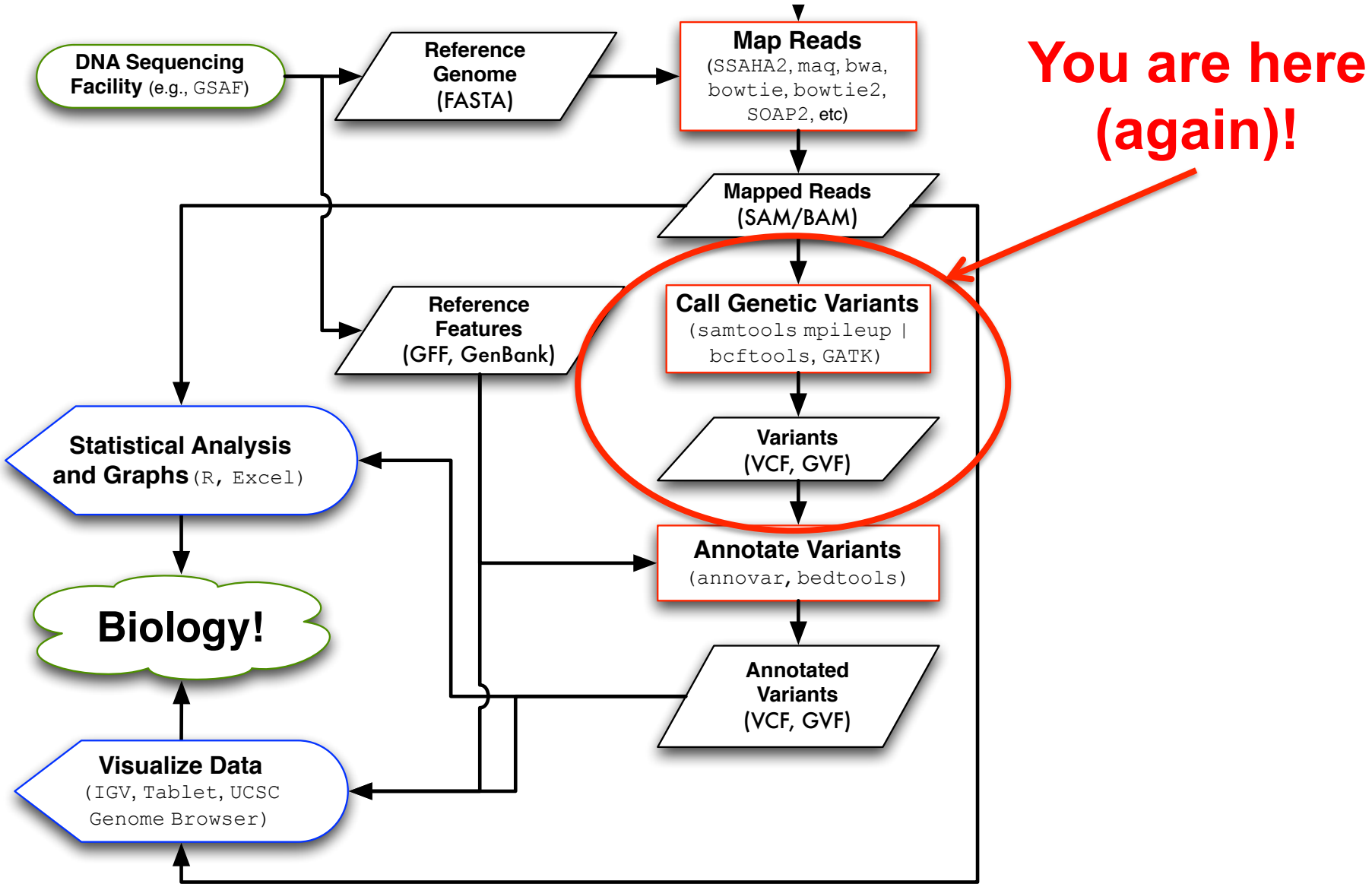
Structural Variant Calling



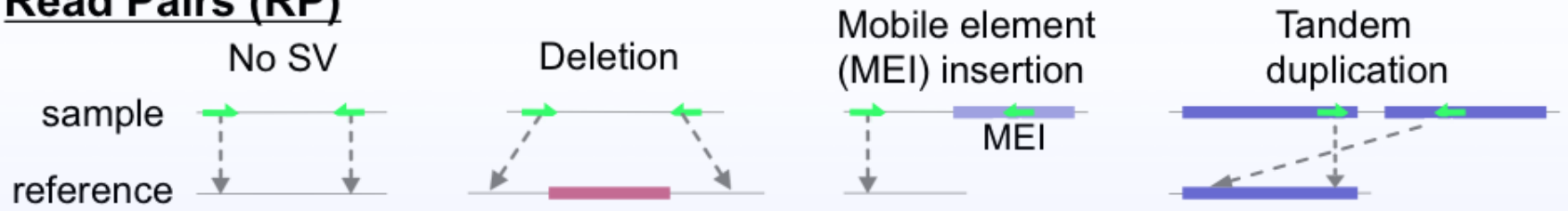
Types of Genome Sequence Variants

1. Single Nucleotide Variants (SNVs) ∗
   – Single base changes, e.g., A→T.
2. Insertions-Deletions (Indels; DIPs) ∗▶
   – Consisting of one or a few bases, e.g., +ATGA, ΔT.
3. Structural Variants (SVs) ▶
   – Everything else: large deletions, insertions, duplications, inversions, translocations, mobile element insertions, horizontal gene transfer

Different sequencing information and different algorithms are used to predict each kind of variant.

Predicting structural variants

Read Pairs (RP)

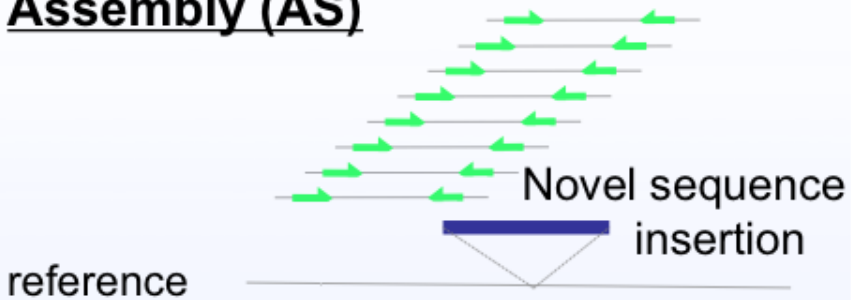


Read Depth (RD)

Unfortunately there is no program or pipeline that does **all** these things!!

Split Reads (SR)

Assembly (AS)



(http://www.genome.gov/Pages/Research/DER/1000GenomesProjectTutorials/StructuralVariants-JanKorbel.pdf)

Known Unknowns

Rumsfield on Genome Variants?

"...as we know, there are **known knowns**; there are things that we know that we know. We also know there are **known unknowns**; that is to say we know there are some things we do not know. But there are also **unknown unknowns**, the ones we don't know we don't know."

Accuracy

True positives
False positives
True negatives
False negatives

How **complete** is your data?

How **complete** is your variant analysis pipeline?

Knowing what you don't know

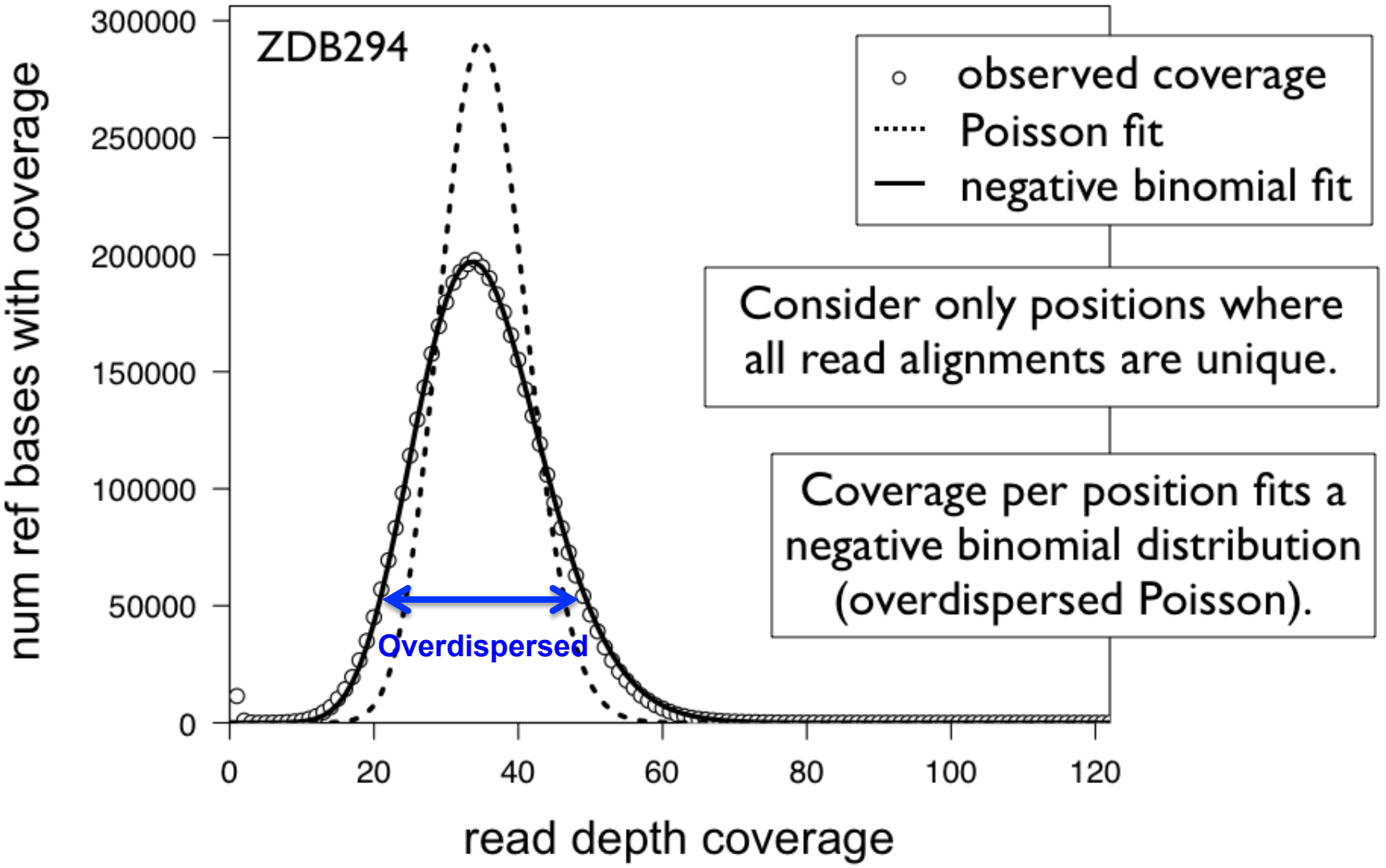
1. Theoretical limits: Read length and pair distance.
2. Practical limits: Base quality and coverage evenness.

	single-end	paired-end	mate-paired
IS insertions	*	*	*
duplications	*	*	*
inversions across IS	–	–	*
SNPs in repeats	–	–	*
insertion of new seq	–	–	–

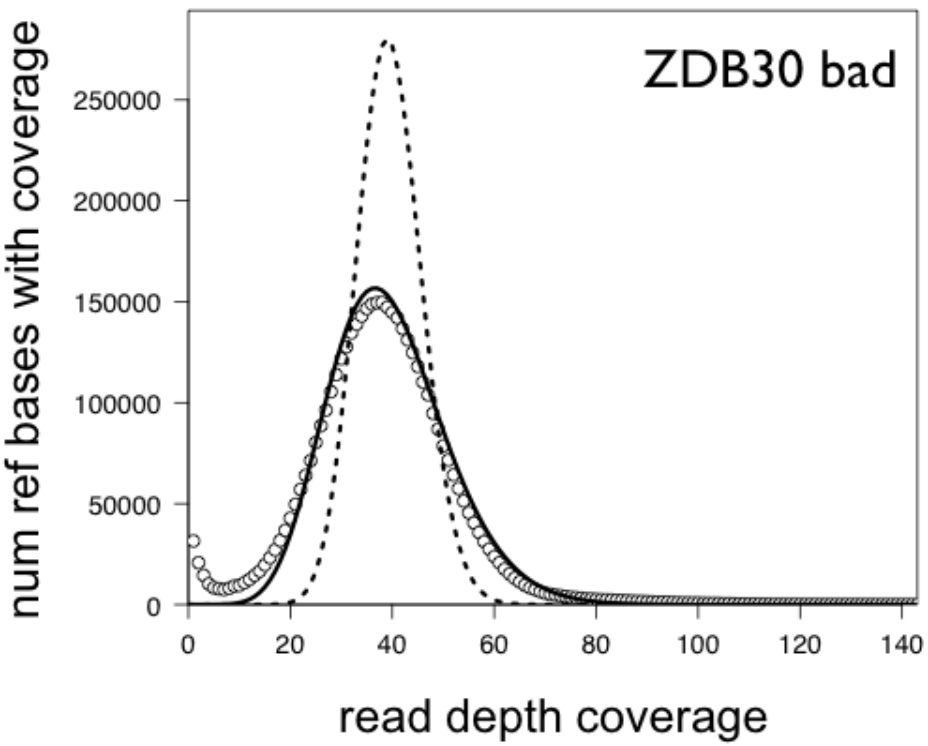
IS = bacterial mobile elements 0.8-1.5 kb in length.

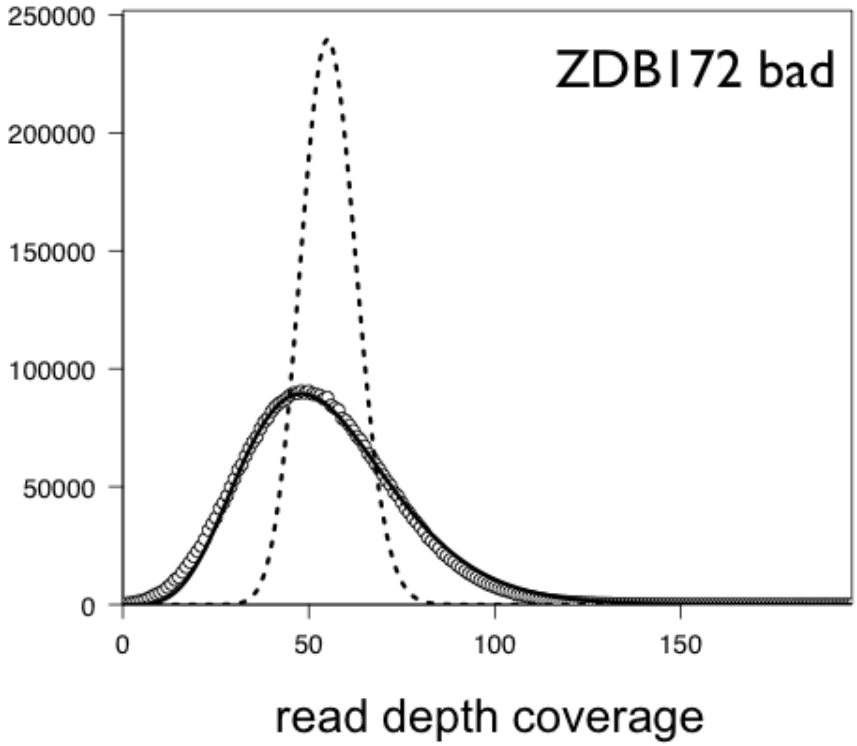
Need standardized metrics to describe completeness of re-sequencing data on a per-base per-genome basis.

Typical Coverage Distribution



Problem Coverage Distributions





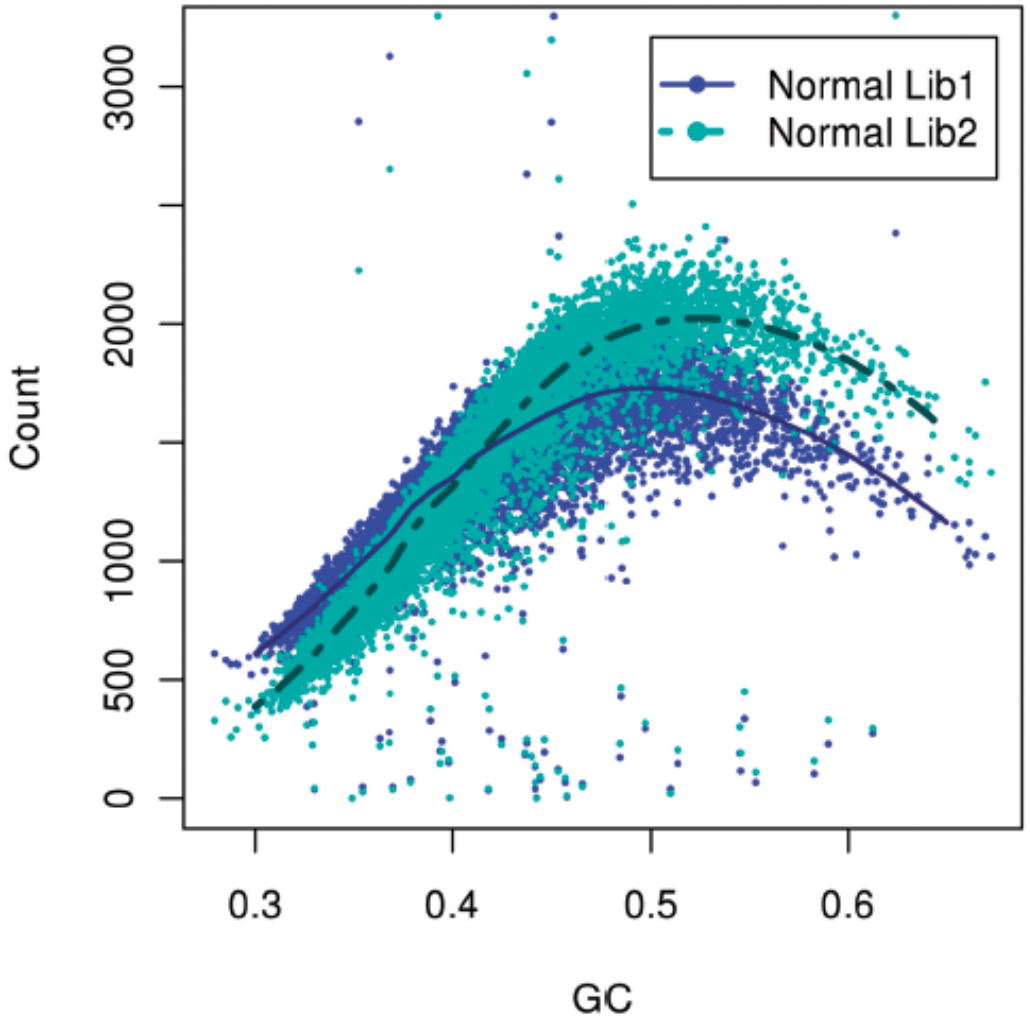
- Contamination with another sample?
- Large variance, missing coverage.

Both apparently from problems with library prep.

Coverage Bias

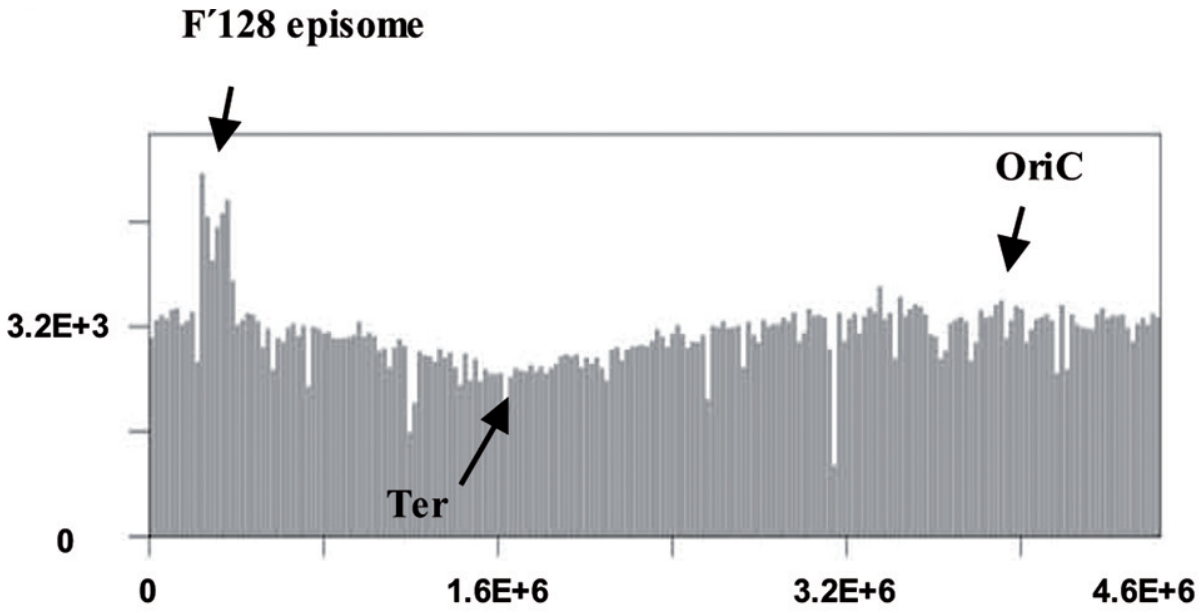
Reproducible bias with GC-content of fragment (mostly related to PCR amplification efficiency).

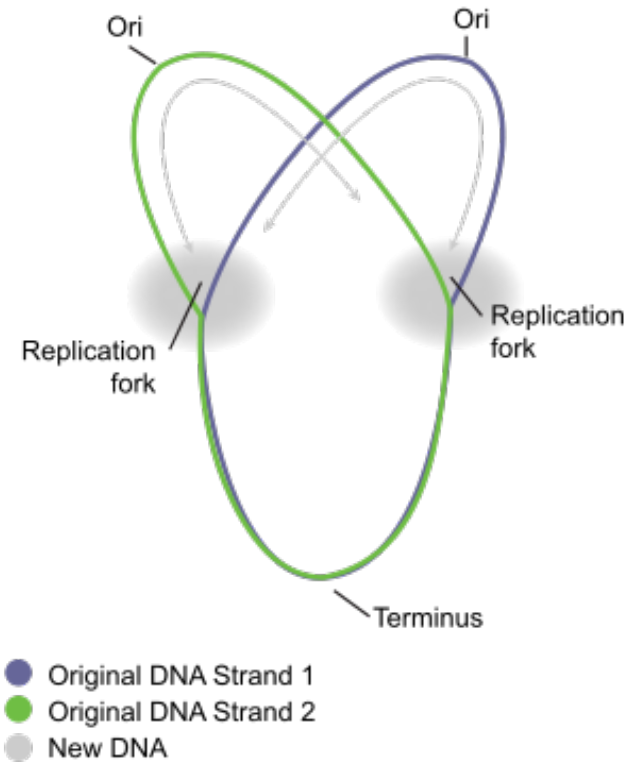
Bias can be very reproducible for a region – like a "fingerprint" between samples.



Benjamini and Speed. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *NAR* **40**: e72.

Coverage Bias





Catching replication in the act.

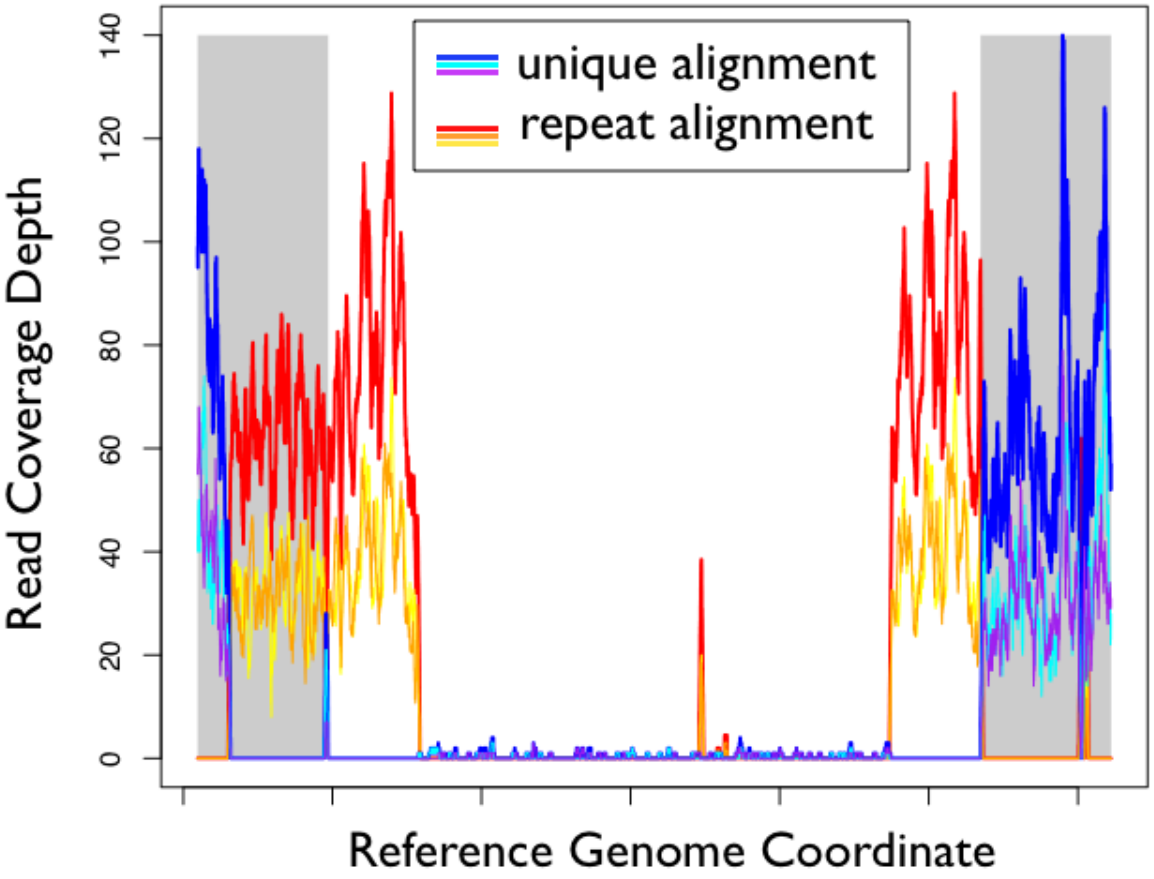
Parkhomchuk et al. (2009) Use of high throughput sequencing to observe genome dynamics at a single cell level. *PNAS* **106**: 20830–20835.

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

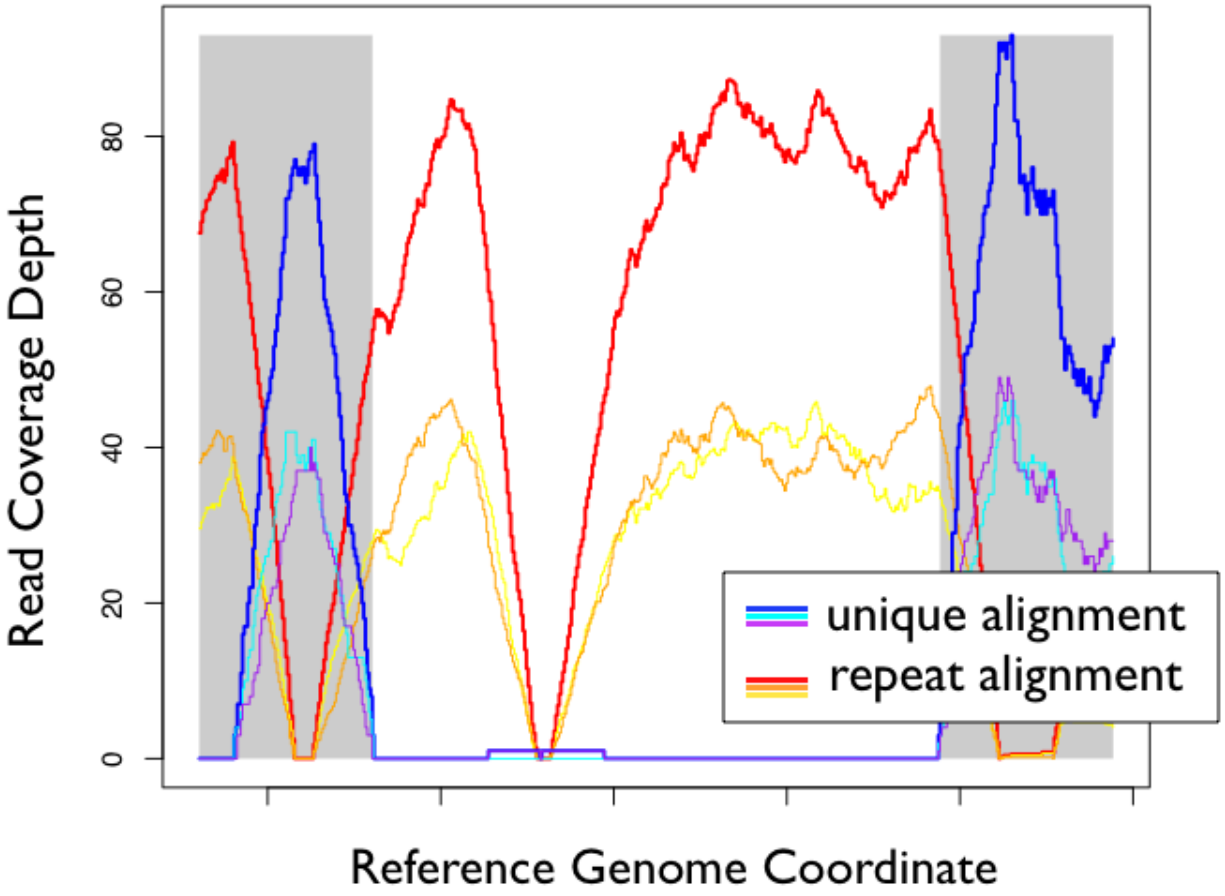
- Sometimes the molecular event is obvious...



- Recombination between nearby IS3 copies.

Example of a ***breseq*** prediction

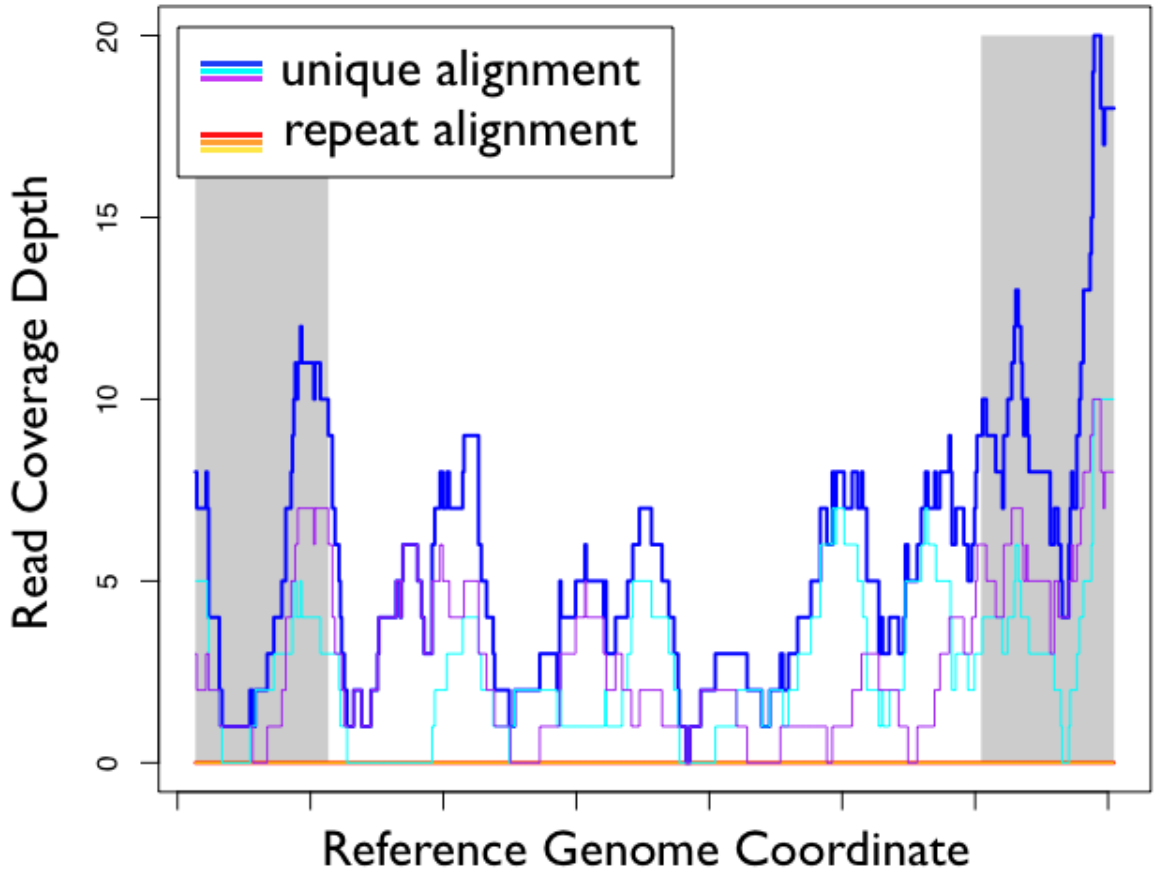
- Sometimes the mutation is not obvious...



- 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

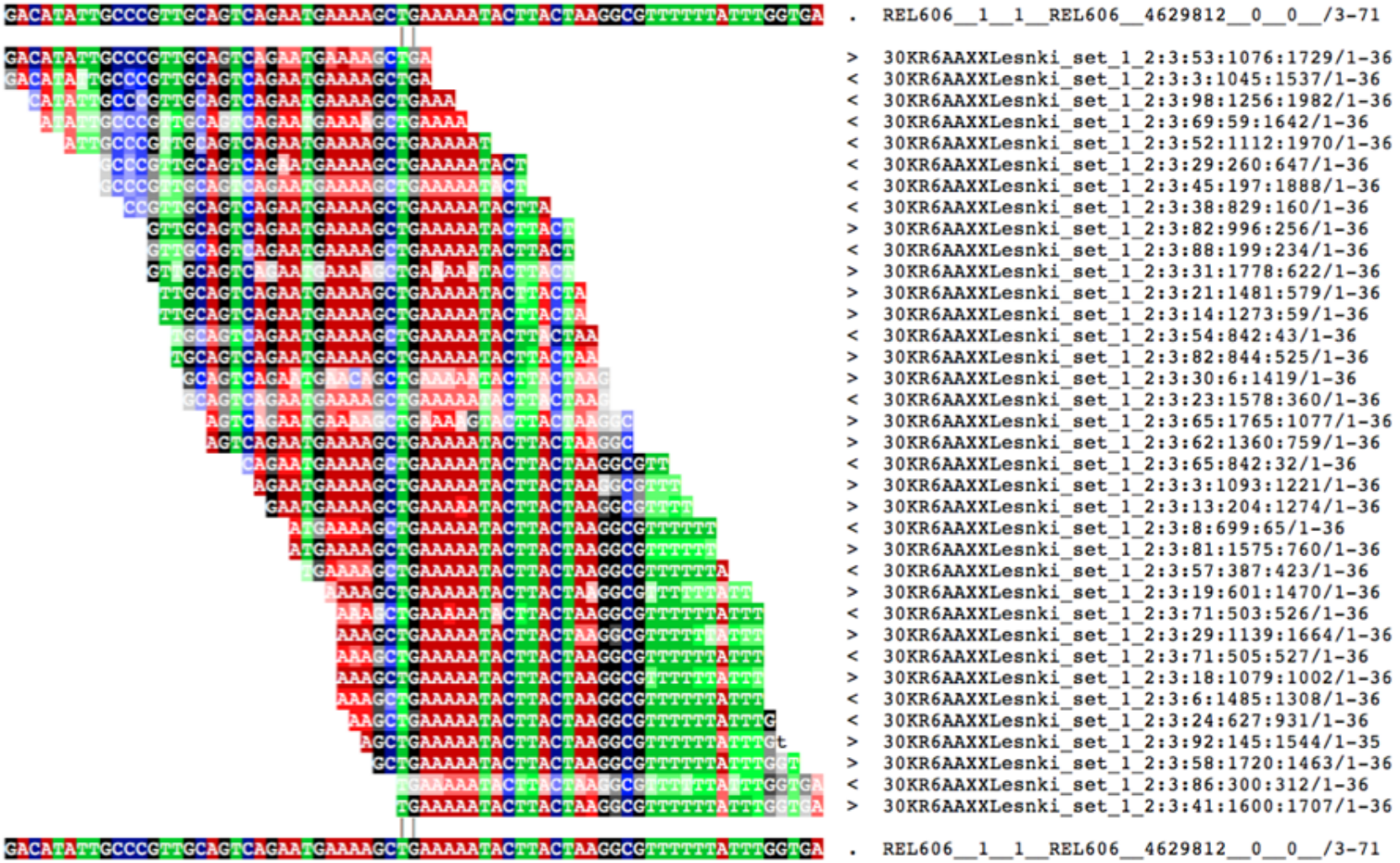
1. Find "mosaic" reads that partially map to two locations in the genome (possibly with overlap).

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 good junction

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

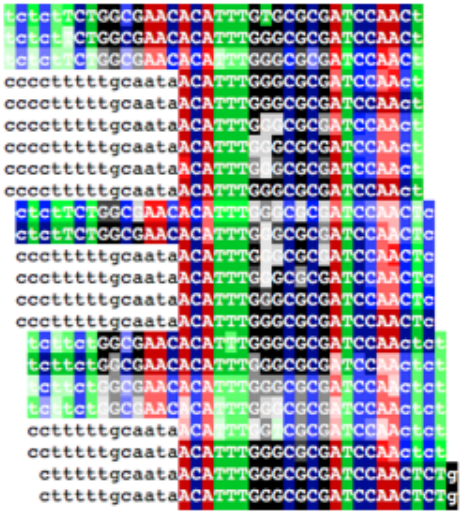


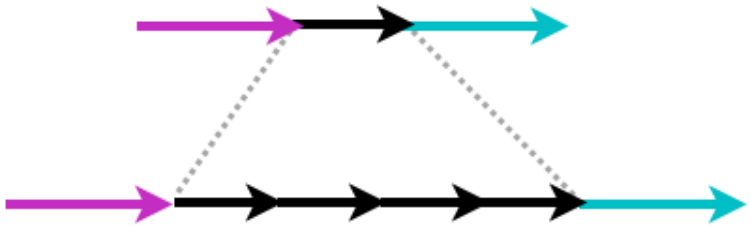
Example of a ***breseq*** prediction

- 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 both new and old junctions exist...

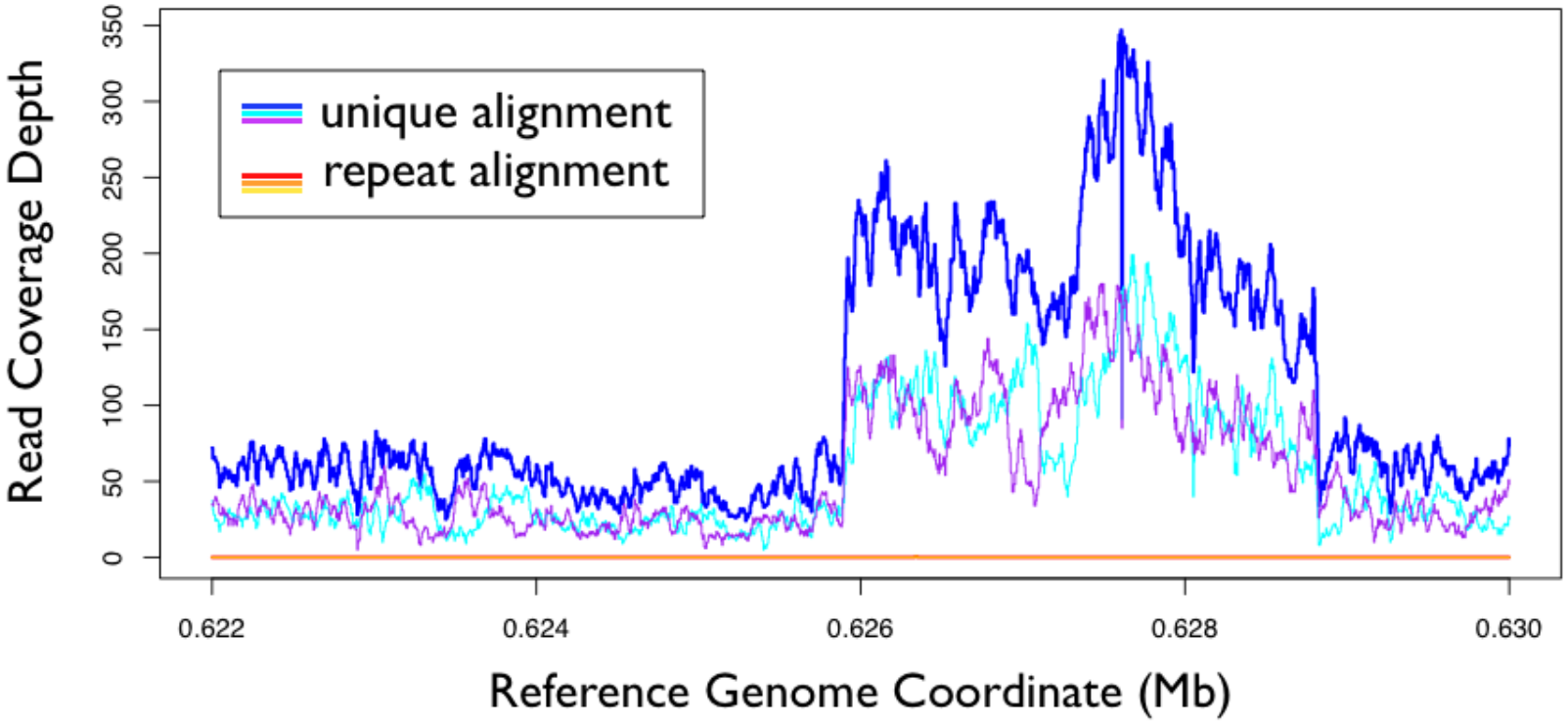




tandem head-to-tail duplications

Identifying copy number variation

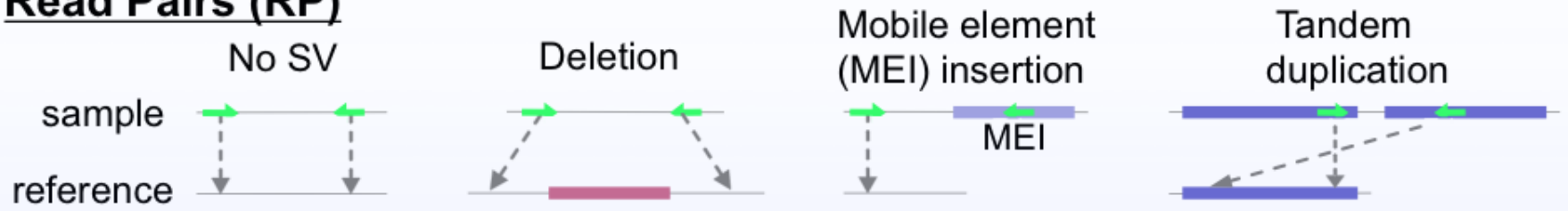
- Coverage is very noisy, but a fingerprint is (somewhat) consistent across runs.



- Tile into segments, train model on many genomes, look for deviation

Predicting structural variants

Read Pairs (RP)

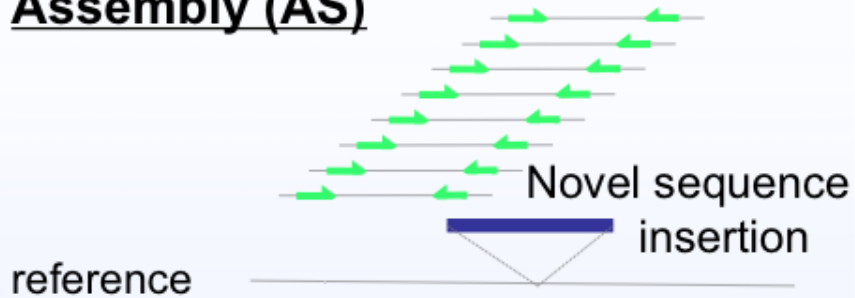


Read Depth (RD)

Unfortunately there is no program or pipeline that does **all** these things!!

Split Reads (SR)

Assembly (AS)



(http://www.genome.gov/Pages/Research/DER/1000GenomesProjectTutorials/StructuralVariants-JanKorbel.pdf)