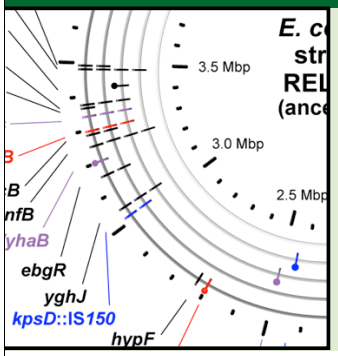
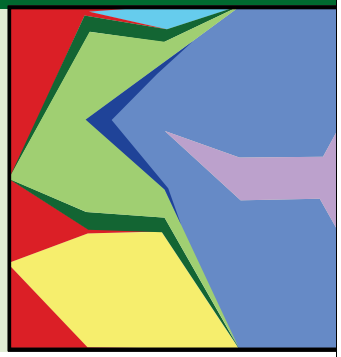
Re-sequencing hundreds of evolved *E. coli* genomes:



Finding non-SNP mutations, analyzing mixed populations, and knowing what you don't know



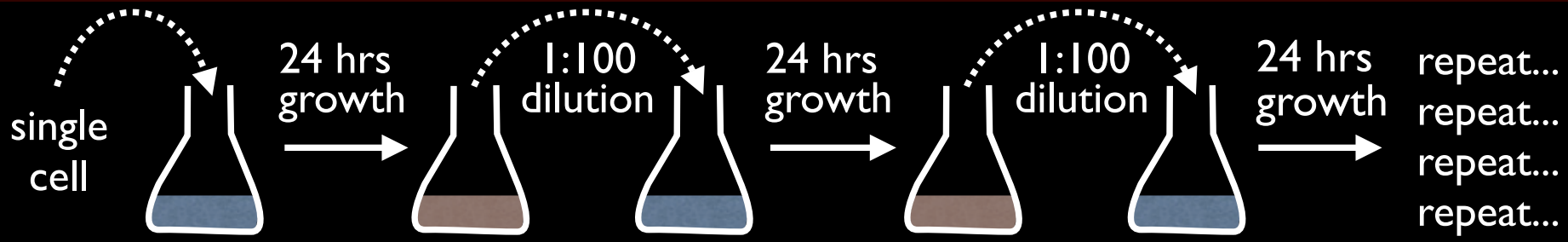
Jeffrey E. Barrick

Laboratory of Richard E. Lenski

Microbiology and Molecular Genetics

MICHIGAN STATE UNIVERSITY

Lenski long-term evolution experiment



- 12 independent populations evolved >20 yrs. Frozen "fossil record" has been archived.
- How many and what mutations?
- Compare rates of genomic change and fitness increase, monitor diversity in the population, understand molecular basis of adaptation.

Overview

- Application: **Re-sequencing** *E. coli* genomes
- Platform: **Illumina Genome Analyzer**

1. Overview of strategies and sequencing data
2. ***breseq*** ("brēz-sēk") bacterial re-sequencing pipeline
    - characteristics of a typical data set
    - identifying different kinds of mutations
    - analysis of SNPs in a mixed population
3. Asking evolutionary questions

Before I forget...

MSU has great resources for genomics

- RTSF – thanks Kevin, Shari, Jeff, ... !
- HPCC – thanks Ed, Bill, ... !

Thanks to the Lenski lab, particularly
Brian, Neerja, and Zach.

Thanks to collaborators
Jihyun Kim et al. (KRIBB)
Dom Schneider et al. (Grenoble)
Genoscope

Strategies for finding mutations

re-sequencing

- map reads to known reference genome (ssaha2, maq, ...)

- infer mutations

de novo assembly

- assemble reads by overlap (velvet, ...)

- map contigs to reference genome

- infer mutations

Strategies for library preparation

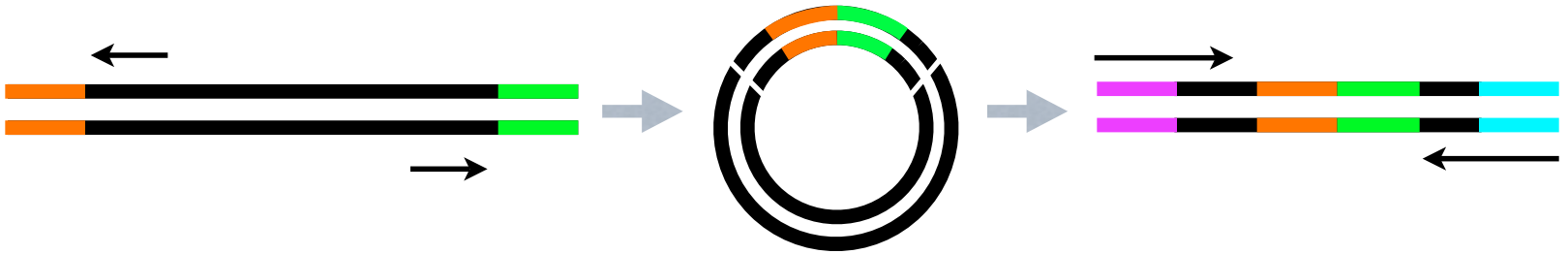
single-end

independent reads

paired-end

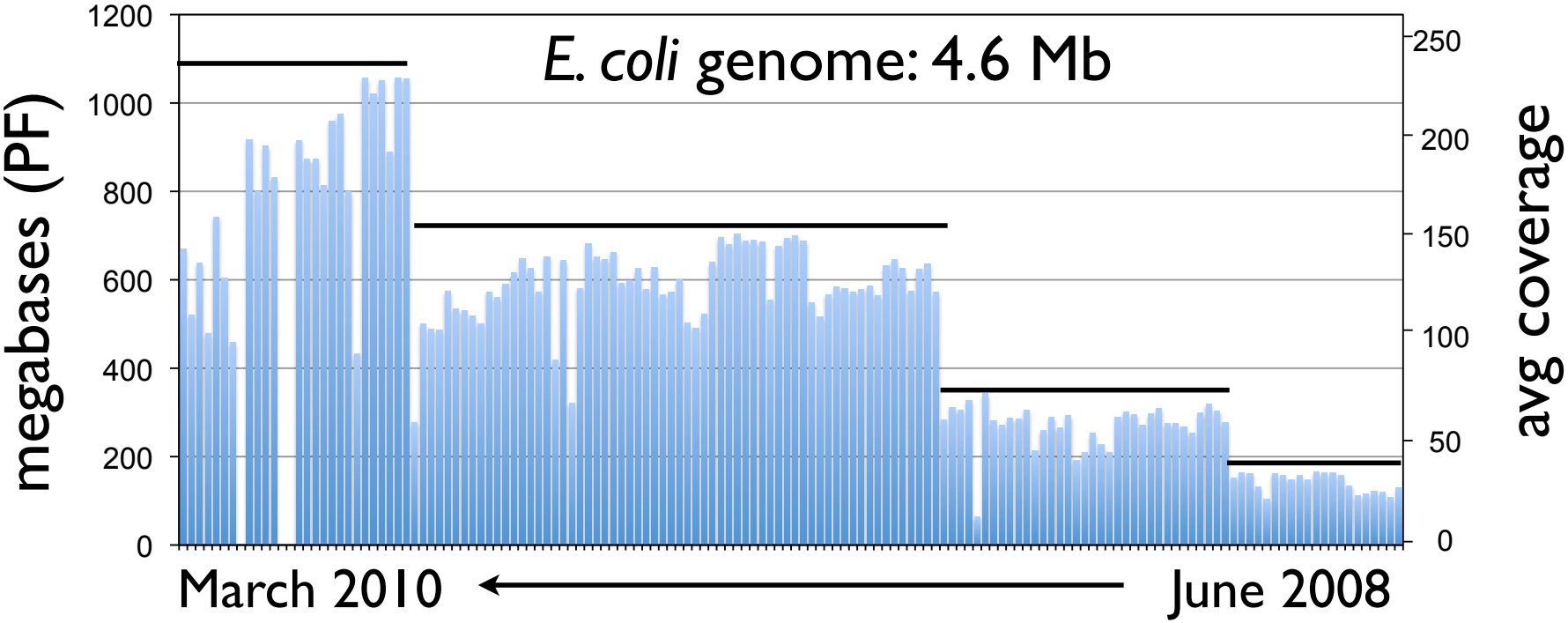
two inwardly oriented reads separated by ~200 nt

mate-paired



two outwardly oriented reads separated by ~3000 nt

Sequence Data Sets

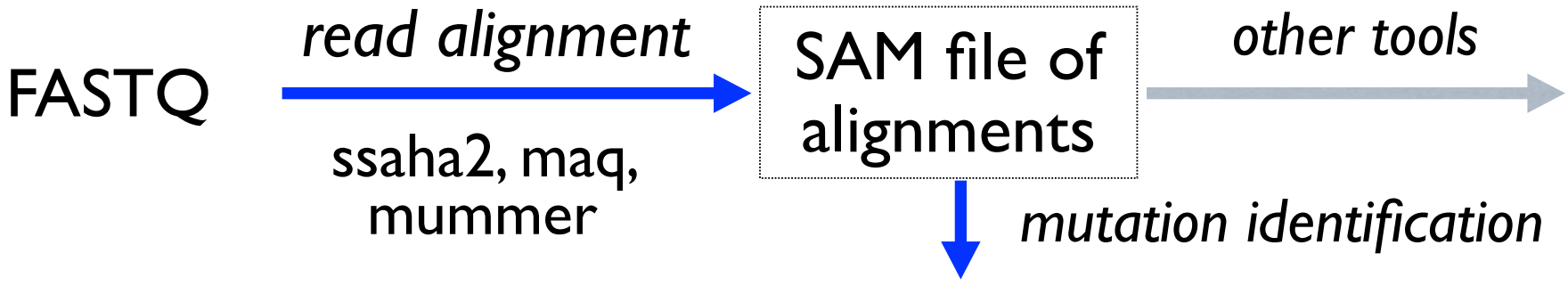


148 samples at RTSF (shown)

120 samples at Genoscope with Dom Schneider et al.

mostly 1 lane per genome, 36-bp single-end reads

breseq re-sequencing pipeline



1. single-base substitutions (SNPs)
2. small within-alignment indels
3. large deletions
4. new junctions (IS insertions)
5. copy number variation
6. mixed population SNP analysis

HTML / PDF / TXT output

breseq

implementation:

- command line tool
- unholy alliance of Perl and R (... and C++)
- emphasis on accuracy over speed
- runs on Unix, HPCC, Mac OS X

SAM: Sequence Alignment/Map format

Text (SAM) and binary (BAM) files organized for quick retrieval of reads aligned to a certain position.

Created to support 1000 Genomes project by a team at the Sanger Center.

SAMtools (http://samtools.sourceforge.net/)

- C library with bindings to Java, Perl, Python, Lisp, etc.
- Command-line tools for manipulation, consensus/ indel calling, viewing alignments as text, ...
- Many aligners output in SAM format (ssaha2, maq)

Knowing what you don't know

1. Theoretical limits: Read length and pair distance.
2. Practical limits: Base quality and coverage evenness.

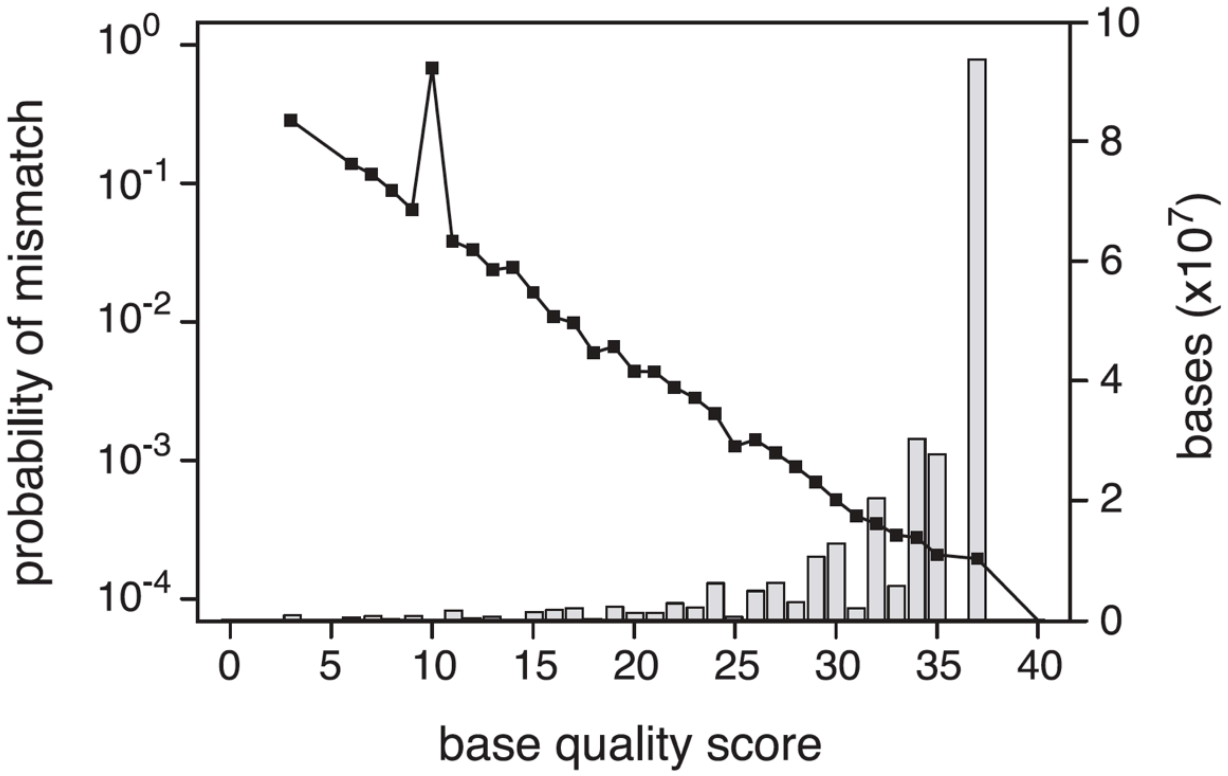
	single-end	paired-end	mate-paired
IS insertions	*	*	*
duplications	*	*	*
inversions across IS	–	–	*
SNPs in repeats	–	–	*
long tandem repeats	–	–	–

IS = bacterial mobile elements 1.0-1.5 kb in length.

Need standardized metrics to describe completeness of re-sequencing data on a per-base per-genome basis.

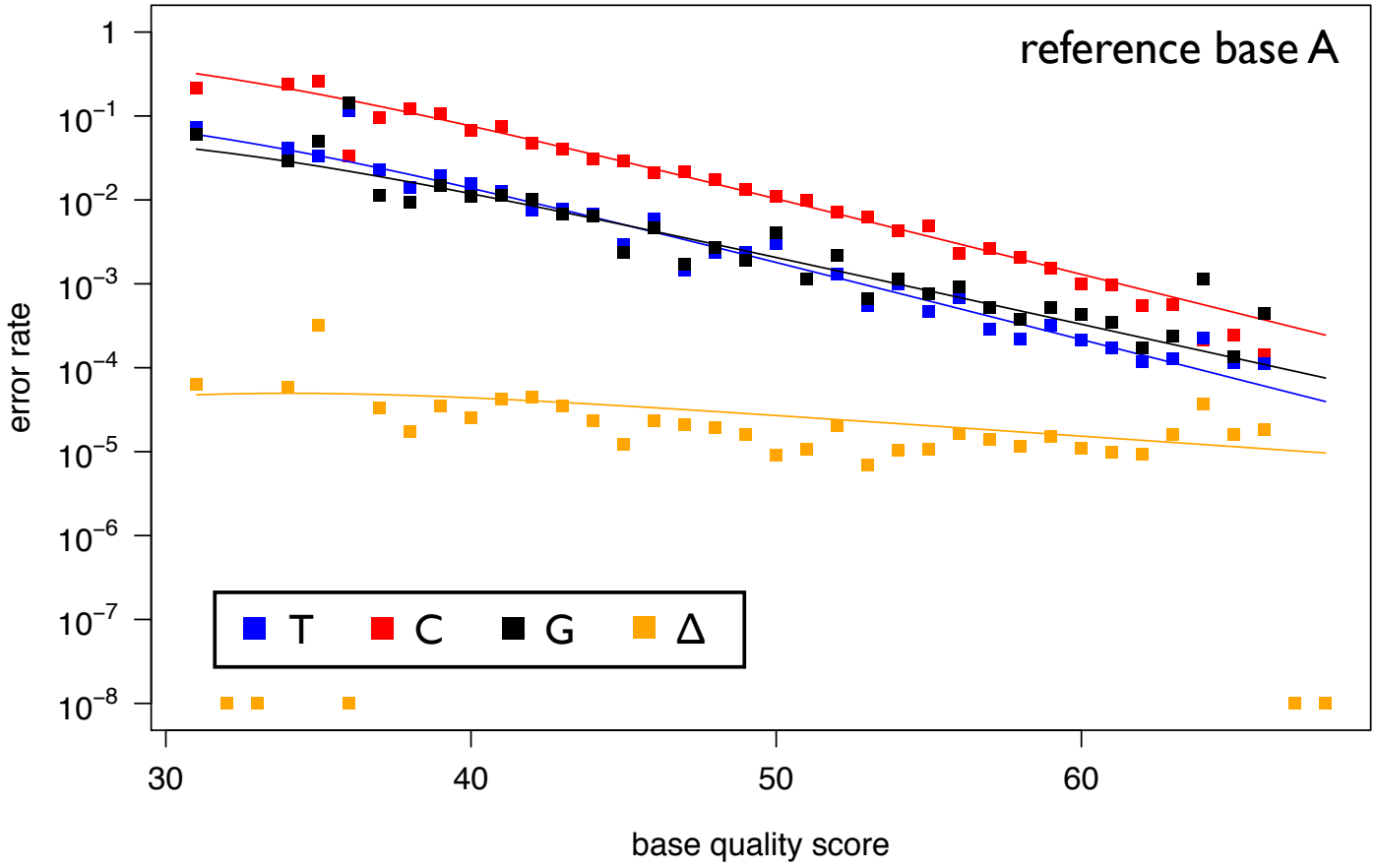
Typical Base Error Rates

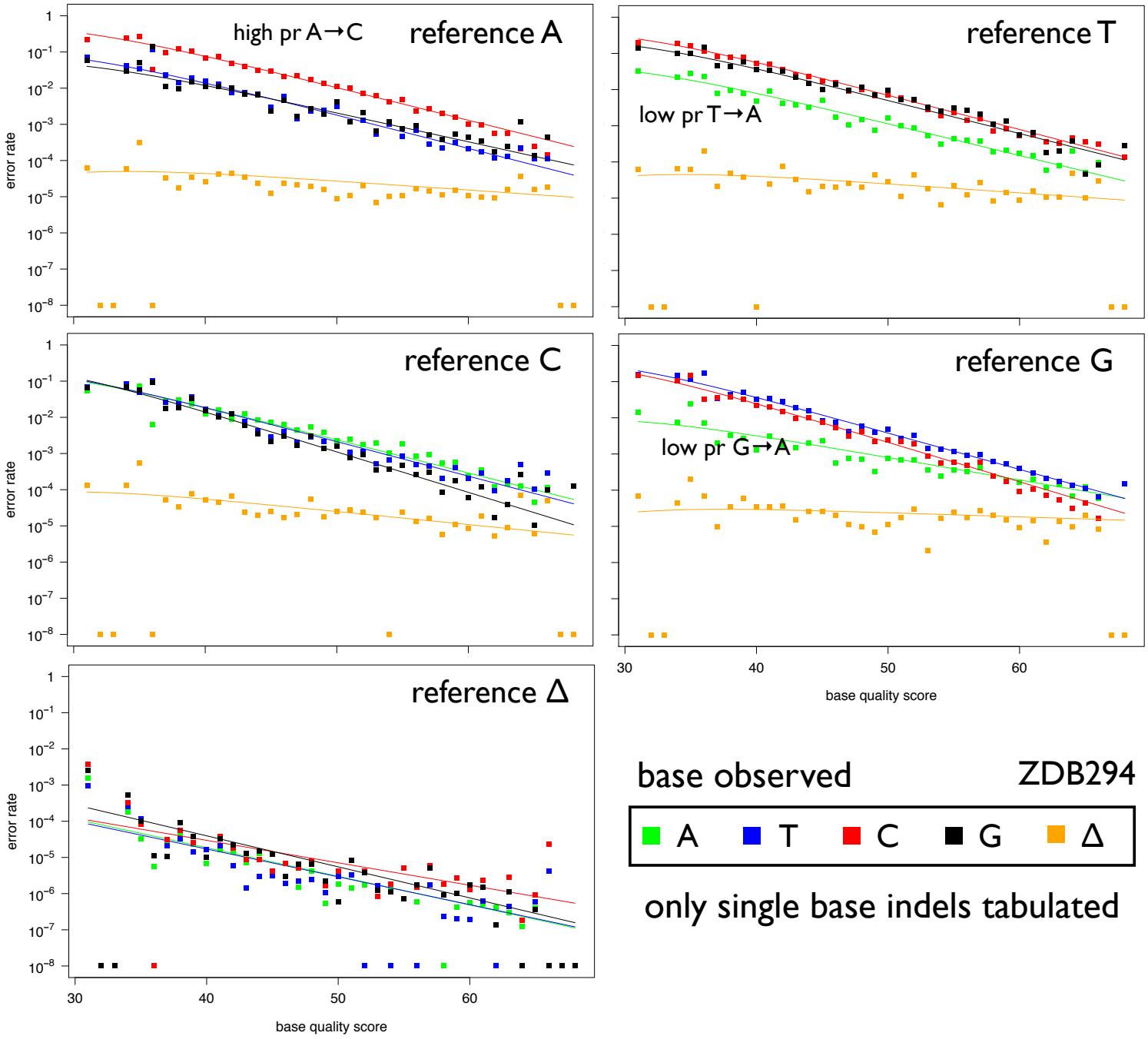
- Most bases in a run have error frequencies between 10^{-4} and 10^{-3}. Overall error rates agree well with Phred quality scores [$E=10^{-(Q/10)}$].



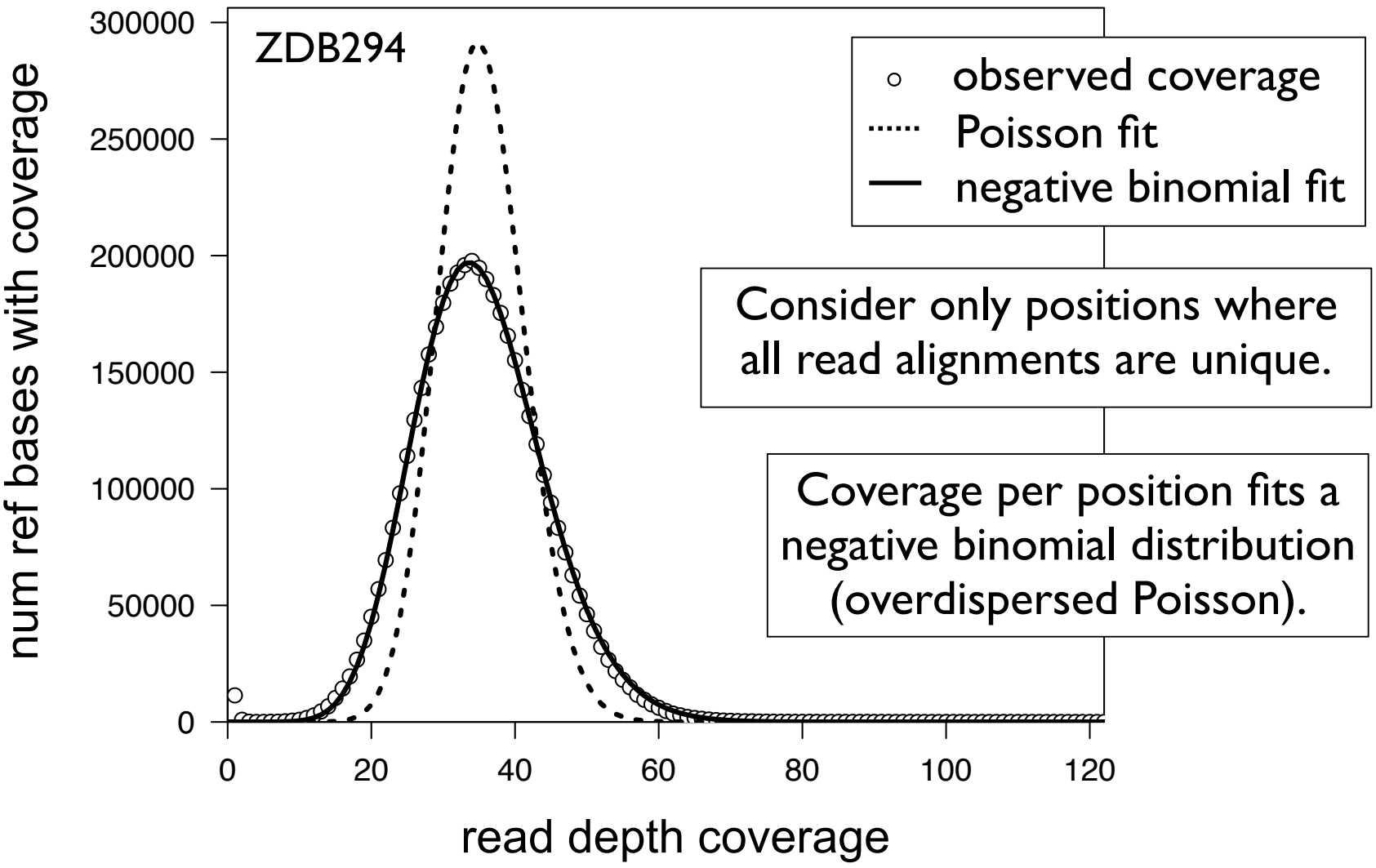
Typical Base Error Spectrum

- There is variation in the frequency at which different base errors occur at a given quality score.

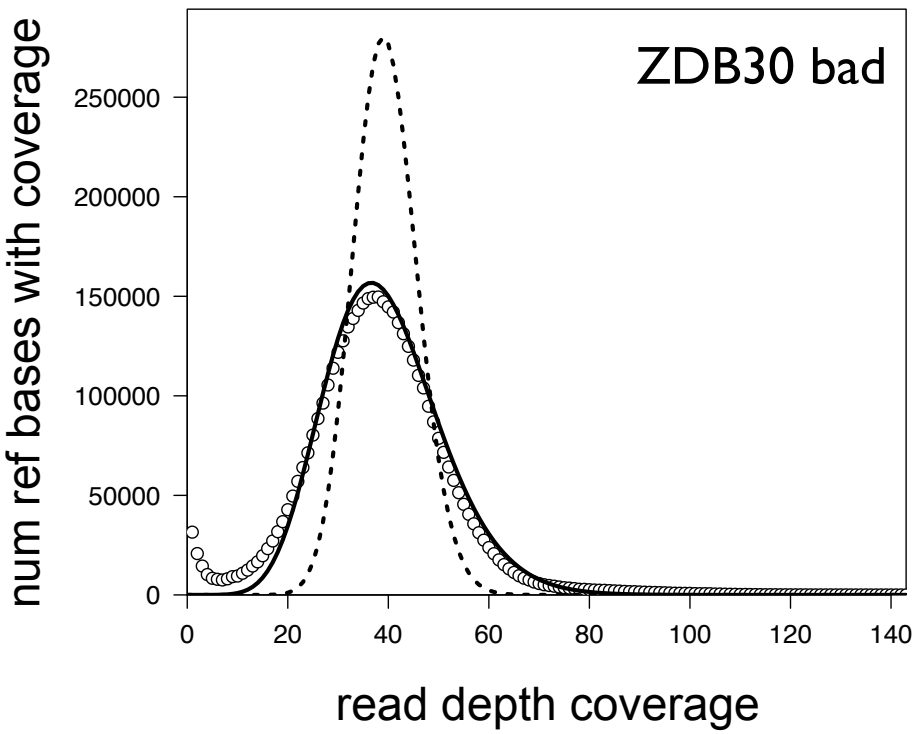


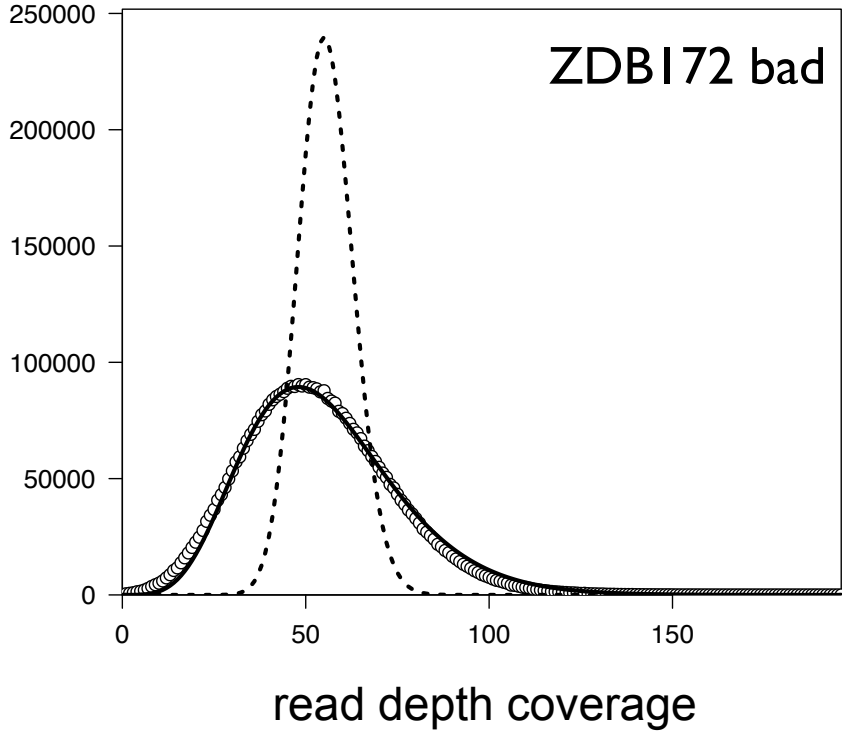

high pr A→C
reference A
reference T
low pr T→A
reference C
reference G
low pr G→A
reference Δ
error rate
base quality score
base observed
ZDB294
A
T
C
G
Δ
only single base indels tabulated

Typical Coverage Distribution



Problem Coverage Distributions





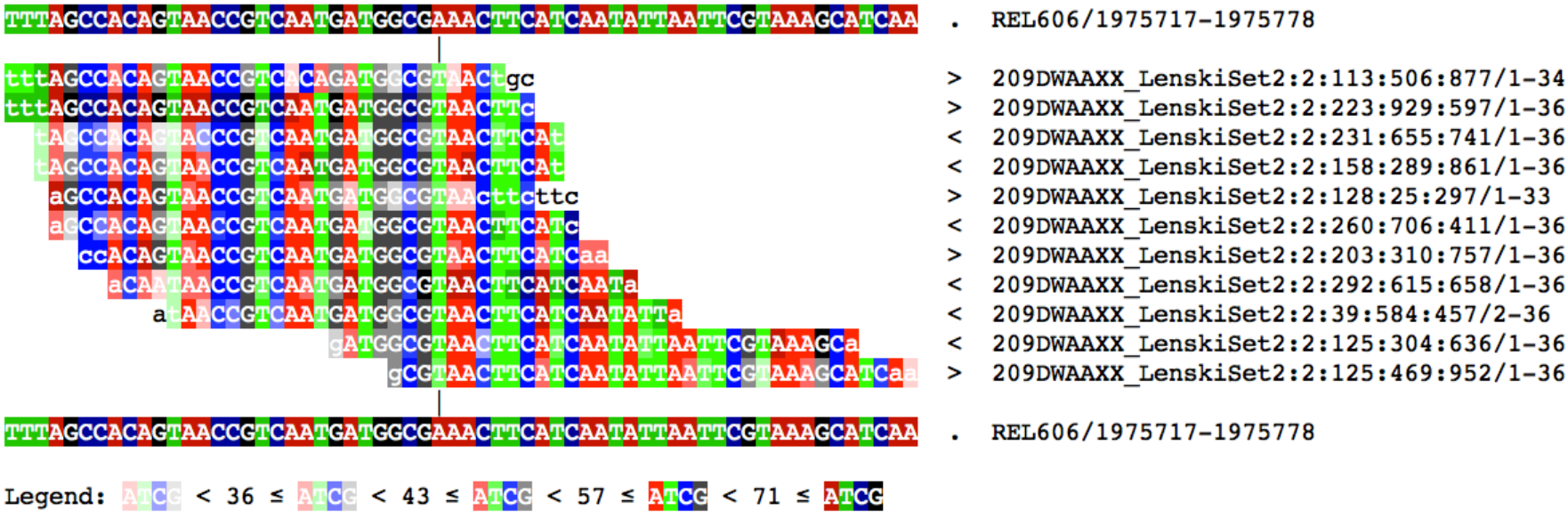
- Contamination with another sample?
- Large variance, missing coverage.

Both apparently from problems with library prep.

Identifying single-base substitutions

- Calculate probability of best base versus other bases given observed bases and error model.
- Accept as consensus if E-value < 0.01.

start	end	ref	change	quality	cov	tot_cov	type	gene position	codon change	aa change	gene	product
1975746	1975746	A	T	22.6	6/5	6/5	substitution	406 (136)	TCG→ACG	S→T	*yedV*	predicted sensory kinase in two-component regulatory system with YedW



Identifying within-alignment indels

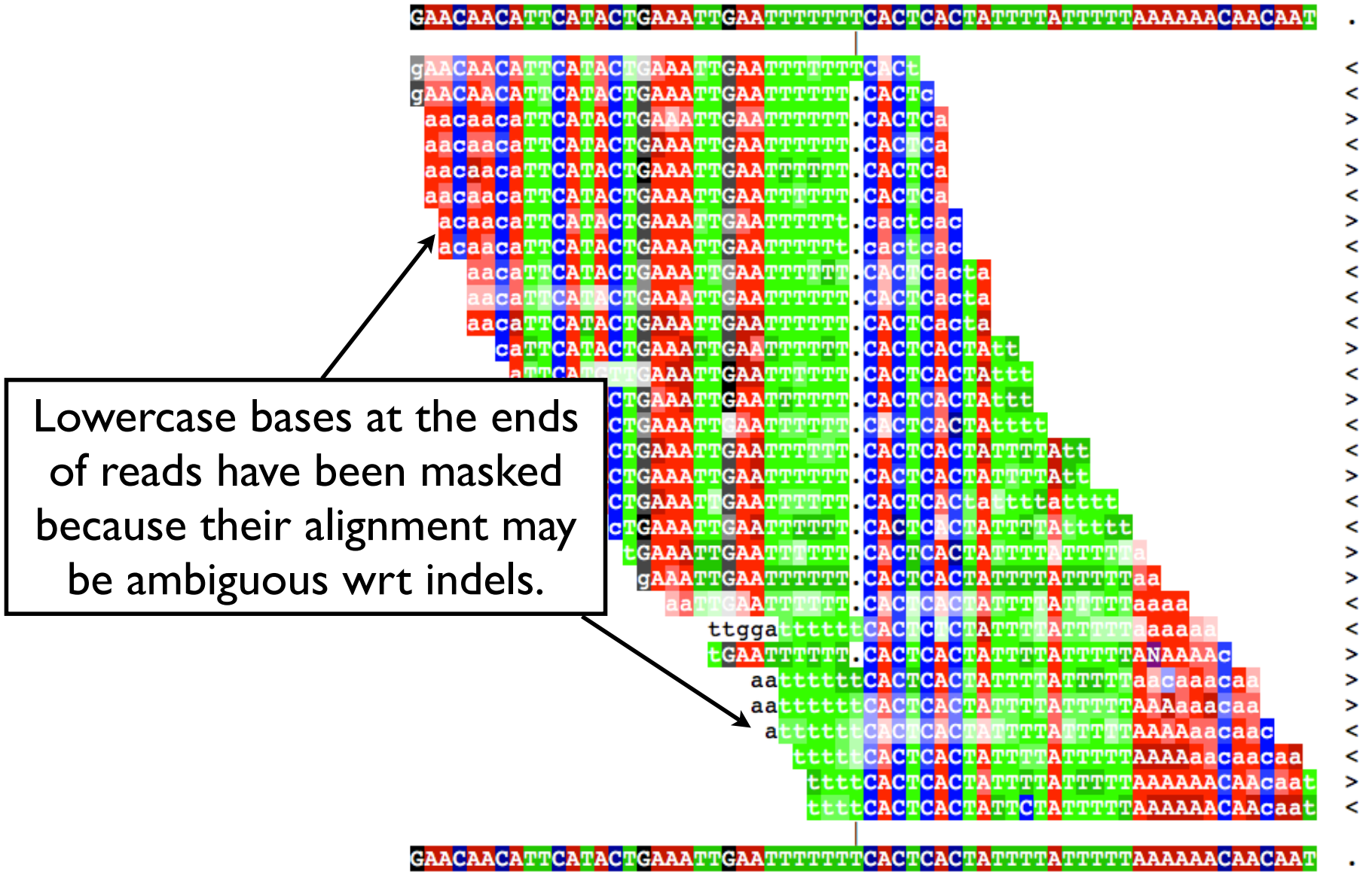
- Need to be careful in repetitive sequences and at the edges of short reads...

```
TATATTAATGCGCGCGCTAGGCTAGCT
TATATTAAT--GCGCGCTAGGCTAGCT <
TATATTAATGCGCGC--TAGGCTAGCT >
TATATTAATGCGCGC............ >
..........GCGCGCTAGGCTAGCT <
```

...where reads aligned from different directions can be ambiguously aligned.

...where reads from different directions that end in a simple sequence repeat may hide indels.

Example of a ***breseq*** prediction



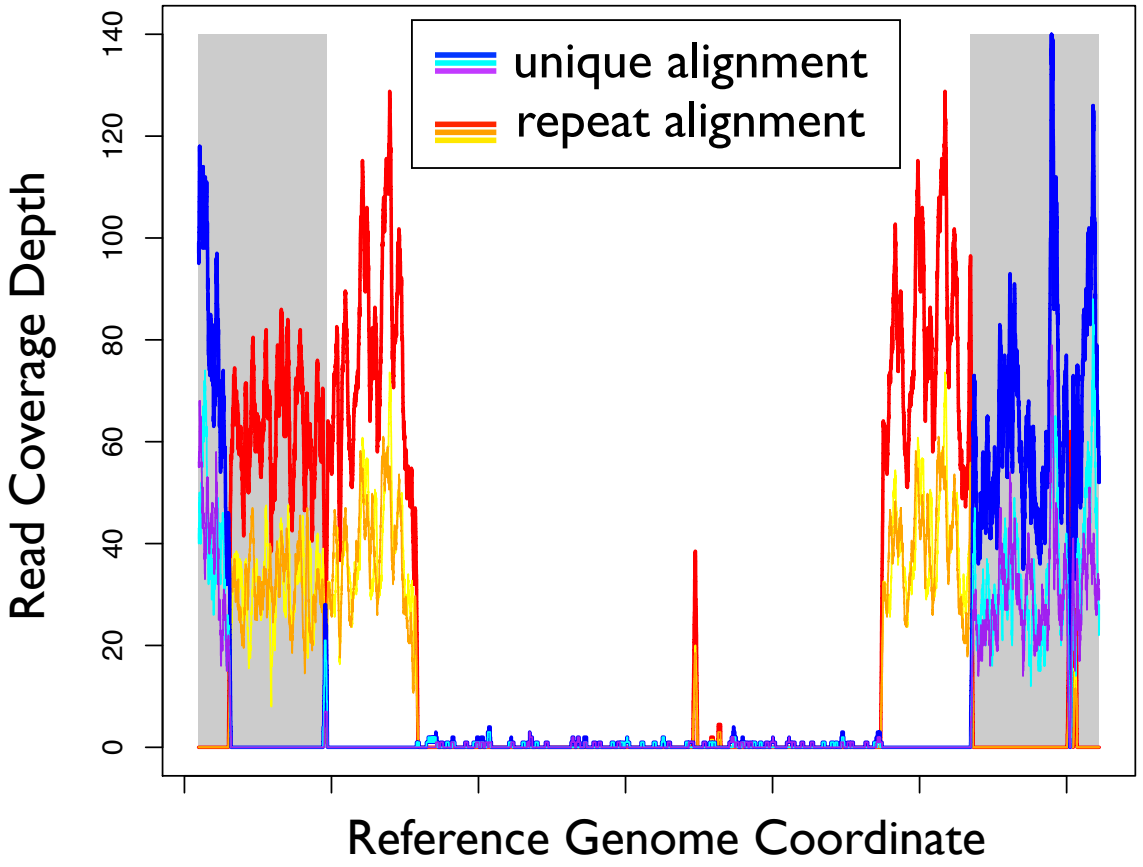
Lowercase bases at the ends of reads have been masked because their alignment may be ambiguous wrt indels.

Identifying large deletions

1. Seed deletions at positions with zero coverage.
2. Propagate boundaries outward until reaching a read-depth threshold based on the overall distribution.
3. Propagate through repeat regions, where a read aligns to multiple places in the genome.

Example of a ***breseq*** prediction

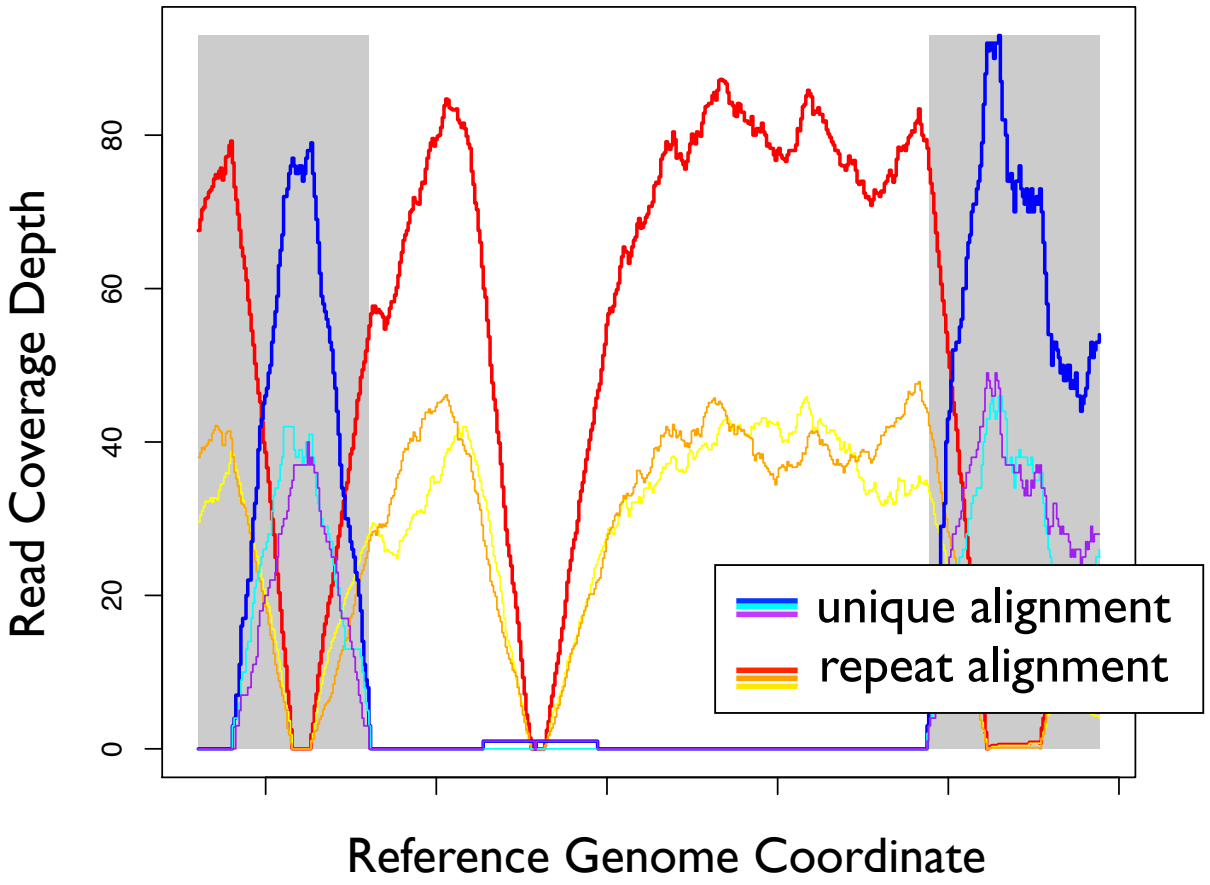
- Sometimes the molecular event is obvious...



- Recombination between nearby IS3 copies.

Example of a ***breseq*** prediction

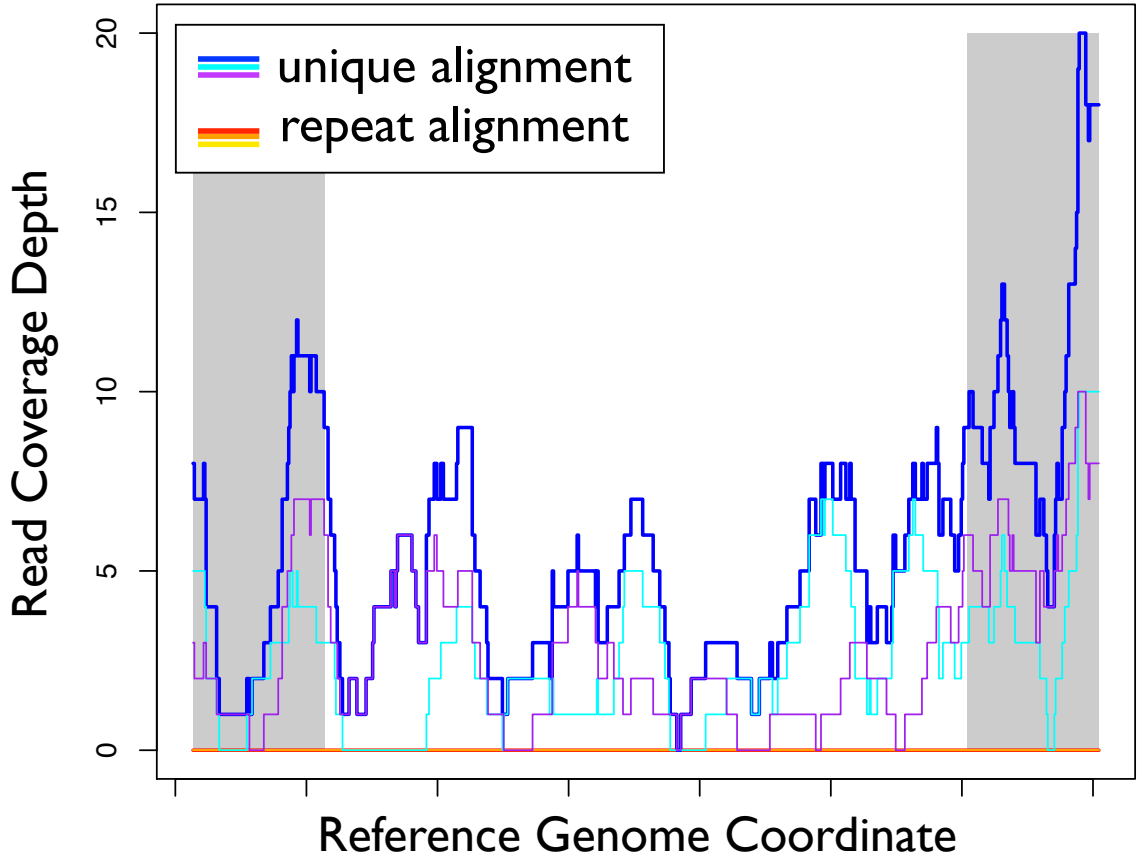
- Sometimes the mutation is not obvious...



- Gene conversion of 23S rRNA copy!!

Example of a ***breseq*** prediction

- Sometimes overall low or biased coverage leads to false predictions of deletions.



- Recognizable by sloped vs. steep edges.

Identifying new junctions

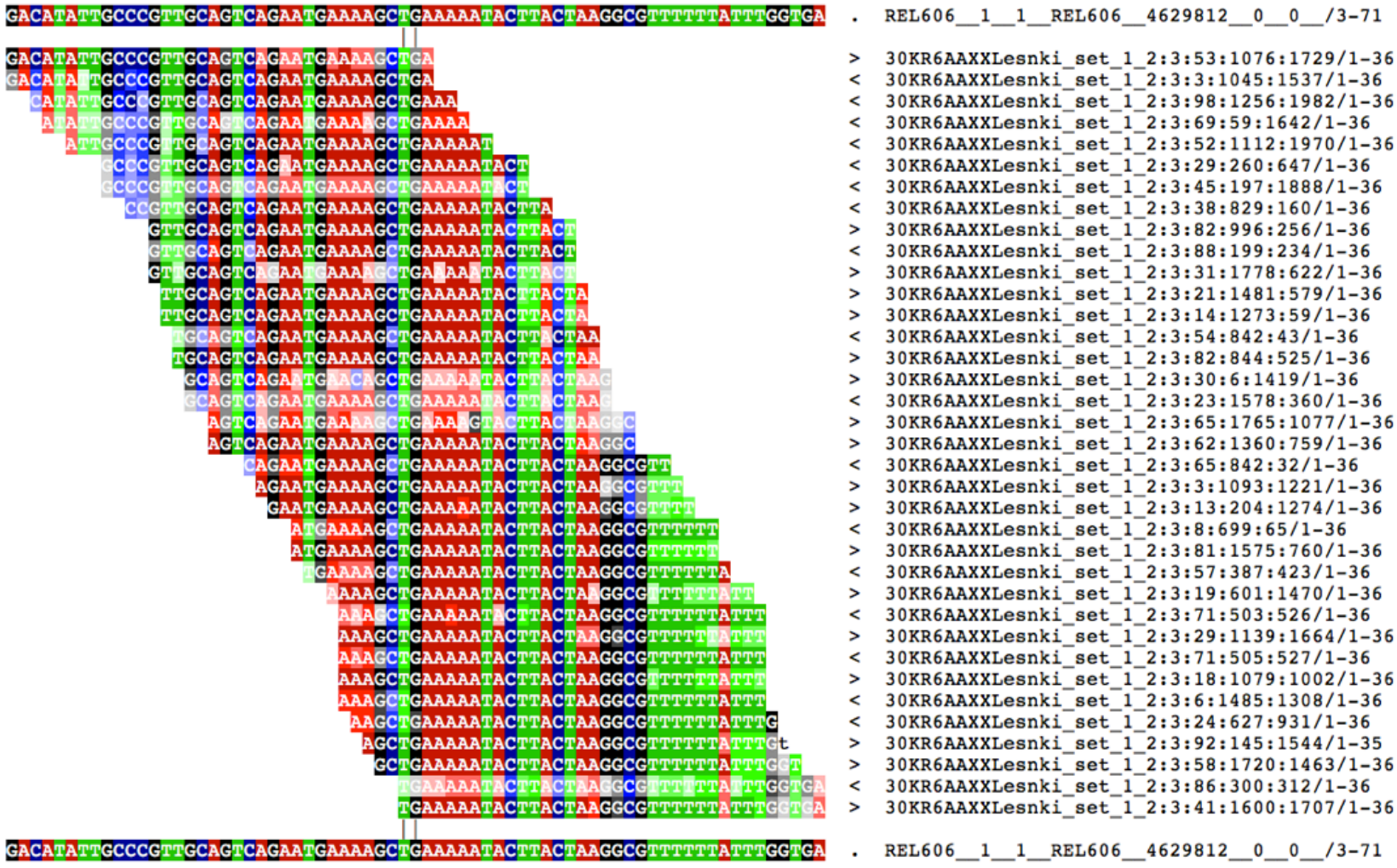
1. Find "mosaic" reads that partially map to two locations in the genome (possibly with overlap).

2. Create consensus list of possible new junctions.
3. Re-align all reads to candidate junctions.

4. Predict a new junction if reads map better to it than to the reference across its whole length.

Example of a ***breseq*** prediction

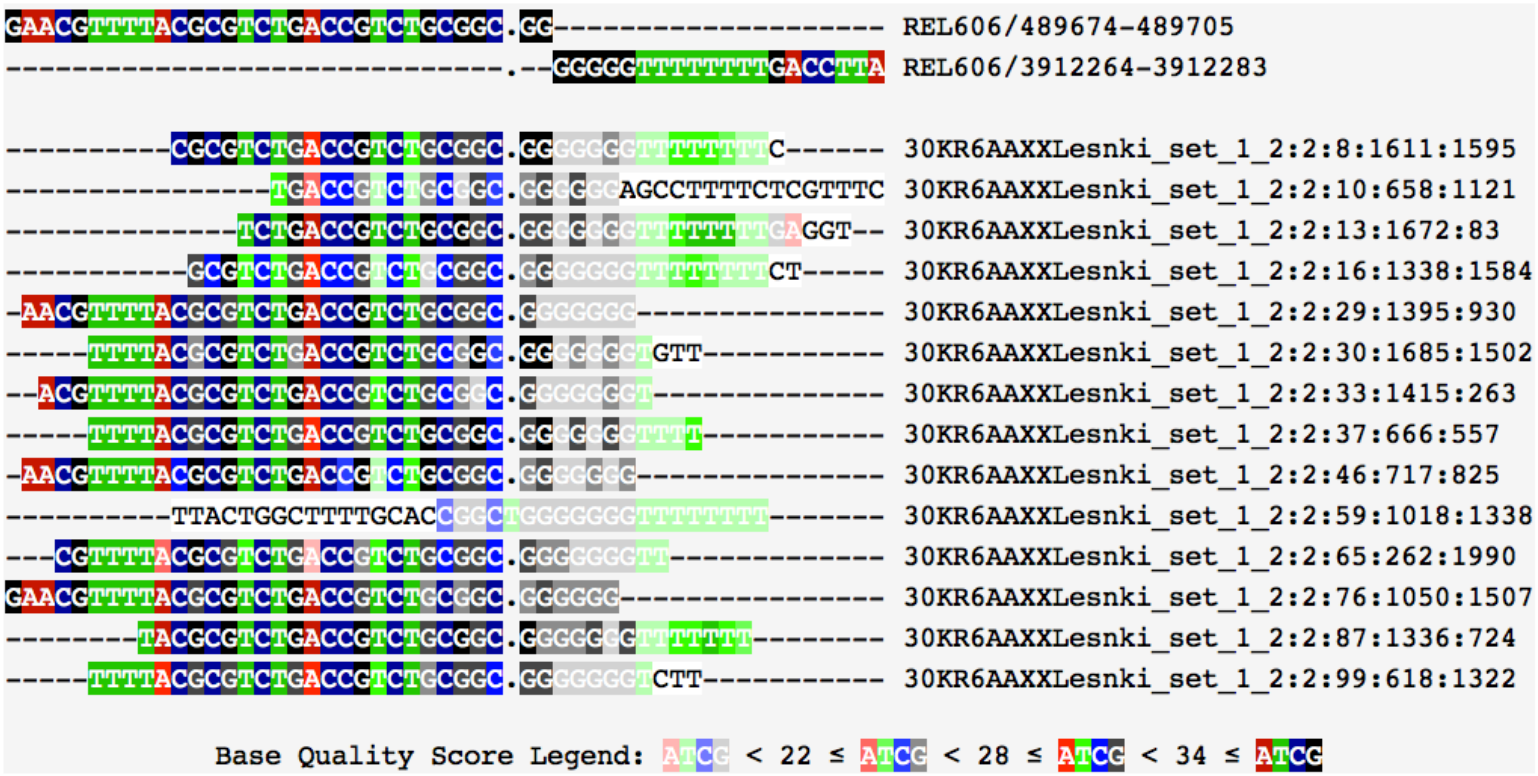
position	overlap	reads	gene	coords	product
1 =	0	36	*-/thrL*	/189	-/thr operon leader peptide
= 4629812			*lasT/-*	4629789/	predicted rRNA methyltransferase/-



Example of a *breseq* prediction

- Beware of reads ending in homopolymer runs!

position	overlap	full / total reads	gene	coords	product
= 489705	0	7 / 14	*ybbN*	490447-489593	predicted thioredoxin domain-containing protein
3912264 =			*ilvL/ilvG*	3912221/3912359	ilvG operon leader peptide/acetolactate synthase II, valine insensitive, large subunit

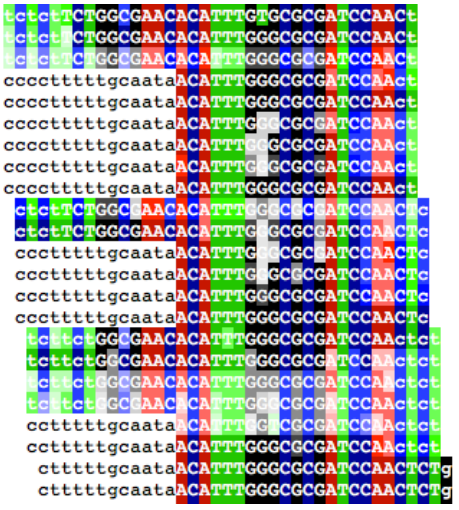


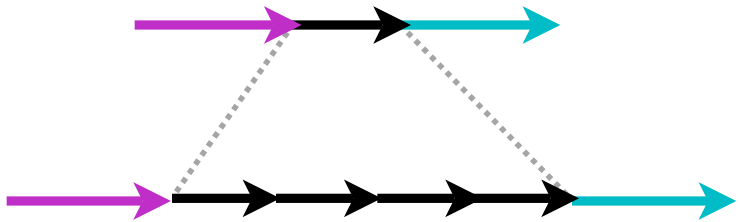
Example of a ***breseq*** prediction

- IS insertions create two new junctions...

		position	overlap	reads	gene	coords	product
		16989			IS150 (+)	+1443 (+3) bp	
*	?	16990 =	0	44	*mokC/nhaA*	16959/17487	regulatory protein for HokC, overlaps CDS of hokC/pH-dependent sodium/proton antiporter
	?	= 3652533			*IS150*	3651091-3652533	repeat region
*	?	= 16992	0	41	*mokC/nhaA*	16959/17487	regulatory protein for HokC, overlaps CDS of hokC/pH-dependent sodium/proton antiporter
	?	3893554 =			*IS150*	3893554-3894996	repeat region

- Sometimes new and old junctions both exist...

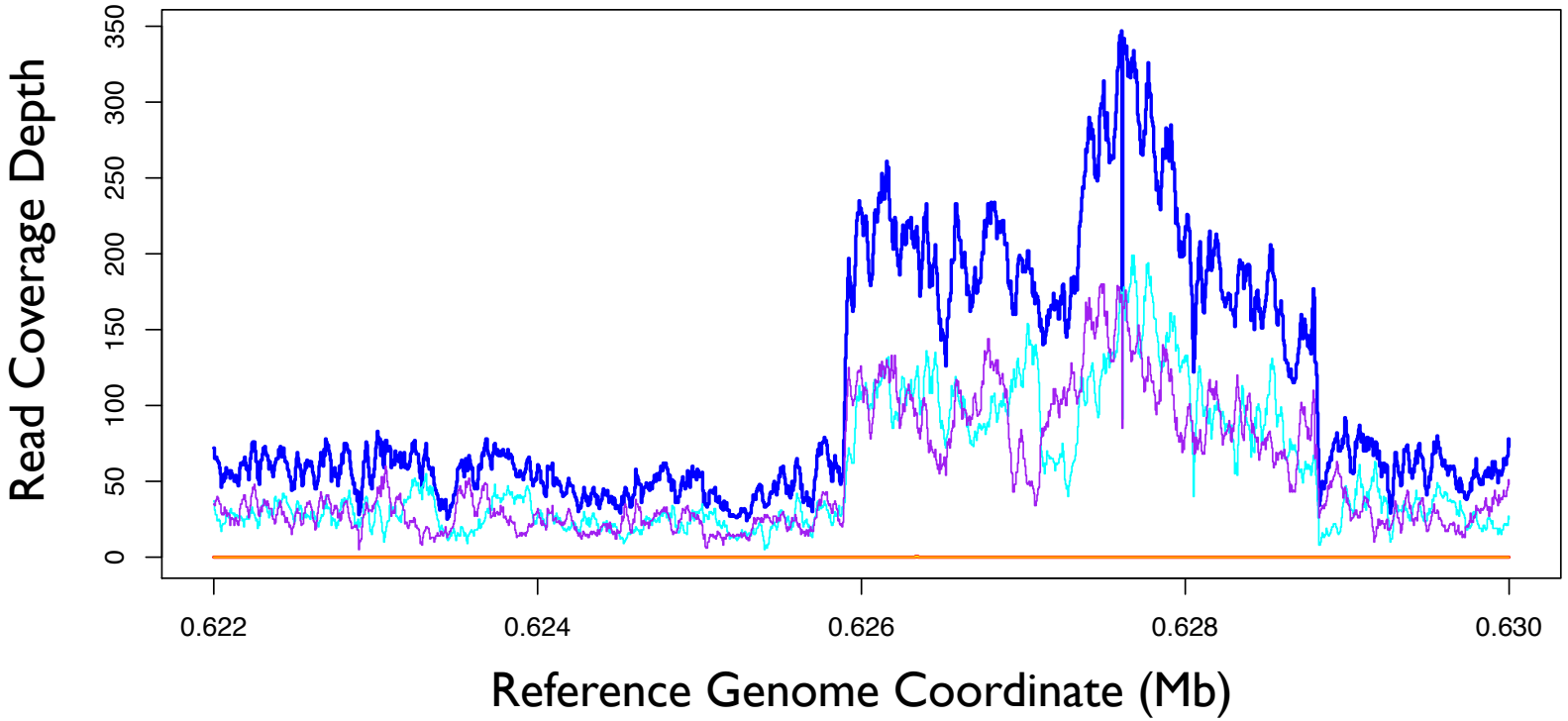




tandem head-to-tail duplications

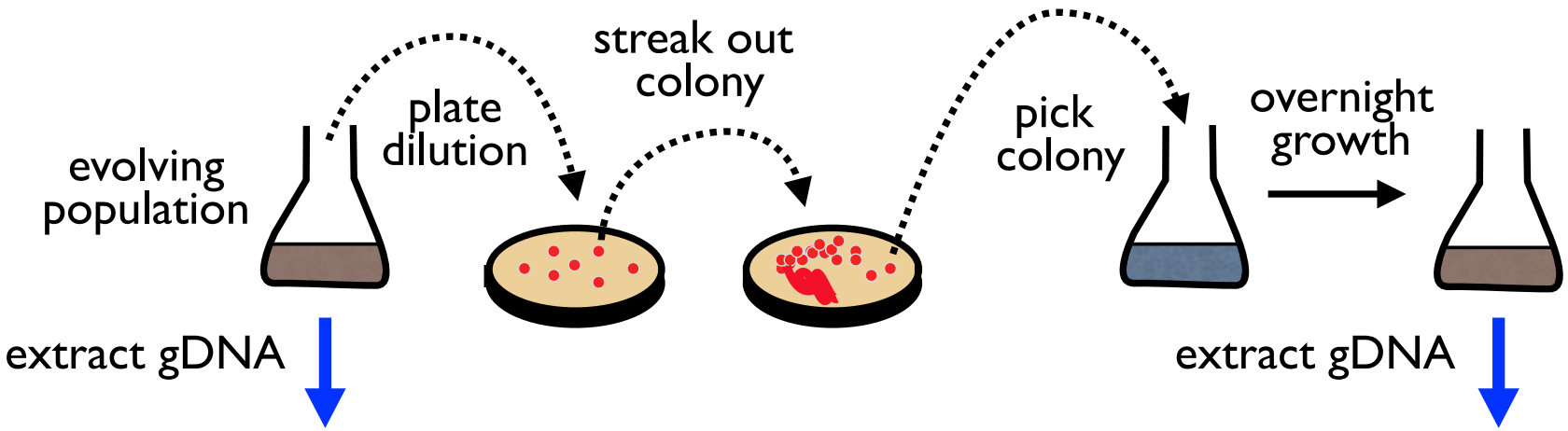
Identifying copy number variation

- Coverage is very noisy, but a fingerprint is (somewhat) consistent across runs.



- Tile into 100 bp segments, train bg model on many genomes, look for deviation (in progress).

Mixed population analysis



Every read could be from any individual.

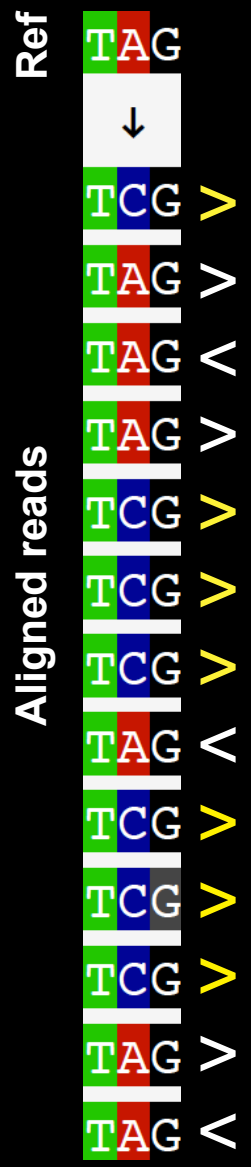
All reads are from a single clone.

Frequencies of mutations competing in population.

No linkage information.

Information about which mutations occur together.

Sequencing error or polymorphism?

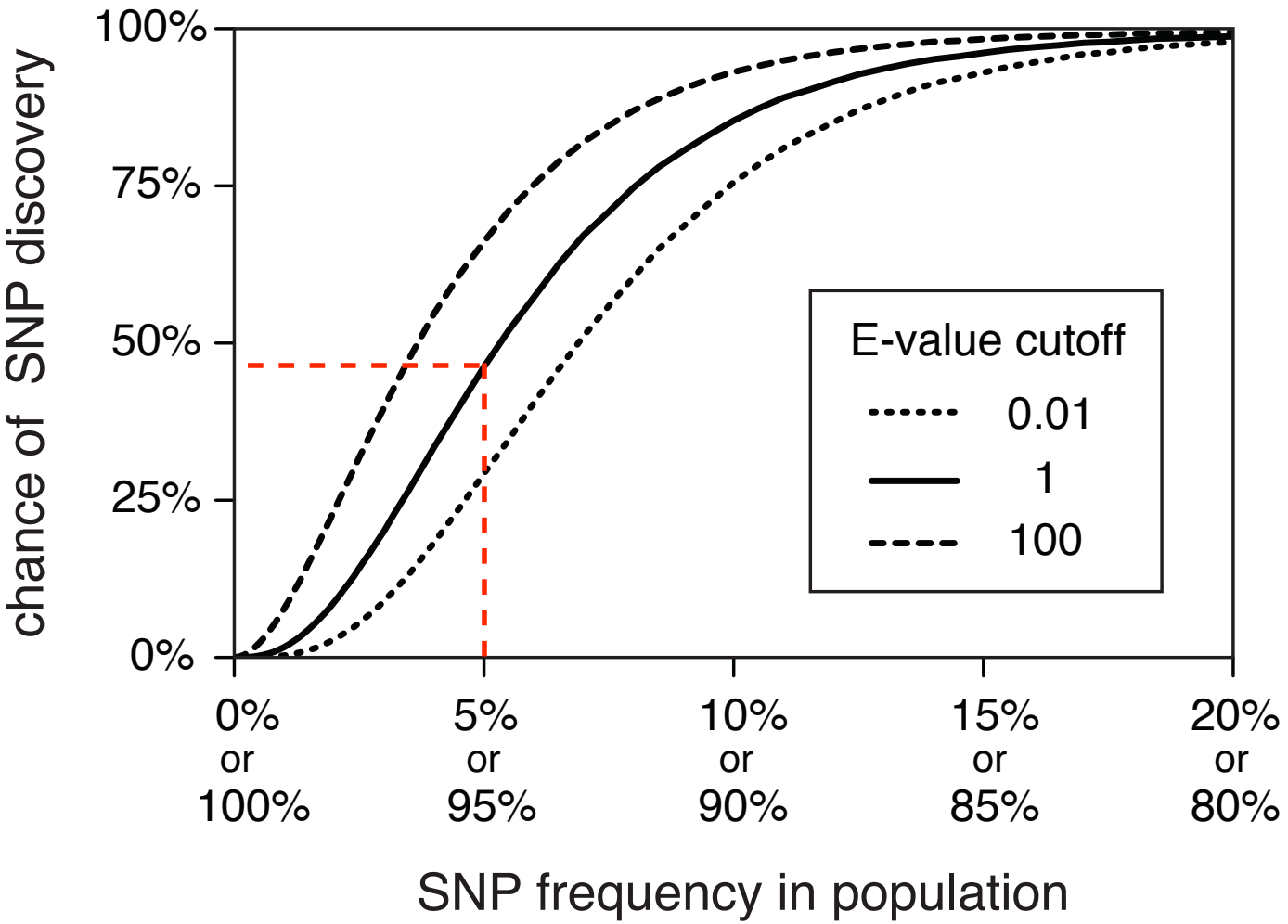


- Map reads to ancestor genome.
  Only consider single-base substitutions.
- Log-likelihood test for polymorphism:

$$D = -2 \ln \frac{\Pr(\text{obs} \mid \text{no polymorphism, i.e. all error})}{\Pr(\text{obs} \mid \text{ML fraction new allele})}$$

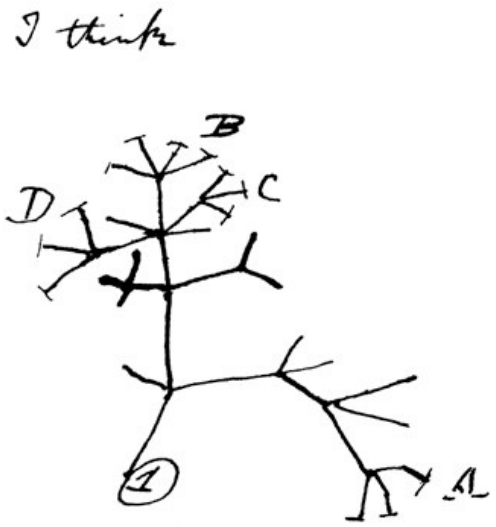
- Clone sequence data serves as a negative control (all errors, no polymorphisms).
- Filter out predictions with other biases:
  strand bias, systematically low quality scores

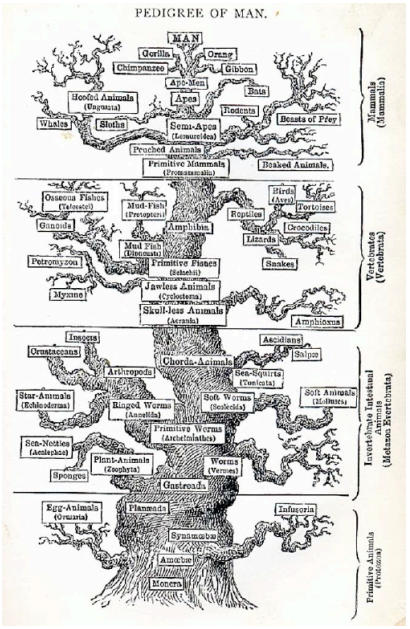
Sensitivity estimate



Asking Evolutionary Questions

- A list of *mutational events* in many clones over time, allows inference of evolutionary history.





- Later events may sometimes hide earlier events. (e.g. SNP in region that is later deleted)

Genome diff files

- For studying evolution we are interested in mutational events (essentially *genome diffs*).
- To submit a changed genome sequence to GenBank you must currently re-submit the entire genome – *even if it has only a one base difference*.
- Supplementary tables are not a sustainable or standardized way to report this data.
- Ideal *genome diff* format would also allow reporting of what is not known, frequency information for mixed population samples, and quality metrics.

Where are things going

- Re-sequencing will be used to routinely check mutant constructs. ($1000 human = $1 *E. coli*)
- *De novo* assembly will become more common, as technologies with longer read lengths come online.
- Studies of within-host diversity of virus populations and genetic diversity of neoplastic tumors.
- Every strain in the Lenski freezer will be sequenced (11,000 to go)...

Please let me know if...

...you have any ideas for better analysis strategies.

...you want to run ***breseq*** on the HPCC.

...you want any sample FASTQ datasets to analyze.

Search "Lenski" in Short Read Archive (SRA)
http://www.ncbi.nlm.nih.gov/sra

For more information...

1. Barrick, J.E. et al. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461**, 1243-1247 (2009).
2. Barrick, J.E. & Lenski, R.E. Genome-wide mutational diversity in an evolving population of *Escherichia coli*. *Cold Spring Harbor Symp. Quant. Biol.* **74**, ePub Sept. 23, 2009 (2009).