1000 Genomes project: from mapping reads to *de novo* mutations

Mark A. DePristo (depristo@broadinstitute.org)
Manager, Genome Sequencing and Analysis Group
Medical and Population Genetics Program
Broad Institute of Harvard and MIT
December 3, 2009

Acknowledgments

Quality-score recalibration	Local realignment	Variation discovery	De novo mutations
• Anthony Philippakis • Andrew Kernytsky • Matt Hanna	• Eric Banks • Andrey Sivachenko	• Jared Maguire • Kiran Garimella • Manny Rivas • Michael Melgar • Eric Banks	• Andrew Kernytsky • Matt Hurles and Philip Awadalla at the Sanger

Other contributors

- The entire genome sequencing and analysis group
    - Especially the GSA software engineering team: Matt Hanna and Aaron McKenna
- MPG directorship: Stacey Gabriel, David Altshuler, Mark Daly
- Carrie Sougnez, production teams and folks at 320 and 7CC
- The SAM/BAM working group: Bob Handsaker, Tim Fennell, Heng Li, and Richard Durbin
- The cancer genome analysis group: Gad Getz, Kristian Cibulskis, Andrey Sivachenko
- The IGV team: Jim Robinson and Helga Thorvaldsdottir
- Production informatics: Tim Fennell and Alec Wysoker
- The 1000 genomes project

Agenda

- Introduction to the 1000 genomes project
- Mapping and alignment
  - SAM/BAM format
  - Visualizing the data
- The Genome Analysis Toolkit
  - The infrastructure supporting our tools for working with next-generation sequencing data
- Tools developed in the GATK for calling SNPs and indels in the 1000 genomes pilot

The 1000 genomes project is characterizing common genetic variation with MAF >1% in three populations

Pilot 1:

Applies a multi-sample generalization of the single sample approach in pilot 2

Method not discussed in detail

~ 2-10M short indels

Pilot 2:
Two children and their parents whole genome sequence to ~70x

Data production and analysis

~ 3-5M SNPs

~ 200-500K short indels

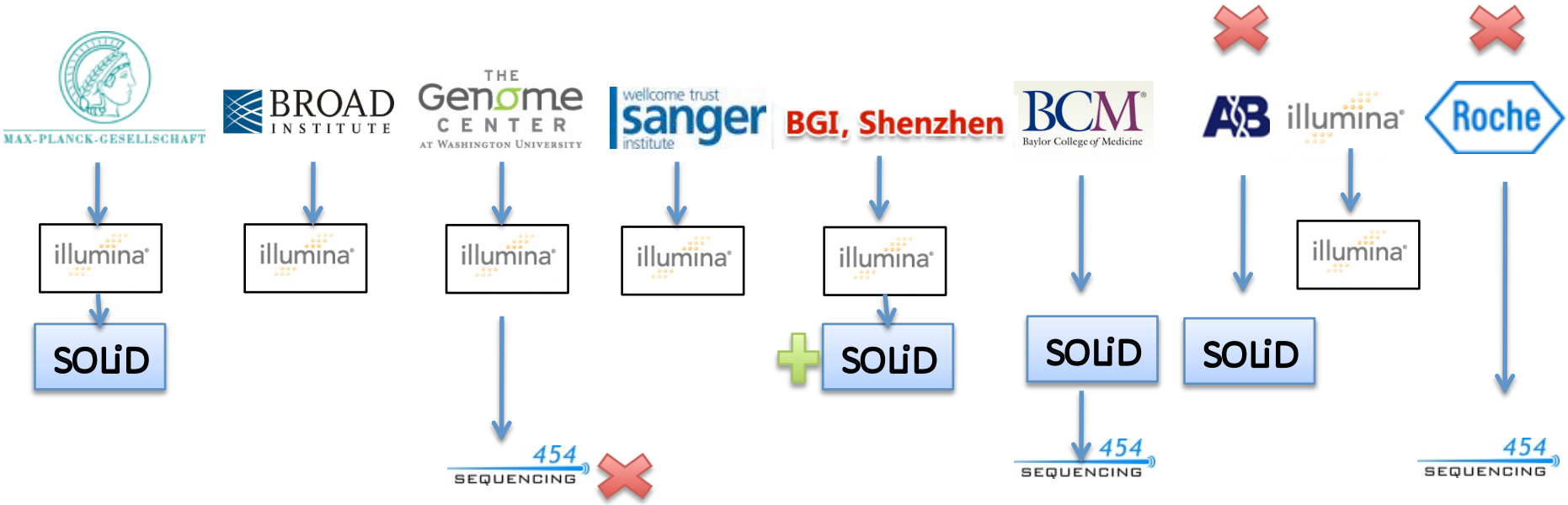
Pilot 3:

Applies the same SNP and indel calling methods as Pilot 2

Method not discussed in detail

~ 1000 short indels

Data for the project comes from many centers and several technologies



Added for production phase

For pilot phase only

Slide courtesy of Carrie Sougnez

The pilot phase alone has generated ~5 Tb of sequence

	Pilot 1	Pilot 2	Pilot 3	Total
Number of Samples	185	6	574	765
Illumina	1.99	0.85	0.54	3.37
SOLID	0.75	0.27	0.0	1.02
454	0.20	0.08	0.14	0.42
Total	**2.93**	**1.19**	**0.69**	**4.81**

Slide courtesy of Carrie Sougnez

Agenda

- Introduction to the 1000 genomes project
- Mapping and alignment
  - SAM/BAM format
  - Visualizing the data
- The Genome Analysis Toolkit
  - The infrastructure supporting our tools for working with next-generation sequencing data
- Tools developed in the GATK for calling SNPs and indels in the 1000 genomes pilot

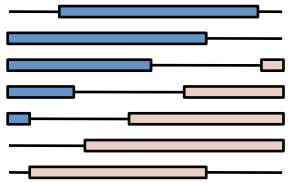
From unmapped reads to true genetic variation in next-generation sequencing data

Solexa

SOLiD

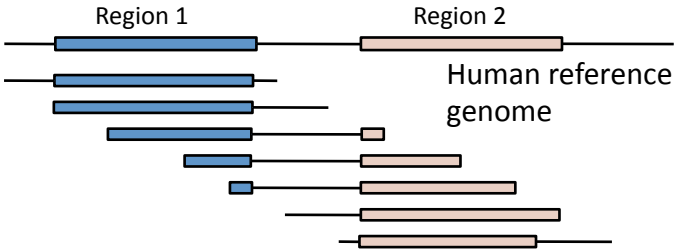
454

Raw short reads



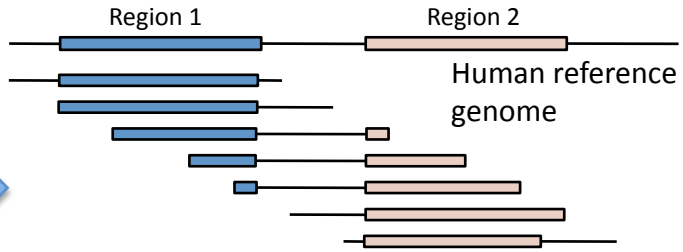
A single run of a sequencer generates ~50M ~75bp short reads for analysis

Mapping and alignment



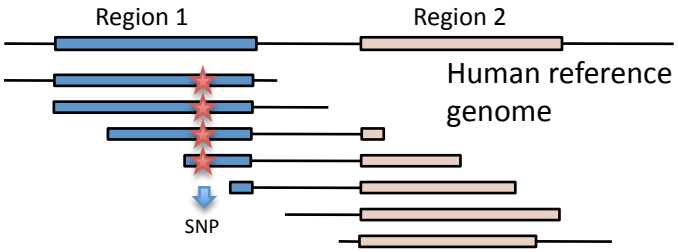
The origin of each read from the human genome sequence is found

Quality calibration and annotation



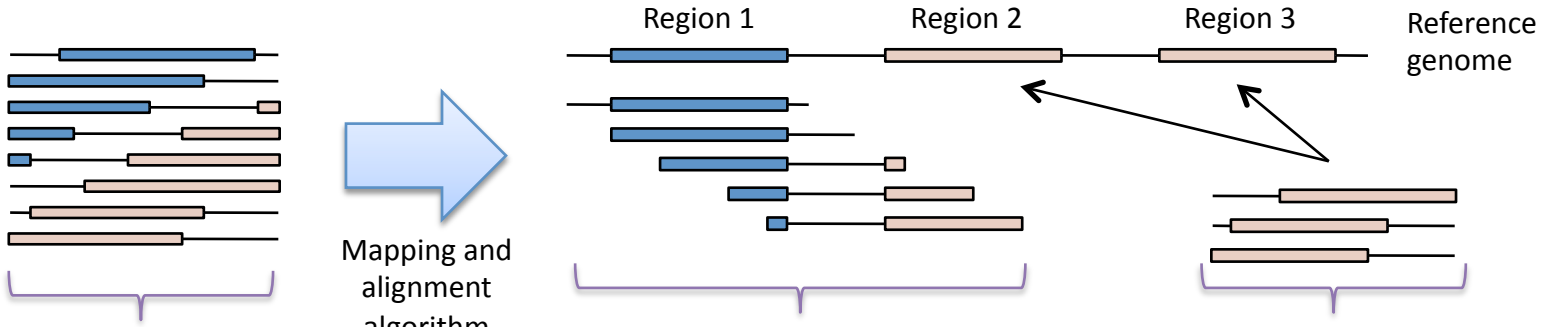
The quality of each read is calibrated and additional information annotated for downstream analyses

Identifying genetic variation



SNPs and indels from the reference are found where the reads collectively provide evidence of a variant

Finding the true origin of each read is a computationally demanding and important first step



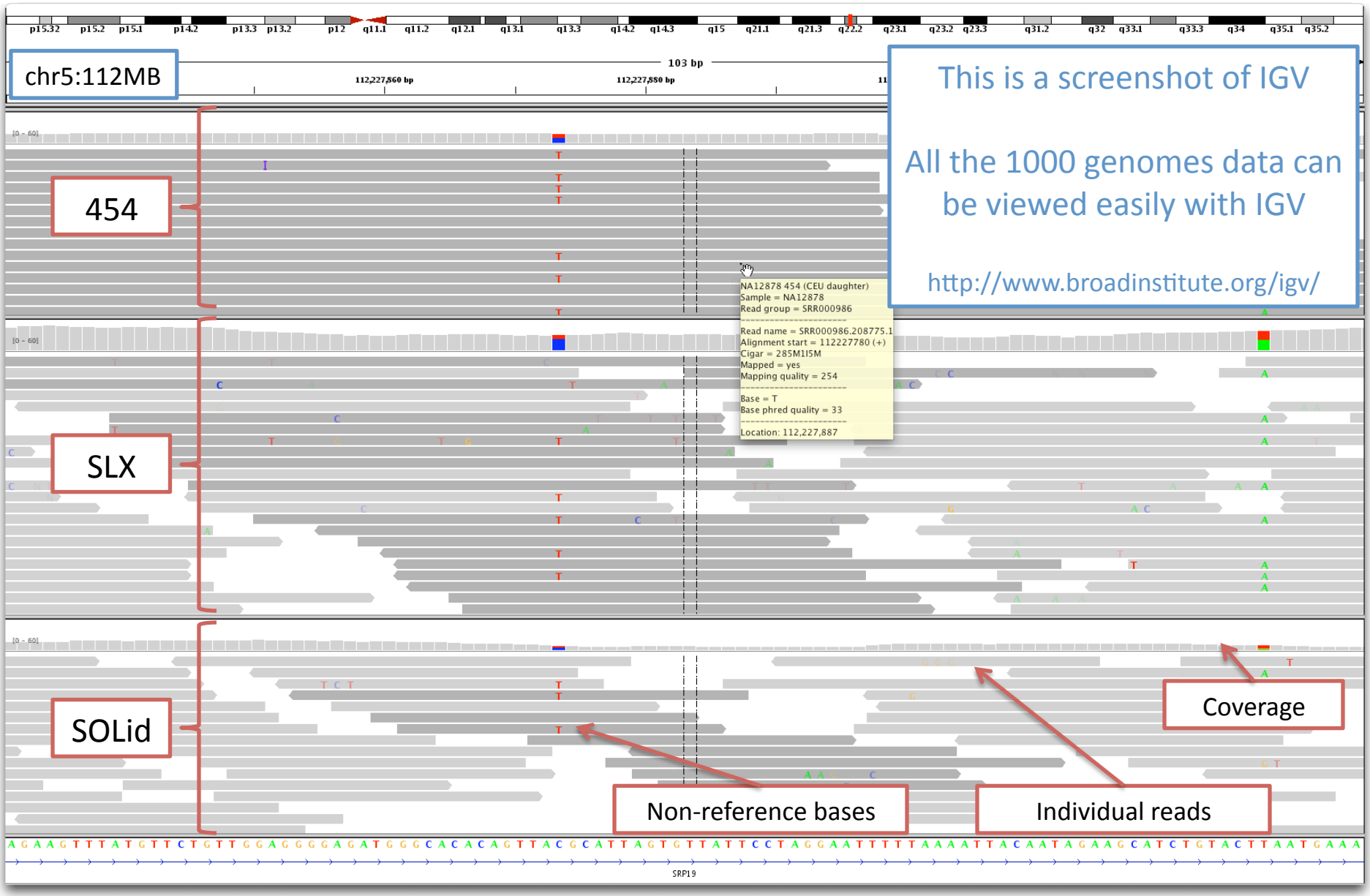
Solexa : MAQ	454 : SSAHA	SOLiD : Corona
• Robust, accurate 'gold standard' aligner for NGS • Developed by Li and Durbin • Soon to be replaced by BWA, also by Li and Durbin	• Hash-based aligner with high sensitivity and specificity with longer reads	• ABI-designed tool for aligning in color-space

SAM/BAM files

The SAM file format

- Data sharing was a major issue with the 1000 genomes
  - Each center, technology and analysis tool used its own idiosyncratic file formats – no one could exchange data
- The Sequence Alignment and Mapping (SAM) file format was designed to capture all of the critical information about NGS data in a single indexed and compressed file
  - Becoming a standard and is now used by production informatics, MPG, and cancer analysis groups at the Broad
- Has enabled sharing of data across centers and the development of tools that work across platforms
- More info at http://samtools.sourceforge.net/

What does the data actually look like?



Agenda

- Introduction to the 1000 genomes project
- Mapping and alignment
  - SAM/BAM format
  - Visualizing the data
- The Genome Analysis Toolkit
  - The infrastructure supporting our tools for working with next-generation sequencing data
- Tools developed in the GATK for calling SNPs and indels in the 1000 genomes pilot

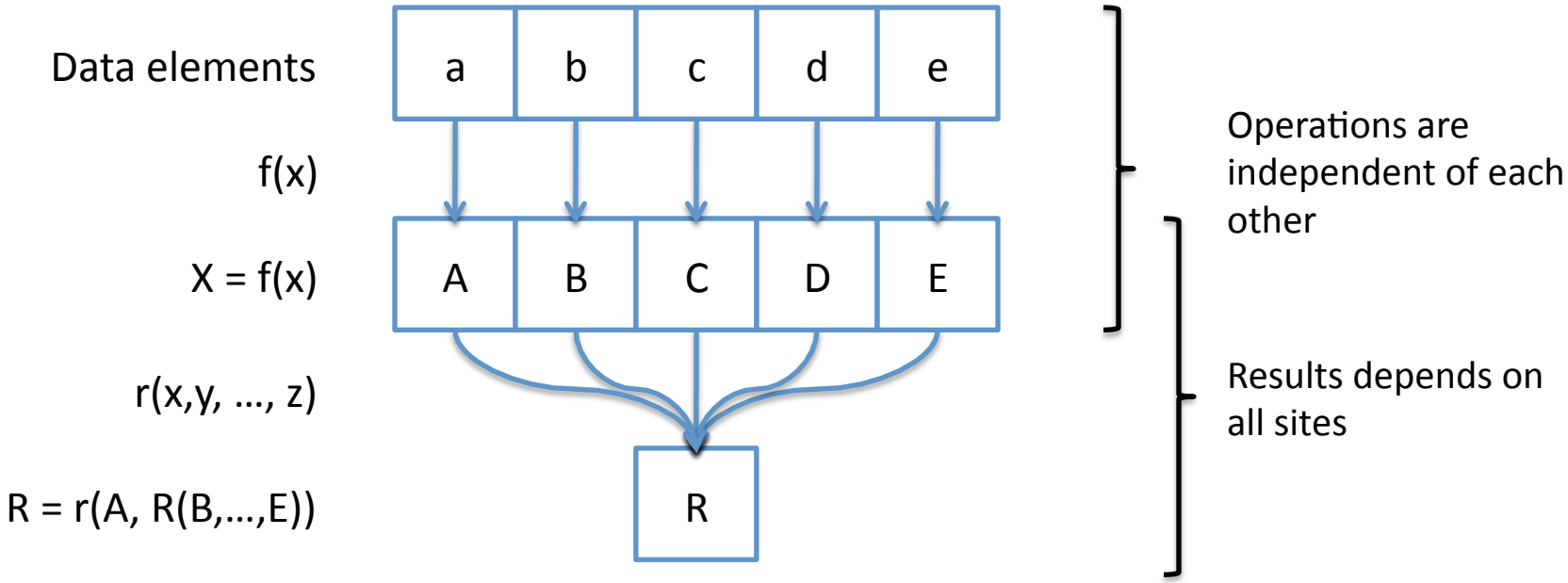
The GATK is a structured programming framework that aims to simplify writing analysis tools for resequencing data

- The framework is designed to support most common paradigms of analysis algorithms
  - Provides structured access to reads in SAM format, reference context, as well as reference-associated meta data
- General-purpose
  - Optimized for ease of use and completeness of functionality within scope
- Efficient
  - Engineering investment on performance of critical data structures and manipulation routines
- Convenient
  - Structured plug-in model makes developing against the framework relatively painfree

The functional programming paradigm

- The GATK follows a common functional programming paradigm called map and reduce
  - `reduce( g, map( f, list ), init )` `## python`
  - ```
    Object result = init;                    // java
    for ( List x: list )
       result = g( result, f(x) );
    ```
  - `(reduce g (map f list))` `;; scheme`
    ```

The map / reduce framework



Result is:

Map — Function f applied to each element of list

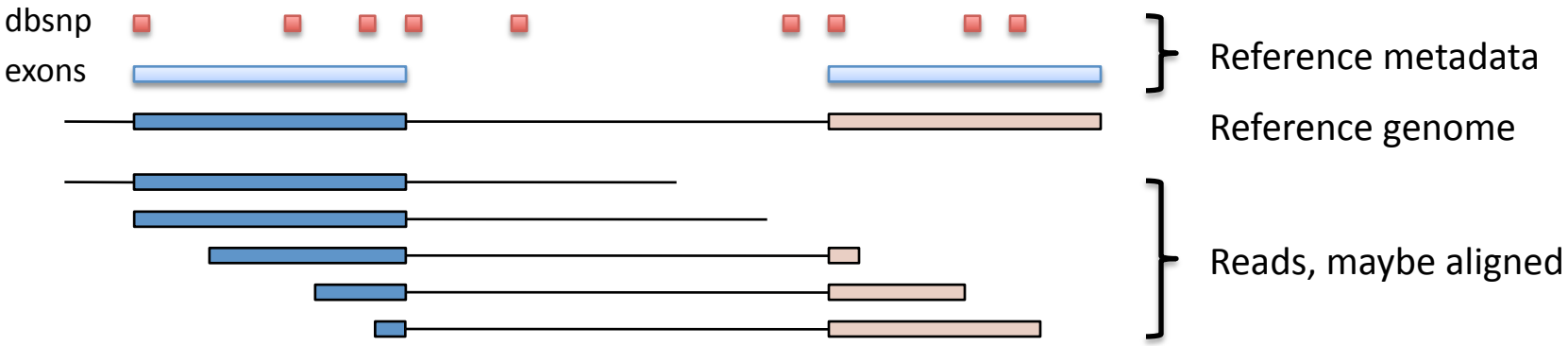
Reduce — Function r recursively reduced over each f(…)

Many algorithms fit within the Map/Reduce framework

- Idea behind Map/Reduce is to provide structured traversal and access to data
  - Separate problems of accessing data from calculations on the elements in the data
  - Developers can provide powerful, intelligent, efficient traversal engines that implement the map operation
  - Analysts can easily write functions to analyze their data, and then map them across the data
- Google popularized map/reduce
  - see Dean and Ghemawat, OSDI'04: Sixth Symposium on Operating System Design and Implementation
  - Becoming so popular there was a New York Times article about it on Tuesday, March 17th, 2009!

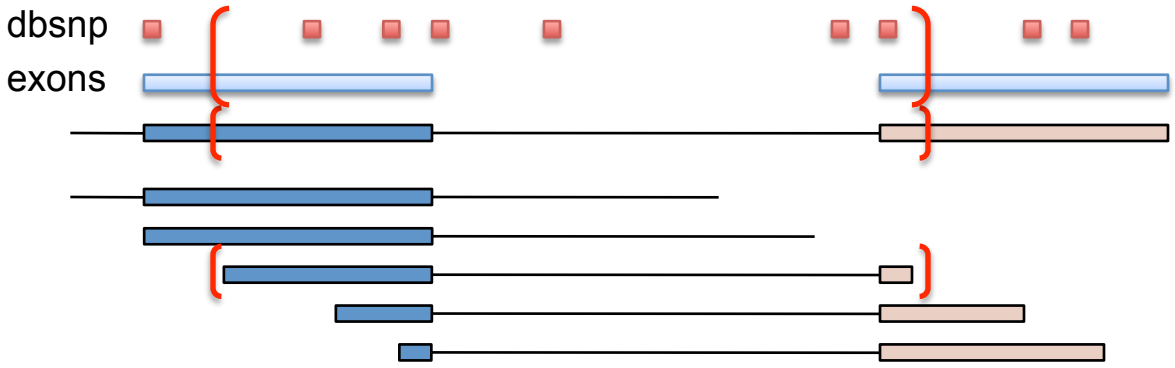
Map/Reduce over the genome

Fundamental data



Reference	Reads	Metadata
• Reference genome in fasta format	• SAM format reads • Some traversal types may required reads to be aligned (by locus, for example)	• Data associated with positions on the reference genome • E.g., dbSNP, exons

Map/Reduce by read



Reference metadata

Reference genome

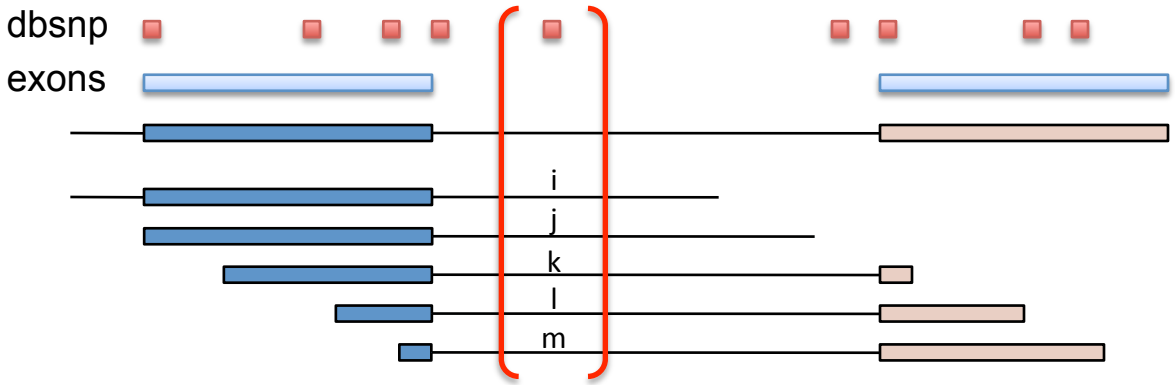
Reads, maybe aligned

```
f (single read,
   covered reference seq,
   covered metadata)
```

x

Evaluated over each read, with reduce accumulating x results at ever read

Map/Reduce by loci



Reference metadata

Reference genome

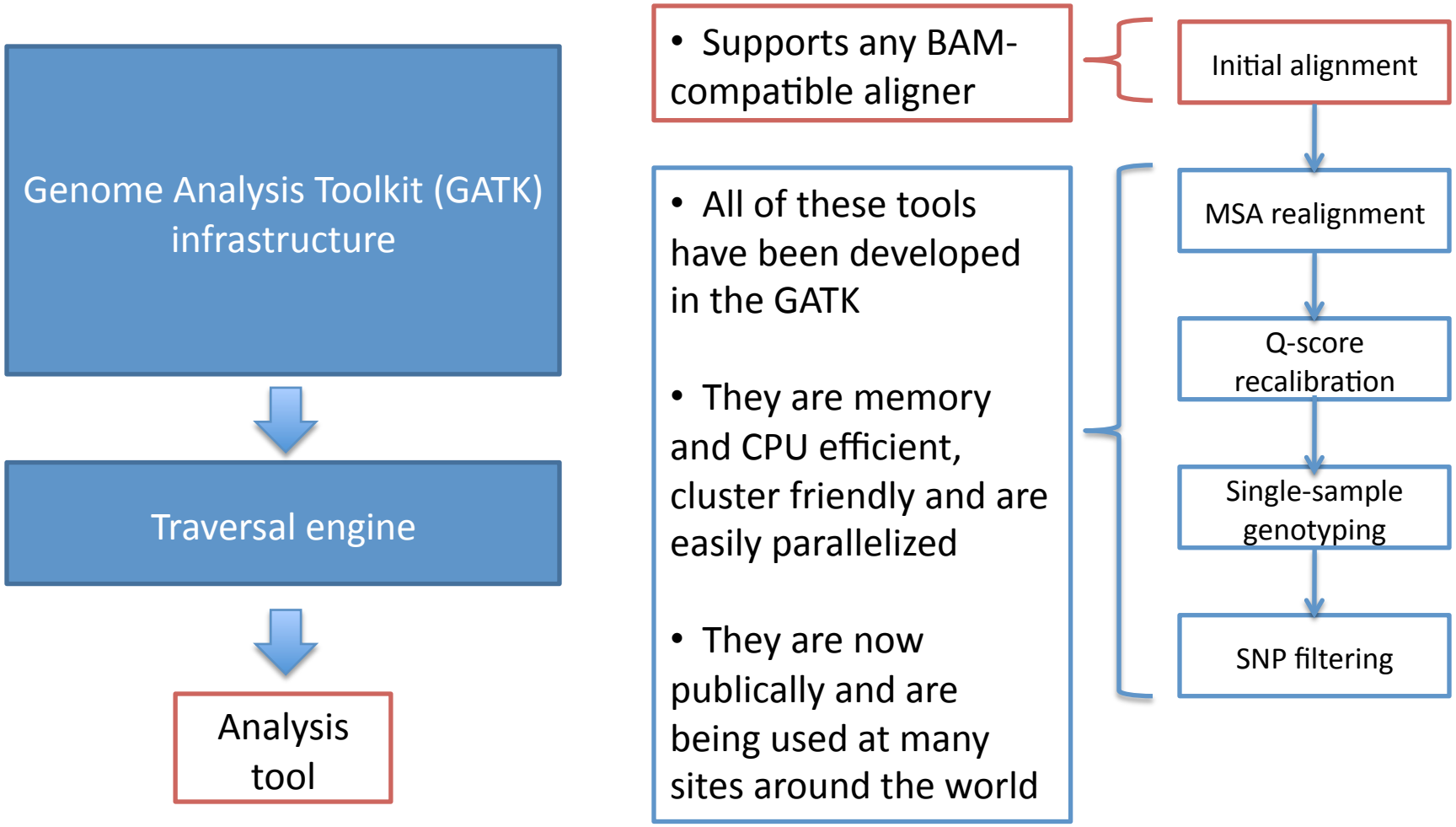
Reads, maybe aligned

```
f (all reads cover locus,
   indices into reads yielding
      equivalent positions
   covered reference seq,
   covered metadata)
```

Evaluated over each locus in the genome, with reduce accumulating x results at ever locus

```
x
```

The Genome Analysis Toolkit (GATK) enables rapid development of efficient and robust analysis tools

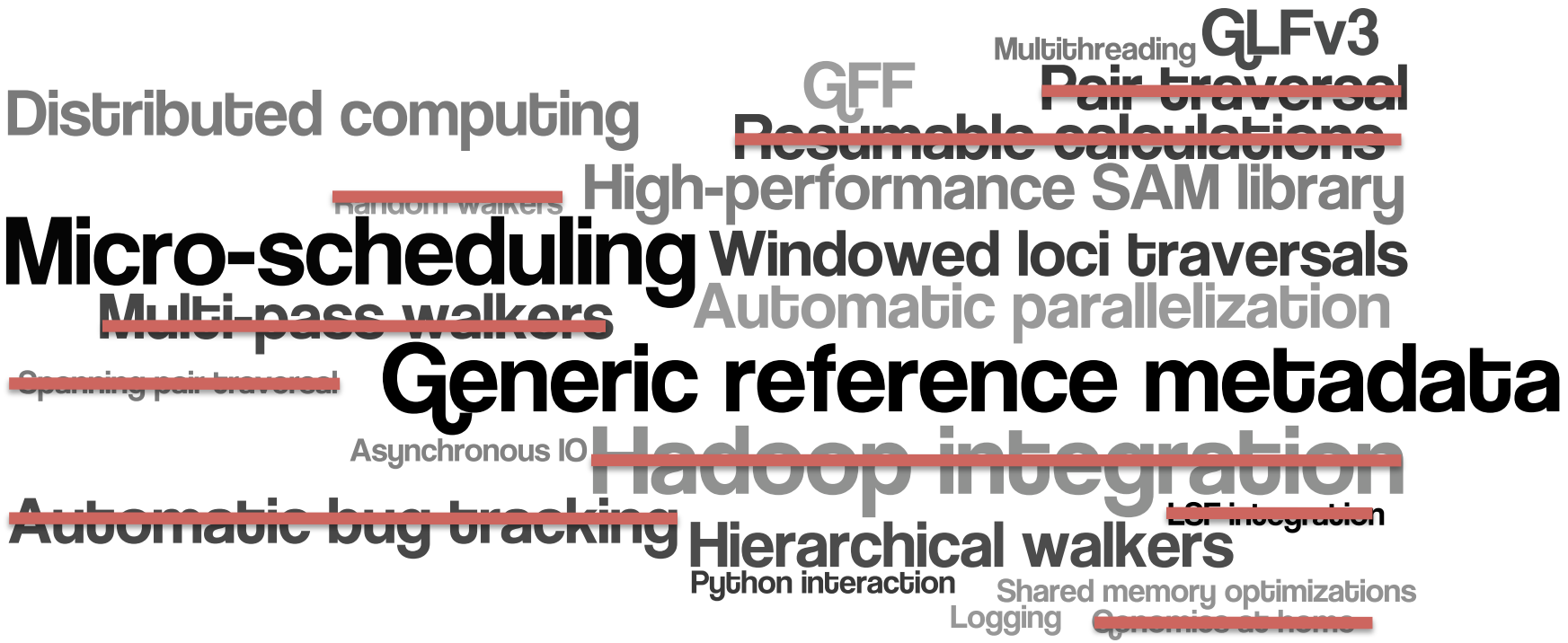


Provided by framework

Implemented by user

More info: http://www.broadinstitute.org/gsa/wiki/

The GATK engine already supports many advanced features



Pileup with dbSNP

Code: org/broadinstitute/sting/gatk/walkers/Pileup.java

package, imports, etc. removed for presentation

```
public class DepthOfCoverageWalker extends LociWalker<Integer, Integer>{
    public Integer map(List<ReferenceOrderedDatum> rodData,
                       char ref, LocusContext context) {
        String bases = "";
        String quals = "";
        for ( int i = 0; i < context.getReads().size(); i++ ) {
            SAMRecord read = context. getReads().get(i);
            int offset = context.getOffsets().get(i);
            bases += read.getReadString().charAt(offset);
            quals += read.getBaseQualityString().charAt(offset);
         }

        String rodString = "";
        for ( ReferenceOrderedDatum datum : rodData ) {
            if ( datum != null && datum instanceof rodDbSNP) {
                rodDbSNP dbsnp = (rodDbSNP)datum;
                rodString = "[ROD: " + dbsnp.toMediumString() + "]";
             }
        }
         System.out.printf("%s: %s %s %s %s%n",
                     context.getLocation(), ref, bases, quals, rodString);
         return 1;
    }
}
```

Build bases and quals strings

Build the dbSNP string

Pileup with dbSNP II

CPU time	10 secs
Max. memory	1 GB

Command

```
java -jar dist/GenomeAnalysisTK.jar –T Pileup
    -I /broad/1KG/legacy_data/tcga-freeze3/tcga-freeze3-normal.bam
    -R /seq/references/Homo_sapiens_assembly18/v0/Homo_sapiens_assembly18.fasta
    -L chr1:559,844-559,848
    -DBSNP /humgen/gsa-scr1/GATK_Data/dbsnp_129_hg18.rod
```

Analysis name

Reads

Ref

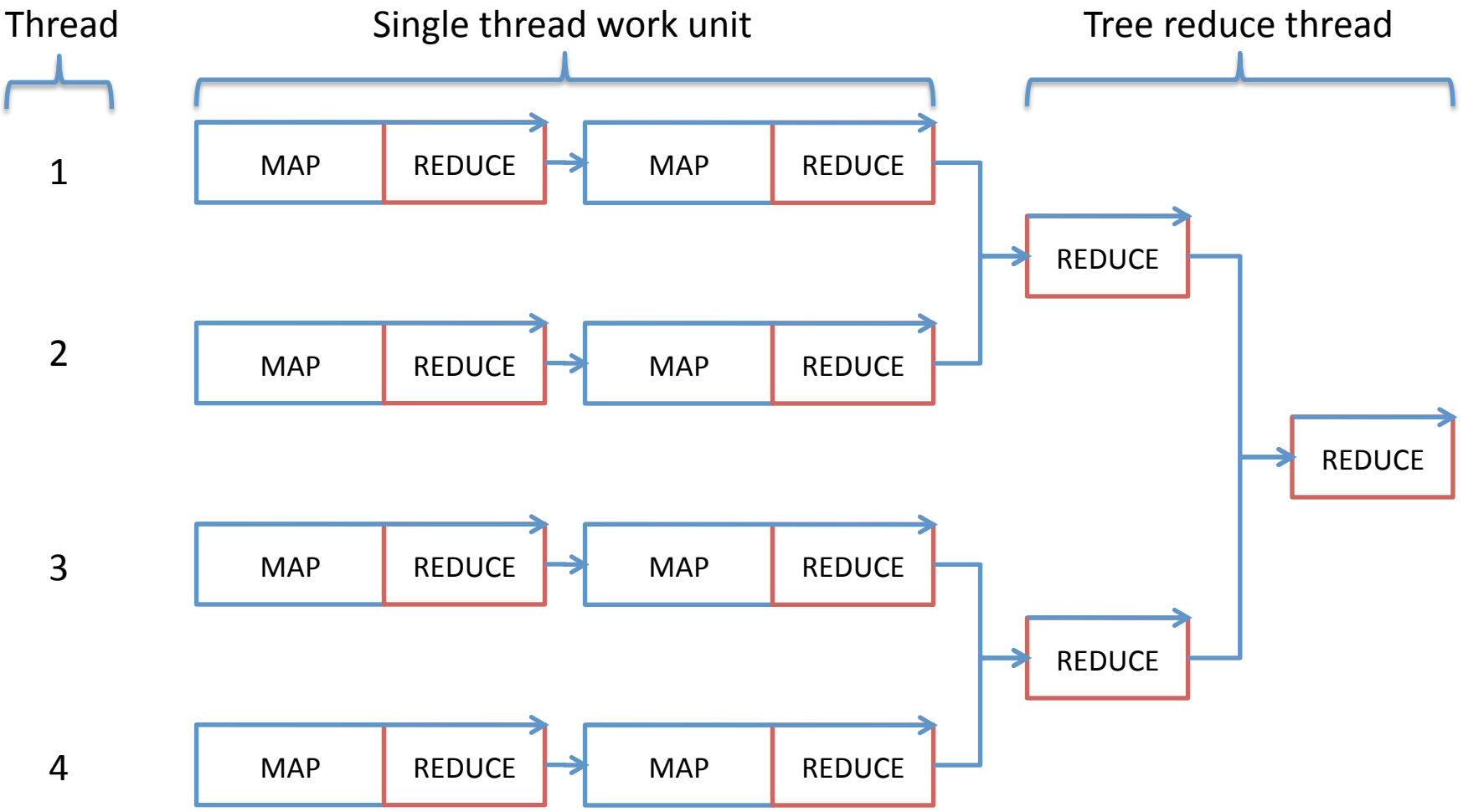
dbSNP track

Output

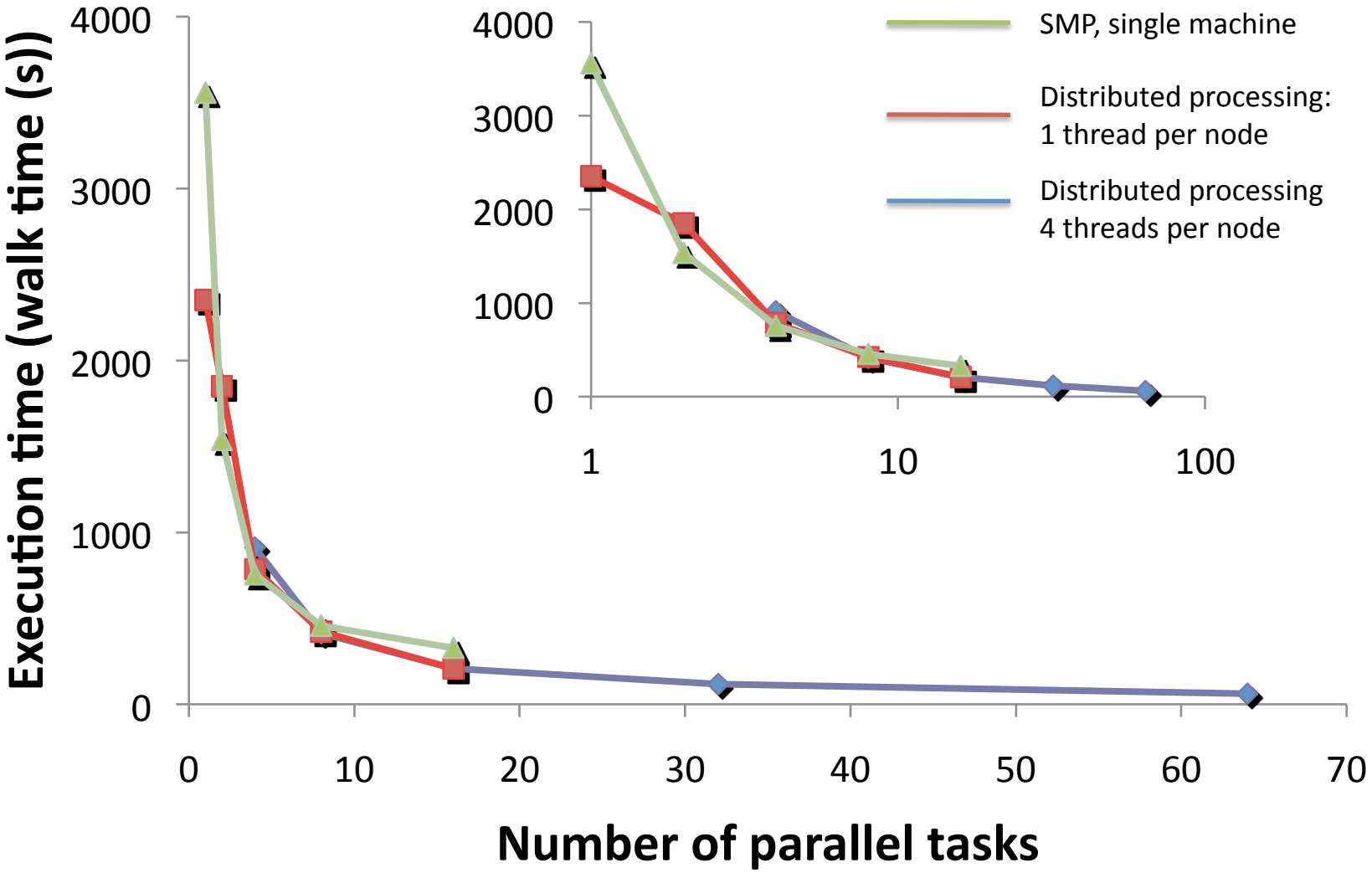
```
Sort order is: coordinate
chr1:559844: C CCCCCCCCTGGCTCCCCCCCCCAGCCCTCCCCCCCACCCCCCCACCCCCCCCCCCCCCC 4;6@@2;?&'(8(-00=??6@31)@)<).@?6?
3/18?(=833.;(<?:@?9?>*95)>
chr1:559845: A AAAGACAAAAAAAAGAAAAAAAAAAAAAACAAAAAAAATAAAAAAAAAAAAAAA ,>?&*(5(((8(??)@(>4@2<,
1>=9;8)30<)463((=,4?;???9>>*:5.>
chr1:559846: G AGAACAAAGAAAAAAACGAAAAGGCTAAGTAAAAAACGGGGGGGGGGGGGG *&((5,((@?)@(5)?1;,.><>:.)50<#7/),(=/
9?:<>8>=3/1(> [ROD: chr1:559846-559847:rs2096047:A/G:SNP:Hapmap:2Hit]
chr1:559847: A AAAAAAAAAAAAAAAAAAAAAAAAAACAAATAAAAAAAAAAAAAA 4:=@?)?(30@);).>>>:81>8<0#>09*>,4?>@>6>=7(3>
chr1:559848: A AAAAAAAAAAAAAAACAAAGAAATAAAAAAAAAAACAAAA )@()0@)=).9>1:7)>-<#4>)(>=/1??<>6>=659)>
[PROGRESS] Traversed 81 loci in 9.98 secs (123222.22 secs per 1M loci)
Traversal reduce result is 5
```

chr1:559846 is a heterozygous A/G site, consistent with hapmap

Tree reduce parallelism framework



Automatic parallelization in the GATK



Single sample genotyper on chr20 30x SLX reads for NA12878 (1000 genomes)

Getting and using the GATK

- Visit our wiki

  http://www.broadinstitute.org/gsa/wiki/

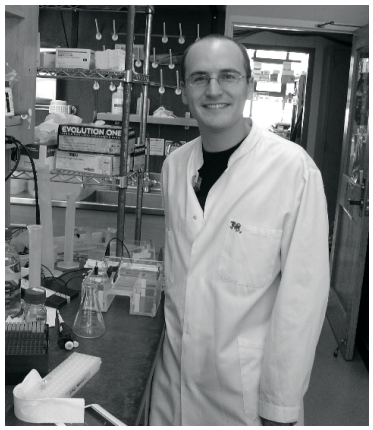
  – Has developer documents describing how to build the system and read the “hello reads” tutorial

  – Download binary Jar as well as publically available tools

- Check out source from SVN repository:

  https://svnrepos.broadinstitute.org/Sting/

Core GATK development team



Mark DePristo
depristo@broad

Matthew Hanna
hanna@broad

Aaron McKenna
aaron@broad

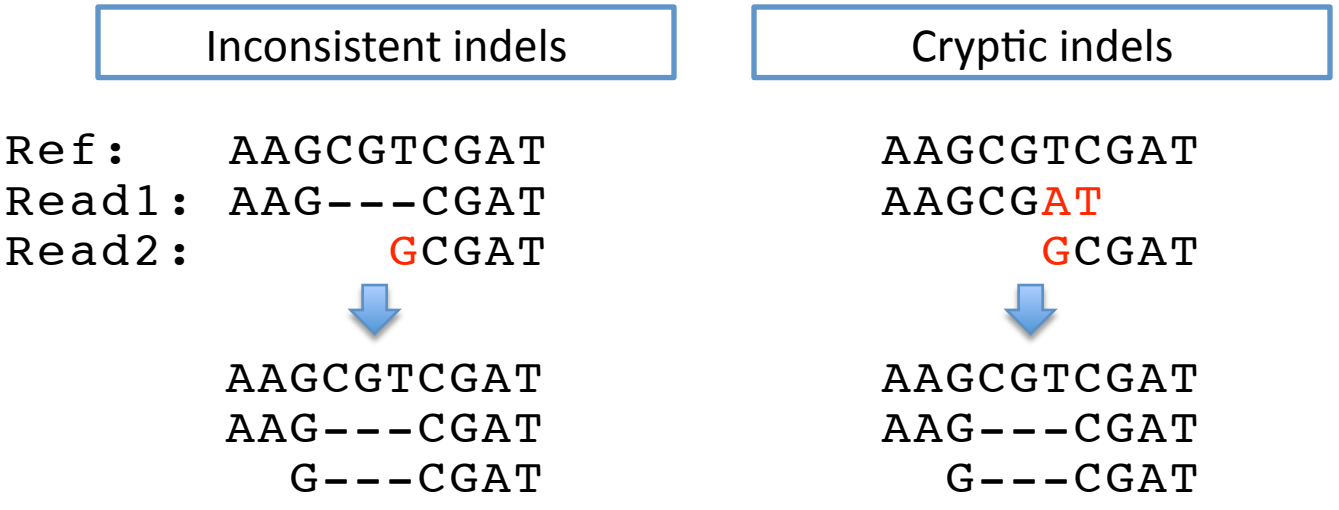
- We are looking for feedback, bug reports, feature requests, brainstorming sessions, etc. to make the system as powerful and easy-to-use as possible
- Please understand that the system is in active development, it's usable but interfaces, functionality, etc., are continuously changing and improving

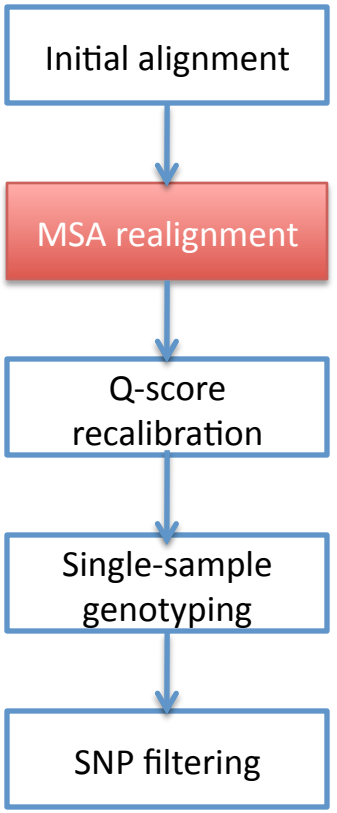
Agenda

- Introduction to the 1000 genomes project
- Mapping and alignment
  - SAM/BAM format
  - Visualizing the data
- The Genome Analysis Toolkit
  - The infrastructure supporting our tools for working with next-generation sequencing data
- Tools developed in the GATK for calling SNPs and indels in the 1000 genomes pilot

Multiple sequence realignment

- Read-by-read mapping introduces artifacts that can only be resolved by examining multiple reads within their local context

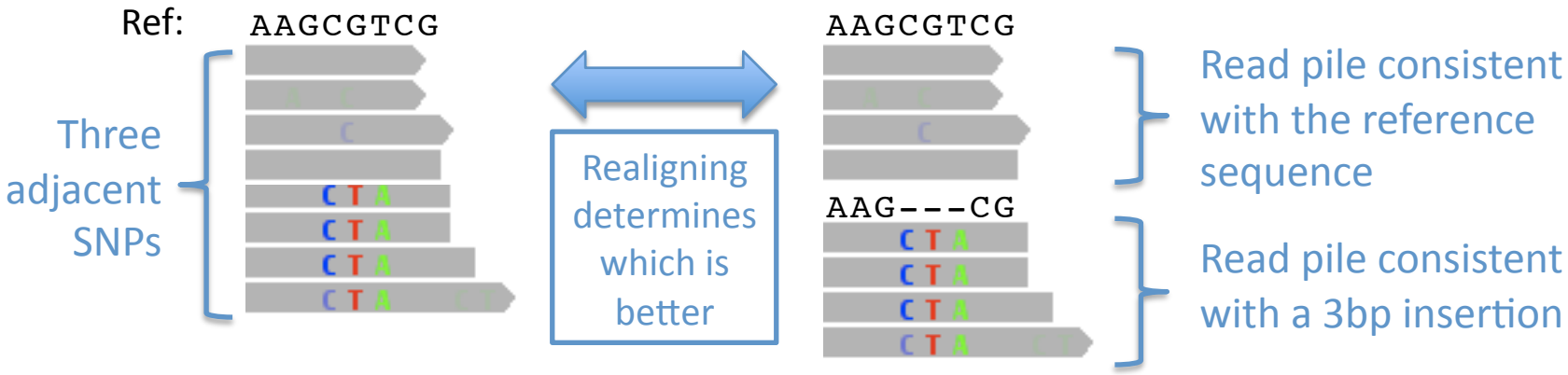




Bases mismatching reference in red

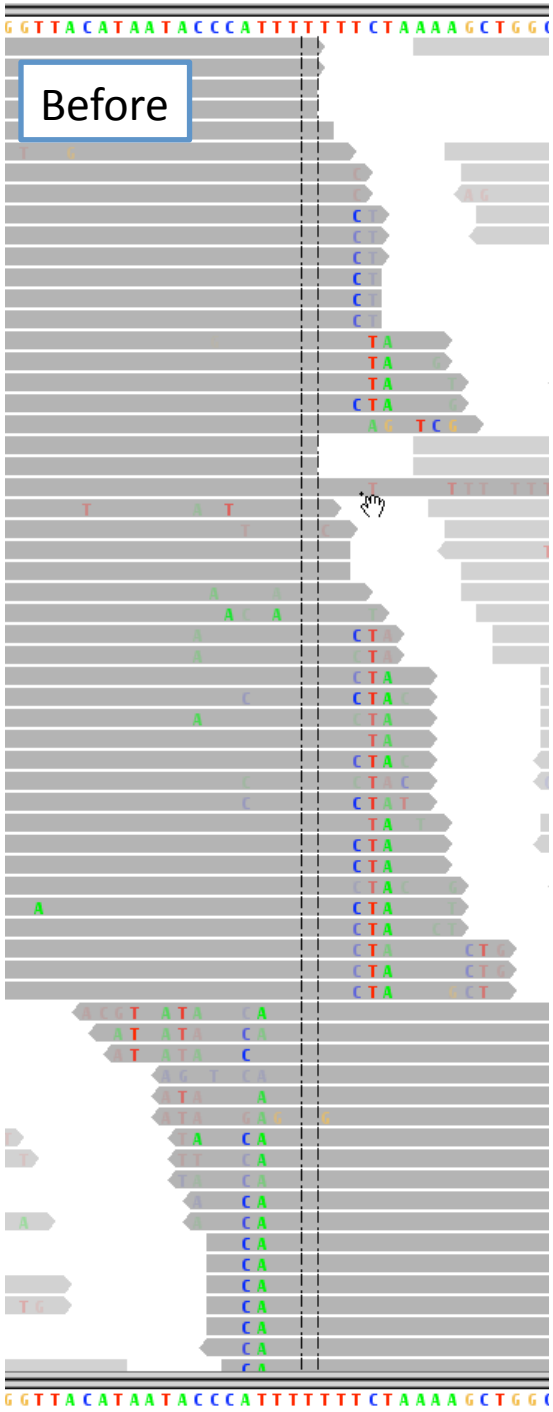
Local realignment identifies the most parsimonious alignment along all of the reads at a problematic locus

1. Find the best alternate consensus sequence that, together with the reference, best fits the reads in a pile (maximum of 1 indel)



2. The score for an alternate consensus is the total sum of the quality scores of mismatching bases

3. If the score of the best alternate consensus is sufficiently better than the original alignments (using a LOD score), then we accept the proposed realignment of the reads



Local realignment uncovers the hidden indel in these reads and eliminates all the potential FP SNPs

Local realignment enabled us to find ~90% of short indels with ~70% specificity in a blind simulation assessment

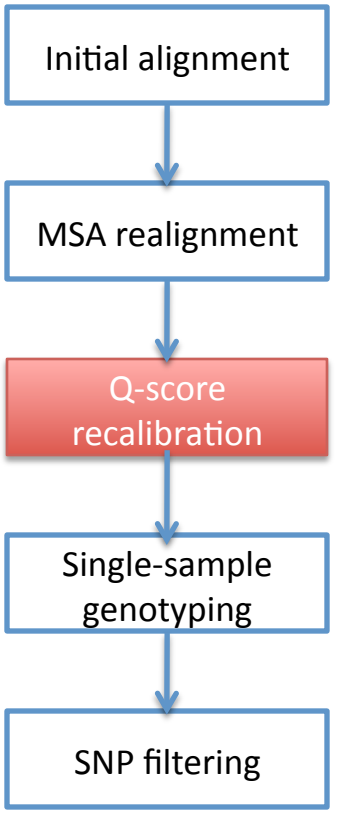
Modeling the error process

- An accurate error model is essential for reliable downstream analyses such as SNP calling

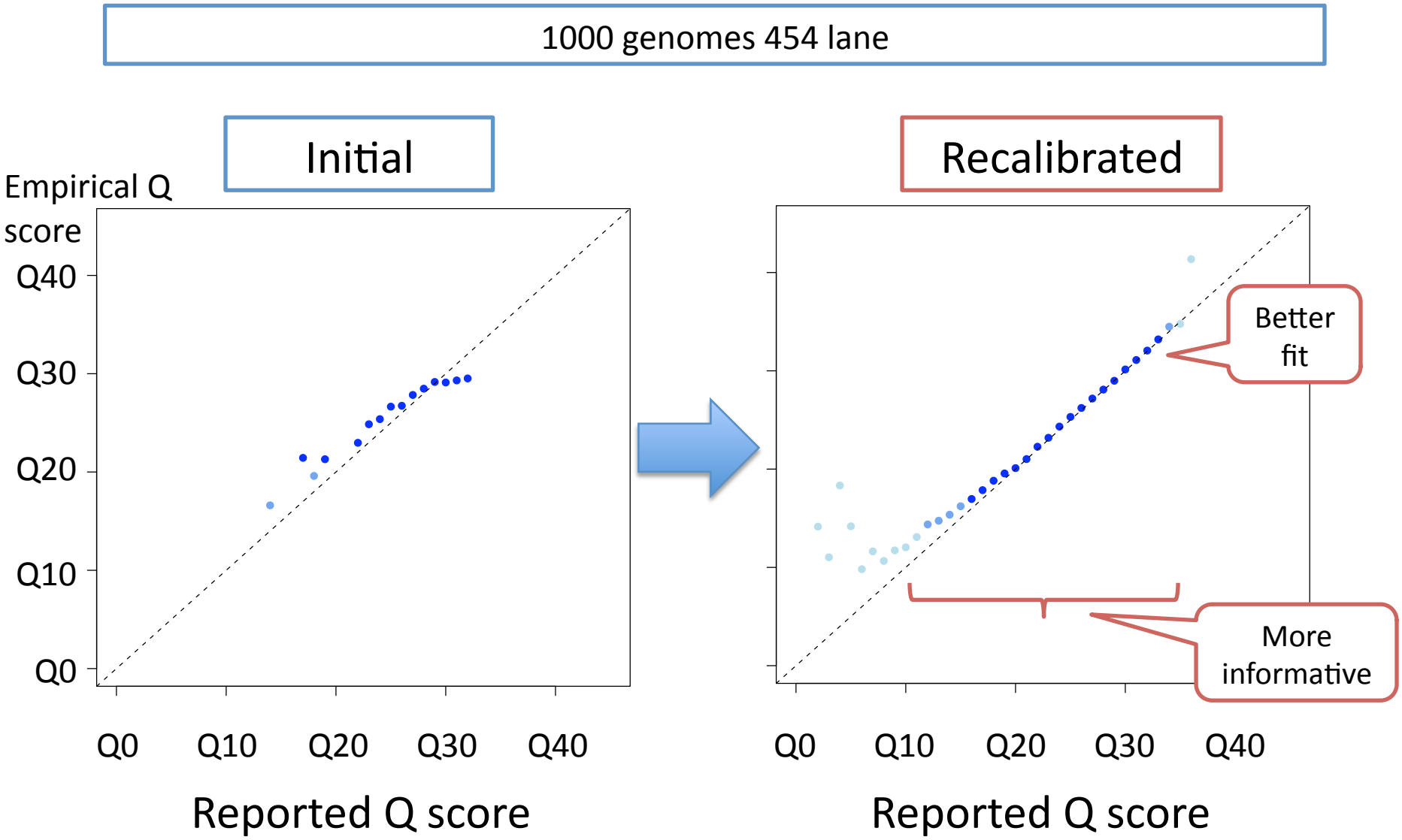
  Pr{ observing base *b* | true genotype is G }

  – What is the probability that *b* (e.g., A) is actually some other base (e.g., either, C, G, or T)?
    - This prob. is encoded by the phred-scaled quality score

- The quality scores reported by the Solexa, SOLiD, and 454 base callers are inaccurate
  – To correct them, we examine the aligned reads and use the reference mismatch rate at non-dbSNP sites to recalibrate the reported quality scores
  – We can also account for covariates of base errors, such as local sequence context and machine cycle, to identify subsets of higher-quality bases



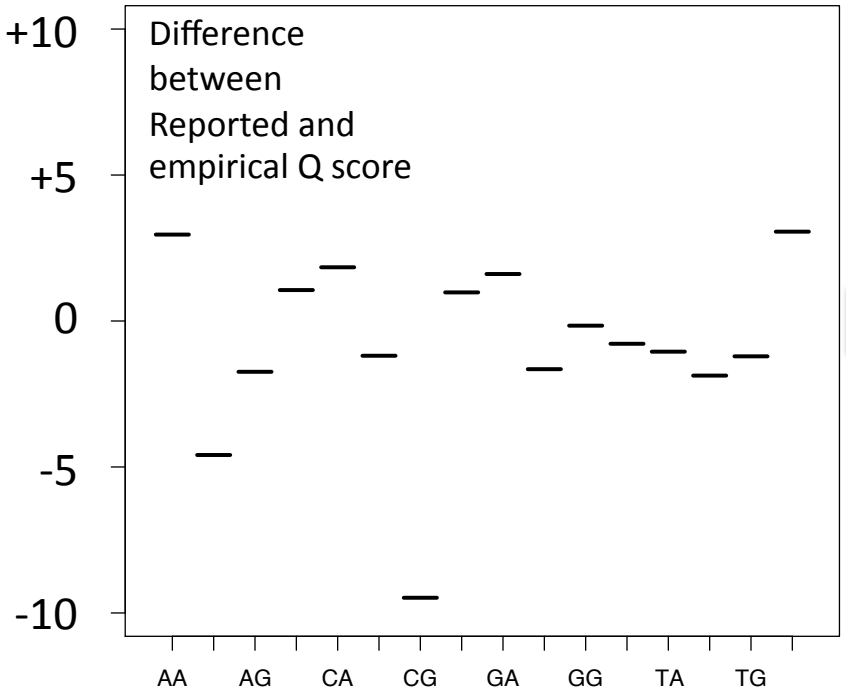
Recalibration make quality scores more accurate



Recalibration removes some error covariates

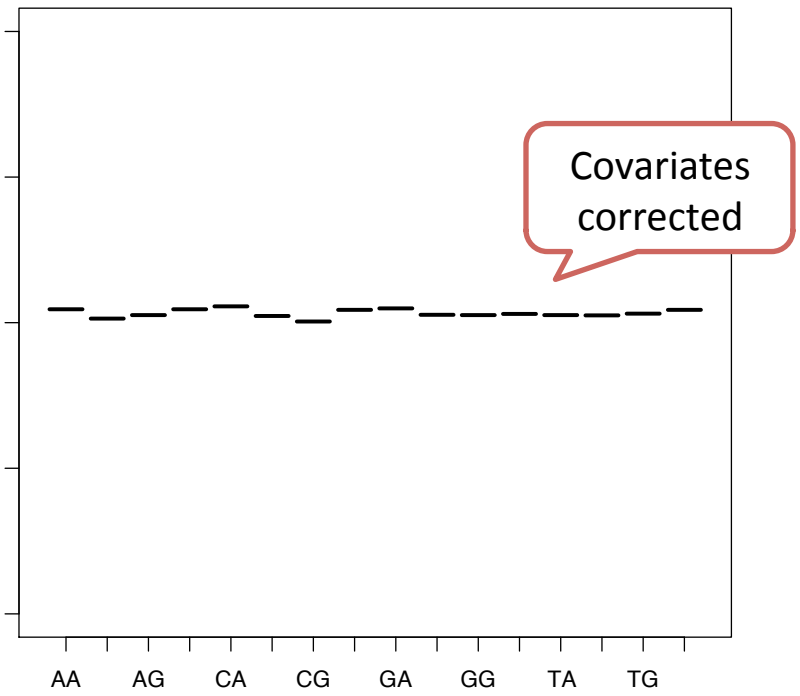
1000 genomes 454 lane

Initial



Dinucleotide context

Recalibrated



Dinucleotide context

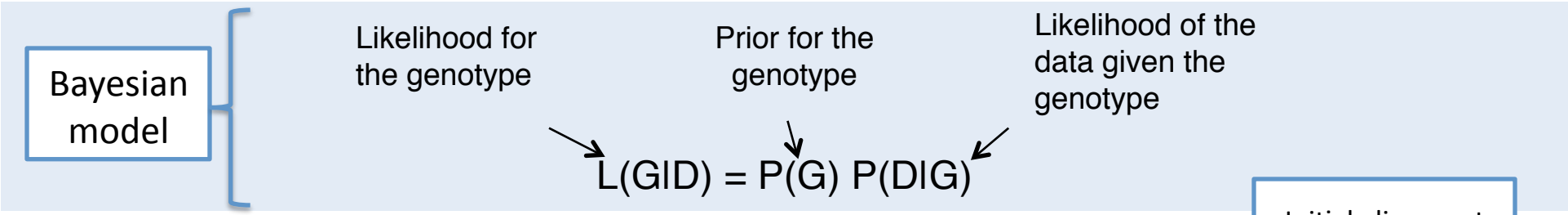
Recalibration identifies high-quality bases and improves SNP calls

1KG 454 lane	Initial	Recalibation
No. bases in lanes	80M	80M
Lane wide reported Q	28.7	28.2
Lane wide empirical Q	28.4	28.4
RMSE between Q reported and empirical	17,554	9,635
% of true Q25 bases	89%	**95%**
% of true Q30 bases	0%	**53%**

Identifies >50% bases as true Q30

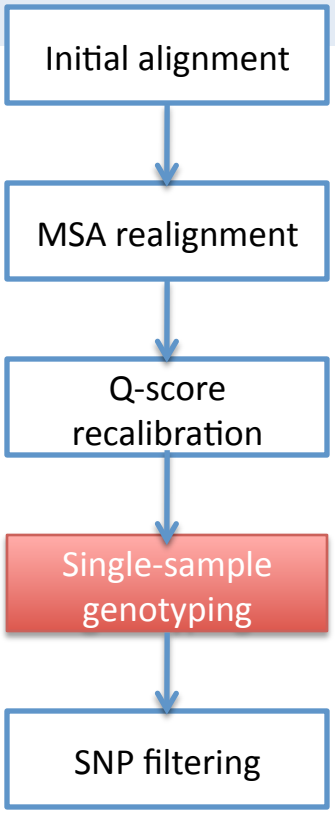
Results in ~10% more SNP calls at same quality compared to unrecalibrated data

Bayesian SNP Caller for Pilot 2



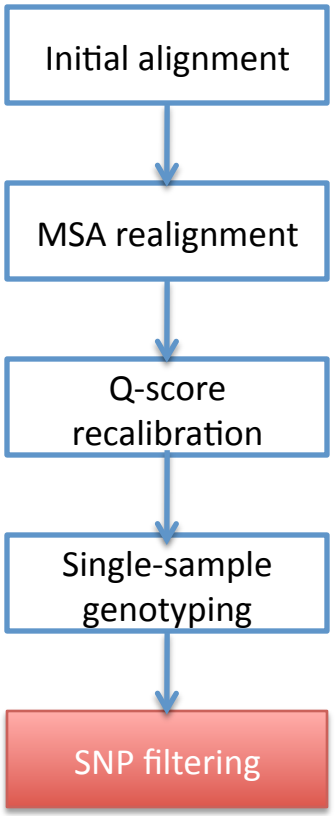
- Prior genotype probabilities enforce variant expectation rates
- Likelihood of data computed using pileup of bases and associated quality scores at given locus
- L(G|D) computed for all 10 genotypes
- Confidence in call given by $lod = \log 10\left(\frac{L(G_{best}|D)}{L(G_{ref}|D)}\right) \geq 5.0$

T=5 is common



Filtering poor SNP calls in pilot 2

- We use a battery of expectation tests to separate likely FP SNPs from our SNP calls
  - This is possible because erroneous SNP calls often result from recurring systematic errors
- We flag a SNP as a likely FP if it exhibits unusual behavior according to:
  - In excessive depth of coverage
  - Occurs preferentially on a single strand
  - Has a skewed allelic imbalance
  - In a region of poor read mapping
  - Occurs in very close proximity to other SNPs



Evaluating SNP call quality

Did I get the right number of calls?

- The number of SNP calls should be close to the average human heterozygosity of 1 variant per 1000 bases
- Only detects gross under/over calling

Concordance with hapmap chip results?

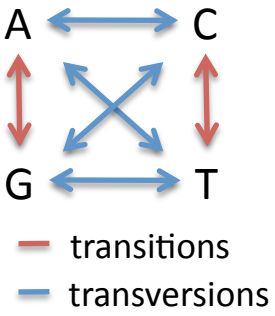
- Often we have genotype chip data that indicates the hom-ref, het, hom-var status at millions of sites
- Good SNP calls should be >99.5% consistent these chip results, and >99% of the variable sites should be found
- The chip sites are in the better parts of the genome, and so are not representative of the difficulties at novel sites

What fraction of my calls are already known?

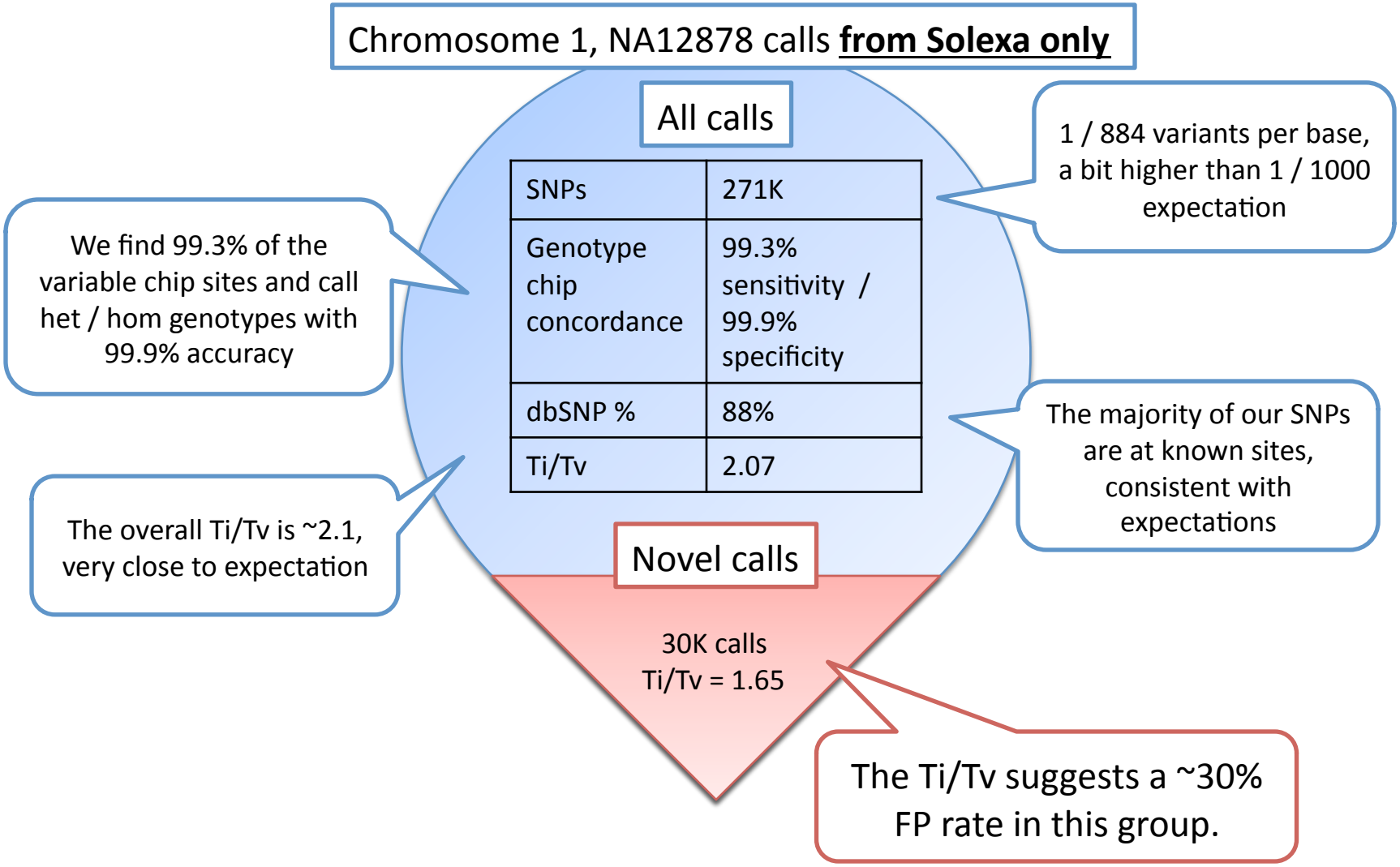
- dbSNP catalogs most common variation, so most of the true variants found will be in dbSNP
- For single sample calls, ~90 of variants should be in dbSNP
- Need to adjust expectation when considering calls across samples

Reasonable transition to transversion ratio (Ti/Tv)?

- Transitions are twice as frequent as transversions (see *Ebersberger, 2002*)
  - Validated human SNP data suggests that the Ti/Tv should be ~2.1 genome-wide and ~2.8 in exons
- FP SNPs should has Ti/Tv around 0.5
- Ti/Tv is a good metric for assessing SNP call quality

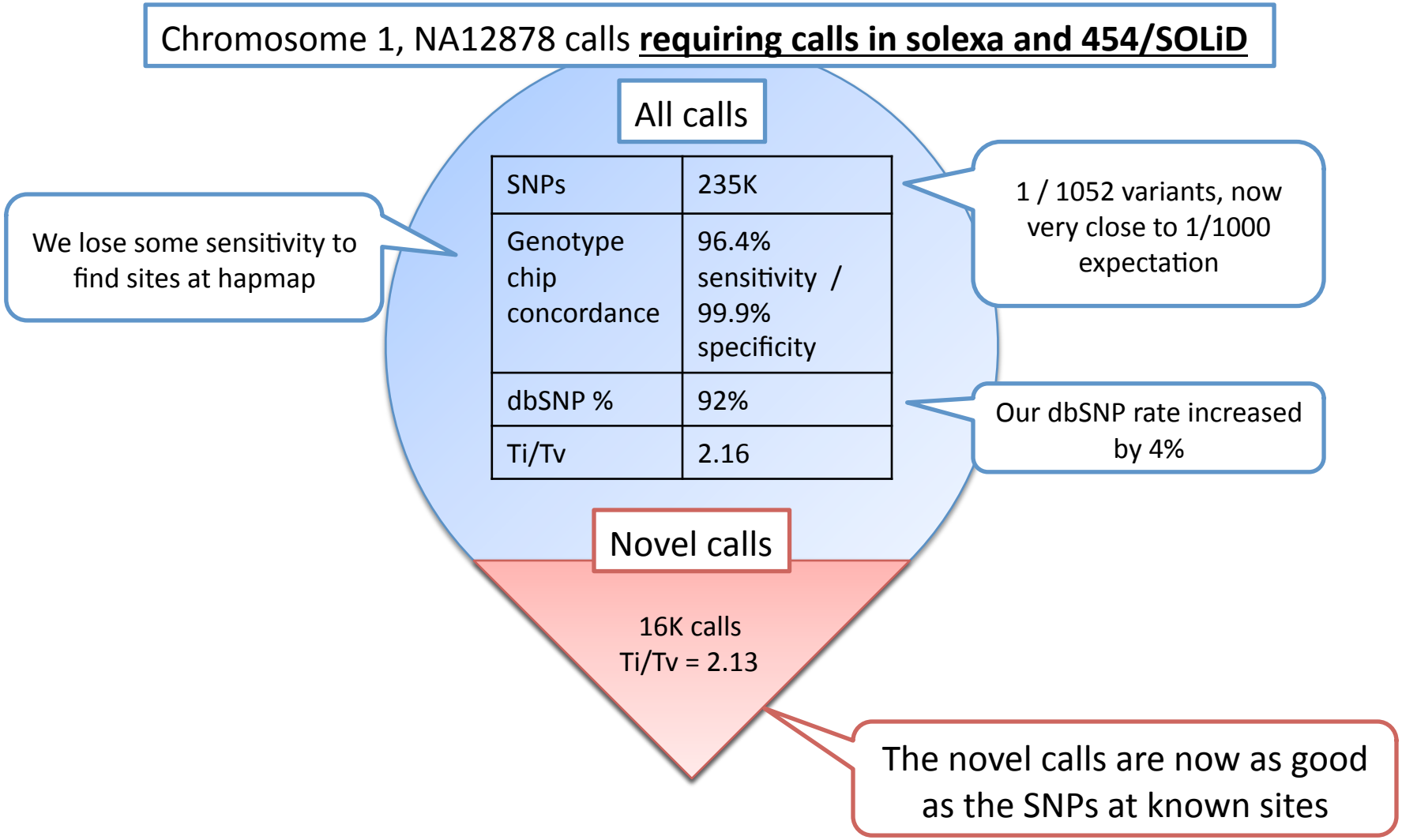


A quality-score aware Bayesian SNP caller produces accurate SNP calls



Calls from recalibrated, indel-realigned Solexa NA12878 with LOD > 5

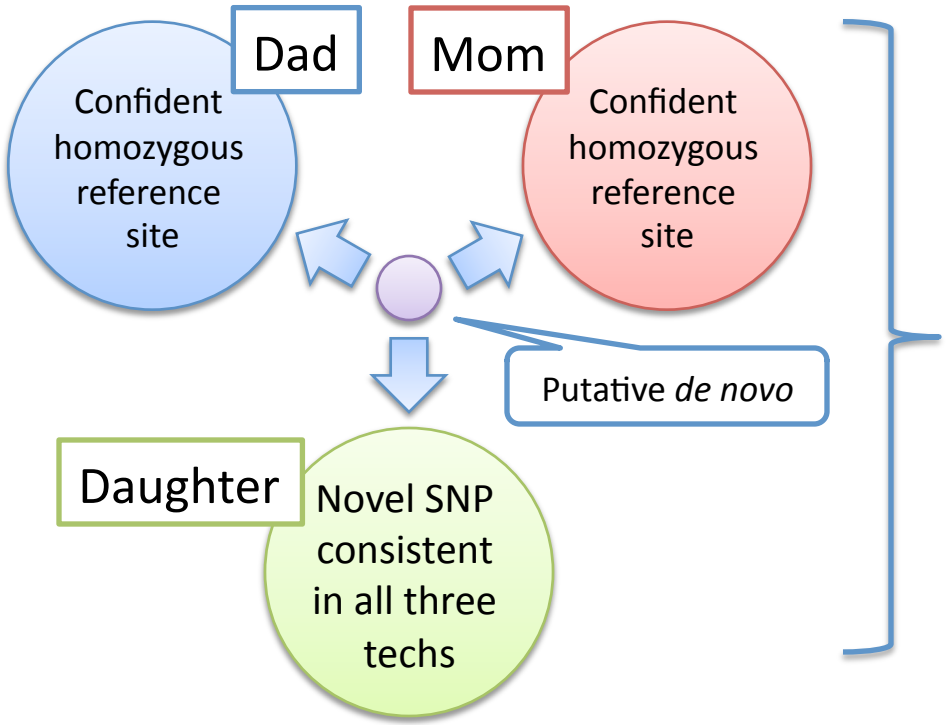
Consistency among SOLiD, 454, and SOLEXA reads enables an even more accurate set of calls



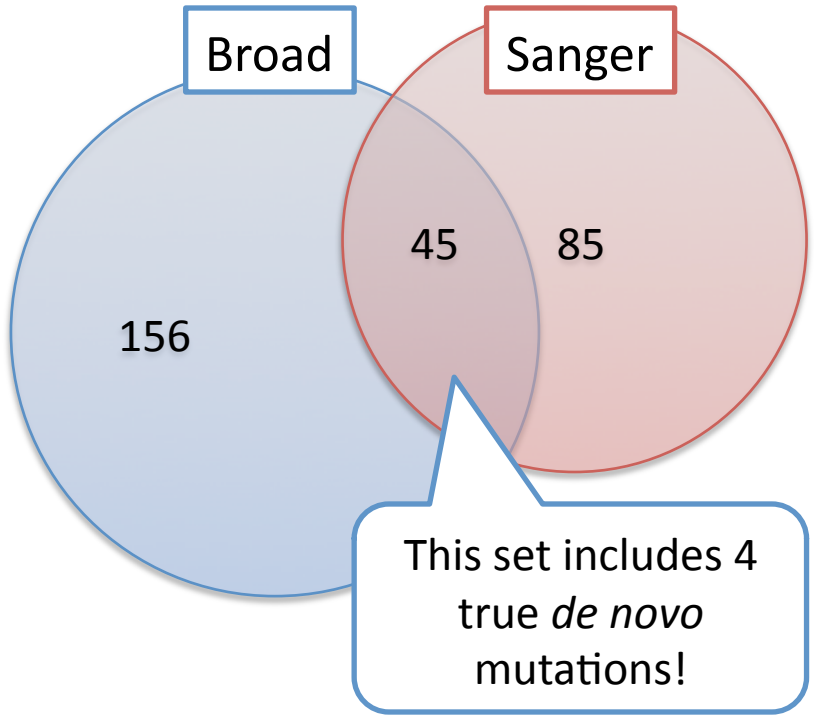
Calls from recalibrated, indel-realigned NA12878 with LOD > 5

Using these concordant calls allows us to identify *de novo* mutations

Algorithm for identifying putative *de novo* mutations

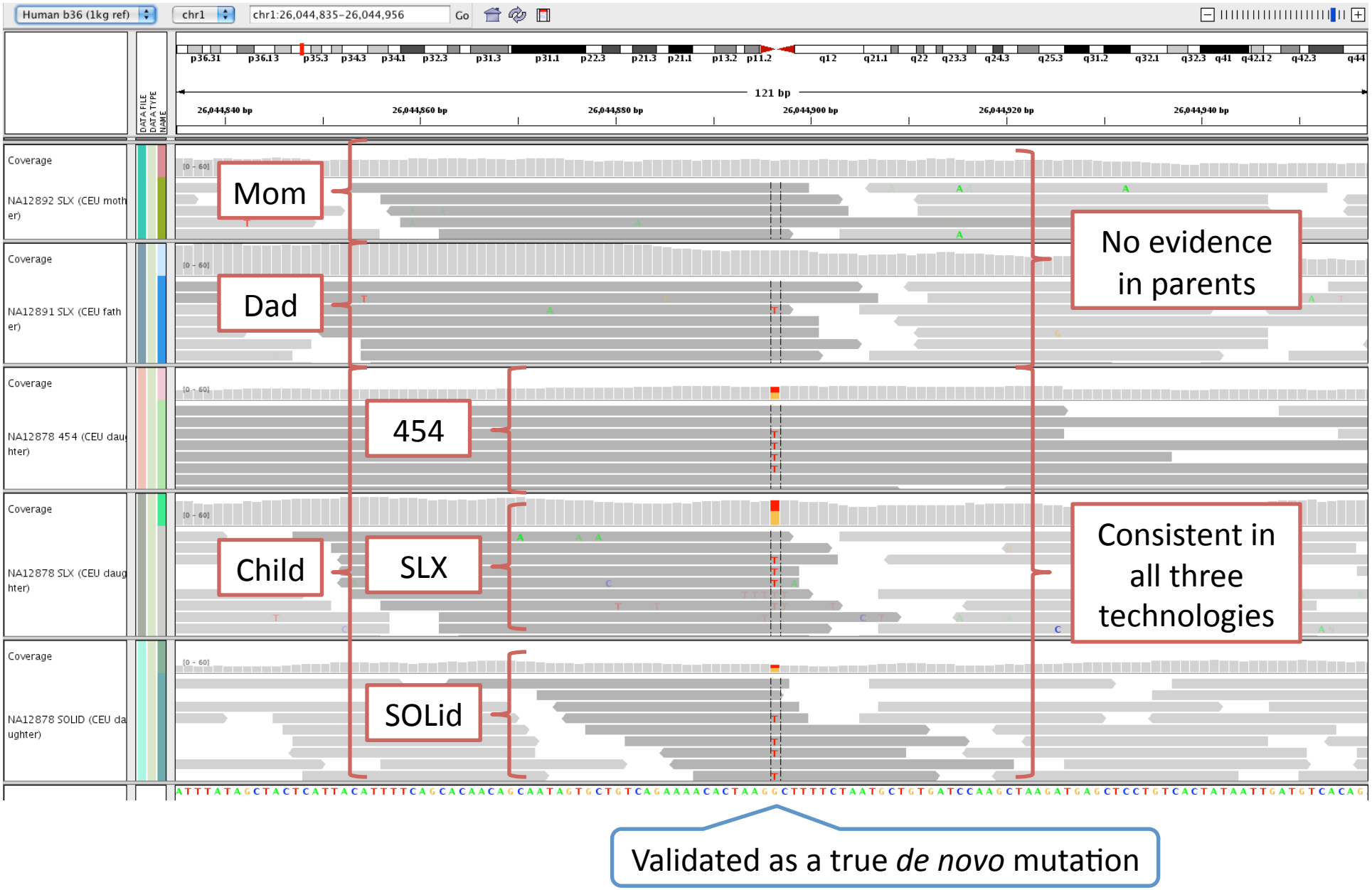


De novo mutation calls from chr1 of NA12878

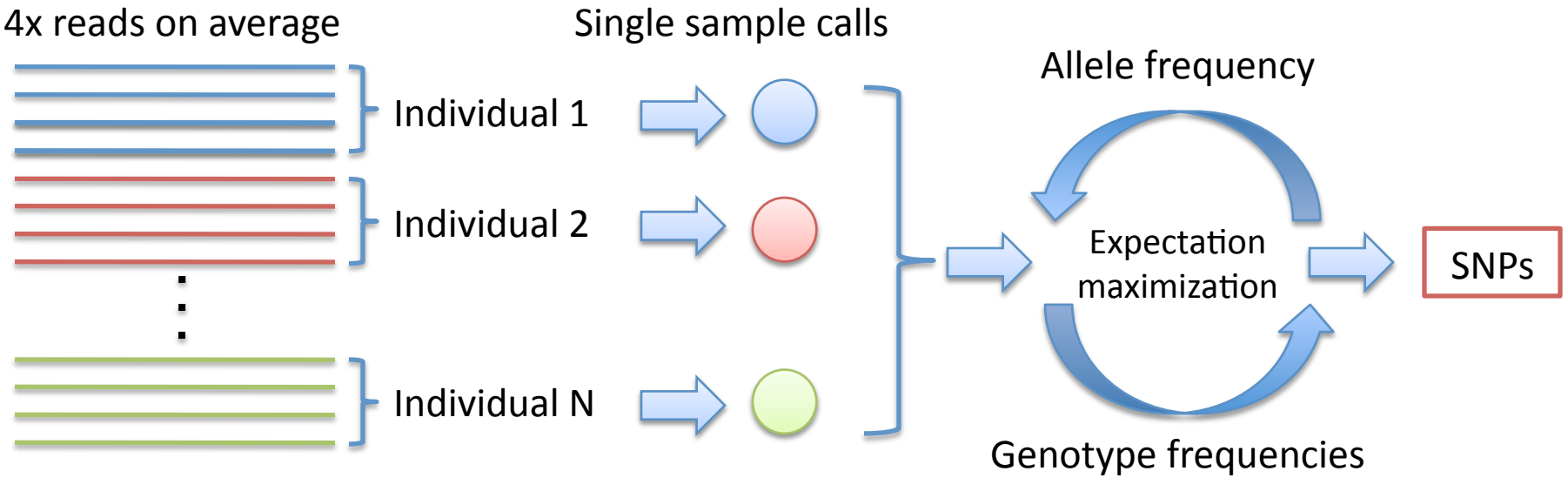


Calls from recalibrated, indel-realigned NA12878, NA12891, NA12892

Validation data courtesy of Matt Hurles and Philip Awadalla


Human b36 (1kg ref)
chr1
chr1:26,044,835-26,044,956
Go
121 bp
Coverage
NA12892 SLX (CEU mother)
Mom
Coverage
NA12891 SLX (CEU father)
Dad
Coverage
NA12878 454 (CEU daughter)
454
Coverage
NA12878 SLX (CEU daughter)
Child
SLX
Coverage
NA12878 SOLID (CEU daughter)
SOLid
No evidence in parents
Consistent in all three technologies
Validated as a true *de novo* mutation

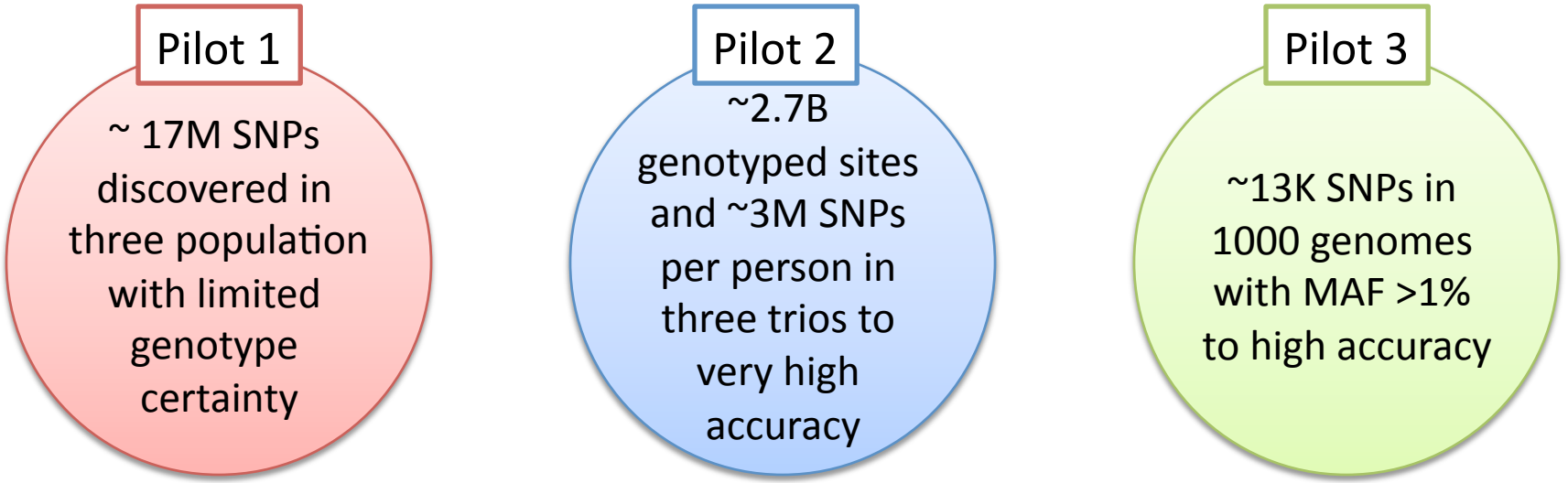
We apply a generalization of the single sample caller to pilot 1



- This approach allows us to combine our poorly determined single sample calls (its 4x after all) to make high-quality population calls
- We have been working with the Sanger (Durbin) and U. Michigan (Abecasis) to make project-wide Pilot 1 calls
  - Other approaches use LD to separate machine errors (which are inconsistent with LD) from true variants (which are)
  - Very powerful but introduces an LD-bias into the call set
  - The best combined approach is still an open question

Work of Jared Maguire and Mark Daly

Available in preliminary form from 1000 genomes



- Preliminary calls have been made for all pilots 1, 2 and 3 by several centers and groups around the world
- All three pilots are proceeding to validation in the next month
  - Final, high-quality calls by November…
  - Publication and public release in December…

Help develop and apply methods in NGS to medical genetics projects

- The Genome Sequencing and Analysis group in Medical and Population Genetics at the Broad Institute is hiring

Computational Biologist	Ph.D.-level research scientist focused on algorithmic R&D
Bioinformatic Analyst	B.A./M.A.-level analyst focused on algorithmic R&D
Senior Software Engineer	B.A./M.A./Ph.D in CS with 5+ years of experience to lead MPG software development projects
Software Engineer	B.A. in CS to develop software throughout MPG

Talk to me for more information or email depristo@broadinstitute.org