Appendix

B2.1—*HybridVar* (*v1.5*)

Aim: Processing of VCF files for heterogenous, hybrid spore read data

Input(s): Reference Genome (FASTA), VCF Files

Output: Dual coordinate variant tables, modified reference genome

Req(s): Perl 5.25, BioPerl

B2.1.1—VCF Processing

HybridVar is designed to work in conjunction with GATK HaplotypeCaller (v3.7), a *de novo* assembly approach to SNP/INDEL discovery, of which a typical run per sample constitutes:

java –jar GenomeAnalysisTK.jar –T HaplotypeCaller –R S288cReference.fa –I Sample_Sorted.bam –o Sample.vcf

For each sample, a Variant Call Format (VCF) file (v4.1) is produced, detailing all discrepancies (SNPs/INDELs) between read and reference. Data throughout this chapter was aligned against S288c, SGD Jan 2015 - R64-2-1. Variant miscalling rates are typically high, and thus further filtering is often required. VCF files adopt a columnar format defined by 9 sections: CHR, POS, ID, REF, ALT, QUAL, FILTER, INFO, GENOTYPE. Initial columns specify the detected variant:

CHR	POS	QUAL	REF	ALT	TYPE
I	27397	.	T	C	SNP
I	27398	.	T	C	SNP
I	27402	.	G	T	SNP
I	27405	.	G	GA	INDEL
I	27408	.	G	A,T	MULTIALLELIC SNP

The variable GENOTYPE section provides delimited information useful for assessing variant quality or confidence, namely (i) AD (Allelic Depth)—the number of reads which support each reported allele e.g. 0,19, denotes that 0 reads match REF, 19 reads match VAR (ii) DP (Read Depth—the number of reads covering this loci (i.e. coverage). Additional information is also specified including GT (genotype of the sample site), GQ (phred scaled confidence) and PL (normalised phred scale likelihood):

TAGS	VALUES
GT:AD:DP:GQ:PL	1/1:1,16:17:48:642,48,0
GT:AD:DP:GQ:PL	1/1:0,16:16:48:642,48,0
GT:AD:DP:GQ:PL	1/1:0,19:19:57:855,57,0
GT:AD:DP:GQ:PL	1/1:0,19:19:57:855,57,0
GT:AD:DP:GQ:PL	1/1:0,15:15:45:392,45,0

HybridVar exploits these scoring parameters, sequentially reading each VCF file provided and calculating (i) call frequency (CF) (% of spores (VCF files) any given allele is present within) (ii) cumulative total read depth (tRD) of each loci, calculated via DP (iii) cumulative allelic read depth (vRD) (% of reads that contain a specific allele at a specific loci), calculated via AD. A typical run of *HybridVar*, which allows user specified filtering based on CF, tRD and vRD, is:

HybridVar.pl –r <ReferenceFASTA> –lf <CallFreqLowerLimit> –uf <CallFreqUpperLim> –trd <MinReadDepth> –vrd <MinVarDepth>

HybridVar.pl –r s288c.fasta –lf 48 –uf 52 –trd 250 –vrd 0.95

Multiallelic sites, with >1 ALTs specified (e.g. A,T as shown above), are split, assessed and filtered separately. INDELs shift the relative positions of all variants. *HybridVar* therefore progressively tracks these changes in order to construct a dual coordinate tab delimited .txt variant file for all SNPs/ INDELs which pass filtering in the following format:

ID	chrom	pos_A	pos_B	seq_A	seq_B	type_A	type_B
1	1	27804	27804	C	A	s	s
2	1	27810	27810	T	C	s	s
3	1	27816	27816	G	A	s	s
4	1	27822	27822	T	C	s	s
5	1	27823	27823	C	A	s	s
6	1	27825	27825	C	T	s	s
7	1	27914	27914	T	C	s	s
8	1	27948	27948	A	G	s	s
9	1	27970	27970	T	G	s	s
10	1	27983	27983	T	A	s	s
11	1	27997	27997	T	C	s	s
12	1	28007	28007	G	C	s	s
13	1	28008	28008	C	C	d	i
13	1	-	28009	-	A	d	i
14	1	28021	28022	G	T	s	s

Pos_A specifies the reference coordinate while pos_B specifies the position of any given variant within a hypothetical genome that contains only these listed variants. Variant ID(13) denotes an insertion relative to the reference (C→CA). All subsequent pos_B positions are thus shifted by 1 bp to account for the additional A base. The terminal columns (type_A, type_B) specify the type of variant relative to each genome—s = SNP, i = insertion, d = deletion.

B2.1.2—Variant Genome

Reads heavily laden with variants relative to the base reference (i.e. S288c) may fail alignment, losing critical event information—a caveat, however, that is bypassed by a dual alignment approach against two references. To accommodate this, *HybridVar* utilises the information stored within filtered variant tables to modify a user provided FASTA file (REF), constructing a novel reference containing all detected variants (VAR), improving the alignment of variant dense reads:

TTGTTC**T**TTTTAAATT**GC**_AATTTAAAGAGCGTACCTGTAAATAAGAAG — *REF* (uIDs 11/12/13)

TTGTTC**C**TTTTAAATT**CCA**AATTTAAAGAGCGTACCTGTAAATAAGAAG — *VAR* (uIDs 11/12/13)

Variants ID(11) (T→C SNP), ID(12) (G→C SNP) and ID(13) (C→CA insertion) are marked above. The modified genome and generated tab delimited variant tables feed directly into the event assignment pipeline.