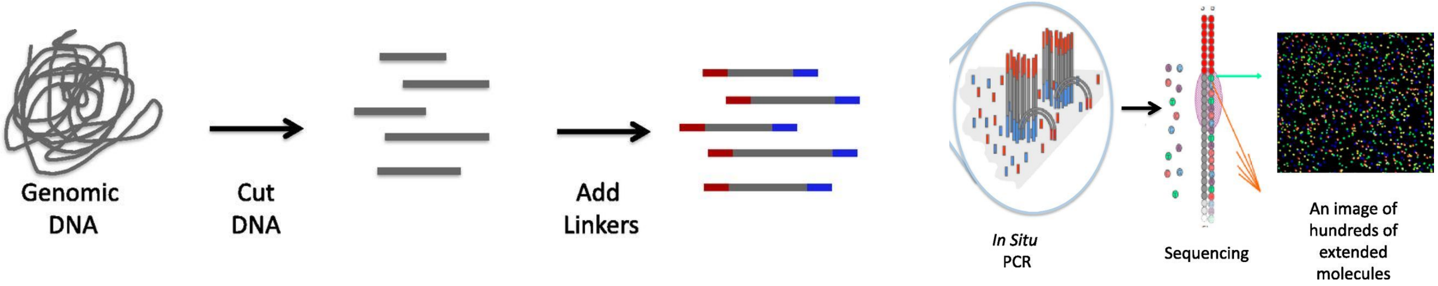
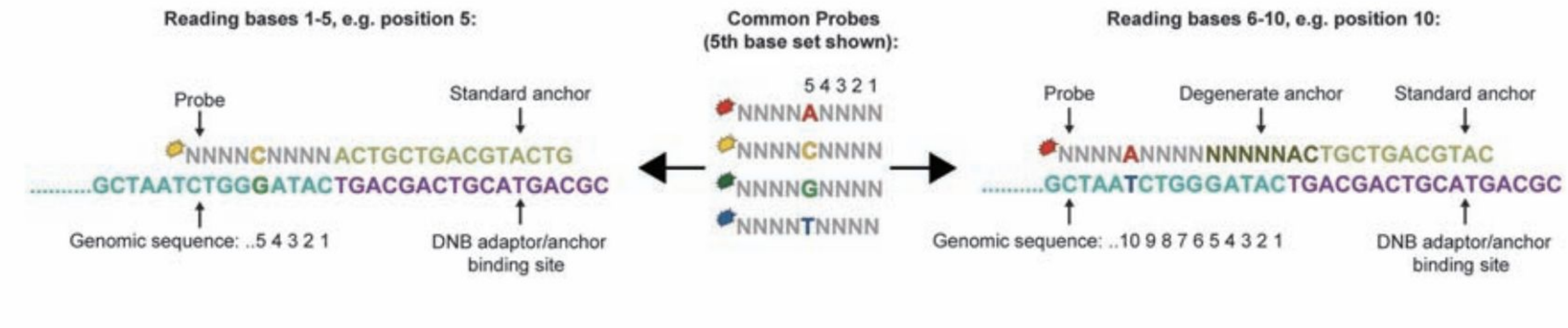
How Effective Are Illumina Methods for BGI-SEQ?

DNAnexus

Illumina - base by base



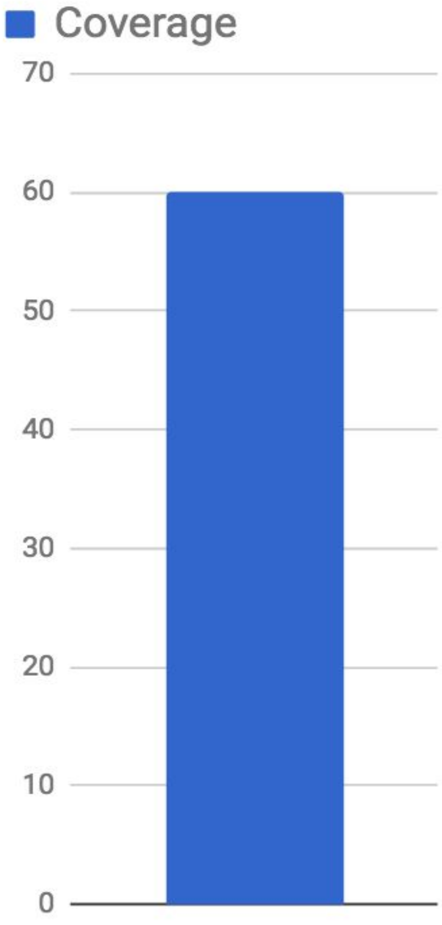
BGI-SEQ (Nanoball) - probe by probe

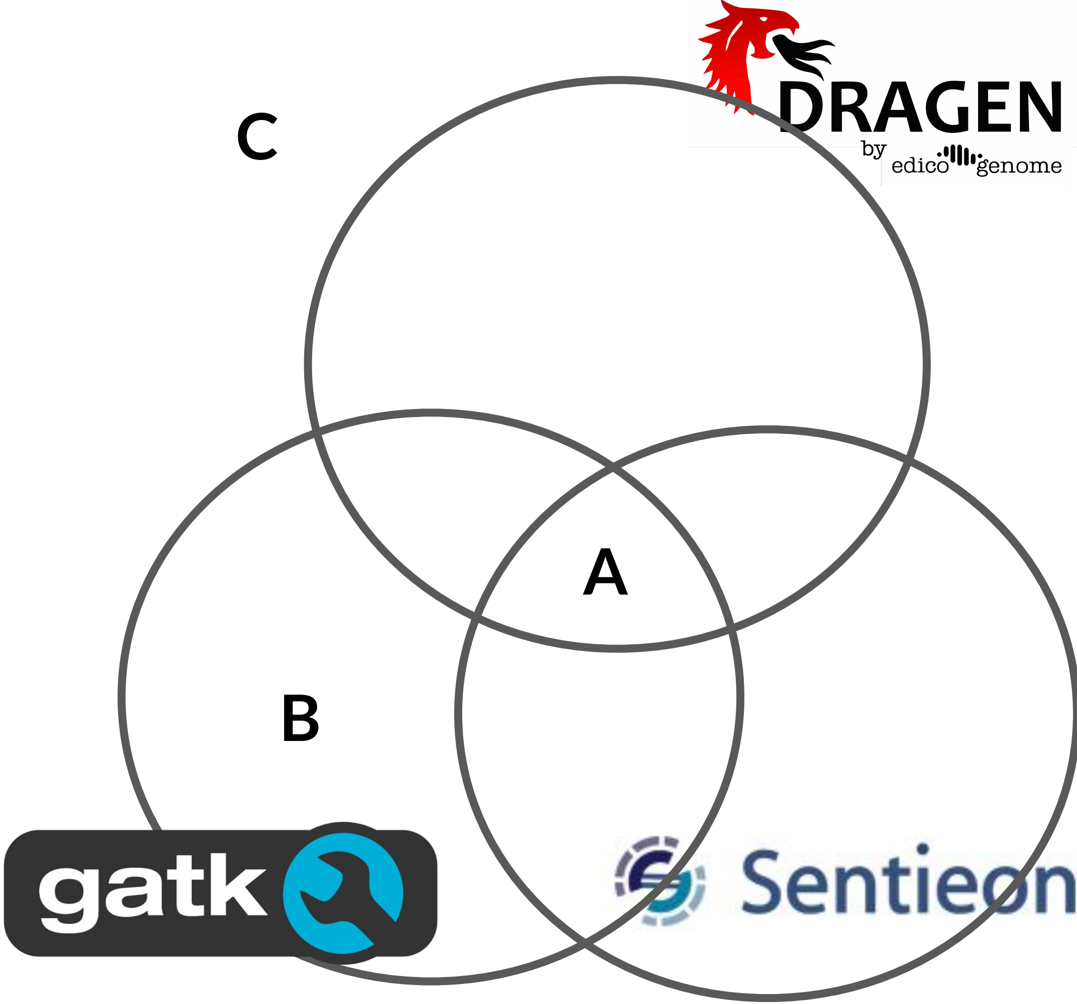

Reading bases 1-5, e.g. position 5:
Common Probes (5th base set shown):
Reading bases 6-10, e.g. position 10:
Probe
Standard anchor
NNNNCNNNNACTGCTGACGTACTG
..........GCTAATCTGGGATACTGACGACTGCATGACGC
Genomic sequence: ..5 4 3 2 1
DNB adaptor/anchor binding site
5 4 3 2 1
NNNNANNNN
NNNNCNNNN
NNNNGNNNN
NNNNTNNNN
Probe
Degenerate anchor
Standard anchor
NNNNANNNNNNNNNNACTGCTGACGTAC
..........GCTAATCTGGGATACTGACGACTGCATGACGC
Genomic sequence: ..10 9 8 7 6 5 4 3 2 1
DNB adaptor/anchor binding site

Incredibly Important Resources

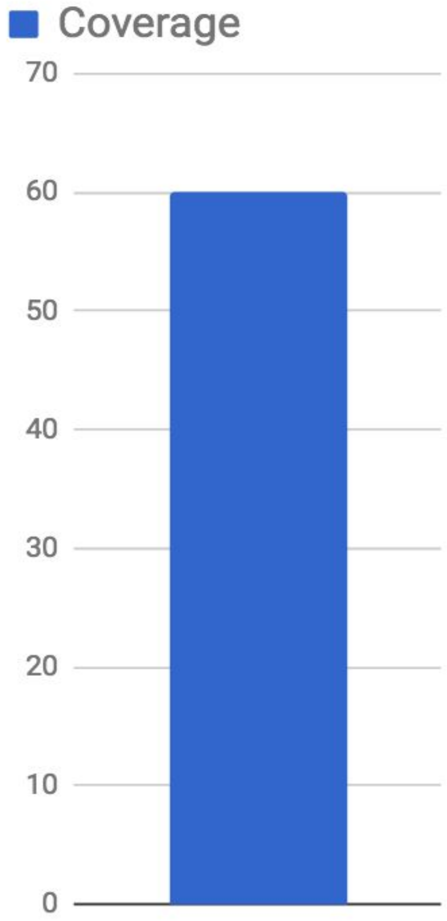
Only 3 Genomes, Not Somatic, Not Bill

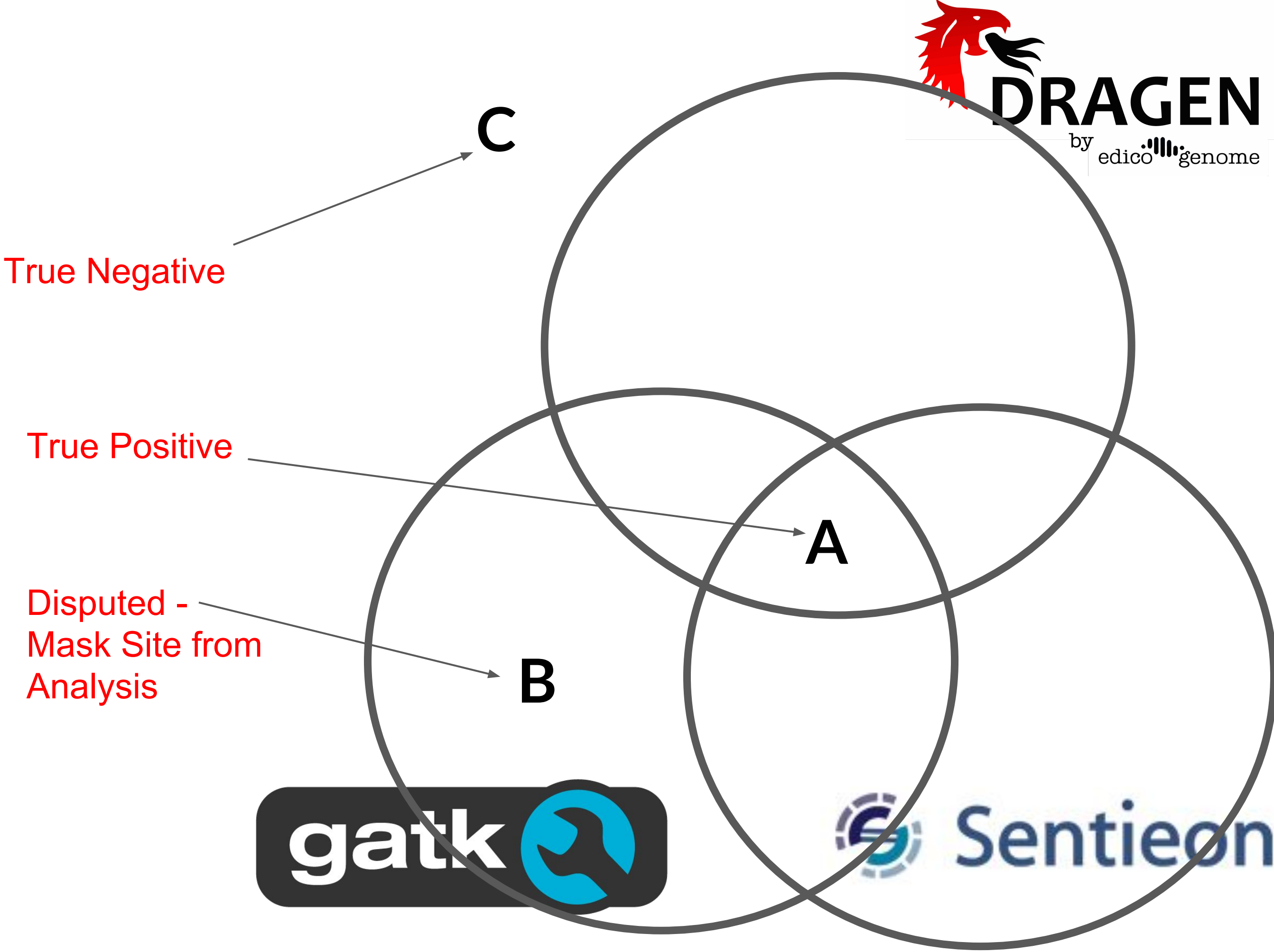
High Coverage - More Reliable





High Coverage - More Reliable



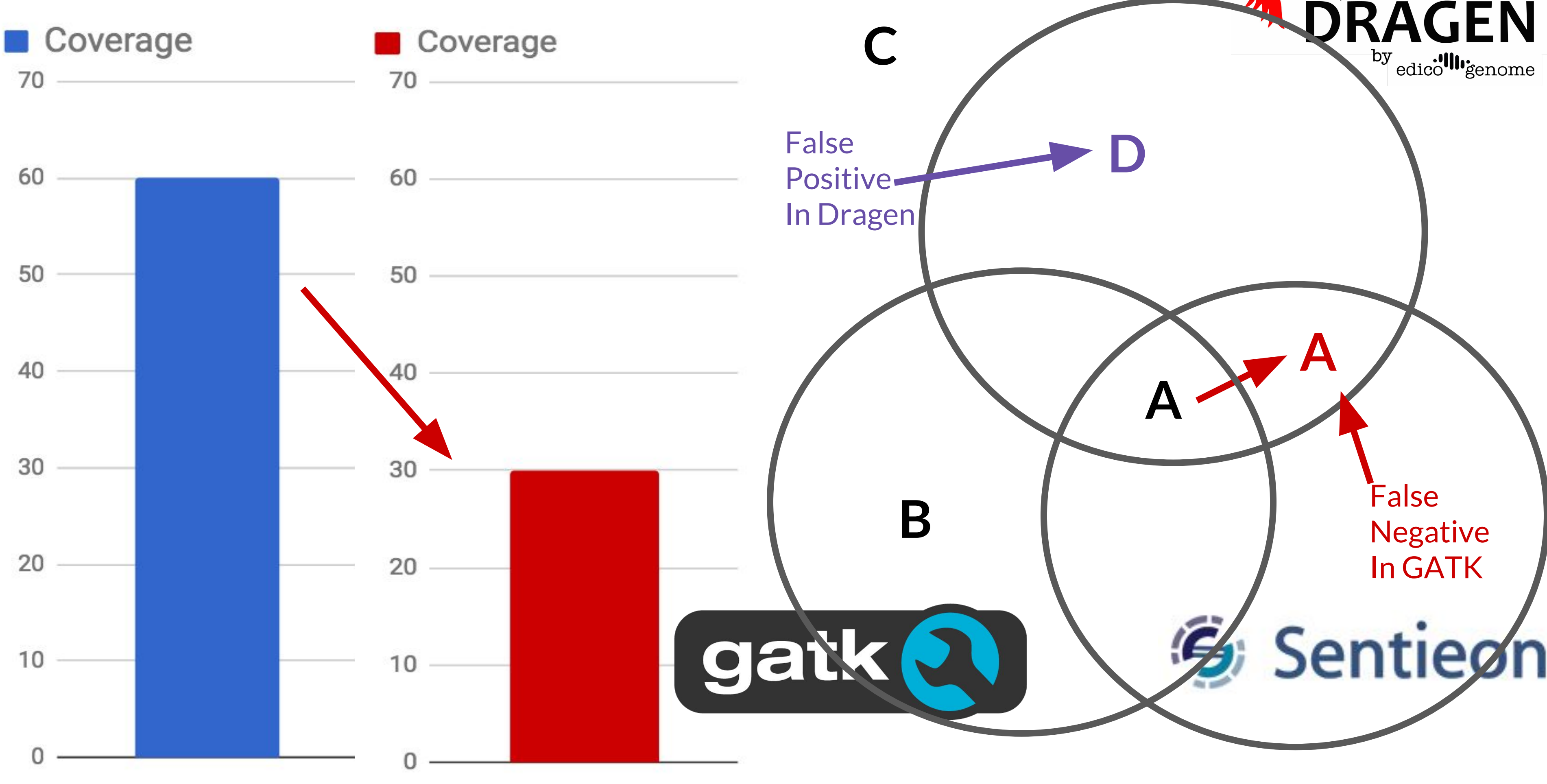


True Negative

True Positive

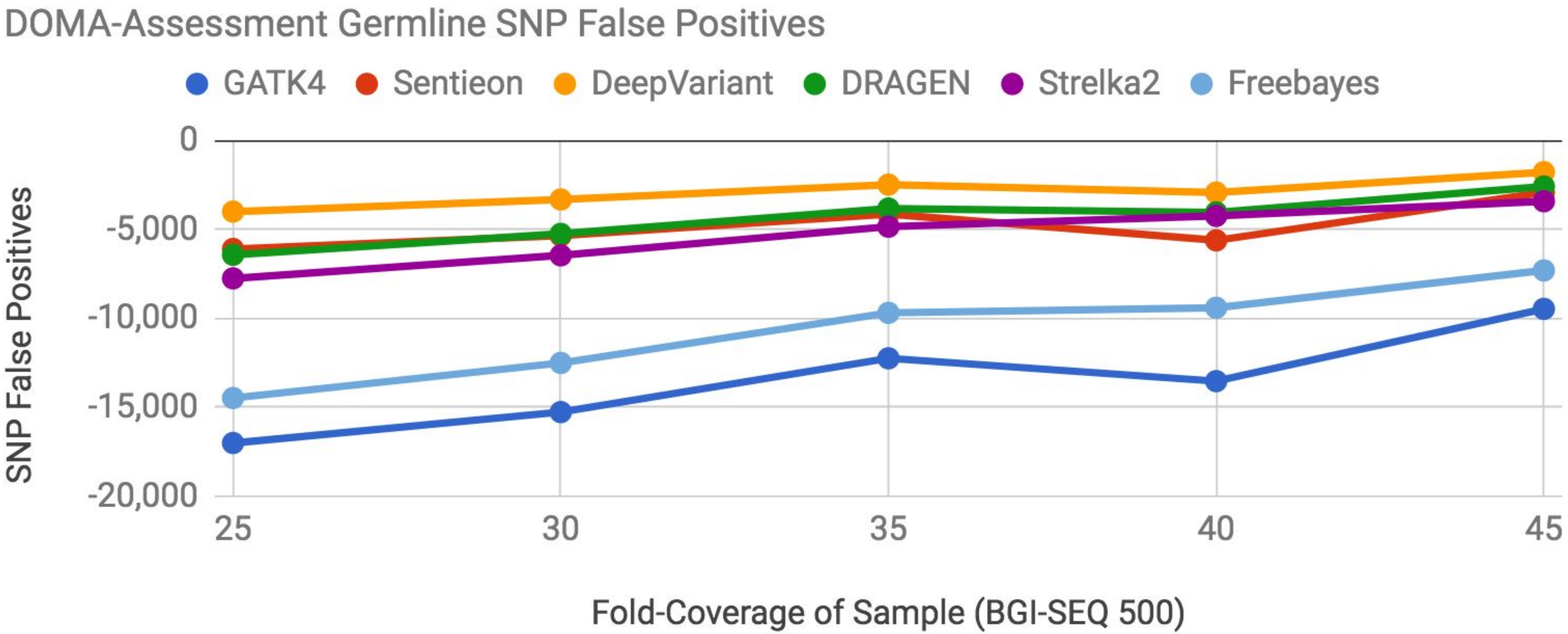
Disputed - Mask Site from Analysis

New Benchmark Method - Drop Out of Mutual Agreement



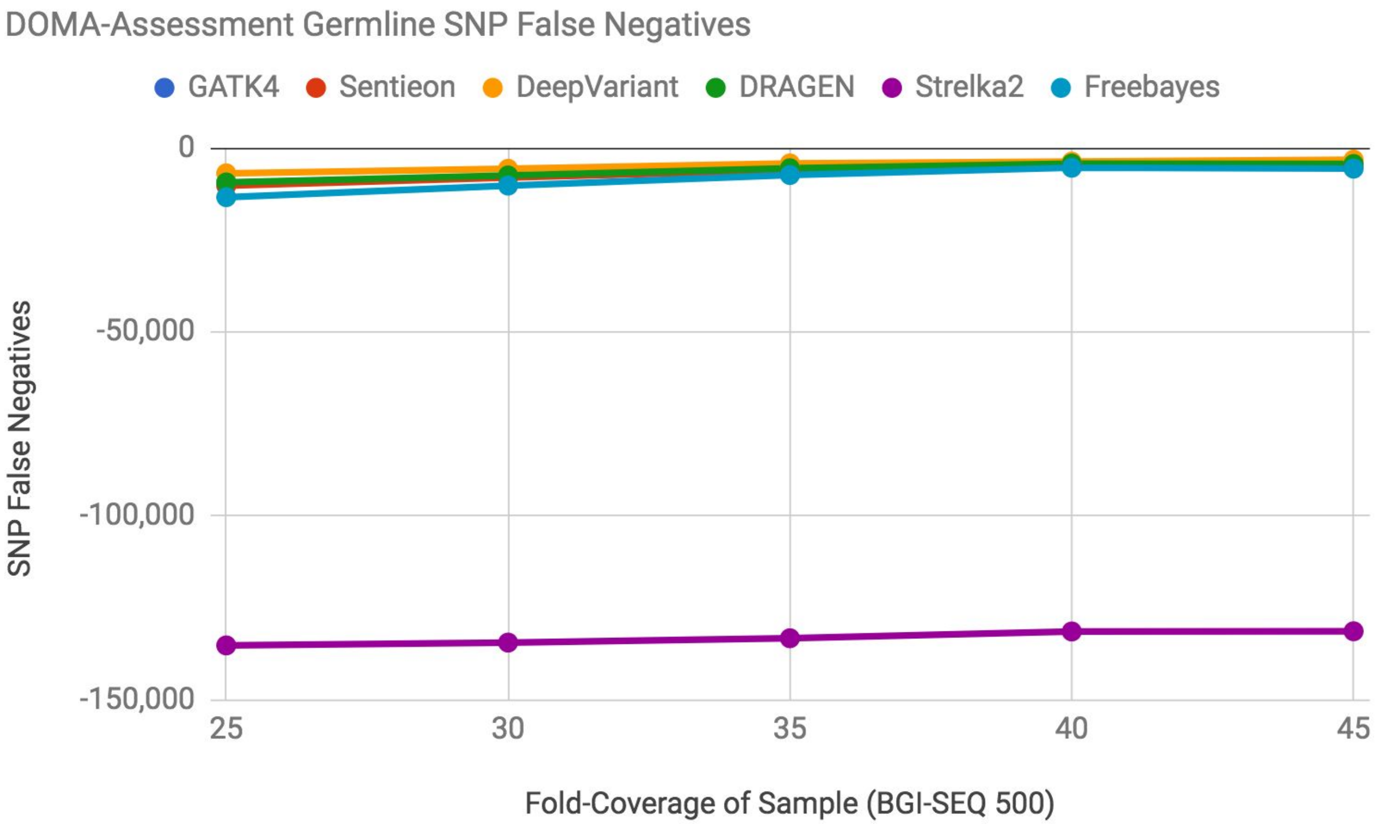
False Positive In Dragen

False Negative In GATK



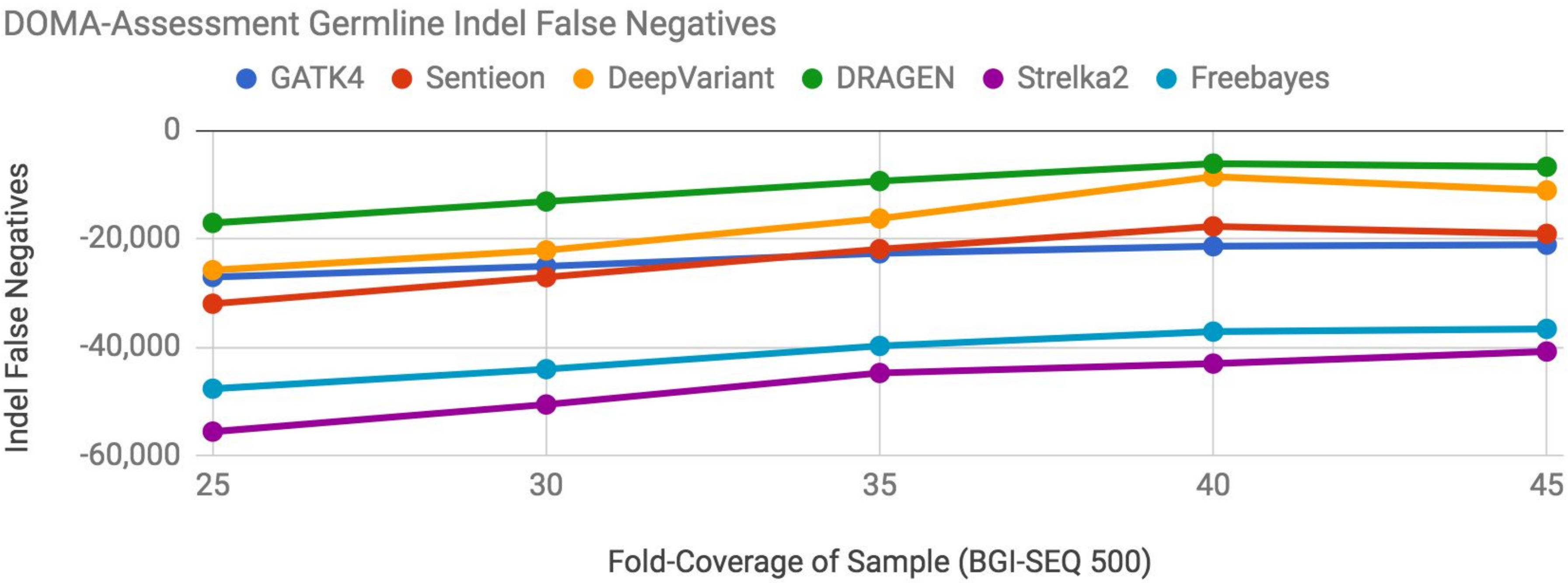
Example Conclusion : GATK4 and Sentieon are "the same" except Sentieon doesn't downsample and is reproducible

This amount of separation is very weird in Illumina benchmarks. Looks like coverage is more variable in BGI-SEQ



We may have observed a negative interaction with Strelka2 germline heuristics and BGI-SEQ data

In retrospect, It may be better to run BGI-SEQ with --exome parameters even for WGS.



For Indels, it seems all methods have issue with Recall. This may be the dominant error mode. (y-axis is 3x SNP)

Train DeepVariant for BGI-Seq data

- DeepVariant is trained on data from **Illumina** sequenceers
- How does it perform on **BGI-Seq** data?
- Can we make it better?

Train DeepVariant for BGI-Seq data

- DeepVariant is trained on data from Illumina sequenceers
- How does it perform on BGI-Seq data?
- Can we make it better? → YES!!
- Indel F1

  94.28% → 98.10%
- SNP F1:

  99.83% → 99.89%

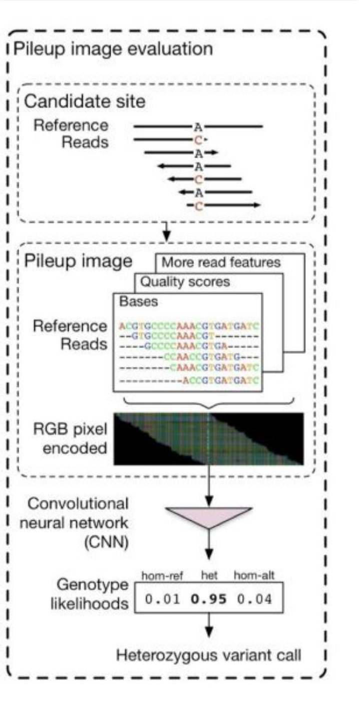
Train DeepVariant for BGI-Seq data

- DeepVariant is trained on data from Illumina sequenceers
- How does it perform on BGI-Seq data?
- Can we make it better? → YES!!
- Indel F1:
  94.28% → 98.10%
- SNP F1:
  99.83% → 99.89%
- Step-by-step (you can do it too!) http://bit.ly/train-deepvariant
- Please also check out: https://github.com/google/**nucleus**
  (the library that supports the genomics file formats for DeepVariant!)

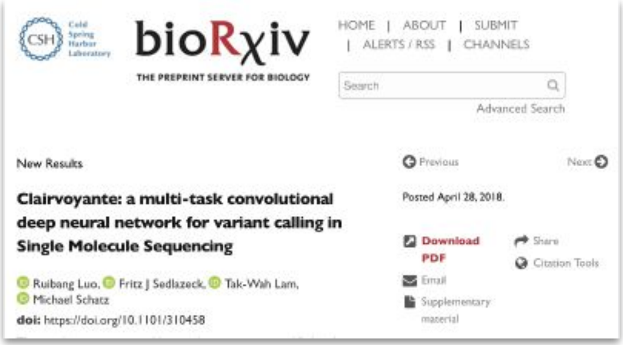
Germline Variant Calling with Various Deep Network Architecture

Different technologies could have different error profiles. Deep learning can help to adapt variant calling for different technologies. We have re-trained DeepVariant and Clairvoyante for BGI-Seq data.

DeepVariant



Clairvoyante


CSH Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY
HOME | ABOUT | SUBMIT
| ALERTS / RSS | CHANNELS
Search
Advanced Search
Previous
Next

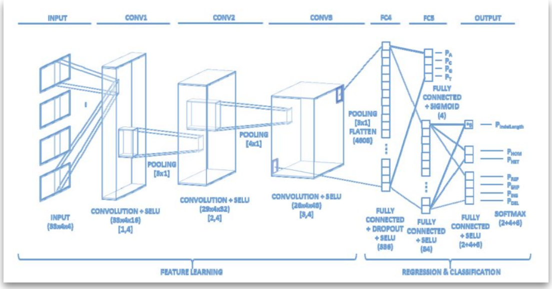
New Results

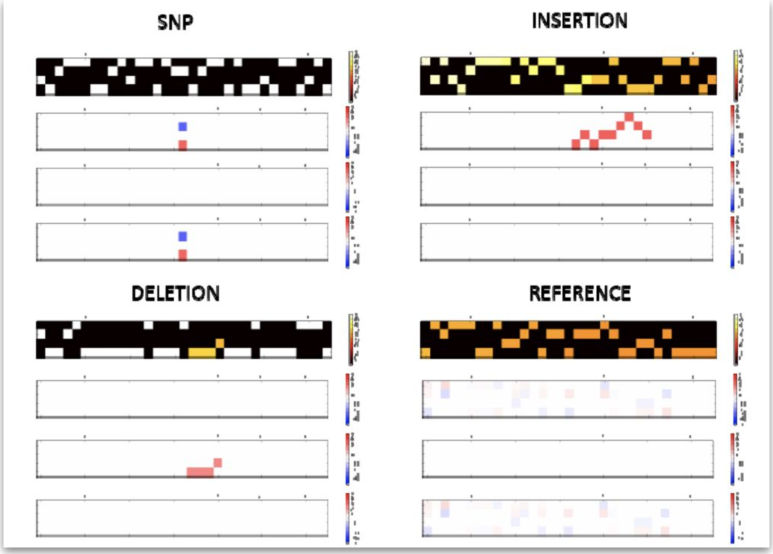
Posted April 28, 2018.

Clairvoyante: a multi-task convolutional deep neural network for variant calling in Single Molecule Sequencing

Ruibang Luo, Fritz J Sedlazeck, Tak-Wah Lam, Michael Schatz

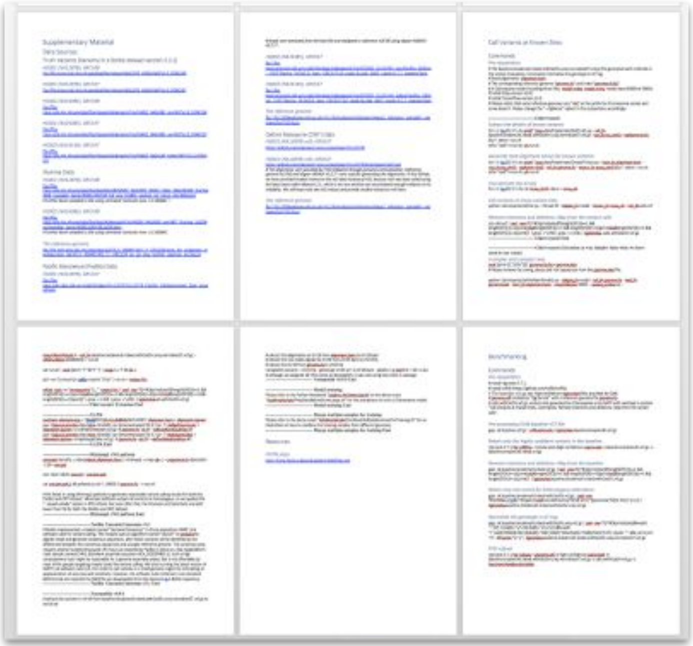
doi: https://doi.org/10.1101/310458


INPUT
CONV1
CONV2
CONV3
FC4
FC5
OUTPUT
INPUT (33x4x4)
CONVOLUTION + SELU (33x4x16) [1,4]
POOLING [3x1]
CONVOLUTION + SELU (29x4x32) [2,4]
POOLING [4x1]
CONVOLUTION + SELU (26x4x48) [3,4]
POOLING [3x1] FLATTEN (4608)
FULLY CONNECTED + DROPOUT + SELU (336)
FULLY CONNECTED + SELU (84)
FULLY CONNECTED + SIGMOID (4)
P_A
P_C
P_G
P_T
$P_{IndelLength}$
P_{HOM}
P_{HET}
P_{REF}
P_{SNP}
P_{INS}
P_{DEL}
FULLY CONNECTED + SELU (2+4+6)
SOFTMAX (2+4+6)
FEATURE LEARNING
REGRESSION & CLASSIFICATION

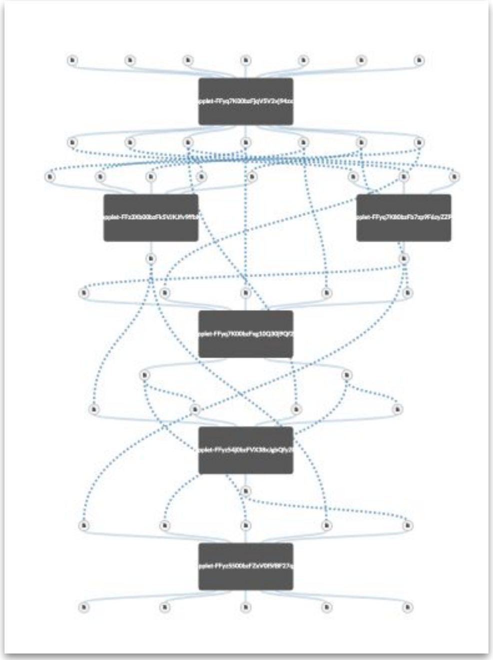


Workflow to create training data for Clairvoyante

“Supplementary Material”

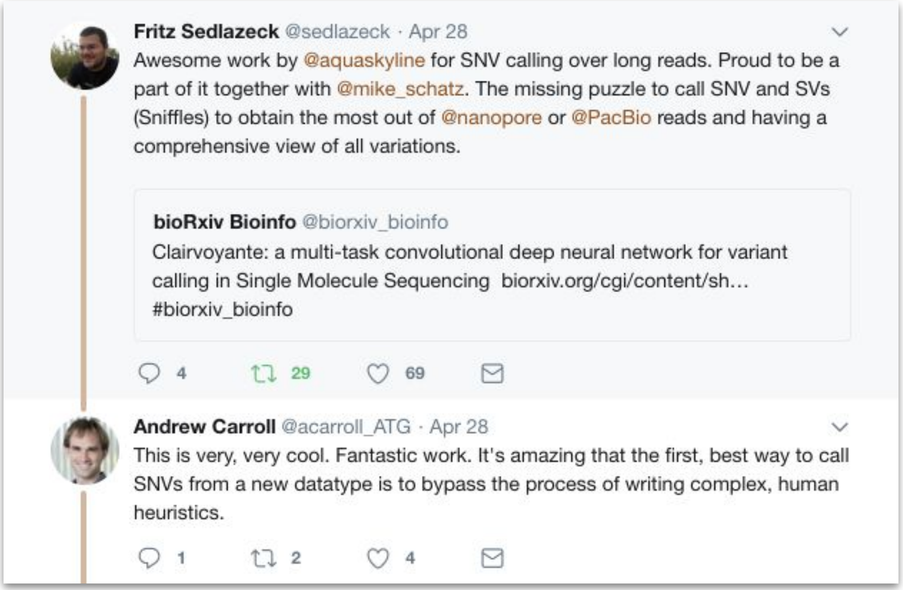


“Crystalize” the workflow/process documented in a “Supplementary Material” to a fully reproducible and re-usable deep-learning-in-cloud workflow.


pplet-FFyq7K00bzFjqV5V2xj94xx
pplet-FFz3Xb00bzFk5VJKJfv9ffb
plet-FFyq7K80bzFb7xp9F6zyZZP
pplet-FFyq7K00bzFxg10Q30j9QY
pplet-FFyz54j0bzFVX38xJgxQfy
pplet-FFyz5500bzFZxV0f5f8F27q

Results: Two new set deep learning based germline variation calls at 45x and 35x generated

Quick evaluations



Fritz Sedlazeck @sedlazeck · Apr 28

Awesome work by @aquaskyline for SNV calling over long reads. Proud to be a part of it together with @mike_schatz. The missing puzzle to call SNV and SVs (Sniffles) to obtain the most out of @nanopore or @PacBio reads and having a comprehensive view of all variations.

> **bioRxiv Bioinfo** @biorxiv_bioinfo
> Clairvoyante: a multi-task convolutional deep neural network for variant calling in Single Molecule Sequencing biorxiv.org/cgi/content/sh…
> #biorxiv_bioinfo

4 29 69

Andrew Carroll @acarroll_ATG · Apr 28

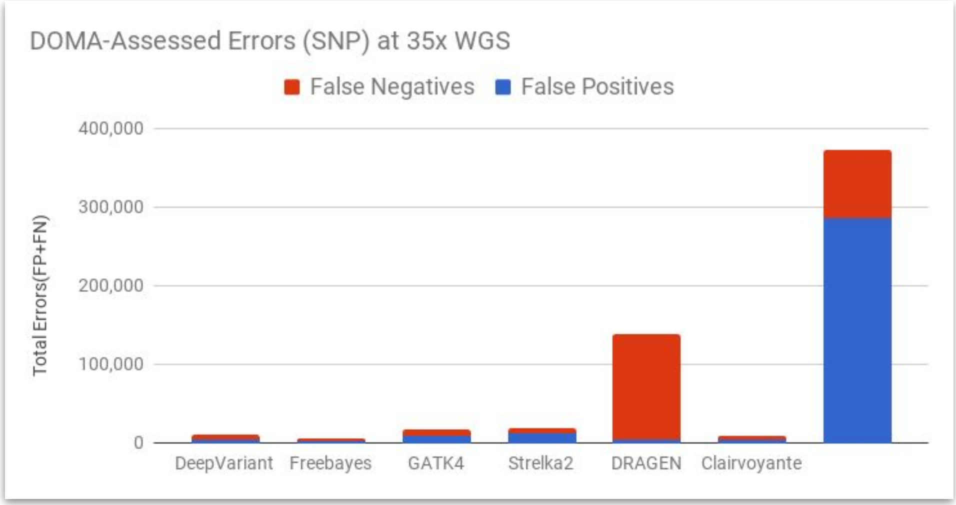
This is very, very cool. Fantastic work. It's amazing that the first, best way to call SNVs from a new datatype is to bypass the process of writing complex, human heuristics.

1 2 4

Open source collaboration:

preprint -> reproduce results on a different technology with reproducible workflow in short time.

Making improve the deep learning model easier.



https://github.com/SVAI/RecausalNucleotideNetworks/tree/master/clairvoyante_workflow

Conclusion

- When using a new technology, you should always carefully understand your data and how methods perform in it.

- These differences represent gaps that current methods can exploit to improve

- The promise of deep learning based methods allows rapid improvement (over 1 weekend)

Andrew Carroll: https://www.linkedin.com/in/andrewcarrolldna/
Pi-Chuan Chang: https://www.linkedin.com/in/pichuan/
Jason Chin: https://www.linkedin.com/in/chenshanchin/ Twitter: @infoecho