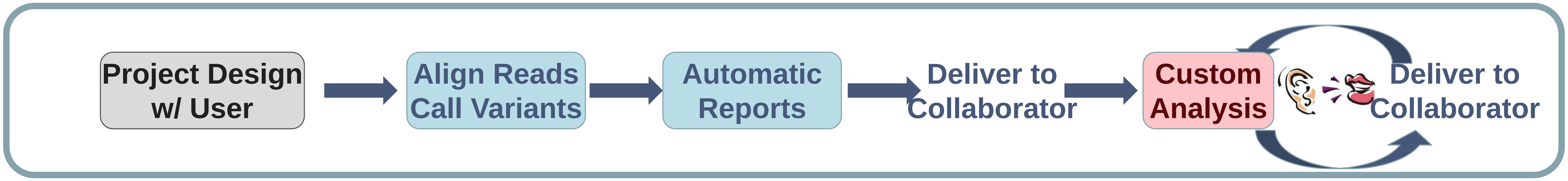
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Resequencing: The Untold Story

Recognizing False Positives, False Negatives and Structural Variation in User Data

Anna Lipzen[1], *Wendy Schackwitz* [1], Joel Martin[1], Len A. Pennacchio[1]

DOE Joint Genome Institute, Walnut Creek, CA USA



Project Design w/ User → Align Reads Call Variants → Automatic Reports → Deliver to Collaborator → Custom Analysis ⇄ Deliver to Collaborator

Project Design

Assist User in Choosing Experimental Design

Before a project begins we have one or more conference calls with the collaborator so we understand the design and goal of their experiment. By understanding the needs of the collaborator we can assist them in choosing the best products the JGI has to offer for their particular experiment.

Tailor Parameters &Tools to Organism & Experiment

To evaluate new tools for implementation and determine optimum parameter settings we have generated test data sets which have distinctive characteristics: haploid/diploid, closely related/divergent, low depth/moderate depth/high depth. Our analysis shows that there is not a single optimal variant caller or parameter setting, rather it depends upon the data and if the collaborator is more sensitive to false positive or false negative calls. We therefore, customize the caller and parameters to the data and the collaborator's specific needs.

Haploid - divergent

	bcftools -B -w0W0	bcftools default	maq default
Found	98%	90%	94%
False Positive	53%	26%	2%
False Negative	2%	10%	6%

Diploid - conserved

	bcftools -B -w0W0	bcftools default	maq default
Found	95%	75%	60%
False positive	5%	2%	2%
False negative	5%	25%	40%

Automatic Pipeline

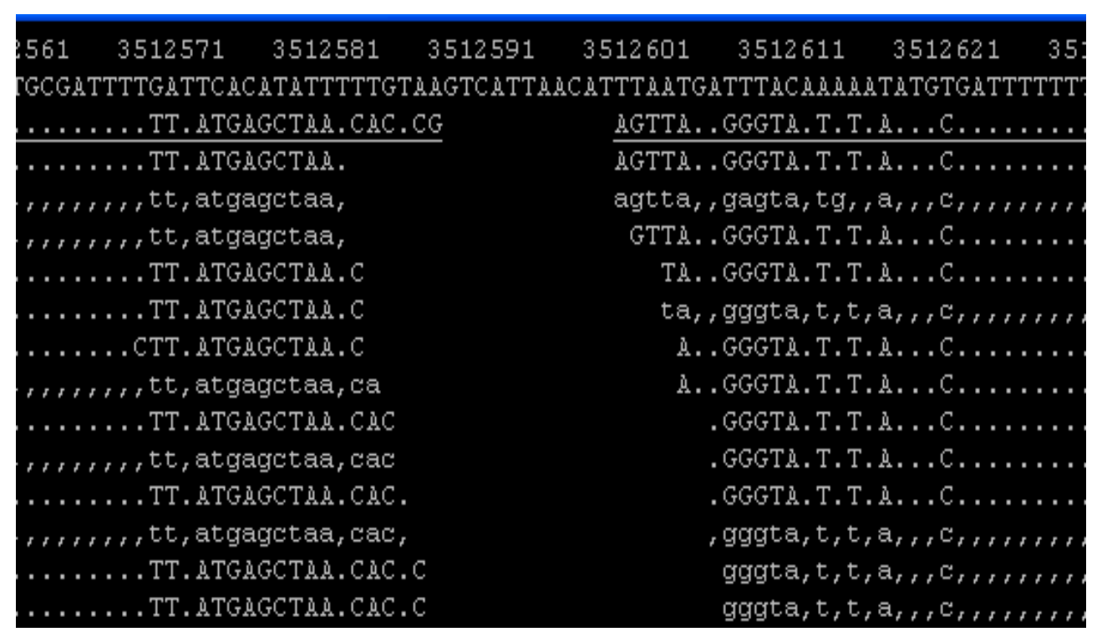
High Throughput, Fast Turn Around

To make high-throughput analysis possible, we created a pipeline that automatically generates analysis reports and files. These are uploaded to the collaborator's website giving them immediate access to their data so they can begin their analysis. The reports and files are explained in a detailed "README" file. An example of one type of report is shown below.

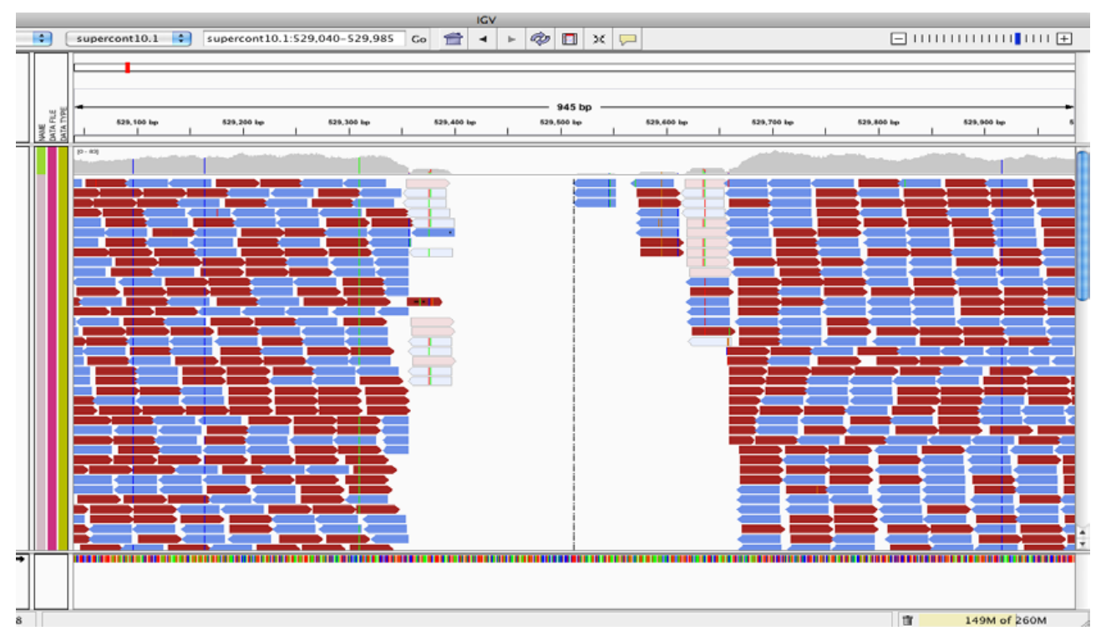
#contig	pos	type	name	strand	ref_nt	cds_nt	ref_codon	cds_aa	C110	C149
ctg1	70473	Int	GENE1	+	T	NC	NC	NC	SNP/99/24/11.69/C/C:25/T:0/NC	SNP/105/26/12.56/C/C:27/T:0/NC
ctg1	193822	NC	NC	NC	C	NC	NC	NC	SNP/44/70/6.50/Y/C:57/T:15/NC	SNP/111/29/10.62/Y/C:20/T:9/NC
ctg1	4113261	CDS	GENE4	+	C	28	Q:CAG	10	/132/35/1.00/C/C:36/N:0/Q:CAG	SNP/84/19/0.75/T/T:21/C:0/-:TAG

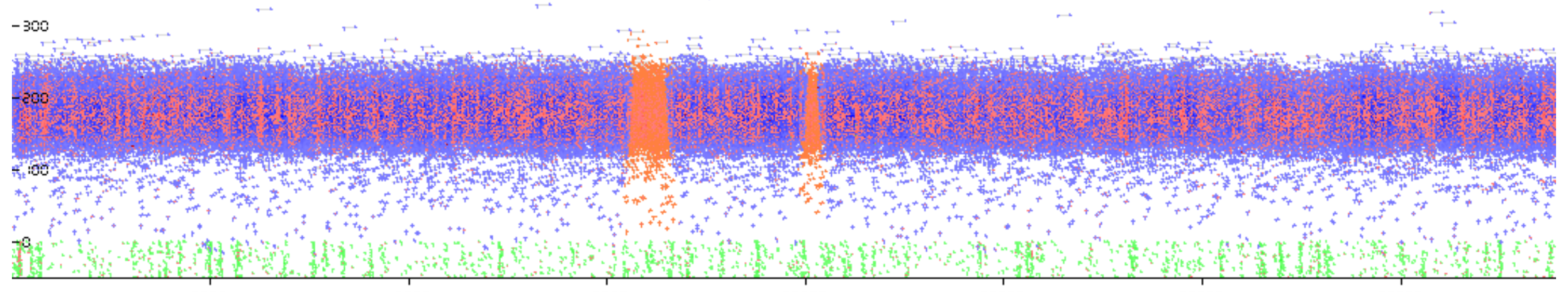
Standard Output Allows User to Plug & Play

The Variant Call Format (vcf), originated by the human 1000 genome project, is quickly becoming the standard for variant calls. By providing our variant calls in this format, it is possible to leverage the many tools the community is developing to work with vcf. For read alignments, bam is the standard format. For each experiment we provide the collaborator the bam file, which they can then load into their favorite tool to visualize their data.



```
561    3512571    3512581    3512591    3512601    3512611    3512621    35
TGCGATTTTGATTCACATATTTTTGTAAGTCATTAACATTTAATGATTTACAAAAATATGTGATTTTTTT
........TT.ATGAGCTAA.CAC.CG             AGTTA..GGGTA.T.T.A...C........
........TT.ATGAGCTAA.                   AGTTA..GGGTA.T.T.A...C........
,,,,,,,,tt,atgagctaa,                   agtta,,gagta,tg,,a,,,c,,,,,,,,
,,,,,,,,tt,atgagctaa,                    GTTA..GGGTA.T.T.A...C........
........TT.ATGAGCTAA.C                     TA..GGGTA.T.T.A...C........
........TT.ATGAGCTAA.C                     ta,,gggta,t,t,a,,,c,,,,,,,,
.......CTT.ATGAGCTAA.C                      A..GGGTA.T.T.A...C........
,,,,,,,,tt,atgagctaa,ca                     A..GGGTA.T.T.A...C........
........TT.ATGAGCTAA.CAC                      .GGGTA.T.T.A...C........
,,,,,,,,tt,atgagctaa,cac                      .GGGTA.T.T.A...C........
........TT.ATGAGCTAA.CAC.                     .GGGTA.T.T.A...C........
,,,,,,,,tt,atgagctaa,cac,                     ,gggta,t,t,a,,,c,,,,,,,,
........TT.ATGAGCTAA.CAC.C                     gggta,t,t,a,,,c,,,,,,,,
........TT.ATGAGCTAA.CAC.C                     gggta,t,t,a,,,c,,,,,,,,
```



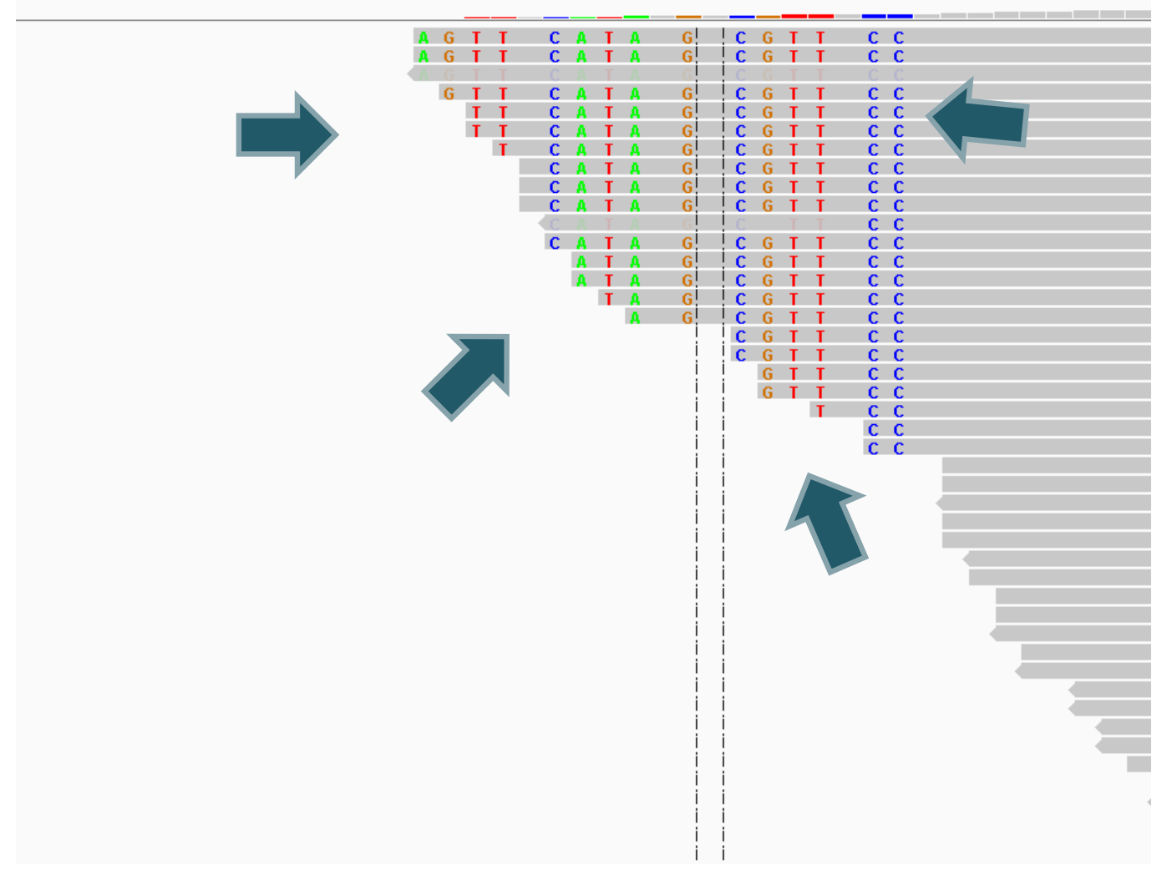


Custom Analysis

One of the biggest values our team brings is the 20 years of combined experience analyzing Re-Seq data. Additionally, the JGI has worked on a huge variety of projects, giving us unmatched exposure to Re-Seq data. This experience is used to assist the collaborator with interpreting their results. Below are several examples of false calls that we can identify. Common sources of false positives include: edges of structural variation, Illumina sequence specific errors, collapsed repeats & ambiguously mapped reads. Sources of false negatives include: library bias and sequence divergence.

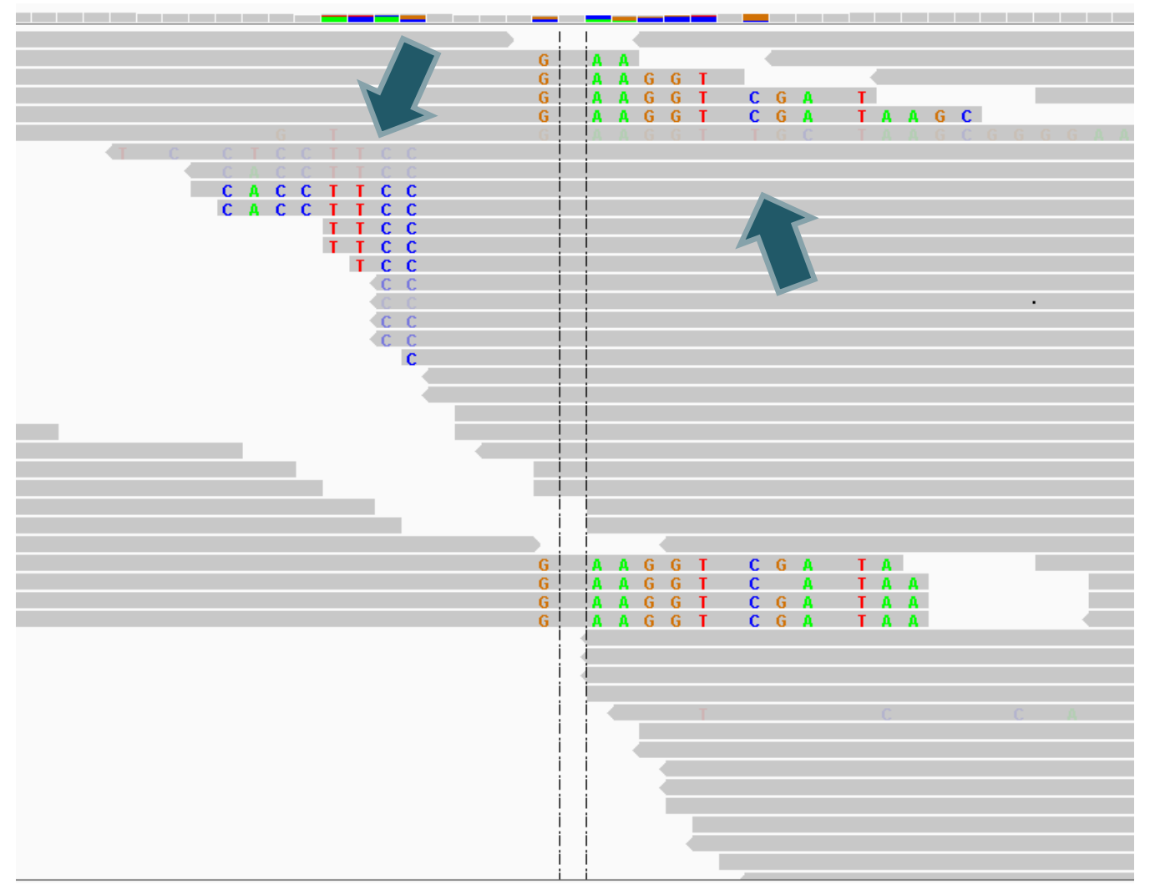
False Positives

Edge of Large Deletion



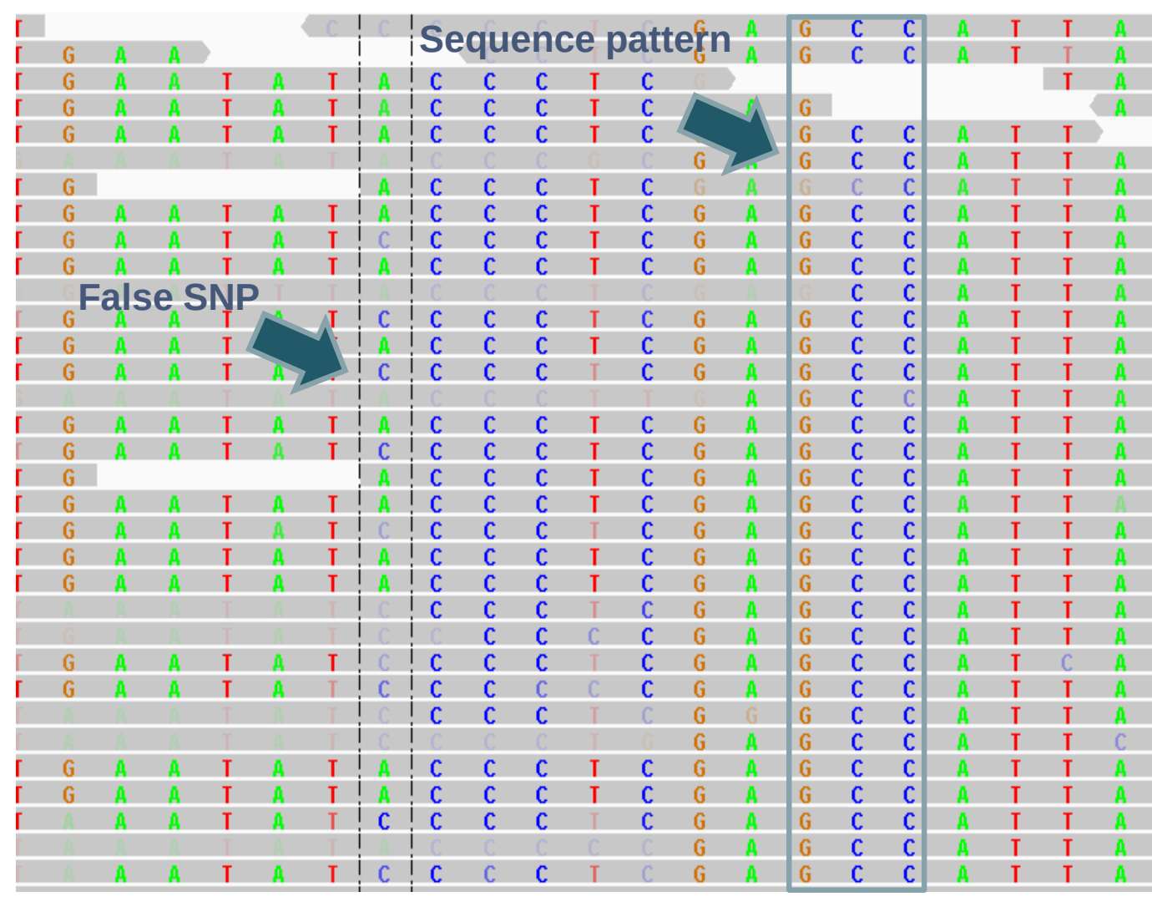
This is a pattern of a large deletion. Mismatched bases match upstream, at the start of this structural variation. All these were falsely called as SNPs.

Edge of Large Insertion



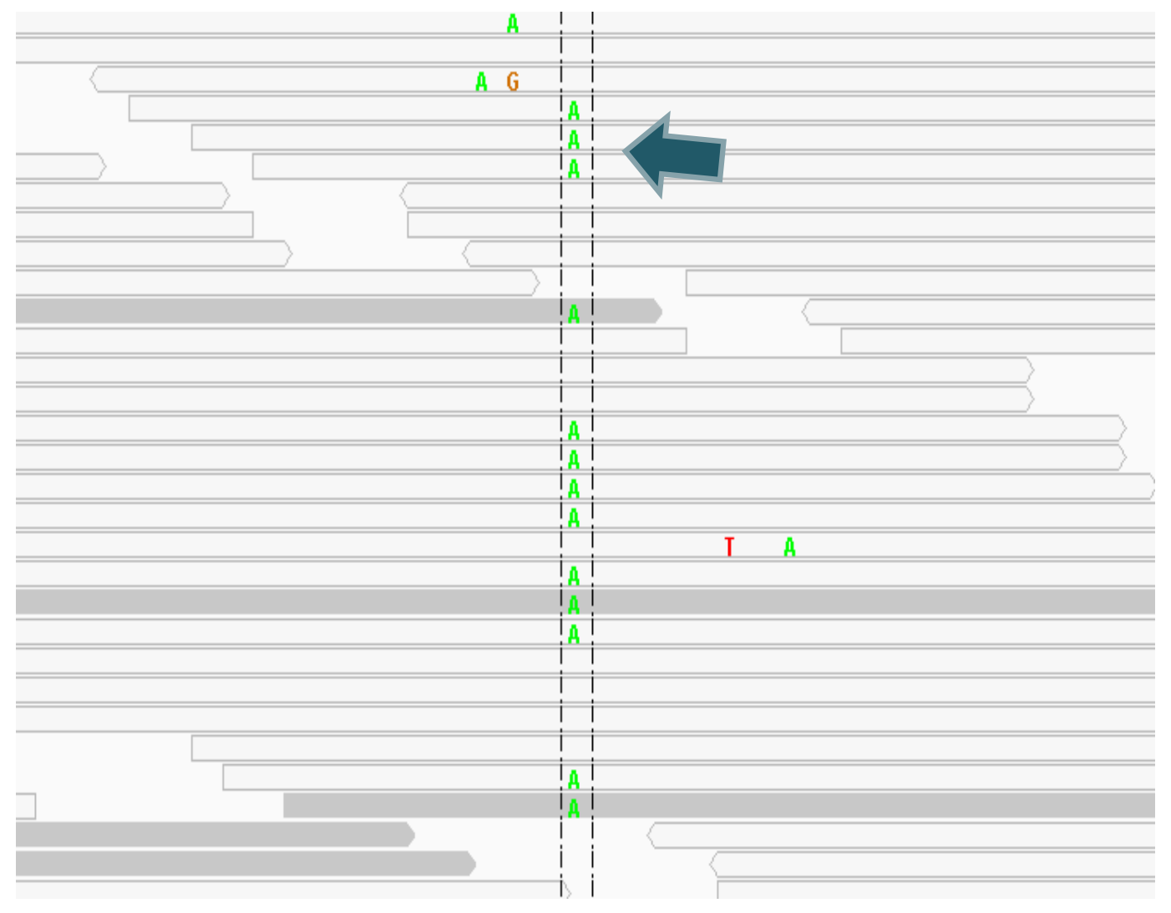
An insertion of transposable element occurred at this position. Misalignments at the edge of structural variations trick variant callers into calling SNPs.

Sequencer-originated miscalls



Certain sequence context can make reads prone to Illumina sequence-specific error[1]. This error results in strand-biased false calls.

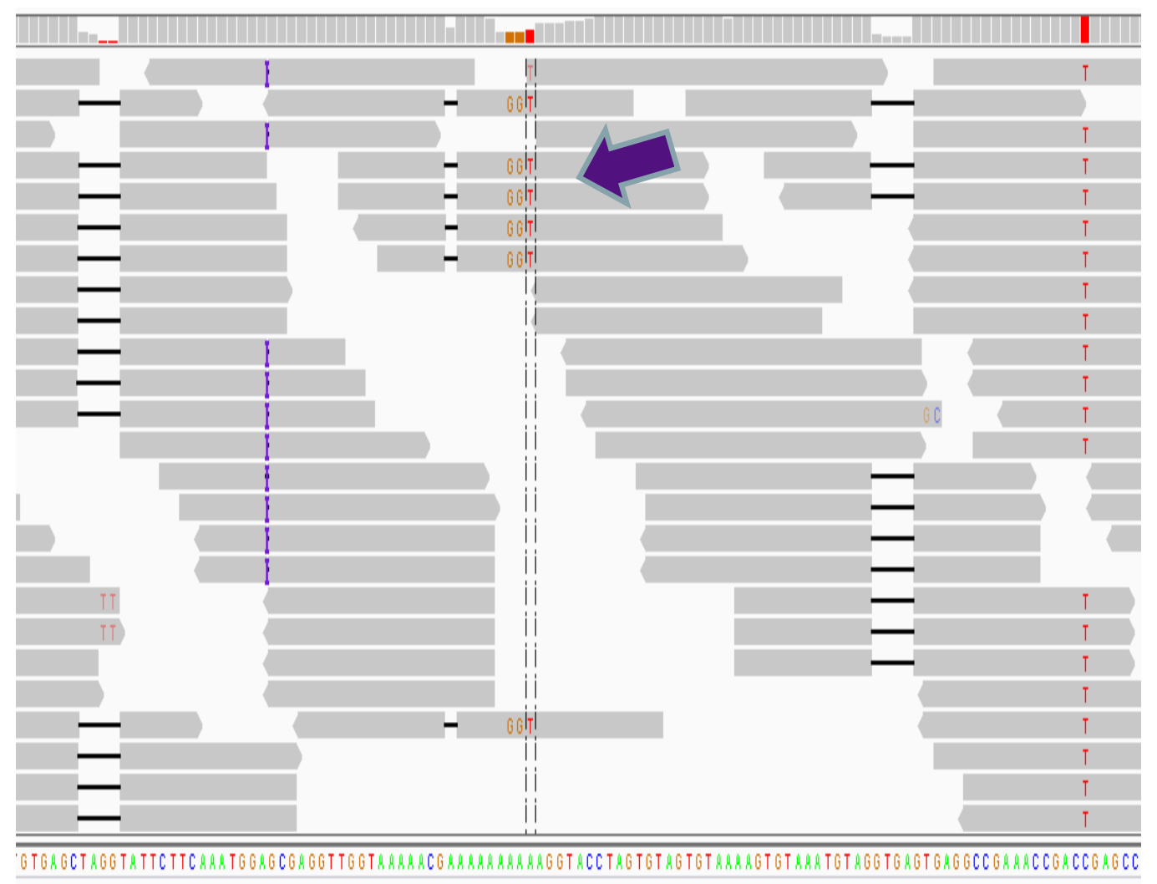
Ambiguously mapped reads



This was a multi-allelic call in a haploid genome. This is likely a real variant and incorrect call is due to reads mapping ambiguously in a repetitive region.

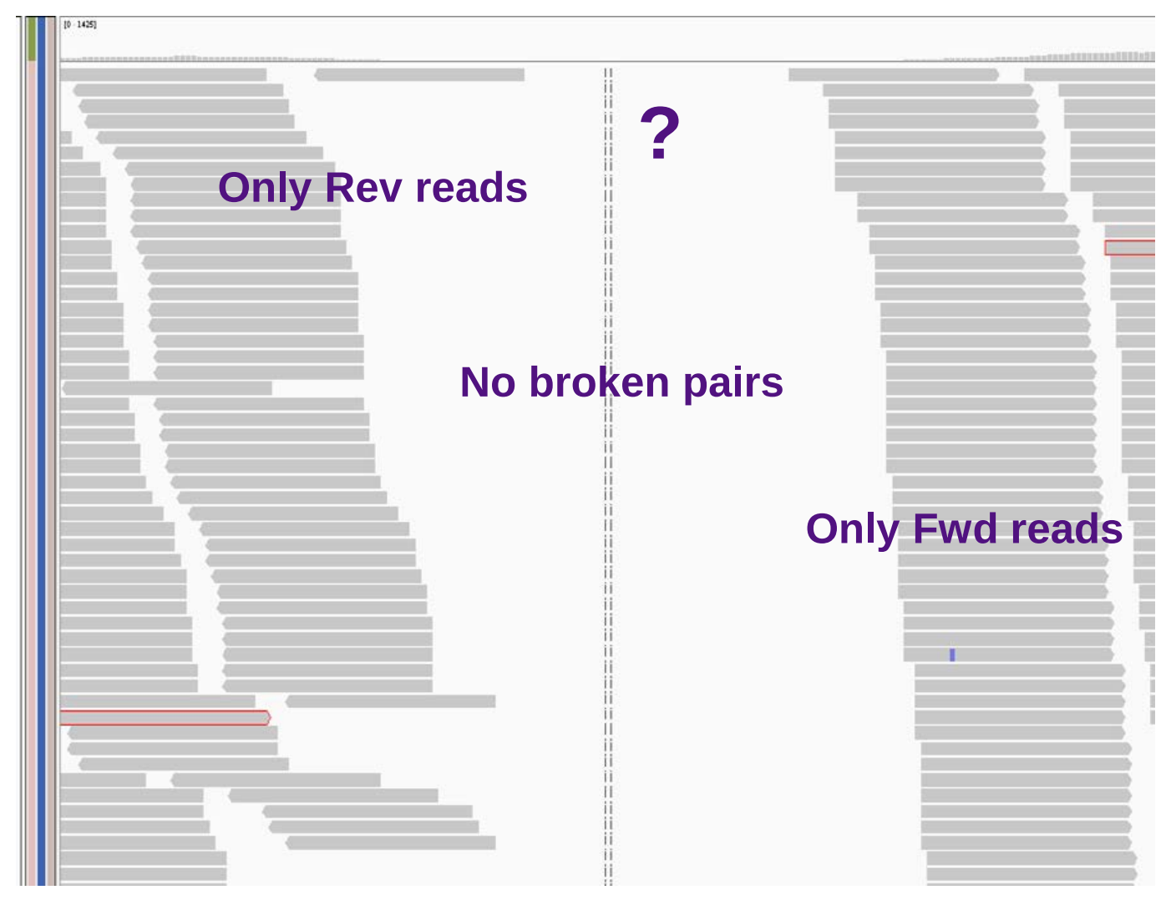
False Negatives

SNP dense regions



In divergent regions, real calls may go undetected. Here, variant caller missed the A>T SNP.

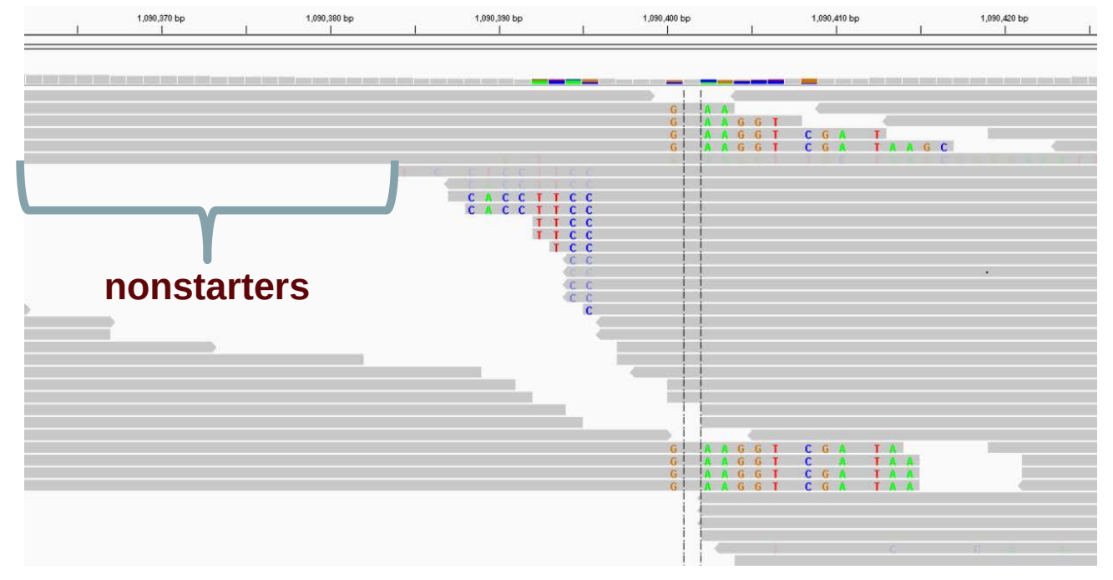
Library bias



This region simply did not get covered by sequencing, therefore we cannot determine anything about variation there.

Structural Variation

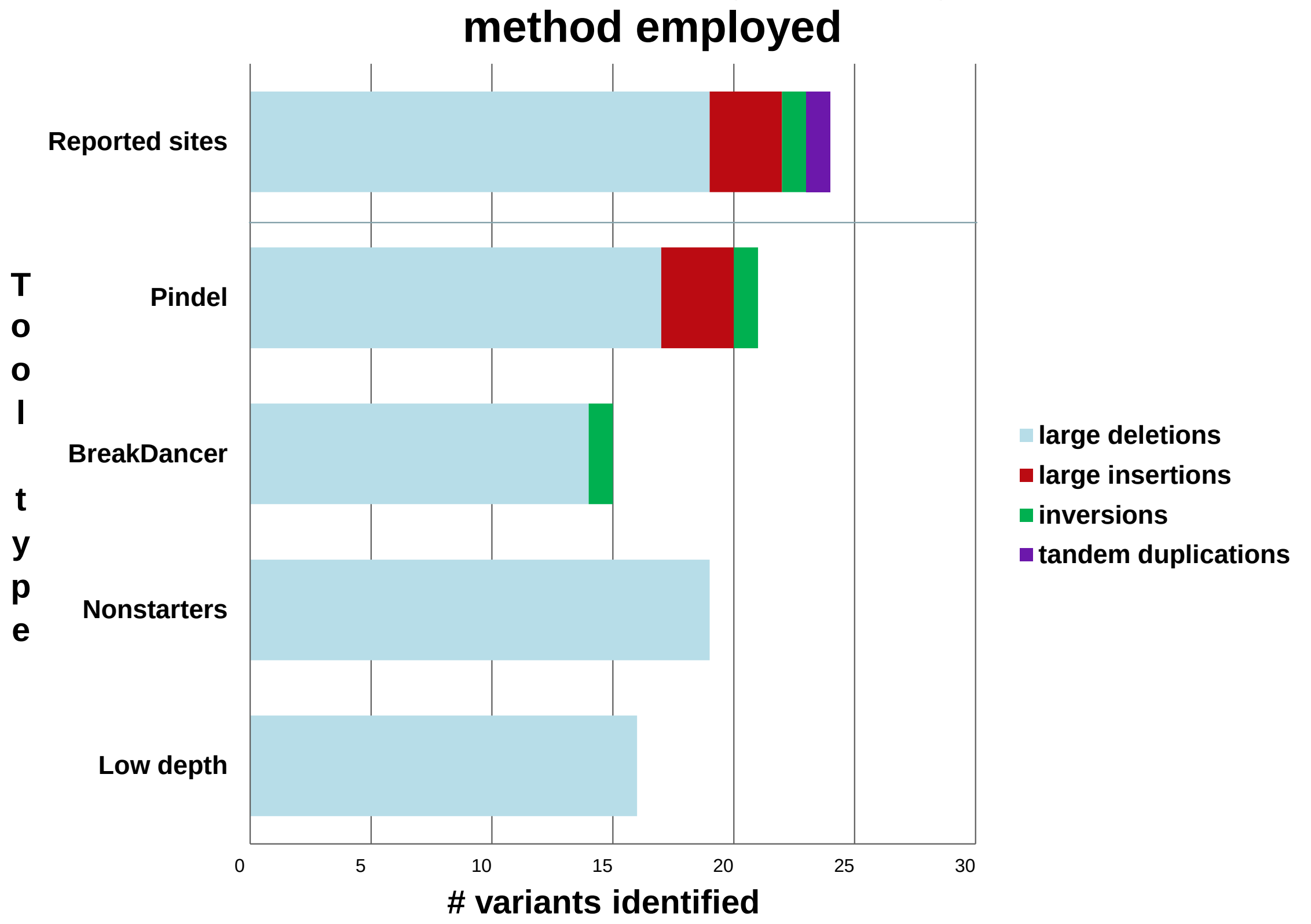
We use several methods for detecting structural variants.   BreakDancer[2] and Pindel[3] compute the SV breakpoints based on read mapping results and the reference genome. For projects with overall high sequence coverage, low depth regions and regions where no reads begin ("nonstarters") often flag certain SV events.   Some tools are quite good at identifying that SV exists, but they are unable to pin point the precise location of the event. We manually examine these sites to attempt to give an exact result.



To test how well we do, we aligned Escherichia coli MG1655 reads to Escherichia coli DH10B reference[4], and called the structural variants between the two strains. No single tool found all the expected variants. By using a combination of methods, with exception of one tandem duplication, we identified all the expected large deletions, large insertions and inversions.

```
3512570 Deletion 3512573-3512629. -56bp.
ref     AGTTAGAGGGTAATATAAATGCGATTTTGATTCACATATTTTTGTAAGTCATTAACATTTAATGATTTACAAAAATATGTGATTTTTTTATGAGCTAATCACTCG
read 1  AGTTAGagggtaatataaatgcgatttt**********************************************************tttatgagctaatcactcg
read 2  AGTTAGAGGGTAATATAAATGCGATTTT**********************************************************TTTATGAGCTAATCACTCG
```

Success rate of SV discovery varies by detection method employed



References:
1. Nakamura, K. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Research, 39(13), 1–13 (2011).
2. Chen, K. et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature Methods, 6, 677-681 (2009).
3. Ye, K et al. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics, 25(21), 2865-71 (2009).
4. Durfee T et al. The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. Journal of Bacteriology, 190(7), 2597-606 (2008).

LBNL-5874E-Poster

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.