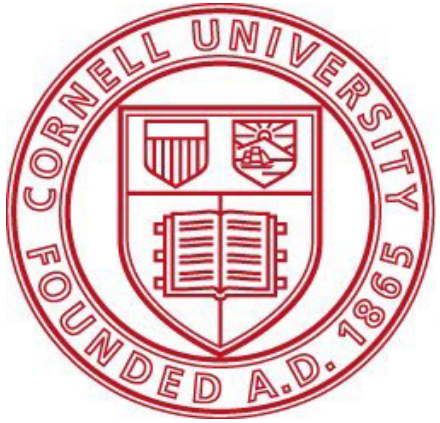


Session 2, Lecture 2:

Variant Detection by DNA Sequencing

Charles G. Danko

Lab of: Adam Siepel and W. Lee Kraus
Biological Statistics and Computational Biology
Molecular Biology and Genetics

March 16th 2010

Lecture Outline

- **Illumina platform quality scores.**
- **Tutorial for the SAMtools SNP Caller.**
- **Other SNP Callers.**
- **Depth & SNP sensitivity.**
- **Tools for calling Structural Variants.**

Types of Human Genetic Variation

- **Single nucleotide differences (SNPs).**
- **Small insertions and deletions (Indels).**
- **Copy number variants (CNVs).**

Determining Variants w/ Next-Gen Tech

Advantages

- Whole-genome coverage.

Weaknesses

- Significant error rates.
- Expensive; hard to do many individuals.
- Data files are quite large, and somewhat difficult to work with.

Software For Calling SNPs & Indels

Illumina:

- **MAQ/SAMtools (Heng Li & R. Durban)**
- SOAPsnp (BGI Shenzhen,China)
- GigaBayes (Marth Lab)

Roche/454:

- Native 454 SNP Caller (Roche/454)
- PyroBayes (Marth Lab)
- ProbHD (Blanchette Lab)

Quality Scores & Variant Detection

Confidence determined by the quality scores.

Quality Scores in SNP Calling

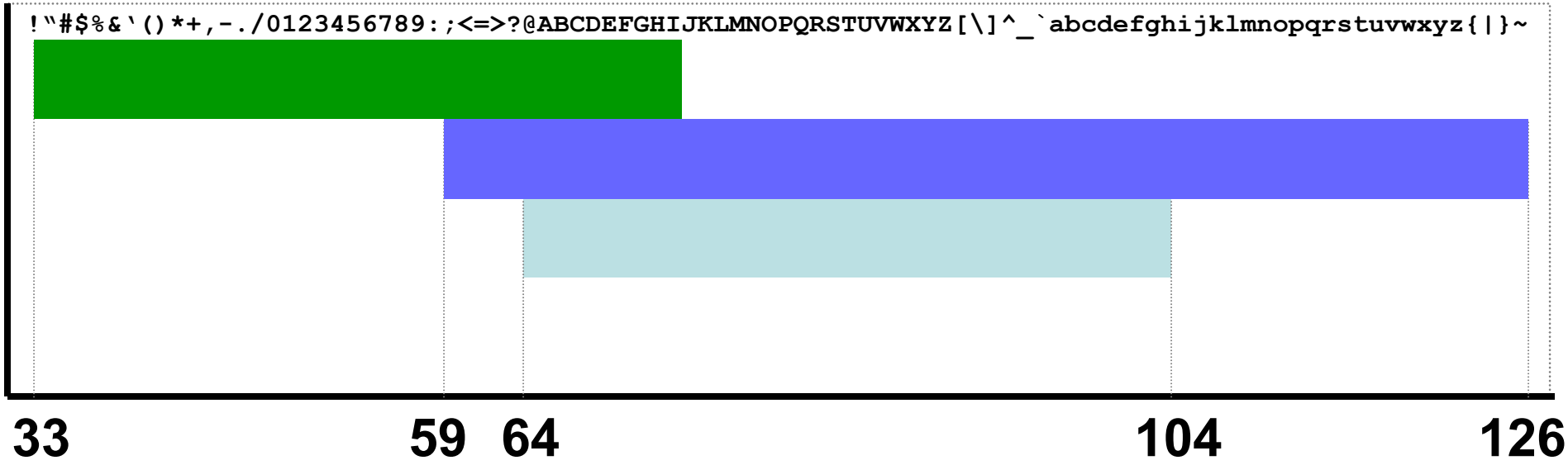
FastQ file

```
@SRR001666.1 071112_SLXA
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh[]M_^O
```

IMPORTANT:

- Base quality is an integral part of SNP calling.
- Most software assumes Sanger format.

Illumina Quality Score Formats



ASCII Value Range

Key	Type	Score	Typical Values
	Sanger	Phred+33	41 values (0, 40)
	Solexa	Solexa+64	68 values (-5, 62)
	Illumina 1.3+	Phred+64	41 values (0, 40)

http://en.wikipedia.org/wiki/FASTQ_format

Converting Quality Scores to Sanger

Using MAQ:

- ***Solexa pre-1.3***

  MAQ: http://maq.sourceforge.net/

```
$ maq sol2sanger solexa.txt sanger.fastq
```

- ***Illumina 1.3+***

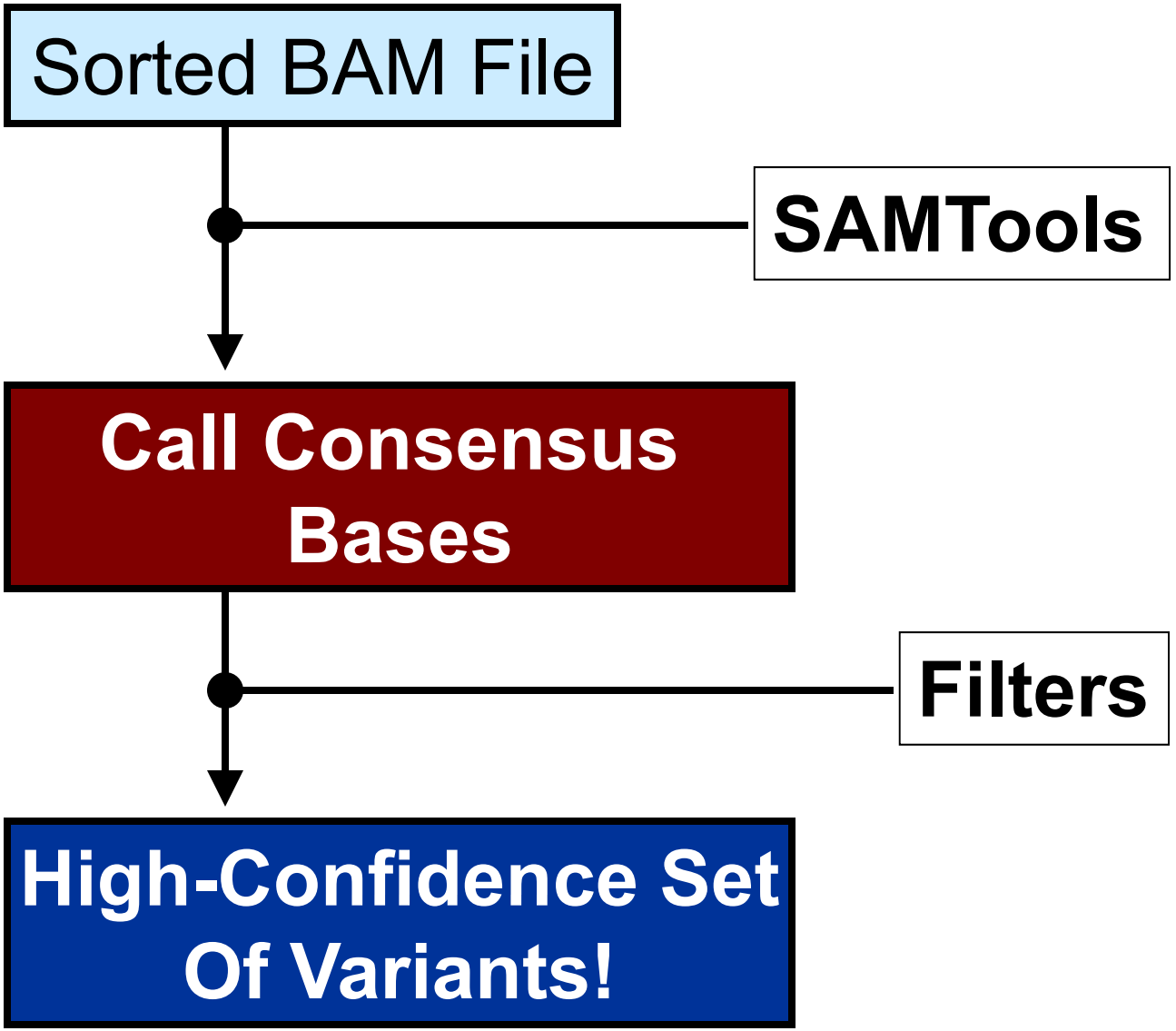
  MAQ patch at: http://tiny.cc/uPiO8

```
$ maq ill2sanger illumina.txt sanger.fastq
```

Other tools (for programmers):

BioPython, BioPerl, and BioRuby

Calling Variant Positions

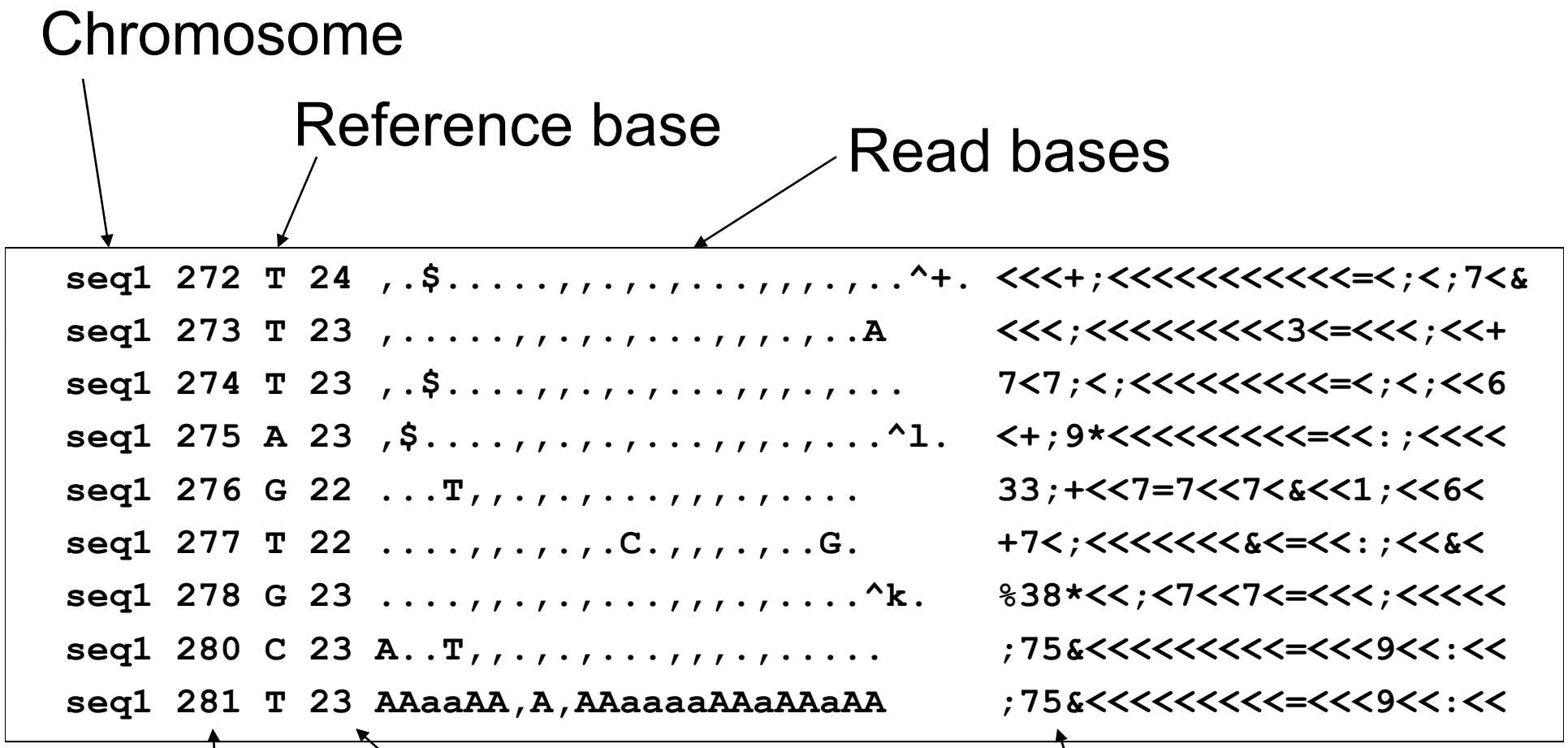


The Pileup Format

Pileup format –

Text based representation in which each row represents all information about a unique position in the reference genome.

Example of a Pileup file



http://samtools.sourceforge.net/pileup.shtml

Characters in the Base Column

Bases:

.	Match to ref genome, forward strand.
,	Match to ref genome, reverse strand.
[ACGTN]	Mismatch on the forward strand.
[acgtn]	Mismatch on the reverse strand.
^{Qual}	The next base is the first in a read; [Qual] denotes alignment quality.
$	The next base is the last base in a read.

Indels:

+[0-9][ACGTNacgtn]	Insertion between this position, and the next.
-[0-9][ACGTNacgtn]	Deletion between this position, and the next.

http://samtools.sourceforge.net/pileup.shtml

Using SAMTools to Extract Variant Calls

SAMTools uses the SNP model in MAQ.

```
$ samtools pileup -cvf ref.fa aln.bam > raw.pileup
```

ref.fa	Fasta formatted file of the reference genome.
aln.bam	Sorted BAM formatted file, from the alignments.
raw.pileup	Output pileup formatted, with consensus calls.
-c	Calls the consensus base at each position.
-v	Show positions that do not agree with ref.fa.
-f	Reference sequence, ref.fa (in FastA format).

http://samtools.sourceforge.net/cns0.shtml

Pileup File With SNP Calls

```
$ cat raw.pileup | more
   ...
chr1 2043842 t G 20 20 15  6   .$,,..G     !@AE)I
chr1 2043906 g A 4  4  0   1   ^!A         E
chr1 2043917 c T 4  4  0   1   T           I
chr1 2044043 t C 10 10 23  1   C           +
chr1 2044047 g T 5  5  23  1   T           &
chr1 2044182 g A 12 12 14  5   A.,^>A^!.   7I'5I
chr1 2044233 g T 27 28 26  4   T,T^!,      %%II
   ...
```

Consensus Base

Phred scaled consensus quality

Maximum Mapping Quality

Phred scaled probability of difference from reference base

Understanding Phred Quality Scores

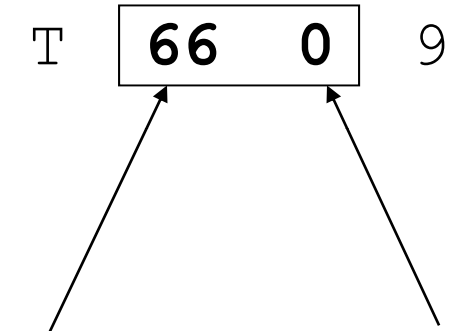
$$Q = -10log_{10}P$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

http://en.wikipedia.org/wiki/Phred_quality_score

Pileup File With SNP Calls

```
$ zcat raw.pileup | more
...
seq1   60   T   T   66   0   99   13   ...
...
```



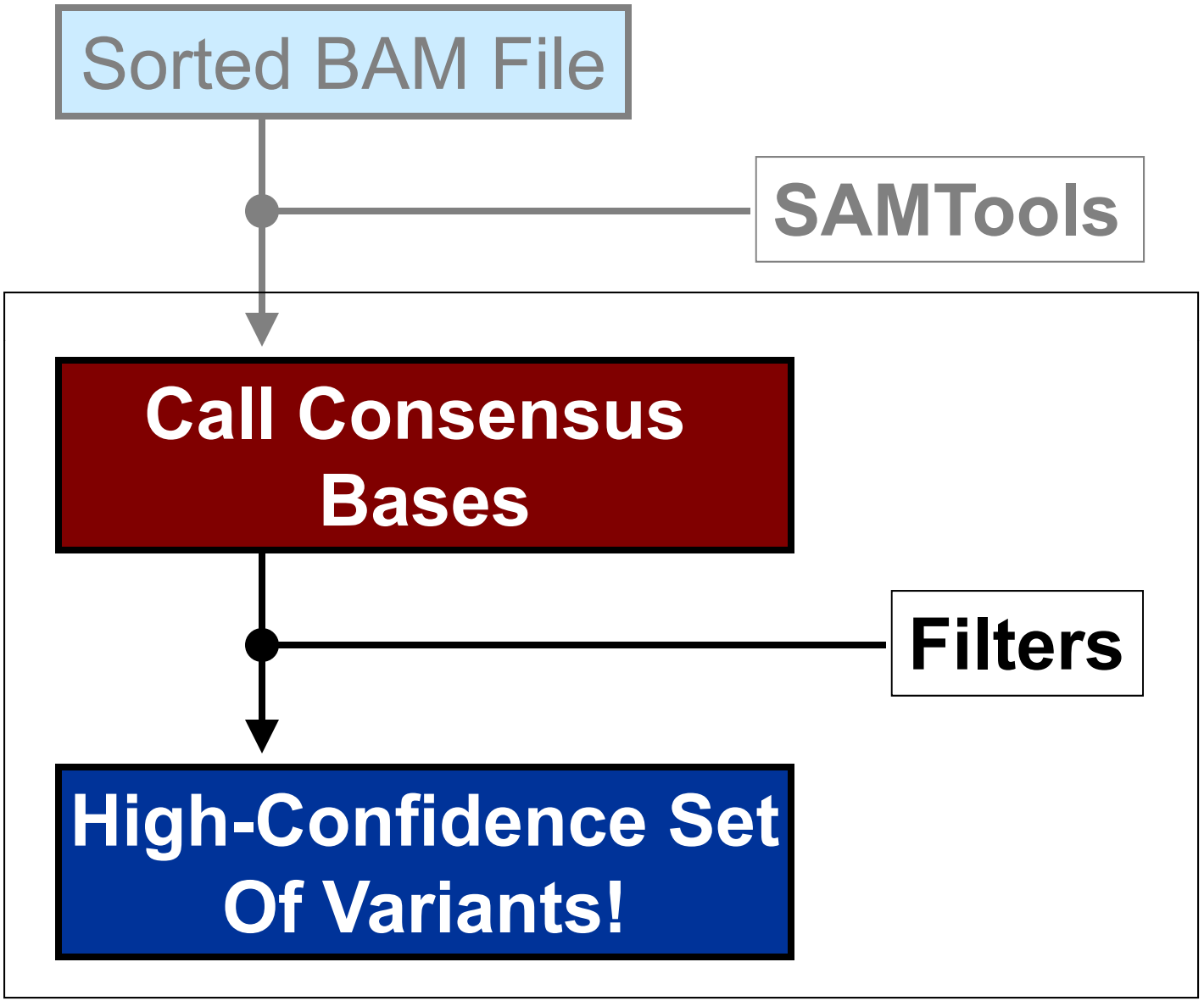
Phred scaled consensus quality

(i.e. Probability that the consensus is wrong)

Phred scaled probability that the position is the reference base

(i.e. 1 - probability of *any* SNP)

Calling Variant Positions



Removing low-confidence SNPs

1. **Apply heuristic filters to remove problematic positions.**

```
$ samtools.pl varFilter raw.pileup | more
```

2. **Threshold the *probability of difference* from the reference base.**

```
$ cat raw.pileup | awk '$6>=20' | more
```

```
$ samtools.pl varFilter raw.pileup [options]
```

Opt	Def	Description
-Q	25	minimum RMS mapping quality for SNPs
-q	10	minimum RMS mapping quality for gaps
-d	3	minimum read depth
-D	100	maximum read depth
-G	25	min indel score for nearby SNP filtering
-w	10	SNP within INT bp around a gap to be filtered
-W	10	window size for filtering dense SNPs
-N	2	max number of SNPs in a window
-l	30	window size for filtering adjacent gaps

Text form: "*samtools.pl varFilter*" output

Thresholding Probability

```
$ cat raw.pileup | awk '$6>=20' | more
```

Returns all lines in raw.pileup for which column 6 >= 20

$1	$2	$3	$4	$5	**$6**	$7	$8	
seq1	60	T	G	66	**21**	99	13	...

Sample Filtering Command

```
$ samtools.pl varFilter raw.pileup | \
    awk '$6>=20' > final.pileup
```

samtools.pl	Perl script in samtools package.
varFilter	Filters variants based on quality filters.
awk '$6>=20'	Simple program; Thresholds the output of the varFilter command to SNPs >= 20.

http://samtools.sourceforge.net/cns0.shtml

Recap

1. Convert quality scores, run alignments & convert to a BAM-formatted file.
2. **Generate pileup file using SAMTools.**
3. **Filter pileup file**
   - **Filter SNPs in error-prone positions using *'samtools.pl varFilter'*.**
   - **Threshold the probability of the reference base.**

Limitations of the current paradigm

Pros & Cons of Threshold Probability

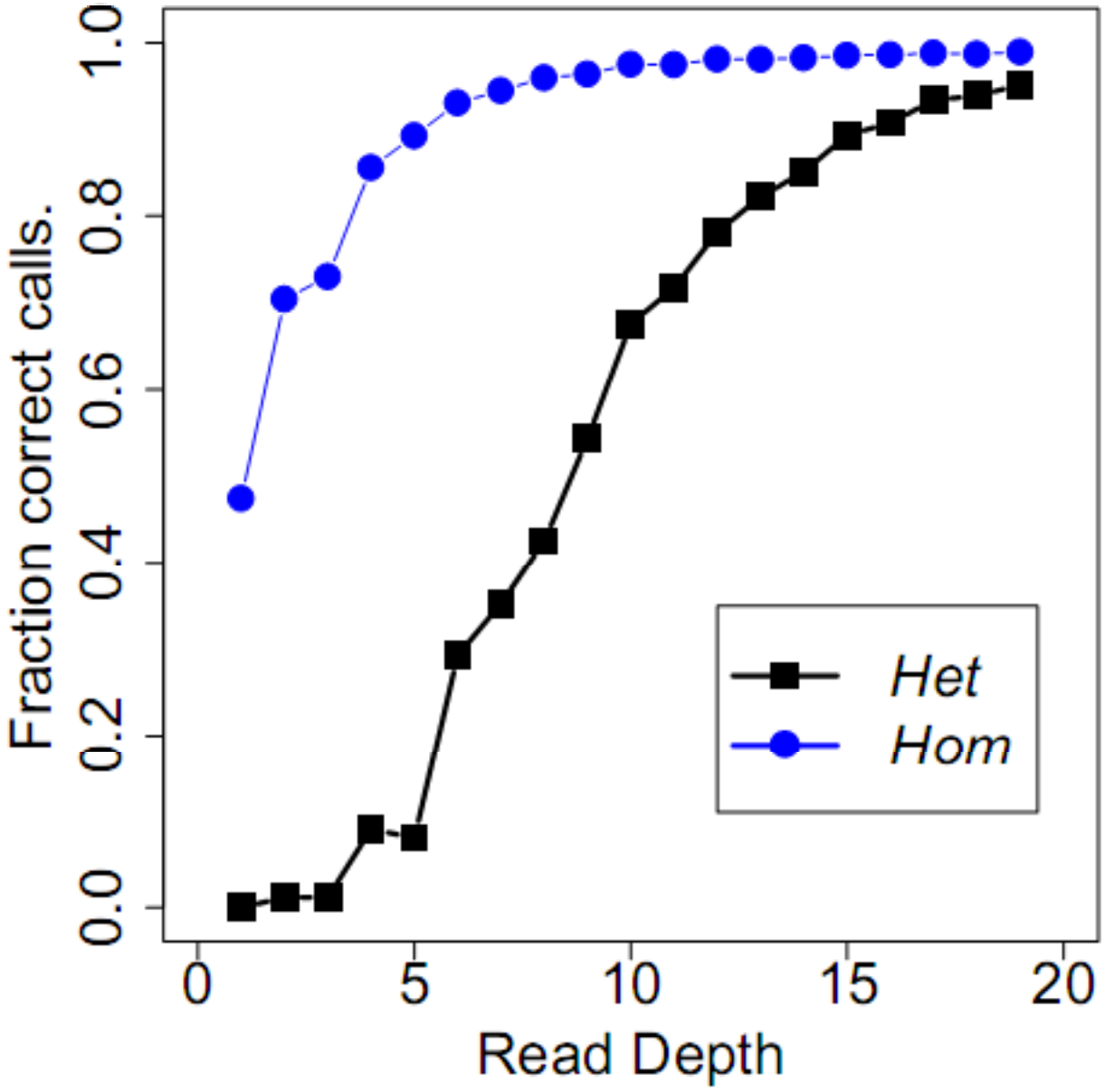
Advantages:

- Limit your false discovery rate.

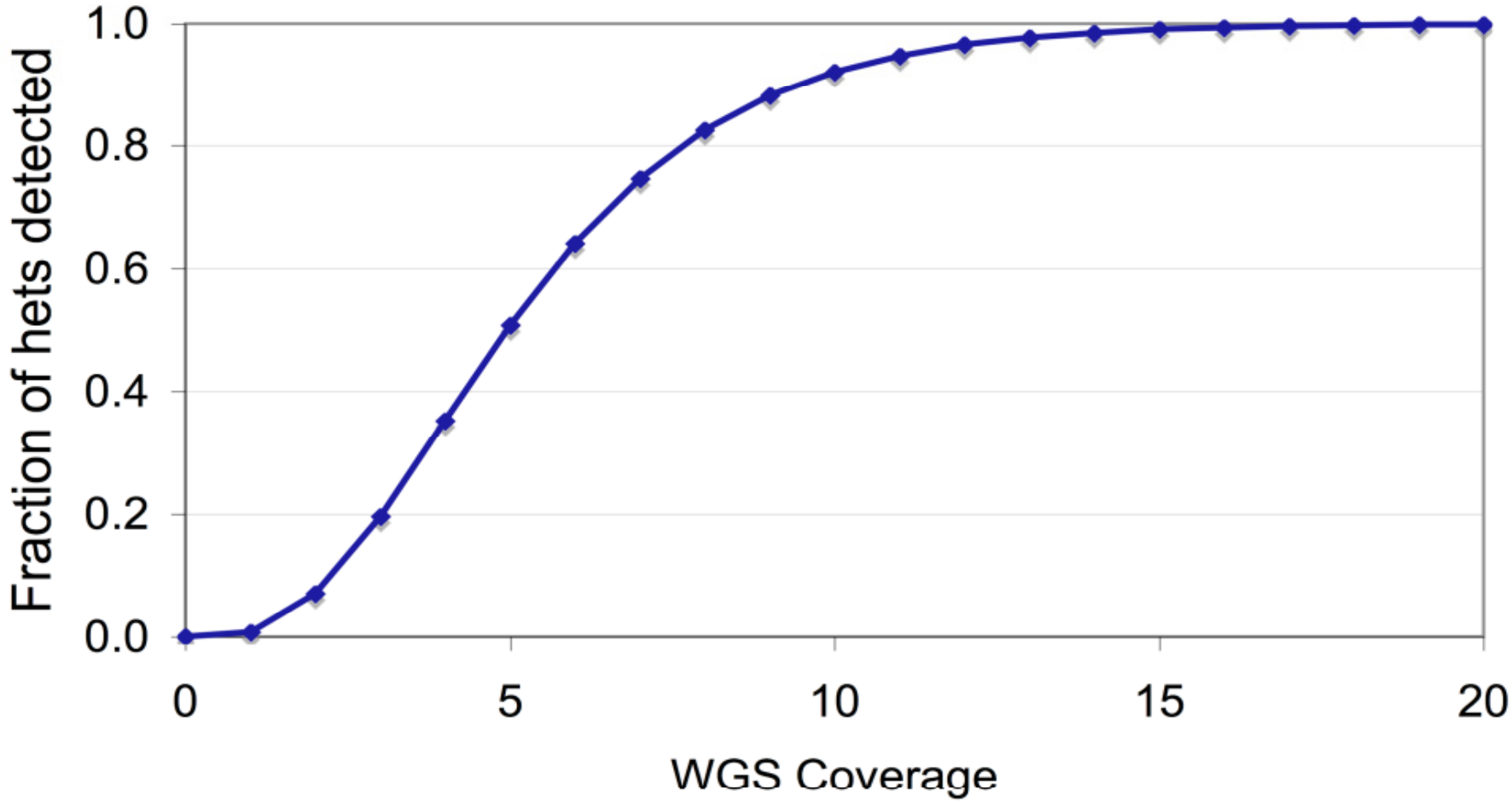
Disadvantages:

- False negatives are difficult to control.
- Implicitly assigning the allele as similar to the reference genome.
- Different use of heuristic filters can create odd biases when comparing genomes.

Effect of Read Depth on Sensitivity



What is your expected sensitivity?



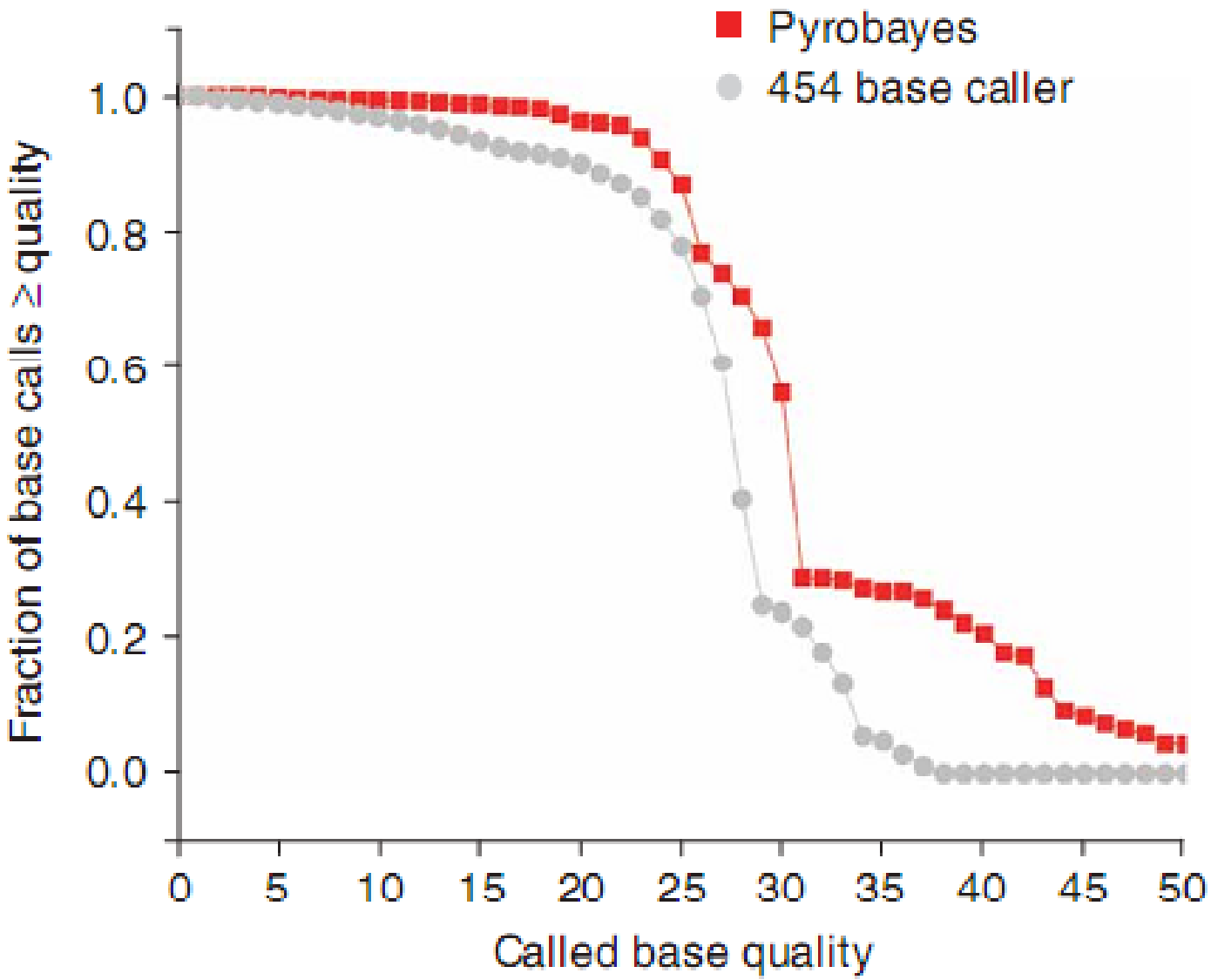
Wheeler DA, et. al. Nature 2008

Other SNP Callers

PyroBayes (454)

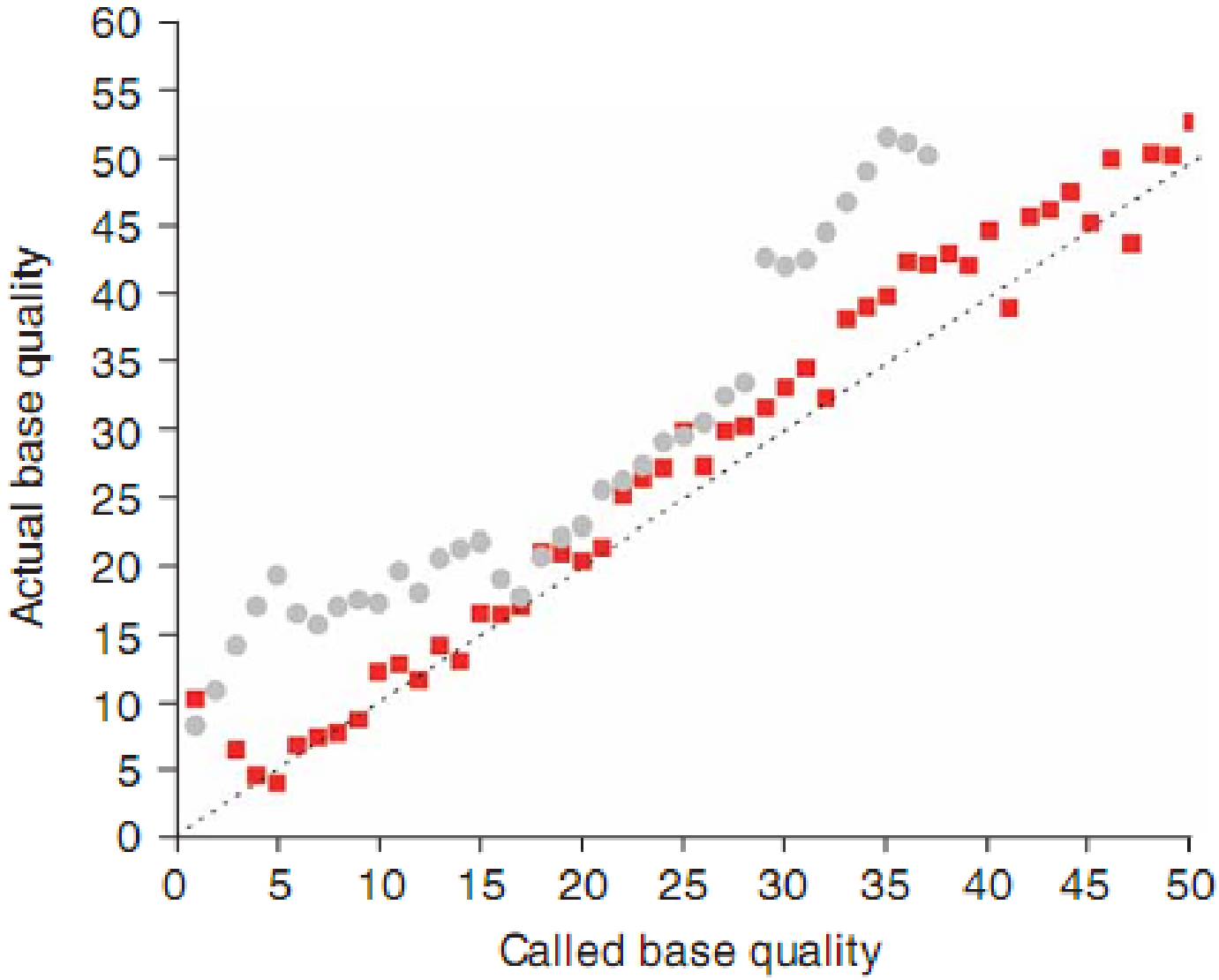
ProbHD (454)

Pyrobayes (454) – Higher Base Quality



Quinlan AR et. al. Nature Methods. 2008

Pyrobayes (454) – Model Matches Data



Quinlan AR et. al. Nature Methods. 2008

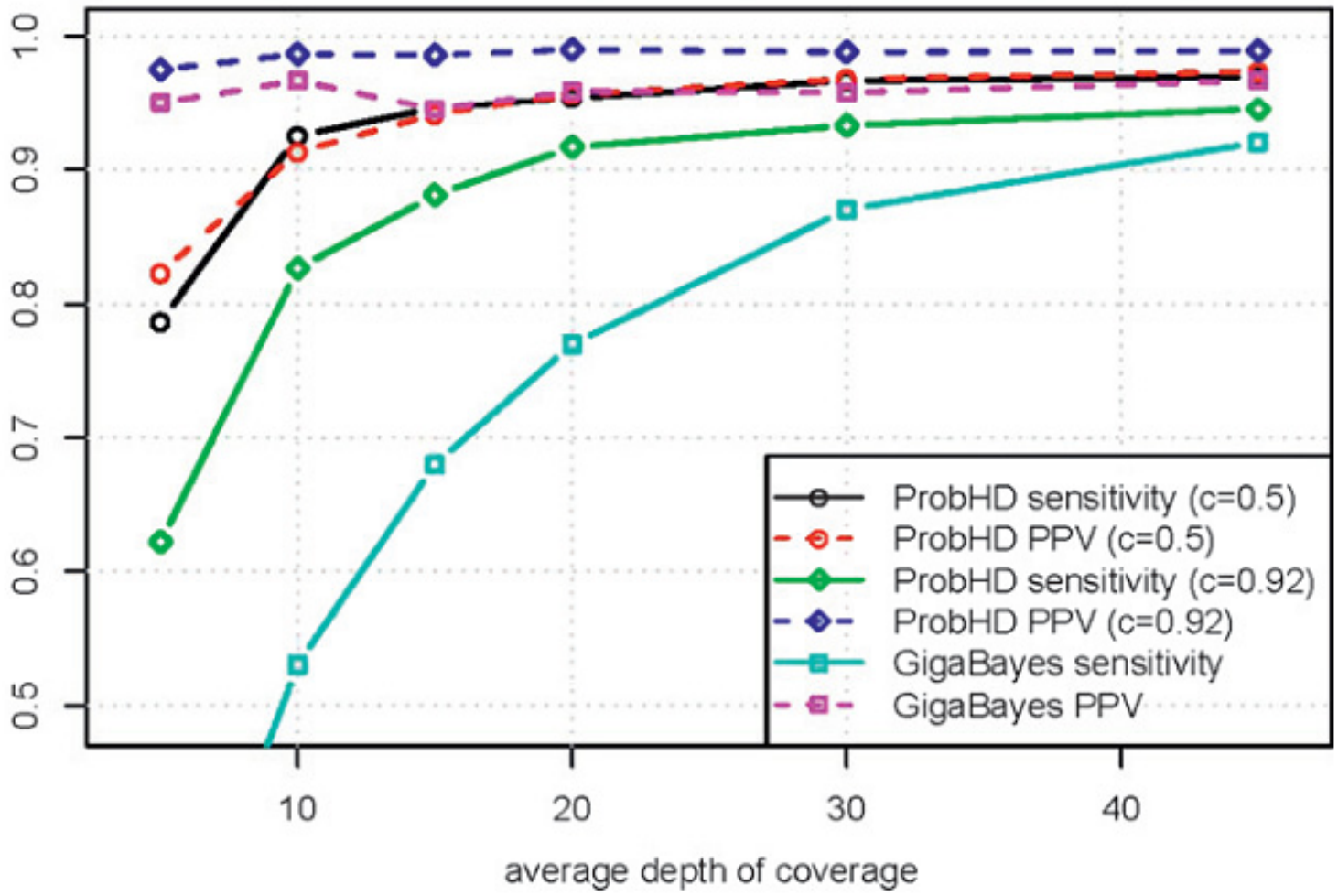
Getting Pyrobayes (454)

Marth Lab Web Site:

http://bioinformatics.bc.edu/marthlab/PyroBayes

Have to register, and then can download a 32 or 64-bit executable for linux.

ProbHD (454) – Higher Accuracy of Het



Hoberman R, *et. al.* Genome Res. 2009

Getting ProbHD (454)

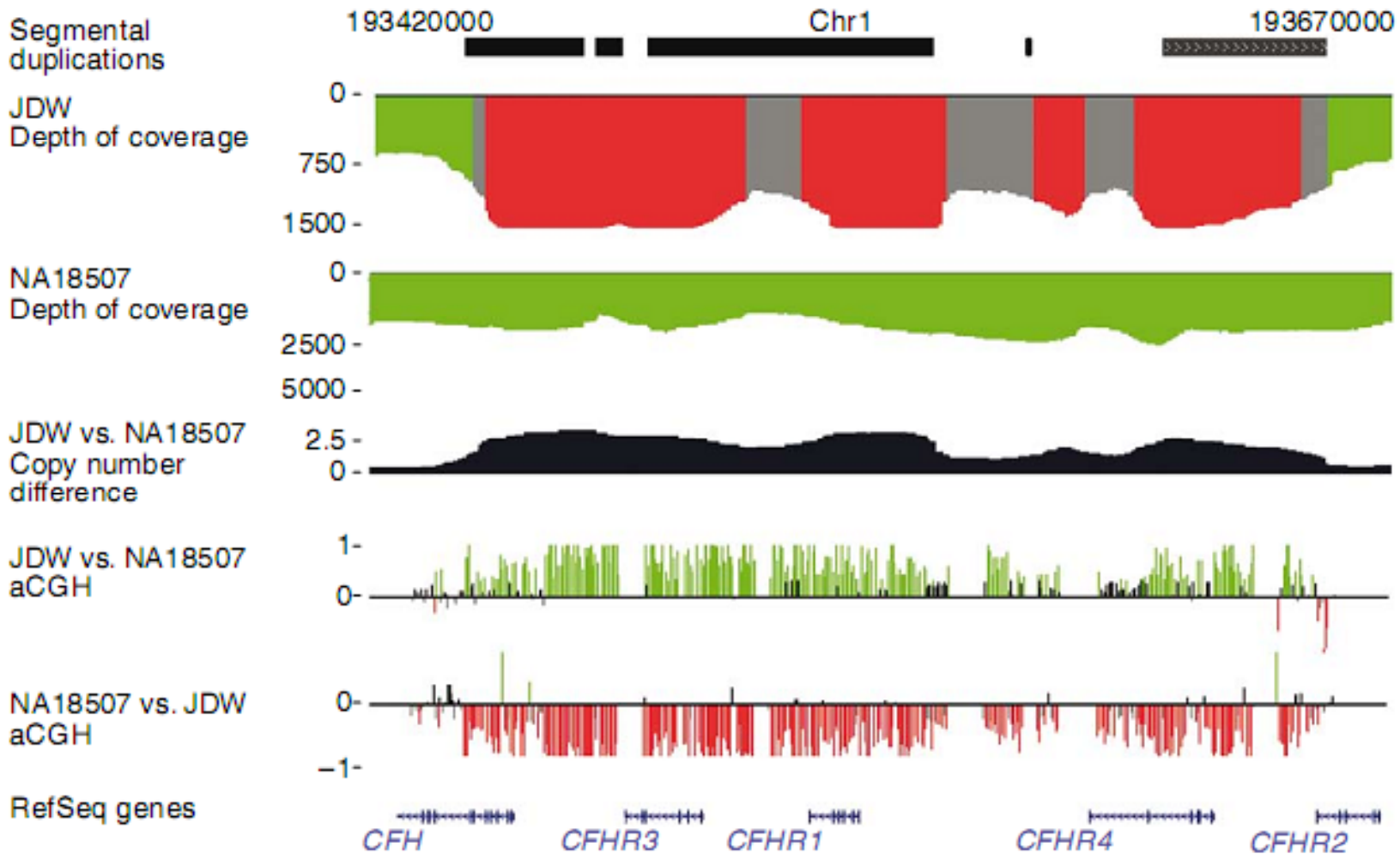
- **Probabilistic representation of different genotypes – really a new approach!**
- **Makes balancing FDR and FNR more straightforward.**

ProbHD:

http://www.mcb.mcgill.ca/~blanchem/reseq/

Calling Structural Variants

Copy Number Variation by Read Depth



Alkan C, *et. al. Nature Genet*. 2009

mrFAST

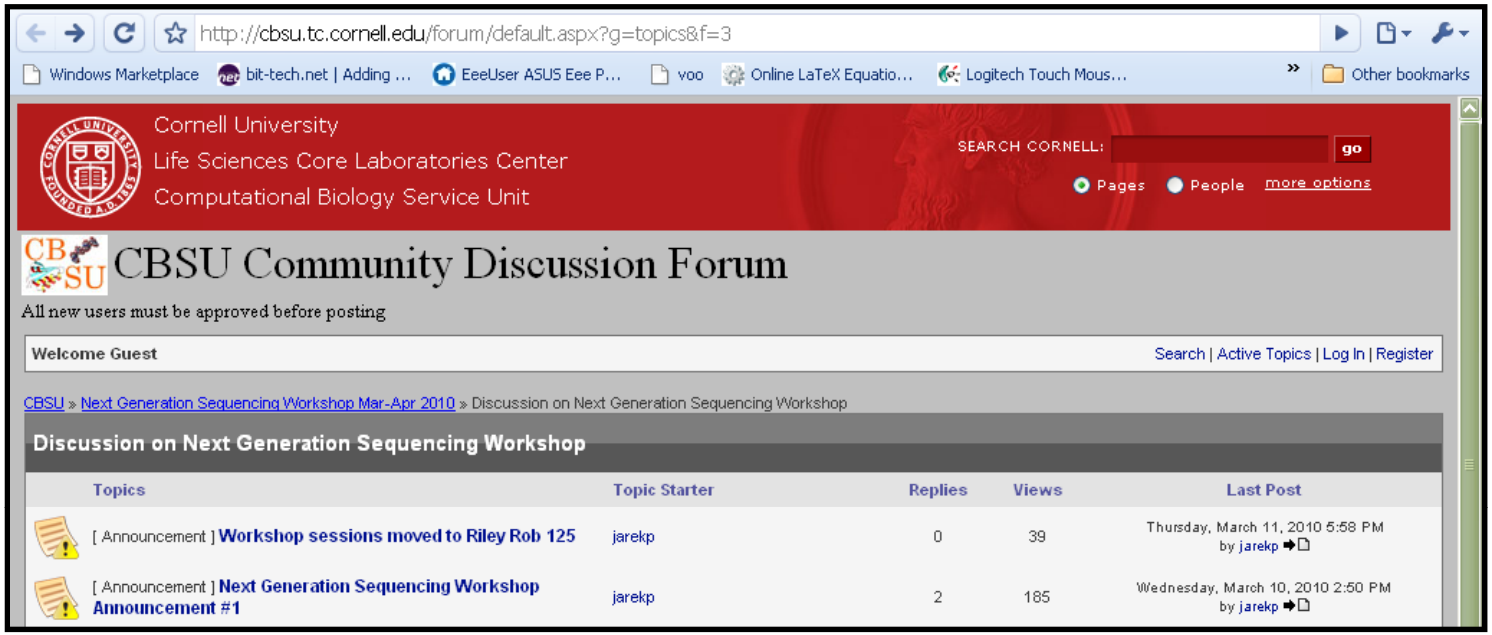
- **Maps structural variation using read depth.**

mrFAST:

http://mrfast.sourceforge.net/

Exercises

1. Align reference sequence (na18507, Yoruban HapMap individual) to the reference human genome.
2. Use SAMTools to build a sorted BAM file.
3. Call SNPs and short indels.
4. Filter this list to derive a high confidence set of SNPs.
5. View BAM files using IGV.



http://cbsu.tc.cornell.edu/forum/default.aspx?g=topics&f=3

Office hours: *Friday at 3pm in Weill 102*