

# Resequencing: The Untold Story – Recognizing False Positives, False Negatives and Structural Variation in user Data

Anna Lipzen<sup>1Δ</sup>, Wendy Schackwitz<sup>1</sup>, Joel Martin<sup>1</sup>, Len A. Pennacchio<sup>1</sup>

<sup>1</sup> Genomics Division, Lawrence Berkeley National Lab, Berkeley, CA / Department of Energy  
Joint Genome Institute, Walnut Creek, CA

<sup>Δ</sup>To whom correspondence may be addressed. E-mail: [alipzen@lbl.gov](mailto:alipzen@lbl.gov)

March 25, 2012

## **ACKNOWLEDGMENTS:**

*The work conducted by the US Department of Energy (DOE) Joint Genome Institute is supported by the Office of Science of the DOE under Contract Number DE-AC02-05CH11231. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government, or any agency thereof, or the Regents of the University of California.*

## **DISCLAIMER:**

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California



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

#### Diploid - conserved

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

## Automatic Pipeline

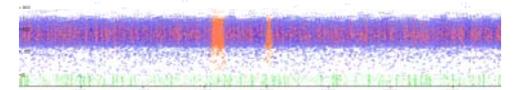
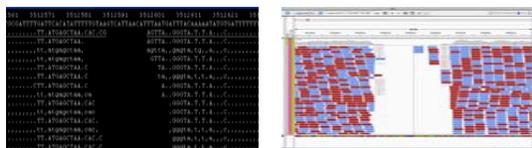
### High Throughput, Fast Turn Around

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	alt	ref	cds	ref	codon	cds	aa	C110	C149
clg1	70473	Int	GENE1	+	T	NC	NC	NC	SNP9924/11.69/C/C:25/T/0/NC	SNP10526/12.56/C/C:27/T/0/NC				
clg1	193822	NC	NC	NC	NC	NC	NC	NC	SNP44/70/6.50/V/C:5/7/1/15/NC	SNP11129/10.62/V/C:20/T/9/NC				
clg1	413261	CDS	GENE4	+	C	28	O:CAG	10	/132735/1.00/C/C:36/N/0/C:CAG	SNP847/10.75/T/121/C/0/1-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.



## Custom Analysis

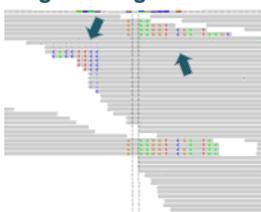
One of the biggest values our team brings is the 20 years of combined experience analyzing Re-Seq data. Additionally, the JGI has worked on a huge variety of projects, giving us unmatched exposure to Re-Seq data. This experience is used to assist the collaborator with interpreting their results. Below 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

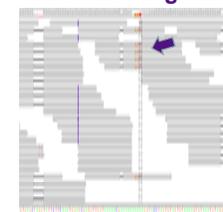


#### Edge of Large Insertion

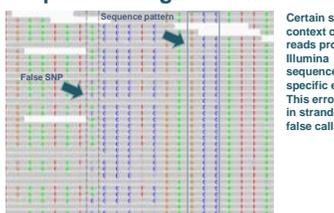


### False Negatives

#### SNP dense regions



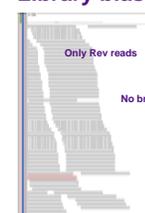
#### Sequencer-originated miscalls



#### Ambiguously mapped reads



#### Library bias



## Structural Variation

We use several methods for detecting structural variants. BreakDancer<sup>2</sup> and Pindel<sup>3</sup> 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.



3512570 Deletion 3512573-3512629. -56bp.  
ref AGTTAGAGGGTAATAAAGGCGATTTTGGTACACATATTTGCTAGCCTTTAATCAATTTAAGCATTTAATAAATAAGCGATTTTATATGAGCTAATCACTCG  
read 1 AGTTAGAGgggtAaataaAaagGogAtttt\*\*\*\*\*tttAagctAaAtCAaCTCG  
read 2 AGTTAGAGGGTAATAAAGGCGATTTT\*\*\*\*\*TTTATGAGCTAATCACTCG

References:  
1. Nakamura, K. Sequence-specific error profiles of Illumina sequencing. *Nucleic Acids Research*, 38(13), 1-13 (2010).  
2. Chen, X. 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).

### Success rate of SV discovery varies by detection method employed

