Semester 6

Evaluation of Variant Callers for Structural Variant Detection and Examination of Copy Number Variations using Long Read DNA Sequencing Data

Honours Project Report

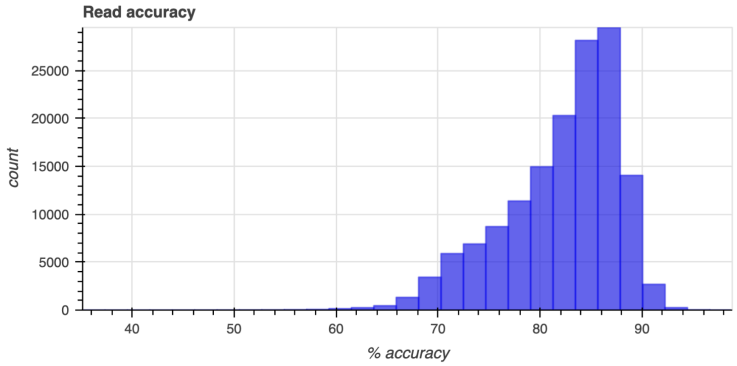
Gaurav Bhole - 2021113015

Objective:

Evaluating Variant Callers for Structural Variant Detection and examining Copy Number Variations using Long Read DNA Sequencing Data

Pre-processing and Visualization of Real Data:

Different tools were used and evaluated for preprocessing and visualizing the real data in its raw form. Pomoxis and Nanoplot are the 2 most useful tools for the task. The “Pomoxis" toolkit is a very recent development and a useful toolkit for performing data analysis on Nanopore Data.
Link: https://nanoporetech.github.io/pomoxis/



The plot for read length vs number of reads is attached above and the plot for read accuracy vs number of reads is attached above.

After this, a FASTQ file was obtained from the above BAM file using samtools. This file was mapped to the respective reference genome, and a SAM file was obtained as a result. Reads having mapQ score <30 were trimmed and the rest of the reads are converted to BAM file. This file was further sorted and indexed. This file is now ready to be analyzed using different variant callers.

Methods and Specifications of the Variant Callers evaluated:

CuteSV, SVIM, SVsearcher, Sniffles2, Nanovar, SVision, SVcnn and Dysgu were the variant callers that were considered for evaluation.

CuteSV, SVIM, SVsearcher and Sniffles2 use the familiar statistical approach, which is made up of the phases - Collect - Cluster - Combine. The first phase is made up of identification of the candidate regions which contain the potential variants. This includes signatures of the variants in and between the read alignments. This is followed by adaptive or hierarchial clustering of the variants on some statistical basis. The genomic location and the span of the variant play a major role in the clustering process. The final phase includes combining and classifying the variant signature clusters in the major variant categories. This phase also includes filtering the false variants. Finally after the variants are detected, some of these toolkits perform genotyping as well.

SVision and SVcnn are the latest tools in the realm of variant calling. They use deep learning based methods incorporating different computer vision techniques. While SVision is made specifically for detection of complex structural variants, SVcnn works well with common structural variations.

Nanovar characterizes the variants based on read depth-calculation. This is followed by inferencing from a simulation-trained model using Neural Networks. Dysgu identifies SVs from alignment cigar information as well as discordant and split-read mappings. It employs a fast consensus sequence algorithm, inspired by the positional De Brujin graph, followed by remapping of anomalous sequences to discover additional small SVs. A classifier is then employed to generate a useful quality score which can be used to prioritize variants.

Real Data:

Below are the results of variant calling on chromosome 21 of NA12878 sample using different caller. All the Structural Variants were obtained by running the respective variant callers on the sorted BAM file of chromosome 21 of the NA12878 sample.

CuteSV:

The output file (vcf) can be seen in the image attached below. A total of 1120 structural variants - 371 insertions, 736 deletions, 10 duplications and 3 inversions were found in this file using CuteSV.

Sniffles2:

A total of 878 structural variants - 372 insertions, 504 deletions, 1 duplication of and 1 inversion were found in this file using Sniffles2.

Dysgu:

A total of 1291 structural variants - 13 insertions, 33 deletions, 80 duplications, 83 inversions and 1082 were found in this file using Dysgu.

SVsearcher:

A total of 375 insertions were found in this file.

Simulated Data:

Homozygous Copy Number Variations:

To mimic the presence of homozygous deletion events, a randomly generated contiguous segment of length 1 Mbp was taken, and the length of deletions was varied from 50 to 30K, simulating 5 deletions in each the following ranges: 50 - 100, 300-500, 500-1000, 1.5K-2K, 4K-5K, 8K-10K and 25K-30K. This modified sequence was added to itself to simulate homozygous deletions, where the first half of the sequence represents one of the pairs, and the 2nd half represents the other half. A similar process was followed for modeling duplications. A randomly generated contiguous segment of length 1 Mbp was taken and the length of duplications was varied from 50 to 30K, adding duplications having a copy number of 3, 4 and 5 in each the following ranges: 50 - 100, 300-500, 500-1000, 1.5K-2K, 4K-5K, 8K-10K and 25K-30K in the modified sequence. This sequence was added to itself to simulate a homozygous duplication.

Heterozygous Copy Number Variations:

To mimic the presence of heterozygous deletion events, a randomly generated contiguous segment of length 1 Mbp was taken, and the length of deletions was varied from 50 to 30K, simulating deletions in each the following ranges: 50 - 100, 300-500, 500-1000, 1.5K-2K, 4K-5K, 8K-10K and 25K-30K. This modified sequence was added to the initially generated untampered DNA sequence. This is done to simulate heterozygous deletions, where the first half of the sequence represents one of the pairs, and the second half represents the other half. A similar process was followed for modeling duplications. A randomly generated contiguous segment of length 1 Mbp was taken, and was appended to itself. Following that, the length of duplications was varied from 50 to 30K, adding duplications having a copy number of 3, 4 and 5 in each the following ranges: 50 - 100, 300-500, 500-1000, 1.5K-2K, 4K-5K, 8K-10K and 25K-30K in the modified sequence.

This modified sequence is now the simulated dataset. Long reads are generated from the simulated data using pbsim3 simulator at 6 different sequencing depths: 5x, 10x, 20x, 30x, 40x. The simulated reads vary in length from

- 15kbp to 25kbp (Mean: 20 kbp) having an error rate of 15% for PacBio CLR Data
- 10kbp to 20kbp (Mean: 15 kbp) having an error rate of 10% for PacBio CCS Data
- 20kbp to 40kbp (Mean: 30 kbp) having an error rate of 20% for ONT Data

As per the data used from the respective platforms.

For each sequencing depth, 35 deletions are carried out over which predictions are averaged. Similarly for Duplications, for each sequencing depth, either of 21, 28 or 35 duplications are carried out based on the copy number being 3,4 and 5, and the predictions are averaged over this.

The reads are mapped to reference using minimap2 aligner to obtain alignment files in SAM format, which are sorted and indexed using samtools. These alignment files along with the reference data were used directly to call for the deletions using SVIM, CuteSV and Sniffles2. Pbsim3 has the ability of simulating reads for data from both PacBio and Nanopore sequencer. In the experiments described above, the respective modes according to the sequencer were used for the simulation. In all the experiments described above, the respective modes according to the sequencer were used for the simulation. These alignment files along with the reference data were used directly to call for the deletions using different variant callers.

Results of the above experiments:

Sniffles2 and SVIM gave an accuracy of greater than or equal to 90% at coverages greater than or equal to 5x for PacBio CCS data for detection of both Homozygous and Heterozygous deletions. CuteSV gave an accuracy of greater than or equal to 90% at coverages greater than or equal to 20x for the same. I am still working on organizing the results obtained from the duplication experiment. The accuracy for the duplications is a complicated measure since a lot of

false positive events are detected while detection of duplications and quite a few of the duplications are detected as insertions by the variant callers.

Performance of different variant callers on simulated dataset:

Homozygous Deletions

CuteSV_summary

Bin	5x	10x	20x	30x	40x
50-100	30	35	49	49	48
300-500	29	38	51	51	48
501-1000	23	36	49	49	47
1500-2000	25	36	50	50	49
4000-5000	25	39	50	50	49
8000-10000	26	37	50	50	49
25000-30000	29	38	50	50	50

Sniffles_summary

Bin	5x	10x	20x	30x	40x
50-100	50	46	50	50	50
300-500	51	50	51	51	51
501-1000	49	48	49	49	49
1500-2000	49	46	50	50	50
4000-5000	50	48	50	50	50
8000-10000	50	47	50	50	50
25000-30000	0	0	0	0	0

SVIM_summary

Bin	5x	10x	20x	30x	40x
50-100	50	49	49	50	50
300-500	51	51	51	51	51
501-1000	49	48	49	49	49
1500-2000	50	48	51	50	50
4000-5000	50	49	50	50	50
8000-10000	50	48	50	51	50
25000-30000	50	49	50	50	50

The results are out of 50 in each (row, column) entry.

Heterozygous Deletions

CuteSV_summary

Bin	5x	10x	20x	30x	40x
50-100	1	24	40	39	44
300-500	1	27	43	41	44
501-1000	0	22	44	39	45
1500-2000	0	25	45	41	45
4000-5000	1	19	44	41	45
8000-10000	0	28	44	42	45
25000-30000	0	22	44	39	45

Sniffles_summary

Bin	5x	10x	20x	30x	40x
50-100	50	50	45	45	45
300-500	50	51	45	46	45
501-1000	48	49	46	44	45
1500-2000	48	49	45	45	45
4000-5000	46	50	45	45	45
8000-10000	48	49	45	45	46
25000-30000	47	50	45	45	45

SVIM_summary

Bin	5x	10x	20x	30x	40x
50-100	50	49	45	46	45
300-500	50	51	45	46	47
501-1000	49	49	46	44	47
1500-2000	49	50	45	46	46
4000-5000	50	50	46	45	45
8000-10000	50	50	46	47	46
25000-30000	50	50	46	46	47

The results are out of 50 in each (row, column) entry.

Homozygous Duplications with Copy Number 3

CuteSV_summary

Bin	5x	10x	20x	30x	40x
50-100	19	24	30	26	27
300-500	7	24	31	29	34
501-1000	7	23	31	27	31
1500-2000	7	19	40	41	52
4000-5000	0	18	42	31	46
8000-10000	11	23	30	26	27
25000-30000	13	23	30	26	27

Sniffles_summary

Bin	5x	10x	20x	30x	40x
50-100	30	27	30	30	27
300-500	40	32	31	30	28
501-1000	36	42	37	41	35
1500-2000	41	60	82	67	60
4000-5000	47	42	72	70	75
8000-10000	29	27	30	31	27
25000-30000	29	27	30	29	28

SVIM_summary

Bin	5x	10x	20x	30x	40x
50-100	30	28	30	31	27
300-500	35	35	42	40	43
501-1000	59	69	79	76	76
1500-2000	112	103	142	128	134
4000-5000	87	105	183	163	186
8000-10000	30	29	31	35	29
25000-30000	30	29	30	29	30

The results are out of 60 in each (row, column) entry.

Heterozygous Duplications with Copy Number 3

CuteSV_summary

Bin	5x	10x	20x	30x	40x
50-100	0	15	27	22	21
300-500	0	14	24	24	21
501-1000	1	12	26	21	21
1500-2000	0	3	25	29	26
4000-5000	0	3	15	23	28
8000-10000	0	8	26	23	21
25000-30000	0	14	27	25	21

Sniffles_summary

Bin	5x	10x	20x	30x	40x
50-100	27	28	29	29	24
300-500	26	29	31	30	24
501-1000	24	30	37	30	26
1500-2000	13	25	29	34	37
4000-5000	31	33	41	41	40
8000-10000	26	27	28	30	22
25000-30000	28	30	27	30	23

SVIM_summary

Bin	5x	10x	20x	30x	40x
50-100	30	30	29	30	28
300-500	34	36	36	37	33
501-1000	56	69	69	72	65
1500-2000	74	108	119	128	111
4000-5000	58	74	109	118	127
8000-10000	30	30	30	31	26
25000-30000	30	30	29	30	26

The results are out of 30 in each (row, column) entry.

Homozygous Duplications with Copy Number 4

CuteSV_summary

Bin	5x	10x	20x	30x	40x
50-100	22	31	39	36	38
300-500	18	31	36	38	46
501-1000	15	28	35	43	43
1500-2000	7	27	50	68	65
4000-5000	5	24	41	55	64
8000-10000	10	34	37	36	36
25000-30000	17	33	38	36	37

Sniffles_summary

Bin	5x	10x	20x	30x	40x
50-100	40	40	40	36	40
300-500	53	46	44	36	42
501-1000	63	62	58	44	52
1500-2000	73	85	101	90	88
4000-5000	70	82	75	91	95
8000-10000	40	40	40	36	40
25000-30000	40	40	40	36	40

SVIM_summary

Bin	5x	10x	20x	30x	40x
50-100	40	40	40	36	41
300-500	44	48	56	61	63
501-1000	87	99	109	106	106
1500-2000	148	166	176	168	181
4000-5000	130	161	215	258	263
8000-10000	39	40	40	36	40
25000-30000	39	39	39	36	40

The results are out of 80 in each (row, column) entry.

Heterozygous Duplications with Copy Number 4

CuteSV_summary

Bin	5x	10x	20x	30x	40x
50-100	2	19	35	38	40
300-500	0	15	35	38	42
501-1000	0	17	35	37	39
1500-2000	0	7	29	40	52
4000-5000	0	2	28	33	40
8000-10000	2	15	34	39	38
25000-30000	0	21	35	37	37

Sniffles_summary

Bin	5x	10x	20x	30x	40x
50-100	37	39	38	40	40
300-500	39	50	42	45	44
501-1000	24	49	56	46	48
1500-2000	20	38	42	59	68
4000-5000	36	49	58	65	64
8000-10000	35	39	40	40	40
25000-30000	37	39	40	40	40

SVIM_summary

Bin	5x	10x	20x	30x	40x
50-100	40	40	40	40	40
300-500	46	48	49	49	58
501-1000	73	87	105	105	113
1500-2000	108	144	161	175	177
4000-5000	89	117	174	186	216
8000-10000	36	40	40	40	40
25000-30000	37	40	40	40	40

The results are out of 40 in each (row, column) entry.

Homozygous Duplications with Copy Number 5

CuteSV_summary

Bin	5x	10x	20x	30x	40x
50-100	30	44	46	46	45
300-500	25	49	51	60	56
501-1000	14	48	49	50	50
1500-2000	6	42	67	91	86
4000-5000	3	36	55	76	85
8000-10000	22	49	50	50	45
25000-30000	24	49	50	50	45

Sniffles_summary

Bin	5x	10x	20x	30x	40x
50-100	50	50	50	50	48
300-500	57	54	50	50	51
501-1000	83	76	67	55	56
1500-2000	85	112	134	123	108
4000-5000	76	102	111	131	132
8000-10000	49	50	50	50	50
25000-30000	45	46	46	46	46

SVIM_summary

Bin	5x	10x	20x	30x	40x
50-100	47	46	47	46	48
300-500	57	59	67	72	78
501-1000	113	125	133	133	136
1500-2000	196	213	234	232	227
4000-5000	153	216	284	337	346
8000-10000	50	50	51	52	51
25000-30000	50	50	50	50	50

The results are out of 100 in each (row, column) entry.

Heterozygous Duplications with Copy Number 5

CuteSV_summary

Bin	5x	10x	20x	30x	40x
50-100	6	29	46	41	33
300-500	3	29	49	44	39
501-1000	1	17	50	45	40
1500-2000	0	14	45	47	47
4000-5000	0	3	33	42	37
8000-10000	0	19	48	44	36
25000-30000	3	20	50	44	38

Sniffles_summary

Bin	5x	10x	20x	30x	40x
50-100	42	50	50	49	47
300-500	40	50	50	47	46
501-1000	47	63	60	57	51
1500-2000	28	47	70	66	58
4000-5000	51	70	78	72	75
8000-10000	47	49	50	48	48
25000-30000	43	46	46	44	42

SVIM_summary

Bin	5x	10x	20x	30x	40x
50-100	43	46	47	46	43
300-500	53	62	58	60	57
501-1000	88	101	123	125	114
1500-2000	136	188	216	206	200
4000-5000	120	141	219	234	228
8000-10000	48	50	51	48	49
25000-30000	50	50	50	49	47

The results are out of 50 in each (row, column) entry.

Goals for the Summer:

The goal for the summer is to develop a formal pipeline for Long Read DNA sequencing data from different sequencing platforms and to expand the number of variant callers used to the state of the art ones across different Structural Variant detection methods. This will be followed by attempts to model the Structural Variations using images of the candidate variant regions, which would be used for developing a novel tool for variant detection.