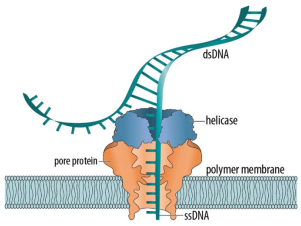
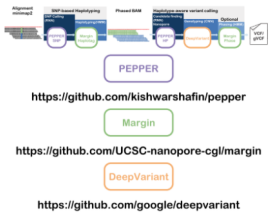




ONT variant Analysis using PEPPER - Margin - DeepVariant and ONT pipeline-structural-variation

Stéphane Plaisance [VIB - Nucleomics Core, nucleomics@vib.be]

July 13th, 2021 - version 1.0

Contents

last edits: Tue Jul 13, 2021

Aim

Call Yeast small variants from ONT reads using the **PEPPER - Margin - DeepVariant** workflow as detailed in a recent report (Shafin et al. 2021). The method and code were obtained from the paper **github page**

Also call wider Structural variants using the ONT dedicated **ipeline-structural-variation** workflow.

The yeast reference genome version **R64-1-1** was **obtained from ensembl**

- fasta file: Saccharomyces_cerevisiae.R64-1-1.dna.toplevel.fa.gz

The reads were obtained from a *Nucleomics Core* Saccharomyces cerevisiae Adaptive sequencing GridION flow-cell loaded twice, once for even chromosomes enrichment and once for odd chromosomes enrichment.

The necessary software was obtained as a Docker image as described in github.com/kishwarshafin/pepper

Small Variant analysis with Pepper-Margin-DeepVariant

This first worfklow finds single nucleotide to ~50bps variants (short variants).

Mapping reads to the Yeast reference genome using Minimap2

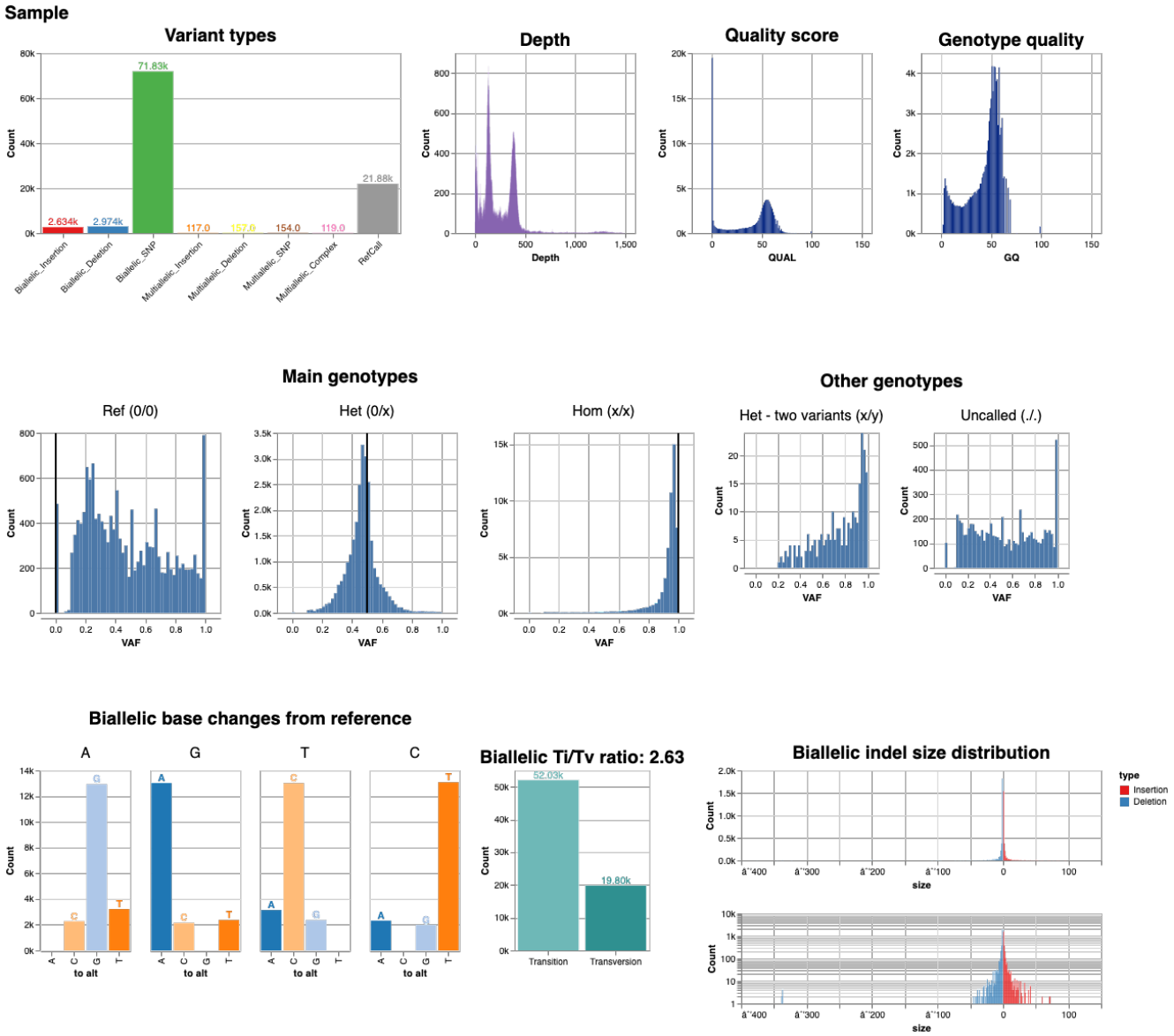
The merged fast5 data from 2 GrdION runs and four barcodes where basecalled and demultiplexed using the configuration and Guppy_basecaller version 5.0.11+2b6dbff. The obtained fastq reads were aligned to the reference using **Minimap2** (Li 2018) and standard parameters. The SAM alignments were sorted and the header corrected using samtools to add RG metadata.

Running the Docker 'pepper_deepvariant' workflow

The **pepper_deepvariant** docker image was obtained and used as described in *github.com/kishwarshafin/pepper*. The tutorial code was edited to match the Saccharomyces genome files and local Minimap2 mappings.

The pipeline runs and generates several output files among which a phased VCF file that can be annotated using SNPSift to annotate gene regions and variant effects.

The DeepVariant workflow also outputs a summary plot (*Saccer.visual_report.html*) with a number of important metrics, reproduced next.



REM: The bi-modal plot for the depth is due to a difference in sequencing depth between the two Flow-cell runs (>2x more data for run#1 on even chromosomes).

Annotate VCF results with genome features

We used **SNPEff** (v5.0e; 2021-03-09) (Cingolani et al. 2012) to add genomic annotations to the VCF results and allow the identification of variants potentially affecting gene function.

Review SnpEff results

The summary of the SNPEff annotation can be found as *snpEff_summary.html* of which some key content is reproduced below.

Summary

Genome	R64-1-1.99
Date	2021-07-13 09:02
SnpEff version	SnpEff 5.0e (build 2021-03-09 06:01), by Pablo Cingolani
Command line arguments	SnpEff R64-1-1.99 Saccer.phased.vcf.gz
Warnings	53,531
Errors	0
Number of lines (input file)	99,863
Number of variants (before filter)	100,479
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	100,479
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	616
Number of effects	673,554
Genome total length	12,157,105
Genome effective length	12,157,105
Variant rate	1 variant every 120 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
I	230,218	4,464	51
II	813,184	5,850	139
III	316,620	3,048	103
IV	1,531,933	11,583	132
IX	439,888	5,337	82
Mito	85,779	3,984	21
V	576,874	5,551	103
VI	270,161	4,215	64
VII	1,090,940	7,806	139
VIII	562,643	4,493	125
X	745,751	6,581	113
XI	666,816	4,997	133
XII	1,078,177	7,463	144
XIII	924,431	5,460	169
XIV	784,333	6,593	118
XV	1,091,291	7,710	141
XVI	948,066	5,344	177
Total	12,157,105	100,479	120

Number of variants by type

Type	Total
SNP	89,022
MNP	0
INS	4,561
DEL	6,896
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	100,479

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	1,780	0.264%
LOW	27,869	4.138%
MODERATE	18,883	2.803%
MODIFIER	625,022	92.795%

Number of effects by functional class

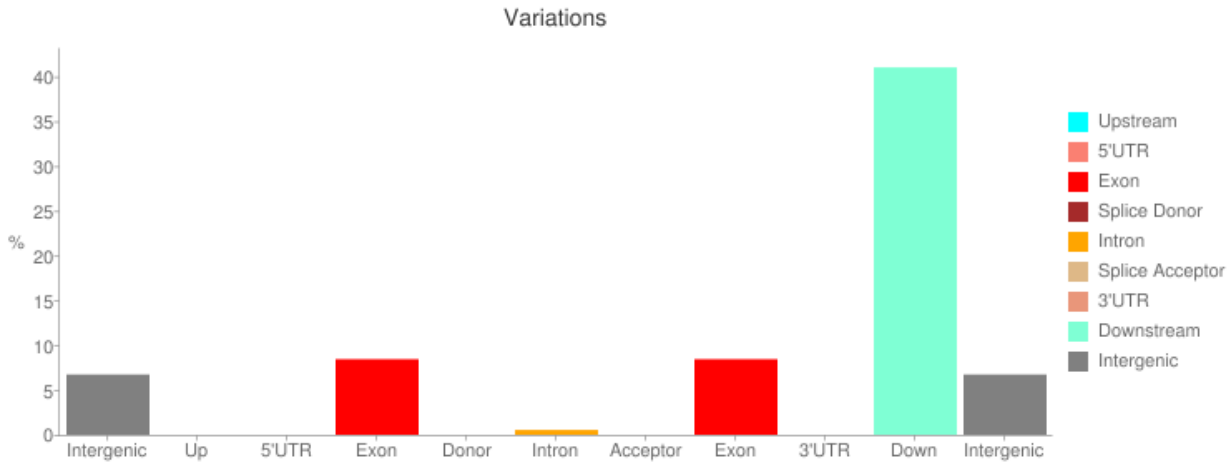
Type (alphabetical order)	Count	Percent
MISSENSE	18,339	39.536%
NONSENSE	223	0.481%
SILENT	27,823	59.983%

Number of effects by type and region

Type (alphabetical order)	Count	Percent
bidirectional_gene_fusion	1	0%
conservative_inframe_deletion	93	0.014%
conservative_inframe_insertion	123	0.018%
disruptive_inframe_deletion	217	0.032%
disruptive_inframe_insertion	200	0.03%
downstream_gene_variant	276,255	40.997%
frameshift_variant	1,451	0.215%
intergenic_region	45,338	6.728%
intragenic_variant	2	0%
intron_variant	3,674	0.545%
missense_variant	18,266	2.711%
non_coding_transcript_exon_variant	8,421	1.25%
splice_acceptor_variant	18	0.003%
splice_donor_variant	6	0.001%
splice_region_variant	169	0.025%
start_lost	56	0.008%
stop_gained	272	0.04%
stop_lost	60	0.009%
stop_retained_variant	53	0.008%
synonymous_variant	27,77	4.121%
upstream_gene_variant	291,398	43.244%

Variants by genomic location

Type (alphabetical order)	Count	Percent
DOWNSTREAM	276,26	41.015%
EXON	56,822	8.436%
GENE	1	0%
INTERGENIC	45,338	6.731%
INTRON	3,621	0.538%
SPLICE_SITE_ACCEPTOR	14	0.002%
SPLICE_SITE_DONOR	2	0%
SPLICE_SITE_REGION	101	0.015%
TRANSCRIPT	2	0%
UPSTREAM	291,4	43.263%



Filter annotated VCF results

SnpSift can be used to filter the variants with most predicted impact and review them manually to confirm their presence.

As illustration, we filtered Hom SNP calls introducing Stop codons in coding regions but many other filtering can be done in order to focus on other variant types or location.

The table of Homozygous 'stop_gained' variants obtained with *SnpSift extractFields* is reproduced below

CHROM	POS	REF	ALT	ANN[0].AA	ANN[0].GENE	GEN[*].GQ	GEN[*].DP	GEN[*].AD	GEN[*].VAF	GEN[*].PL
I	25442	G	A	p.Gln843*	FLO9	7	10	1,8	0.8	6,31,0
I	27506	T	A	p.Lys155*	FLO9	47	107	1,105	0.981308	50,50,0
I	36524	C	T	p.Arg6*	ECM1	58	172	1,167	0.97093	57,76,0
I	186754	G	A	p.Gln34*	YAR030C	57	167	2,163	0.976048	57,70,0
I	186841	G	A	p.Arg5*	YAR030C	67	167	1,164	0.982036	66,99,0
I	186936	G	A	p.Trp34*	PRM9	58	170	3,166	0.976471	58,64,0
I	187569	C	G	p.Ser245*	PRM9	14	160	5,15	0.9375	14,56,0
I	203865	A	T	p.Lys155*	FLO1	7	16	2,14	0.875	6,23,0
I	208511	C	T	p.Arg49*	YAR053W	45	72	1,62	0.861111	48,47,0
II	426936	G	C	p.Tyr41*	YBR090C	61	388	3,37	0.953608	62,70,0
II	465618	G	A	p.Trp18*	YBR113W	45	294	14,278	0.945578	51,45,0
II	465979	G	A	p.Trp138*	YBR113W	49	329	3,318	0.966565	49,57,0
III	264282	C	T	p.Gln103*	YCR087W	46	132	3,127	0.962121	46,59,0
IV	383945	G	C	p.Ser46*	PRM7	15	203	18,162	0.79803	14,26,0
IV	410017	G	A	p.Gln14*	YDL022C-A	58	343	3,337	0.982507	58,73,0
IV	681939	G	T	p.Tyr96*	YDR114C	52	391	1,386	0.987212	52,64,0
IV	1283911	C	A	p.Ser38*	YDR406W-A	62	397	2,389	0.979849	63,68,0
IV	1470784	G	A	p.Gln94*	EMI1	51	390	6,372	0.953846	51,58,0
V	136510	A	T	p.Lys78*	YEL010W	38	144	21,119	0.826389	44,39,0
V	355336	C	G	p.Ser66*	YER097W	49	137	3,122	0.890511	52,52,0
V	569881	G	T	p.Tyr9*	YER188C-A	3	7	2,5	0.714286	6,0,0
VI	18358	G	A	p.Trp452*	AGP3	60	279	5,272	0.97491	60,67,0
VI	106510	C	A	p.Glu152*	YFL015C	58	420	1,395	0.940476	58,69,0
VI	115031	T	A	p.Tyr14*	AUA1	54	401	7,39	0.972569	54,64,0
VI	264305	G	T	p.Tyr7*	YFR056C	10	8	0,8	1.0	9,41,0
VII	627362	A	T	p.Lys94*	YGR069W	52	133	3,128	0.962406	53,57,0
VII	848849	C	T	p.Arg44*	YGR176W	49	125	7,115	0.92	53,50,0
VII	858672	C	T	p.Trp78*	YGR182C	52	139	13,119	0.856115	53,58,0
VII	1075820	C	A	p.Tyr113*	YGR290W	41	57	0,56	0.982456	41,49,0
VIII	7140	T	A	p.Leu247*	COS8	42	224	56,16	0.714286	41,71,0
VIII	167453	G	A	p.Trp34*	YHR028W-A	60	367	1,36	0.980926	59,99,0
VIII	167487	C	T	p.Arg46*	YHR028W-A	58	338	6,325	0.961538	58,68,0
VIII	239033	G	T	p.Tyr66*	YHR071C-A	51	348	5,337	0.968391	51,61,0
IX	247021	C	A	p.Tyr36*	YIL058W	51	107	3,98	0.915888	51,60,0
IX	397547	T	A	p.Lys135*	YIR020C-B	58	131	6,119	0.908397	64,59,0
X	27163	C	T	p.Gln93*	HXT8	34	191	16,17	0.890052	58,34,0
X	85607	G	C	p.Ser49*	YJL182C	59	385	5,373	0.968831	59,66,0
X	158113	C	A	p.Tyr79*	YJL135W	56	377	1,375	0.994695	57,59,0
XI	382710	C	A	p.Cys71*	YKL030W	43	120	1,115	0.958333	42,56,0
XI	549543	G	A	p.Trp32*	TRM2	57	133	8,121	0.909774	57,64,0
XII	22491	G	A	p.Gln164*	YLL059C	52	402	2,397	0.987562	51,99,0
XII	22509	G	A	p.Gln158*	YLL059C	41	402	3,394	0.980099	40,63,0
XII	391658	G	A	p.Trp20*	YLR124W	52	324	7,311	0.959877	56,54,0
XII	683073	A	T	p.Lys113*	CMG1	54	418	16,397	0.949761	55,60,0
XII	704697	A	C	p.Leu46*	YLR280C	49	373	4,366	0.981233	49,56,0
XII	766039	C	T	p.Gln129*	YLR317W	49	337	11,321	0.952522	49,60,0
XII	780772	A	T	p.Lys520*	PEX30	59	357	18,328	0.918768	59,66,0
XII	934221	C	T	p.Trp11*	BLS1	55	410	21,387	0.943902	55,80,0
XII	1023823	C	A	p.Gly56*	YLR444C	50	396	18,366	0.924242	51,54,0
XII	1035751	C	G	p.Ser1042*	HMG2	56	437	6,419	0.95881	56,63,0
XIII	74436	C	T	p.Gln70*	YML099W-A	54	127	1,125	0.984252	53,69,0
XIII	160037	C	A	p.Glu126*	YML057C-A	59	166	1,162	0.975904	59,66,0
XIII	905893	A	T	p.Cys27*	YMR316C-A	47	155	1,148	0.954839	47,55,0
XIII	923615	A	T	p.Leu63*	YMR326C	19	51	0,49	0.960784	18,64,0
XIV	96324	G	A	p.Trp51*	YNL285W	50	365	9,351	0.961644	52,54,0
XIV	144250	G	A	p.Trp2*	YNL266W	62	362	5,354	0.977901	64,66,0
XIV	505410	C	T	p.Gln563*	AQR1	56	353	1,35	0.991501	59,59,0
XV	34772	C	T	p.Gln39*	ZPS1	51	207	4,199	0.961353	51,56,0
XV	70393	C	T	p.Trp51*	YOL134C	53	161	8,152	0.944099	53,66,0
XV	181290	G	A	p.Trp78*	YOL079W	49	143	16,126	0.881119	50,53,0
XV	181452	T	A	p.Tyr132*	YOL079W	51	142	3,135	0.950704	51,57,0
XV	207397	G	A	p.Gln957*	CRT10	63	110	3,106	0.963636	65,66,0
XV	378122	C	G	p.Tyr92*	YOR024W	58	107	0,102	0.953271	60,62,0
XV	389923	G	A	p.Gln385*	HMS1	53	123	5,114	0.926829	55,57,0
XV	631073	C	A	p.Glu227*	PUP1	56	119	0,114	0.957983	61,57,0
XV	654693	G	T	p.Glu162*	YRM1	53	133	1,127	0.954887	54,57,0
Mito	3957	T	A	p.Tyr2*	Q0010	56	1232	155,1046	0.849026	67,55,0
Mito	29207	G	T	p.Gly241*	ATP6	63	1282	3,1246	0.971919	63,68,0
Mito	29218	G	A	p.Trp244*	ATP6	44	1411	160,1212	0.858965	44,68,0

Structural Variant (SV) analysis using the ONT pipeline-structural-variation pipeline

This second workflow is dedicated to long structural variants (including: INS, DEL, SUB, INV, and chromosamal translocations).

The pipeline is wrapped as a Snakemake workflow and can be deployed and ran locally from the source hosted in the **ONT github**.

Running the ONT 'pipeline-structural-variation' workflow

The pipeline performs the following steps:

- Maps reads using **lra** (Ren and Chaisson 2020)
- Produces QC report using **NanoPlot** (Coster et al. 2018) and **Mosdepth** (Pedersen and Quinlan 2017)
- Estimates appropriate parameters for variant calling depending on read depth
- Calls variants using **cuteSV** (Jiang et al. 2020)
- Filters variants by minimum/maximum length, read support, or type (e.g. insertion, deletion, etc.)

The coverage analysis done by Mosdepth retuirned the following metrics.

chrom	length	bases	mean	min	max
I	230218	37370253	162.33	0	463
II	813184	324838628	399.47	0	958
III	316620	53108420	167.74	1	861
IV	1531933	600388088	391.92	0	1343
V	576874	87705989	152.04	0	411
VI	270161	98136318	363.25	0	497
VII	1090940	163490084	149.86	0	784
VIII	562643	228954791	406.93	0	1893
IX	439888	63012817	143.25	0	320
X	745751	285241437	382.49	0	1173
XI	666816	93278072	139.89	0	291
XII	1078177	742025213	688.22	0	27567
XIII	924431	140456819	151.94	0	960
XIV	784333	313106057	399.20	0	1636
XV	1091291	164384118	150.63	0	770
XVI	948066	392069556	413.55	0	1119
Mito	85779	1007818461	11749.01	0	14855
total	12157105	4795385121	394.45	0	27567

The values confirm the higher coverage depth for *even* chromosomes enriched in run#1 as compared to *odd* chromosomes in run#2.

Annotate VCF results with genome features

SnpEff was used here too

Review SnpEff results

Variants rate details

Chromosome	Length	Variants	Variants rate
I	230,218	7	32,888
II	813,184	24	33,882
III	316,620	19	16,664
IV	1,531,933	80	19,149
IX	439,888	34	12,937

Chromosome	Length	Variants	Variants rate
Mito	85,779	42	2,042
V	576,874	29	19,892
VI	270,161	14	19,297
VII	1,090,940	39	27,972
VIII	562,643	24	23,443
X	745,751	31	24,056
XI	666,816	21	31,753
XII	1,078,177	68	15,855
XIII	924,431	39	23,703
XIV	784,333	35	22,409
XV	1,091,291	41	26,616
XVI	948,066	41	23,123
Total	12,157,105	588	20,675

Number variants by type

Type	Total
SNP	0
MNP	0
INS	269
DEL	319
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	588

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	717	14.195%
LOW	7	0.139%
MODERATE	57	1.128%
MODIFIER	4,270	84.538%

Number of effects by type

Type (alphabetical order)	Count	Percent
bidirectional_gene_fusion	7	0.135%
chromosome_number_variation	4	0.077%
conservative_inframe_deletion	37	0.712%
conservative_inframe_insertion	23	0.442%
disruptive_inframe_deletion	17	0.327%
disruptive_inframe_insertion	25	0.481%
downstream_gene_variant	1,811	34.834%
exon_loss_variant	4	0.077%
feature_ablation	110	2.116%
frameshift_variant	193	3.712%
gene_fusion	14	0.269%
intergenic_region	449	8.636%
intron_variant	26	0.5%
non_coding_transcript_exon_variant	30	0.577%
splice_acceptor_variant	1	0.019%
splice_donor_variant	3	0.058%
splice_region_variant	32	0.616%
start_lost	24	0.462%
stop_gained	61	1.173%
stop_lost	23	0.442%
transcript_ablation	340	6.54%
upstream_gene_variant	1,965	37.796%

Type (alphabetical order)	Count	Percent

Number of effects by region

Type (alphabetical order)	Count	Percent
CHROMOSOME	4	0.079%
DOWNSTREAM	1,811	35.854%
EXON	321	6.355%
GENE	131	2.594%
INTERGENIC	449	8.889%
INTRON	21	0.416%
SPLICE_SITE_DONOR	2	0.04%
SPLICE_SITE_REGION	7	0.139%
TRANSCRIPT	340	6.731%
UPSTREAM	1,965	38.903%

Filter annotated VCF results

The table of *Homozygous* SVs is reproduced below as example.

The variant metrics are listed in the VCF header as follows:

- <ID=DR,Number=1,Type=Integer,Description="# High-quality reference reads">
- <ID=DV,Number=1,Type=Integer,Description="# High-quality variant reads">
- <ID=PL,Number=G,Type=Integer,Description="# Phred-scaled genotype likelihoods rounded to the closest integer">
- <ID=GQ,Number=1,Type=Integer,Description="# Genotype quality">

CHROM	POS	ANN[0].GENE	GEN[*].DR	GEN[*].DV	GEN[*].PL	GEN[*].GQ
II	5848	YBL109W	0	67	.,.,.	0
II	5865	YBL109W	0	178	.,.,.	0
II	9089	YBL107W-A	0	317	954,255,0	255
II	29701	YBL100W-A&YBL100W-B	0	295	954,255,0	255
II	221036	YBL005W-A&YBL005W-B	0	263	954,255,0	255
II	259574	YBR012W-A&YBR012W-B&YBR013C	0	278	.,.,.	0
III	268907	FIG2	0	115	954,255,0	255
IV	46355	HO	3	406	954,255,0	255
IV	513684	YDR034C-C&YDR034C-D	1	260	954,255,0	255
IV	645498	YDR098C-A&YDR098C-B	0	297	.,.,.	0
IV	878297	YDR210C-C&YDR210C-D	63	199	496,26,0	26
IV	1095762	YDR316W-A&YDR316W-B	0	274	.,.,.	0
IV	1206700	YDR365W-A&YDR365W-B&YDR366C	0	280	954,255,0	255
IV	1307643	HKR1	0	300	.,.,.	0
IV	1308184	HKR1	0	311	.,.,.	0
IV	1504453	FIT1	0	500	.,.,.	0
IV	1525358	YDR544C	0	53	.,.,.	0
IV	1525371	YDR544C	0	71	.,.,.	0
IX	1801	YIL177C	21	78	544,52,0	52
IX	146944	NUP159	1	136	954,255,0	255
IX	205299	YIL082W&YIL082W-A	6	89	792,185,0	185
IX	246216	YIL059C	0	105	954,255,0	255
IX	382838	YIR016W	0	107	954,255,0	255
IX	390767	YIR019C	8	30	210,21,0	20
Mito	14632	AI1	0	13419	.,.,.	0
Mito	22237	AI5_ALPHA	0	11	.,.,.	0
Mito	22297	AI5_ALPHA	0	532	.,.,.	0
Mito	22816	AI5_ALPHA	0	25	.,.,.	0
Mito	25188	AI5_BETA	23	7960	954,255,0	255
Mito	28061	ATP6	0	11112	.,.,.	0
Mito	46348	OLI1	3	31979	954,255,0	255
Mito	49628	VAR1	3	18951	954,255,0	255
Mito	85586	Q0297	0	114	.,.,.	0
V	4914	YEL076C-A	9	155	859,208,0	207
V	6384	YEL074W	0	63	.,.,.	0
V	175626	TIR1	11	113	802,179,0	178
V	492735	YER159C-A&YER160C	0	32	305,82,0	81
V	571938	YRF1-2	17	63	439,42,0	41
VI	917	YFL067W	0	30	286,77,0	76
VI	30088	YFL051C	0	10	.,.,.	0
VI	107821	YFL013W-A	0	188	.,.,.	0
VI	137956	YFL002W-A&YFL002W-B	49	180	553,55,0	54
VII	8631	YPS5	3	249	935,246,0	245
VII	280086	NAB2	5	137	897,227,0	226
VII	561966	YGR038C-A&YGR038C-B	0	39	372,100,0	99
VII	811442	0	87	830,222,0	222	830,222,0
VII	1048437	CWC22	2	157	935,246,0	245
VIII	85568	YHL009W-A&YHL009W-B	0	109	954,255,0	255
VIII	214132	YHR054C	0	472	.,.,.	0
VIII	215217	RSC30	0	94	.,.,.	0
VIII	543607	YHR214C-B&YHR214C-C	0	259	.,.,.	0
VIII	556852	YHR217C	0	326	.,.,.	0
VIII	560547	YHR219W	0	136	.,.,.	0
VIII	560596	YHR219W	0	393	.,.,.	0
VIII	560652	YHR219W	0	170	.,.,.	0
X	1802	YJL225C	12	46	324,34,0	33

CHROM	POS	ANN[0].GENE	GEN[*].DR	GEN[*].DV	GEN[*].PL	GEN[*].GQ
X	197768	YJL113W&YJL114W	0	17	162,43,0	43
X	293135	PRY3	1	343	954,255,0	255
X	470128	YJR023C	0	200	.,.,.	0
X	472460	RBH2&YJR026W&YJR027W&YJR028W&YJR029W	0	301	.,.,.	0
X	628809	ABM1	0	389	.,.,.	0
X	714387	DAN4	0	415	.,.,.	0
X	715113	DAN4	0	185	.,.,.	0
XI	144777	PIR3	0	109	954,255,0	255
XI	311170	NUP100	35	114	515,36,0	35
XI	380914	IXR1	0	126	954,255,0	255
XI	382551	YKL030W	10	105	802,179,0	178
XI	647365	FLO10	0	74	706,189,0	188
XII	5685	YLL066W-B	0	367	.,.,.	0
XII	7519	YLL066C	0	403	.,.,.	0
XII	103664	SPA2	0	394	.,.,.	0
XII	215074	YLR035C-A	0	279	954,255,0	255
XII	489669	YLR162W	21	137	706,131,0	131
XII	593143	YLR227W-A&YLR227W-B	0	305	954,255,0	255
XII	650821	YLR256W-A	0	254	954,255,0	255
XII	789127	CHS5	0	348	954,255,0	255
XII	941188	YLR410W-A&YLR410W-B	0	295	.,.,.	0
XII	1067108	BSC3	0	36	.,.,.	0
XII	1068289	YRF1-4	0	72	.,.,.	0
XII	1069834	YRF1-4	0	282	.,.,.	0
XII	1071650	YLR466C-B	0	422	.,.,.	0
XII	1073330	YRF1-5	0	124	.,.,.	0
XIII	1841	YML133C	0	214	.,.,.	0
XIII	184163	YML045W&YML045W-A	0	8	.,.,.	0
XIII	196332	YML039W&YML040W	0	104	954,255,0	255
XIII	357003	YMR045C&YMR046C&YMR046W-A	0	99	945,253,0	252
XIII	372695	YMR050C&YMR051C	4	63	563,133,0	132
XIV	519158	YNL054W-A&YNL054W-B&tD(GUC)N	0	266	954,255,0	255
XV	29249	HPF1	0	105	954,255,0	255
XV	29647	HPF1	0	317	954,255,0	255
XV	31216	HPF1	1	228	954,255,0	255
XV	117702	YOL103W-A&YOL103W-B	0	118	954,255,0	255
XV	344992	TIR4	1	87	821,215,0	215
XV	594815	YOR142W-A&YOR142W-B	0	86	821,220,0	219
XV	970284	YOR343W-A&YOR343W-B	0	77	735,197,0	196
XVI	15498	YPL277C	0	390	.,.,.	0
XVI	194007	MF(ALPHA)1	3	428	954,255,0	255
XVI	436887	YPL060C-A	0	289	.,.,.	0
XVI	786432	YPR123C	0	208	.,.,.	0
XVI	804641	YPR137C-A&YPR137C-B	0	303	954,255,0	255
XVI	818558	YPR142C	58	294	649,103,0	102
XVI	844415	YPR158W-A&YPR158W-B	0	107	.,.,.	0
XVI	946259	YPR204W	0	396	.,.,.	0
XVI	946376	YPR204W	0	117	.,.,.	0

Discussion

In this short report, we show how the popular **PEPPER + Margin + DeepVariant** and **pipeline-structural-variation** workflows can be applied to ONT gDNA long read sequencing. The gDNA data was here obtained using Adaptive sequencing the same would work with any other gDNA sequencing approach generating sufficient coverage depth across a genome.

We used here the best basecalling method available to date (guppy +SUP) which is documented by ONT to significantly increase the accuracy with less than 10% residual errors.

When possible, the **DeepVariant** results are phased thanks to long reads support across large regions, which can be very useful to identify linked mutations and have a better view on the ploidy of the affecting variants.

CAUTION: It is well known that the ONT platform suffers from false positive In-Dels which will introduce frame-shifts leading to artefactual stop codons. Therefore this type of results should be considered noisy and validated by re-sequencing before claiming to have found causative variants.

This easy use of the Docker pipeline puts variant analysis at reach with reasonable computer resource needs, especially when dealing with smaller genomes like the yeast and having high read coverage depth as seen in this project (>400x for even chromosomes and >200x for odd chromosomes due to different run throughput).

The **pipeline-structural-variation** results are more difficult to review by nature and require visual inspection as well as validation using custom amplicon PCR.

Additional filtering and variant visualization; using for instance **IGV** or **Ribbon** is not discussed here and will be necessary to produce candidate variant lists.

References

Cingolani, P., A. Platts, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3." *Fly* 6 (2): 80–92.

Coster, Wouter De, Svenn D'Hert, Darrin T Schultz, Marc Cruts, and Christine Van Broeckhoven. 2018. "NanoPack: Visualizing and Processing Long-Read Sequencing Data." Edited by Bonnier Berger. *Bioinformatics* 34 (15): 2666–69. https://doi.org/10.1093/bioinformatics/bty149.

Jiang, Tao, Yongzhuang Liu, Yue Jiang, Junyi Li, Yan Gao, Zhe Cui, Yadong Liu, Bo Liu, and Yadong Wang. 2020. "Long-Read-Based Human Genomic Structural Variation Detection with cuteSV." *Genome Biology* 21 (1). https://doi.org/10.1186/s13059-020-02107-y.

Li, H. 2018. "Minimap2: pairwise alignment for nucleotide sequences." *Bioinformatics* 34 (18): 3094–3100.

Pedersen, Brent S, and Aaron R Quinlan. 2017. "Mosdepth: Quick Coverage Calculation for Genomes and Exomes." Edited by John Hancock. *Bioinformatics* 34 (5): 867–68. https://doi.org/10.1093/bioinformatics/btx699.

Ren, Jingwen, and Mark JP Chaisson. 2020. "Lra: The Long Read Aligner for Sequences and Contigs," November. https://doi.org/10.1101/2020.11.15.383273.

Shafin, Kishwar, Trevor Pesout, Pi-Chuan Chang, Maria Nattestad, Alexey Kolesnikov, Sidharth Goel, Gunjan Baid, et al. 2021. "Haplotype-Aware Variant Calling Enables High Accuracy in Nanopore Long-Reads Using Deep Neural Networks," March. https://doi.org/10.1101/2021.03.04.433952.