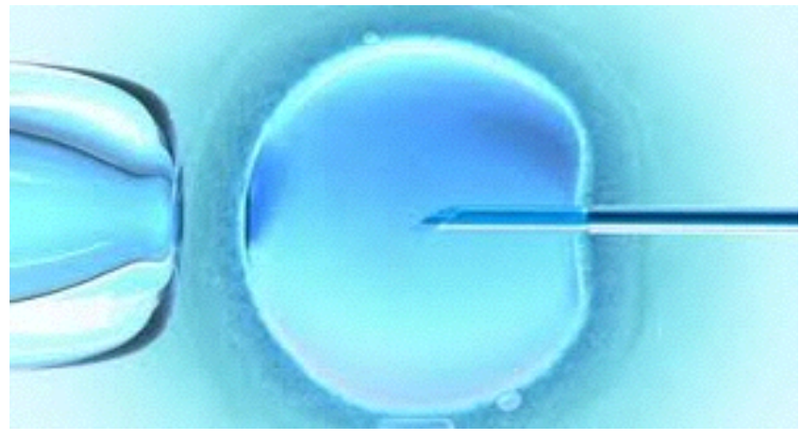
These systematic differences can be identified by comparing affected and unaffected individuals

These associated variants give insight into the biological mechanisms of disease

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

Personalized medicine for rare variants is already a reality

- Couples with rare conditions get referred by local hospitals to local genetics center.



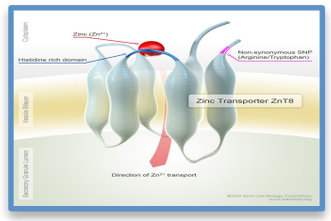
- DNA sequencing can reveal the deleterious mutation causing the condition.

- In vitro fertilization followed by embryo selection can guarantee a disease free baby.

- Limitations of this process are in the size of the control cohort, and the rarity of the condition.

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

Type 2 Diabetes

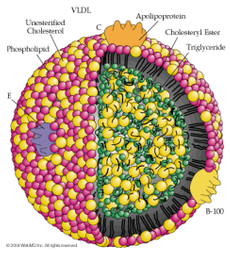


- 13,000 exomes
- SLC30A8
  (Beta-cell-specific Zn^{++} transporter)
- 3-fold protection against T2D!
- ***1 LoF per 1500 people***

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

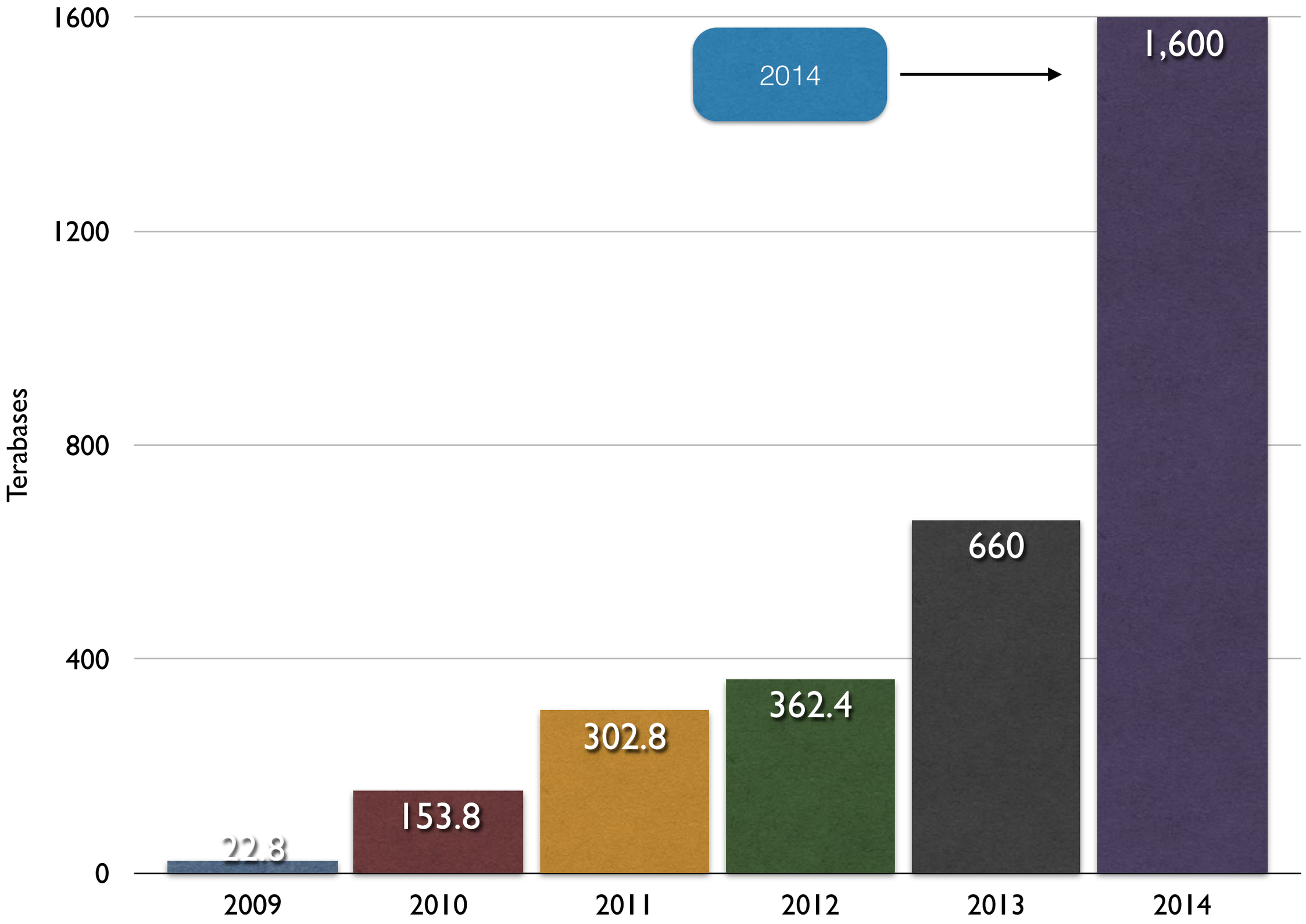


- 3,700 exomes
- APOC3
- 2.5-fold protection from CHD
- ***4 rare disruptive mutations (~1 in 200 carrier frequency)***

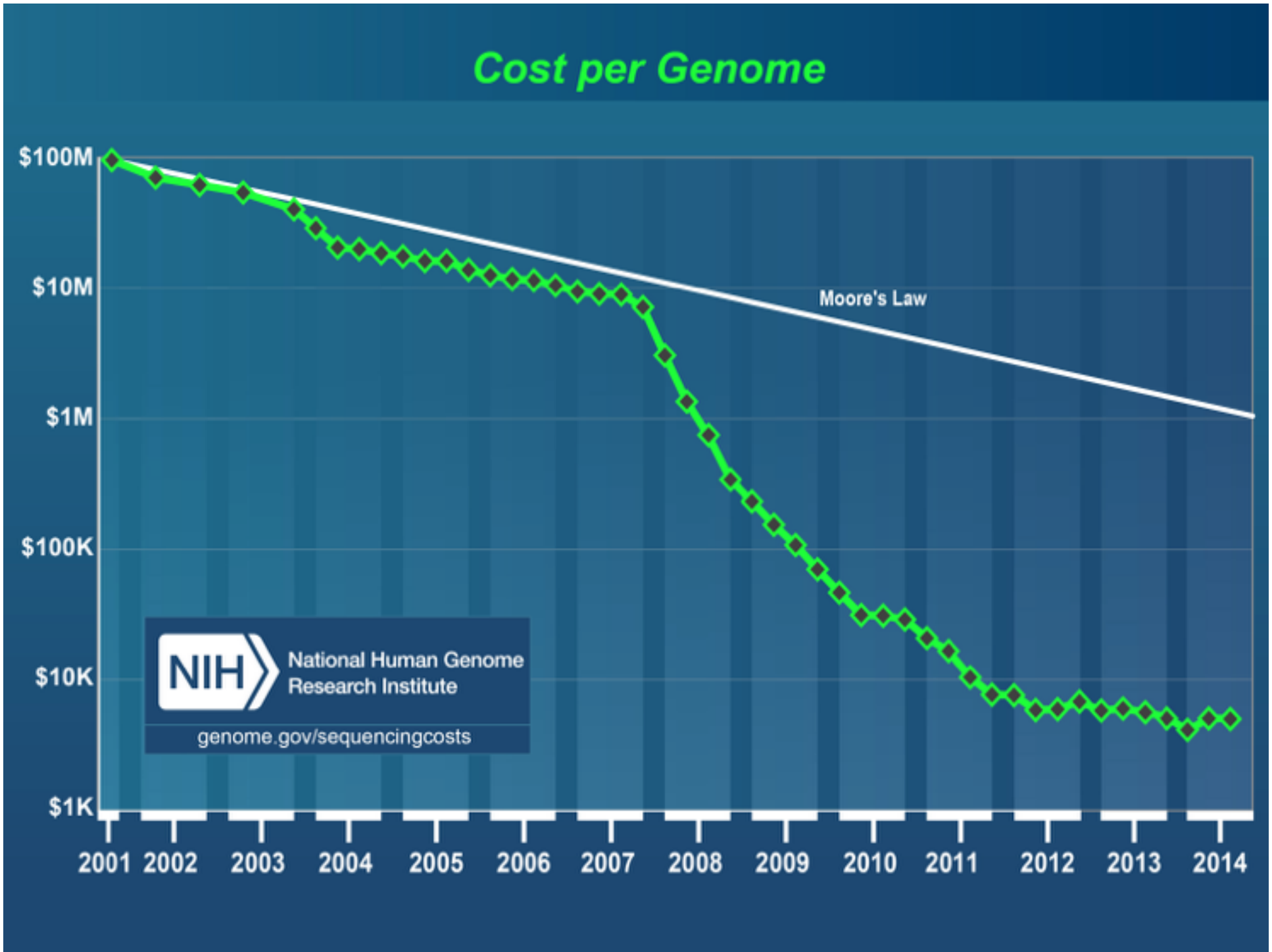
Early Heart Attack

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

Terabases of Data Produced by Year

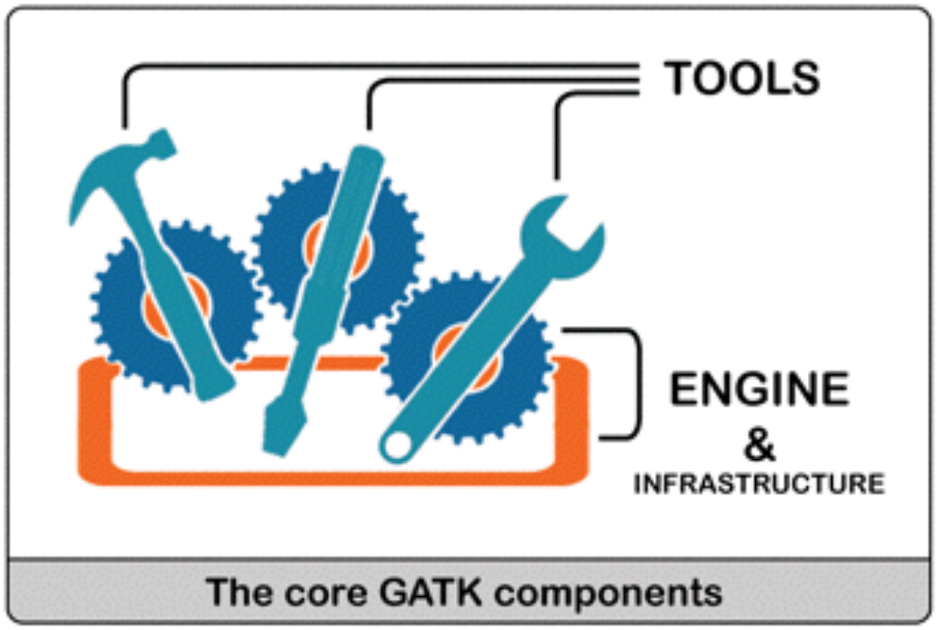


…and these numbers will continue to grow faster than Moore's law



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

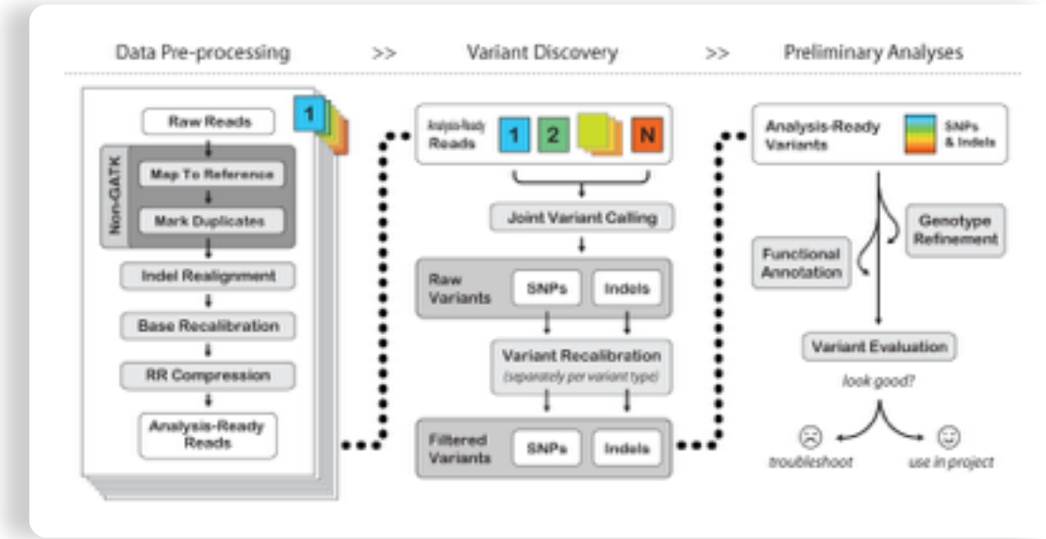
Toolkit & framework packages



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

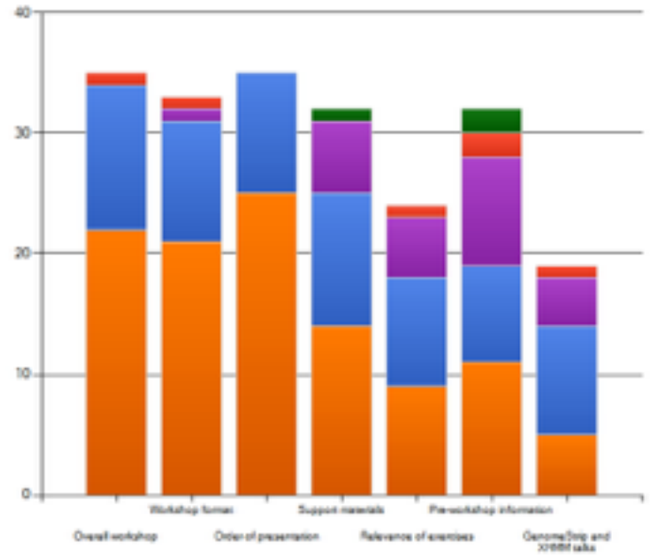
iTunes U Collections

BroadE: GATK
Broad Institute

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

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



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

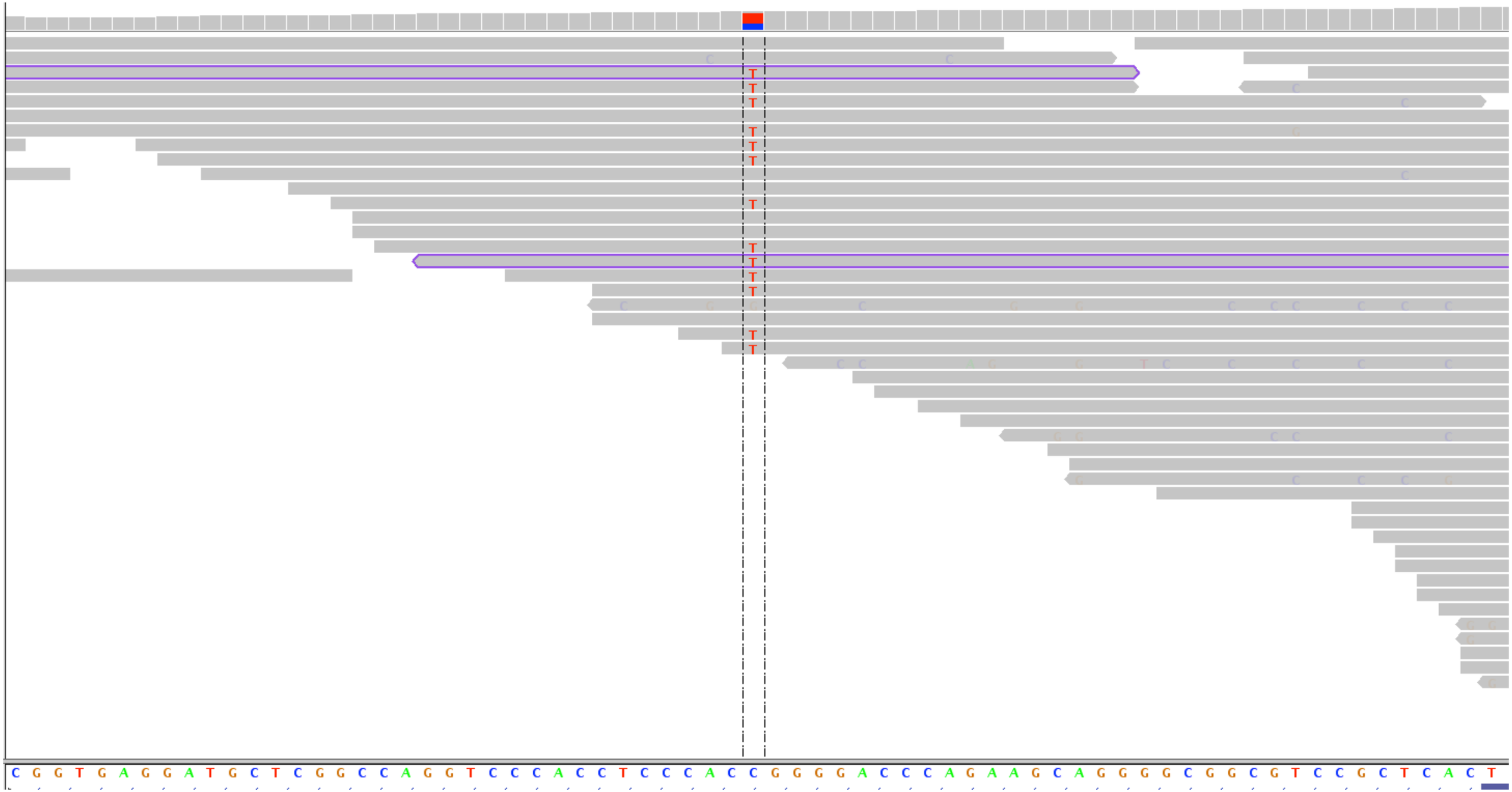
BroadE: Overview of GATK & best practices

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

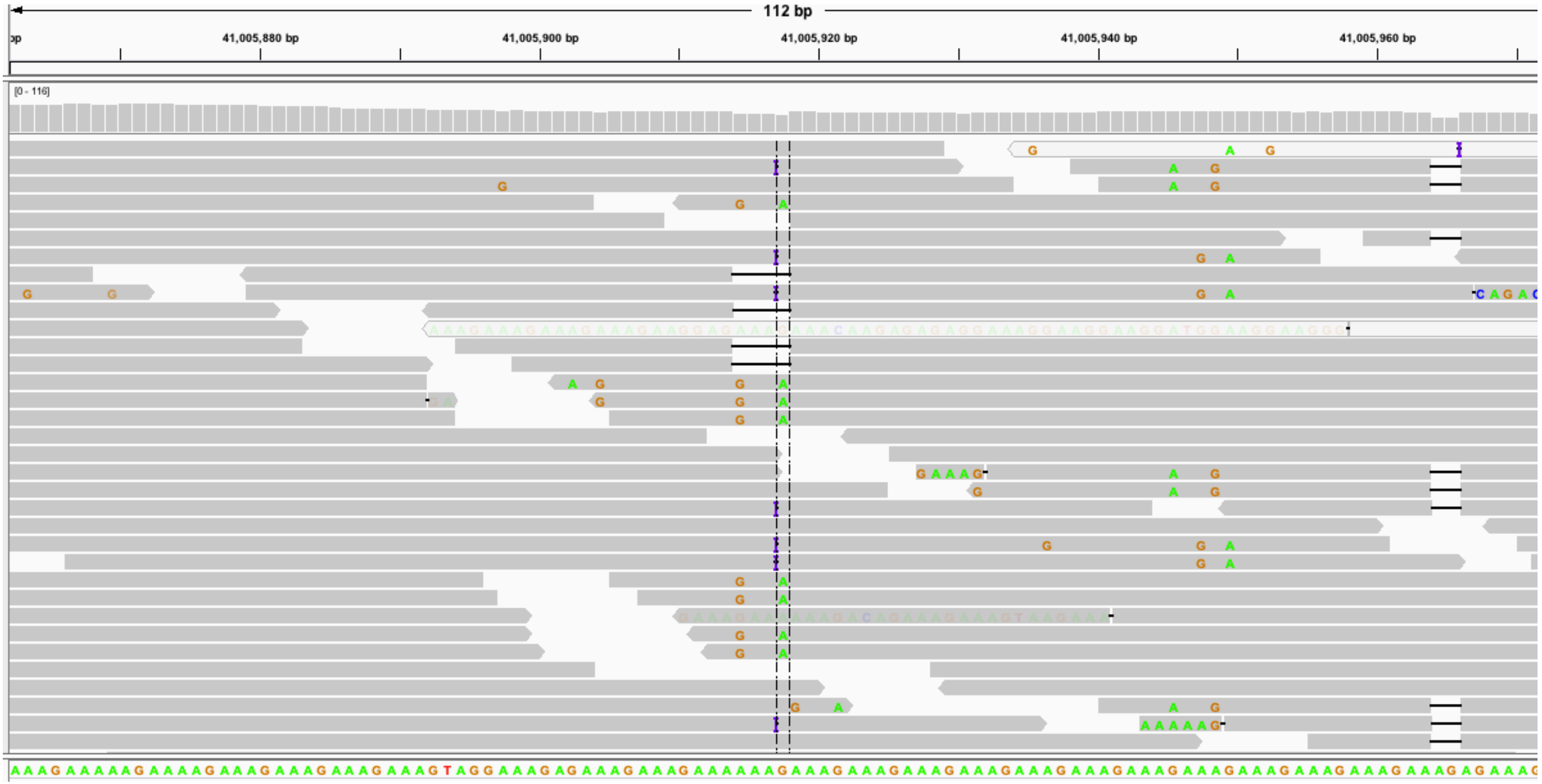
Copyright Broad Institute, 2013. All rights reserved. The presentations below were filmed during the 2013 **GATK** Workshop, part of ...

NEW HD

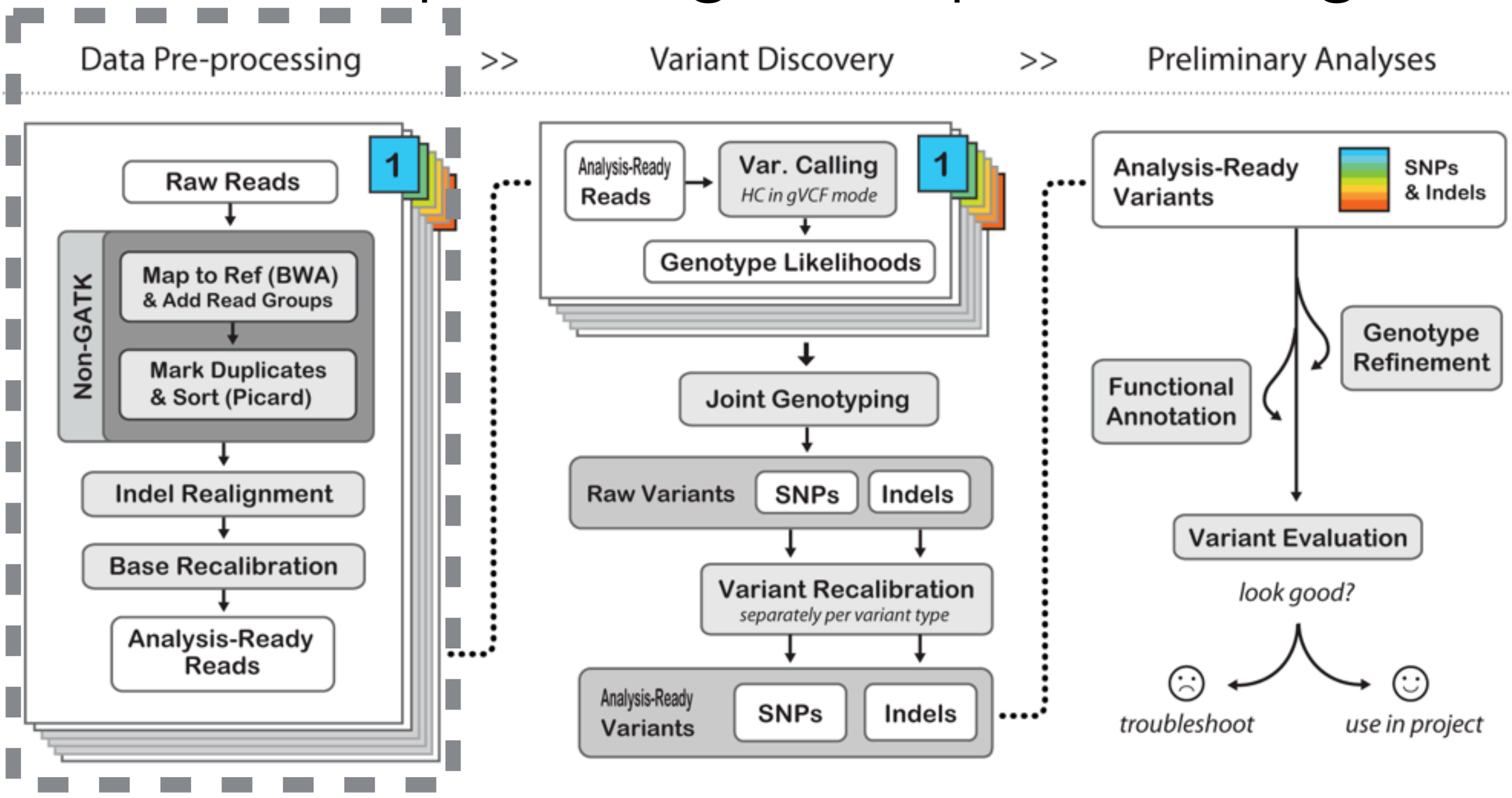
Identifying mutations in a genome is a simple “find the differences” problem



Unfortunately, real data does not look that simple



We have defined the best practices for sequencing data processing

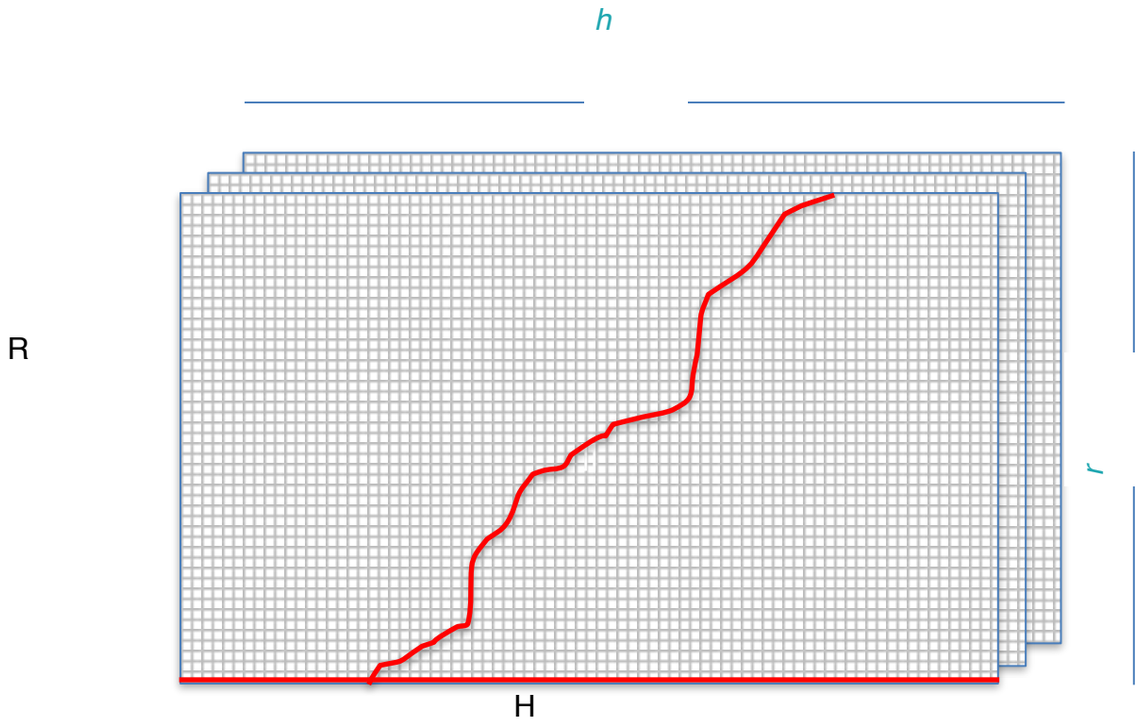


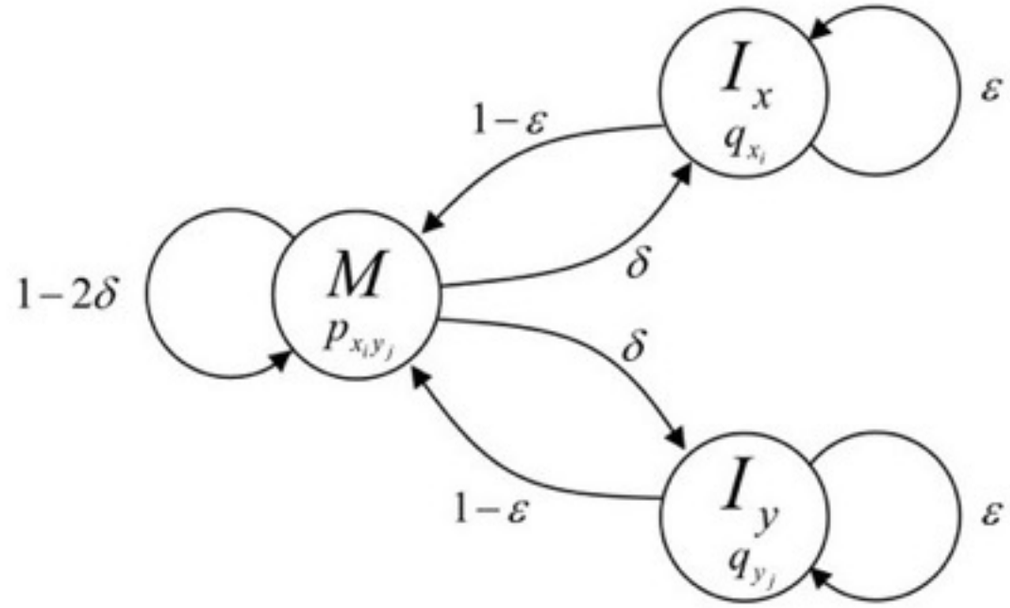
Auwera, GA et al. *Current Protocols in Bioinformatics* (2013)

GPUs have sped 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	-

Variant calling depends heavily on accurate measurements of error





The transition probabilities on this HMM are the

Qualities are the probability measures of error in a read

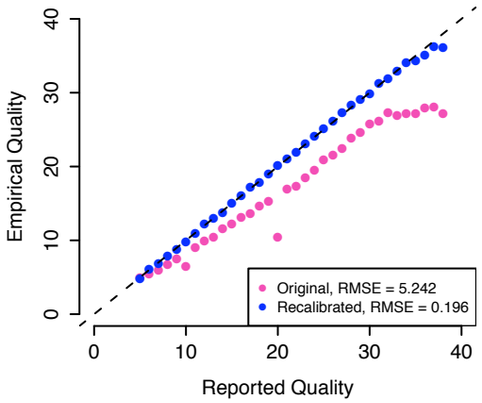
emitted by most instruments

Bases	C	G	G	T	A	C	A	A	T	G
Quals	33	37	29	39	30	32	23	12	2	2
Insertion Quals	43	40	43	42	44	39	22	10	43	40
Deletion Quals	45	45	40	39	42	41	38	32	40	44

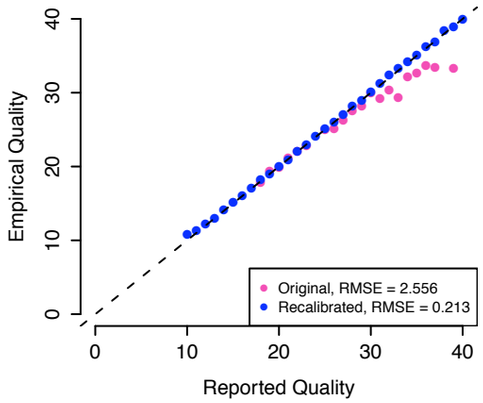
Highlighted as one of the major methodological advances of the 1000 Genomes Pilot Project!

Base Quality Score Recalibration provides a calibrated error model from which to make mutation calls

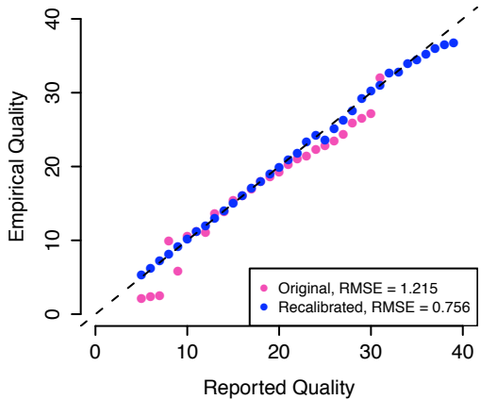
SLX GA



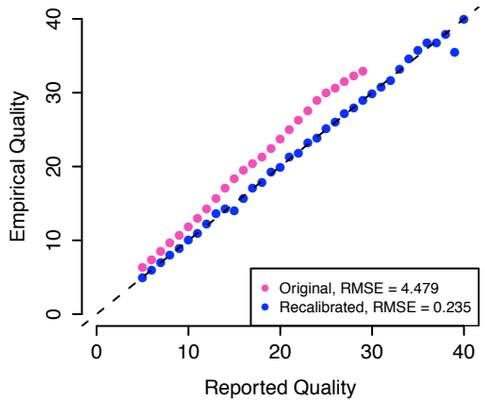
454



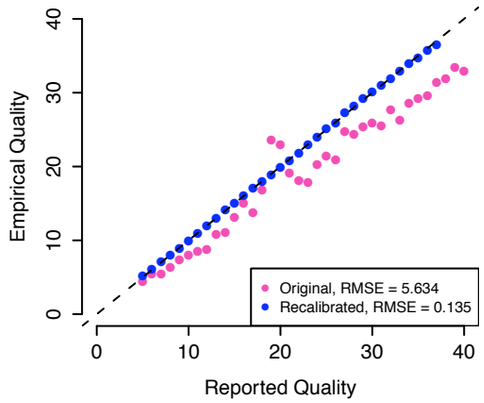
SOLiD

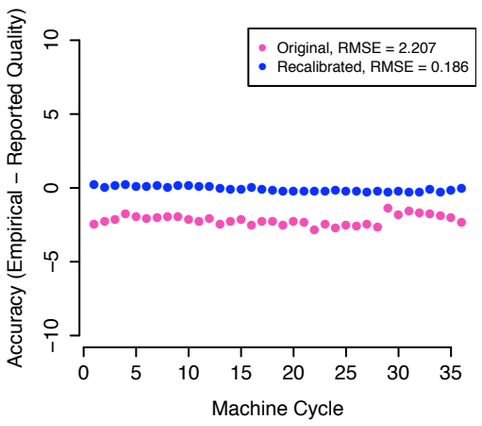


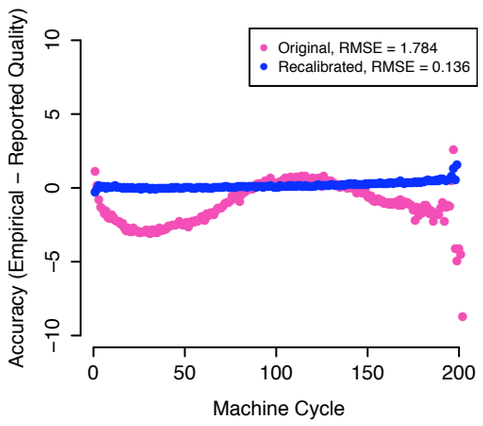
Complete Genomics

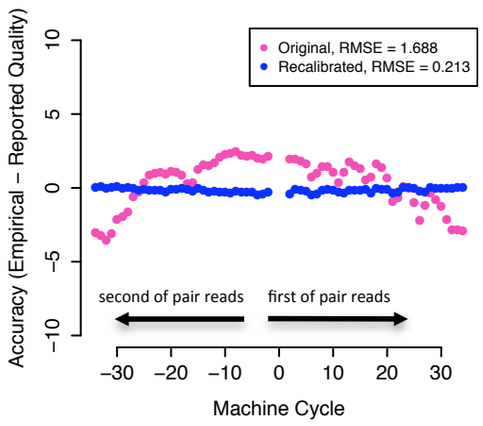


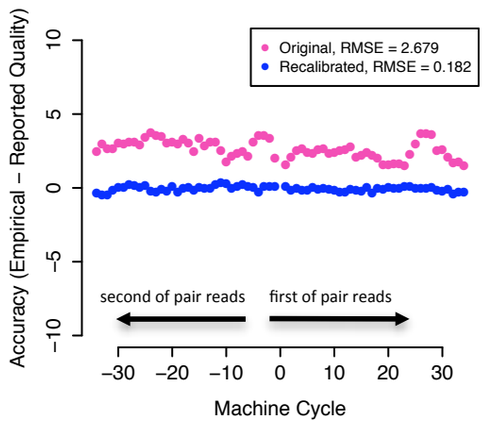
HiSeq

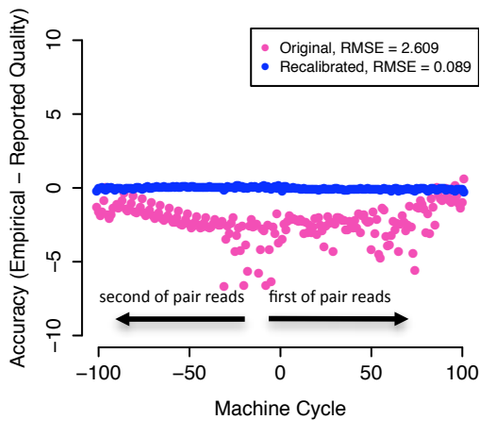


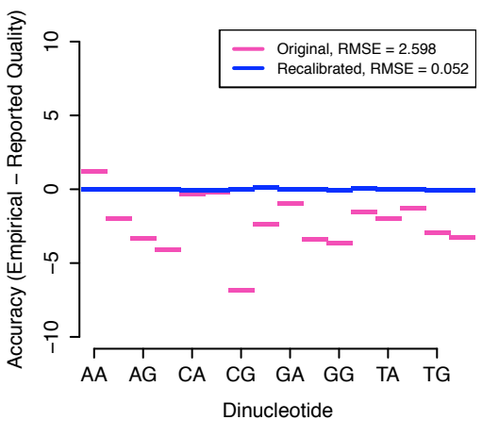


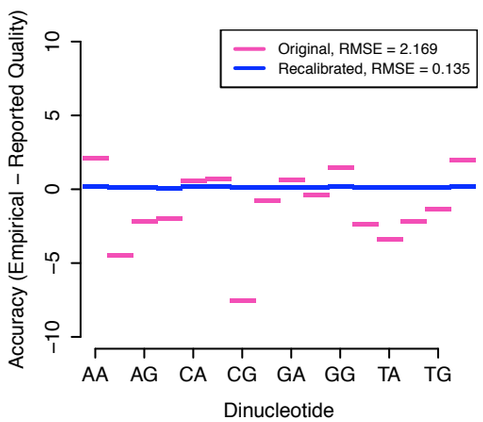


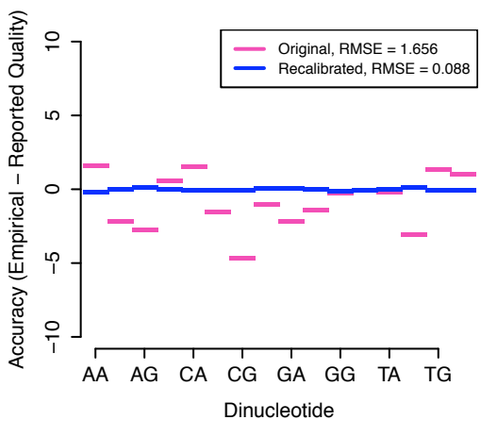


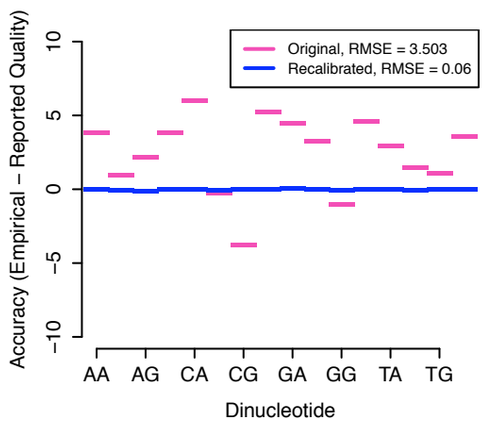


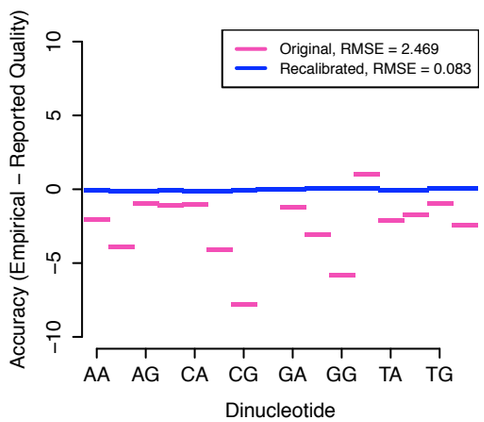












Base recalibration clarifies the unbiased error mode of Pacbio

Carneiro, et al Genome Biology 2014

Processing is a big cost on whole genome sequencing