Structural Variant Combining Illumina and Low-Coverage PacBio

DNAnexus Andrew Carroll, DNAnexus

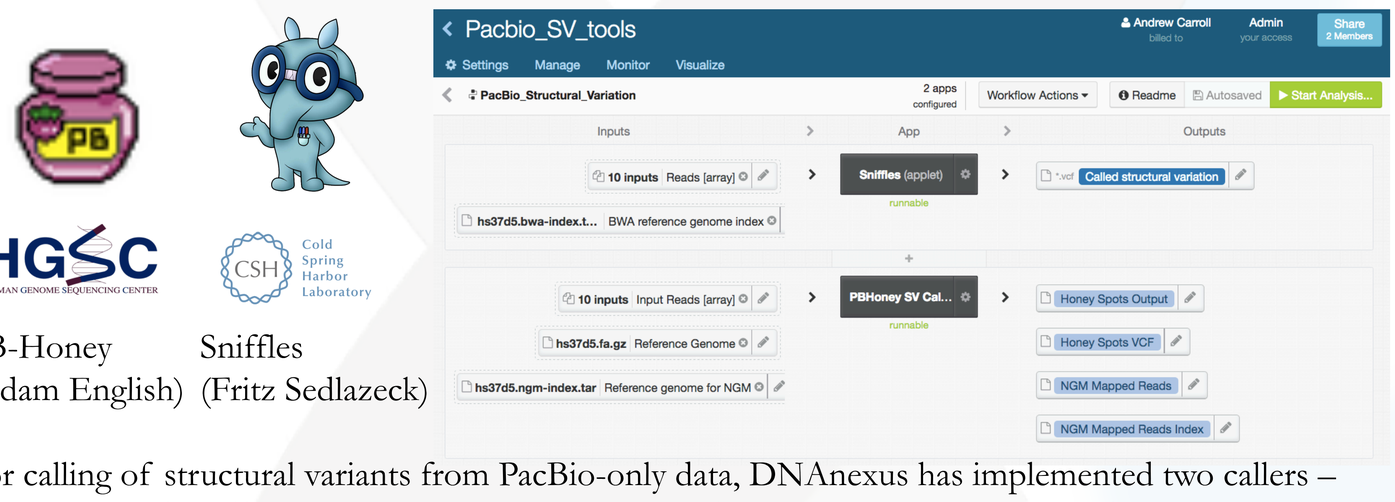
Abstract

Detection of large genomic variation - structural variants (SV) - has proven challenging using short-read methods. Long-read approaches, such as sequencing with Pacific Biosciences (PacBio), can span these large events directly and have shown promise to dramatically expand the ability to call structural variation. This poster investigates the following key questions: What coverages of PacBio sequencing are adequate for SV detection? What are the ability of current approaches to identify SVs in a sample? Can adding Illumina sequencing improve SV discovery at lower PacBio coverages?

Using a new SV benchmark set

For all assessments, we used a preliminary truth set generated by Justin Zook and Genome in a Bottle. To generate this set, calls were submitted from groups using multiple technologies (Illumina, PacBio, Complete, BioNano, Nabsys, and others. Justin generated a set of Deletion calls present in at least two technologies. This set the underpinning of the recall analysis.

Tools for calling SV in only PacBio data

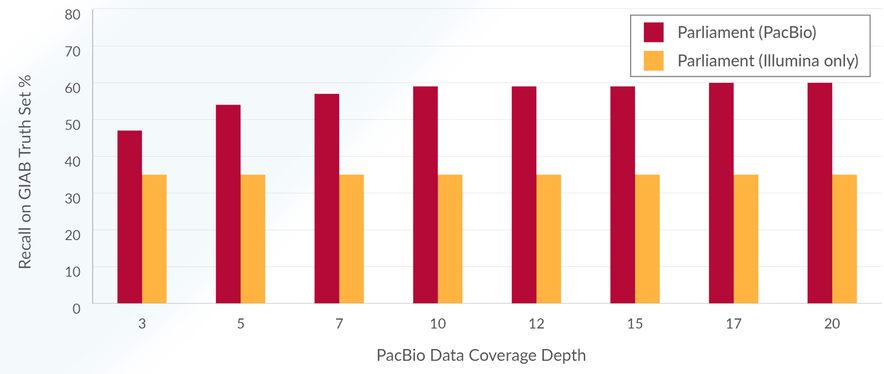


PB-Honey (Adam English) Sniffles (Fritz Sedlazeck)

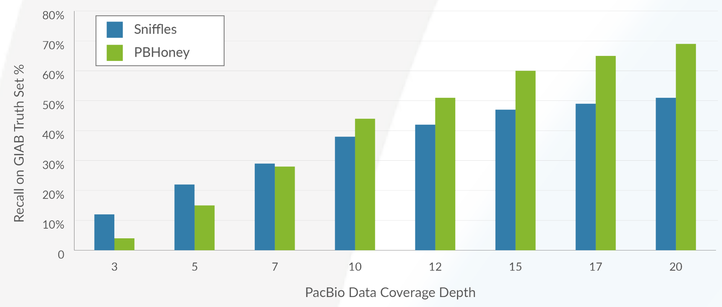
For calling of structural variants from PacBio-only data, DNAnexus has implemented two callers – PB-Honey and Sniffles. These exist as "apps" on DNAnexus – which are packaged tools that can be run either from the web interface (shown above) or through a command-line toolkit. These apps package all required installation and dependencies and allow analyses run in an easy and reproducible manner.

Parliament – an Illumina + PacBio SV caller

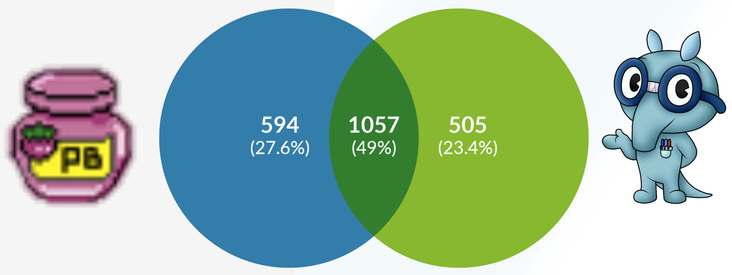
The Parliament pipeline, developed by Adam English with contributions from DNAnexus, combines Illumina and PacBio data to call SVs even at low coverage of PacBio data. Parliament can be run with only Illumina data, or with both Illumina and PacBio data. Parliament is available on DNAnexus.



PB-Honey and Sniffles are complementary to call SV effectively in 10-15 fold coverage



Overlap in True Calls at 10-fold Coverage

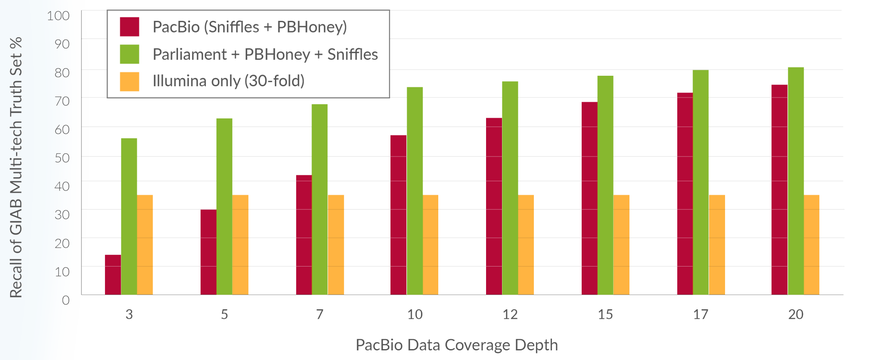


To assess the recall of the tools, we ran both PB-Honey and Sniffles on PacBio data from the Genome in a Bottle AJ-Trio (HG002).

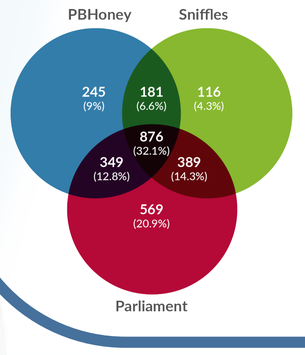
Sniffles is able to recall more SV events at lower coverages (3 fold to 7 fold), while PB-Honey is able to identify more structural variants at high coverages.

To understand the complementarity of the approaches, we looked at the overlap of the call set at 10-fold coverage (left). Because the approaches are different, the call sets can be combined to discover more SV events than by running either individually.

Best Results: ensemble approaches to SV-calling



Overlap in True Calls at 10-fold Coverage



All three methods (PB-Honey, Sniffles, and Parliament) find true calls in different ways.

By combining results, power discovery of structural variants is possible at 5-fold to 7-fold PacBio coverage when Illumina data is available and at 10-fold when not.

All tools can be easily run on DNAnexus.

Thanks:

Aaron Wenger
Tzvetana Kerelska
Luke Hickey
Adam English
Will Salerno
Fritz Sedlazeck
Justin Zook
Singer Ma
Samantha Zarate