VALET: *de novo* pipeline for finding metagenomic mis-assemblies

Supplemental Material

1 VALET Pipeline

VALET flags regions as potential mis-assemblies when the characteristics of the sequence data do not meet the assumptions for the sequence generation process regarding (1) depth of coverage, (2) insert size consistency, and (3) alignability of the sequences (Figure 1). Regions that share multiple mis-assembly signatures may be more likely an *actual* mis-assembly.

1.1 Depth of Coverage Analysis

The sequencing process of randomly-sheared fragments follows a Poisson distribution [Lander and Waterman, 1988]. Regions within assembled contigs that show high variance in coverage may be potential mis-assemblies. Areas of high and low depths of coverage could be compressed/expanded repeats, chimeric contigs, and other types of contamination. The main benefit of this approach is that it is completely data-dependent. No prior assumptions of the distribution of the quality values need to be made.

1.2 Insert size consistency

Although the sequence technology can only give the raw sequence of the first couple hundred basepairs from the ends of the fragment, the distance between the ends of the sequences can be used to aid in resolving repeats, and orienting and scaffolding contigs. Regions of an assembly containing a disproportionate number of mate-pairs (reads from the same fragment) with incorrect insert sizes may be a potential mis-assembly.

1.3 Identifying assembly breakpoints

The reads given to the assembler should be alignable to the resulting assembly. During sequencing, reads are produced starting from random locations within the genome. Thus, the reads must be able to align to the resulting assembly. In practice, however, a read may fail to align for a few reasons. When a read is unable to align to an assembly, there must be some explanation. In metagenomic samples, an unaligned read can be from a rare, low coverage organism, and was

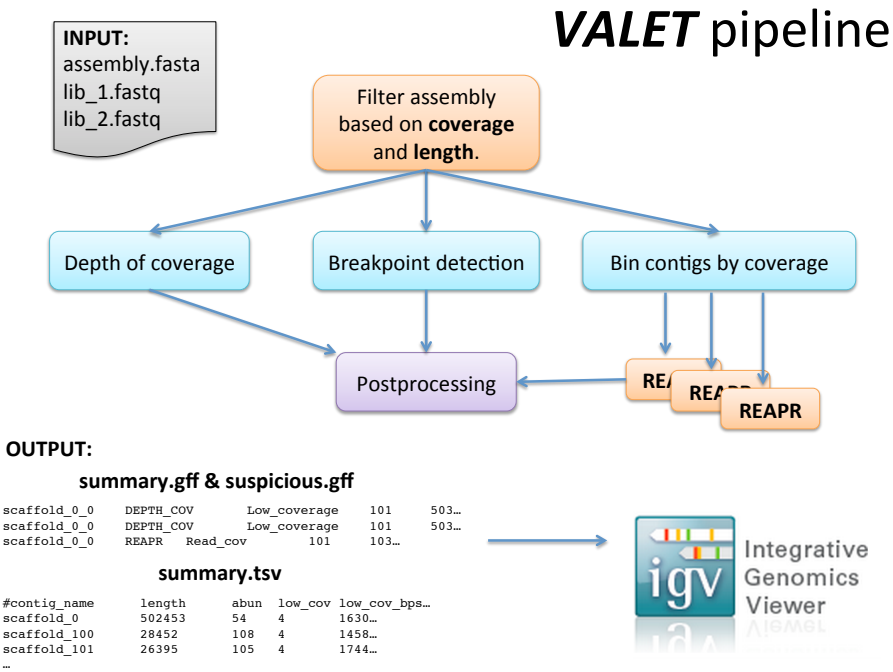

Figure 1: Digram of the VALET pipeline.

never assembled with any other reads. A read with a large amount of errors will be unable to align within a specified similarity to the assembly. A read can come from a unfiltered contaminant or primer. If a read does not fall into one of the above categories, then it may be a sign of a potential mis-assembly.

2 Evaluation

2.1 Definition of a mis-assembly

We compared misassemblies identified by VALET on a simulated and synthetic metagenomic dataset to the reference based metagenome assembly evaluation tool MetaQUAST [Mikheenko *et al.*, 2015]. We ran VALET on the assemblies and compared the errors found with the reference-based mis-assemblies detected by MetaQUAST. MetaQUAST uses the PlantagoraâĂŹs definition of a mis-assembly, i.e., a mis-assembly breakpoint is defined as a position in the assembled contigs where (1) the left flanking sequence aligns over 1kb away from right flanking sequence on the reference, or (2) the sequences overlap by over 1kb, or (3) the right flanking sequence aligns on opposite strands or different chromosomes. As mutations in the strain in a metagenome can contain structural variants relative to the reference genome. MetaQUAST uses the read mapping structural variant caller MANTA [Chen *et al.*, 2015] to filter canidate misassemblies. If any part of a region flagged by VALET overlaps with a mis-assembled region reported by MetaQUAST, we consider it a true positive (mis-assembly correctly identified by our method).

3 Example Misassembly

Misassemblies identified by VALET can be further investigated using a genome viewer. the VALET was used to compare metagenome. For example, lets look at one of the suspicious regions flagged in an assembly of a Vaginal introitus sample (SRS014465) from the Human Microbiome Project [Consortium *et al.*, 2012] (Supplemental Material Figure 2). VALET flags a 1.7 kbp high coverage region flanked by breakpoints. This region BLASTs [Altschul *et al.*, 1997] to *Lactobacillus amylovorus* GRL 1118 plasmid2 (CP002611.1). The right flanking region of the contig from positions 5,362 to 10,839 aligns to *Lactobacillus crispatus* ST1, strain ST1 (FN692037.1) with 98% similarity. The left flanking 3.75 kb region does not have a complete alignment with any sequence in NCBI; the closest alignment being from positions 1 to 2,599 to *Lactobacillus helveticus* strain KLDS1.8701 (CP009907.1).

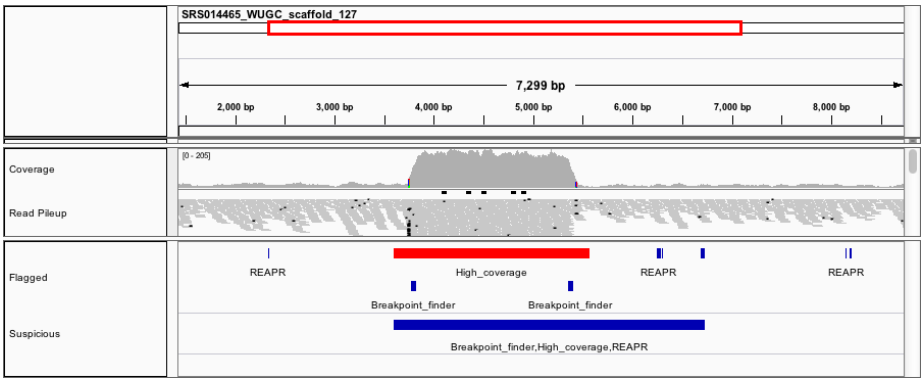

Figure 2: An example suspicious region flagged by VALET in the HMP assembly of of a Vaginal introitus sample (SRS014465) from the Human Microbiome Project [Consortium *et al.*, 2012].

4 Comparison of Valet and MetaQUAST

4.1 Simple Mock Metagenomic Community

We examine the overlap in misassemblies identified by VALET and MetaQUST using a simple mock metagenomic community consisting of four bacteria at varying abundances: 80X *Bacteroides vulgatus* ATCC8482 (Accn. NC_009614.1), 60X *Bacillus cereus* (Accn. NC_003909.8), 40X *Actinomyces odontolyticus* (Accn. NZ_DS264586.1), and 20X *Acinetobacter baumannii* (Accn. NC_009085.1). Approximately four million reads are simulated using WGSIM [Li, 2013] with default parameters and 100 bp forward and reverse reads. The dataset was assembled with Velvet [Zerbino and Birney, 2008], MetaVelvet [Namiki *et al.*, 2012], MaSuRCA [Zimin *et al.*, 2013], and SOAPdenovo2 [Luo *et al.*, 2012] algorithms using MetAMOS [Treangen *et al.*, 2013]. SOME TEXT SUMMARIZING THE COMPARISON RESULTS (Table 1)

4.2 Synthetic Metagenomic Community

A major challenge of evaluating assemblies of environmental datasets is that a sizeable portion of the organsisms are unknown or lack a draft genome to compare against. In silico simulations often lack the complexity and sequencing biases present in real environmental samples. Fortunately, Shakya et al. provide a *gold standard* synthetic metagenomic dataset containing a mixture of 64 organisms (16 members of Archaea and 48 organisms from 18 Bacteria phyla) with complete or high-quality draft genomes and 200-fold differences in abundances [Shakya *et al.*, 2013]. The dataset consists of 53.4 million reads (101 bp in length). We assembled the dataset using Velvet [Zerbino and Birney, 2008], MetaVelvet [Namiki *et al.*, 2012], SPAdes [Bankevich *et al.*, 2012], and SOAPde-

Table 1: VALET results for simulated mock community consisting of four bacteria at varying abundances. Reads were assembled using the four provided assemblers. General assembly statistics include length in Mbp (Len), number of contigs (Ctgs), N50 contig size in kbp (N50), N50 of contigs after broken at mis-assemblies in kbp (NA50) and mismatches per were calculated for contigs >5 kbp. The number of errors detected by MetaQUAST (MQ Errs), number of flagged regions by VALET (Flg), number of flagged regions that overlap an error found by MetaQUAST (Flg Overlap), number of valid misassemblies or flagged regions with multiple misassembly signatures (Vld), number of valid revions that overlap an error found by MetaQUAST (Vld Overlap).

Metric	MaSuRCA	MetaVelvet	SOAPdenovo2	Velvet
Len	16,491,755	12,162,955	8,317,492	1,006,455
Ctgs	343	312	841	149
N50	83,189	67,835	10,424	6,603
NA50	81,327	65,716	10,424	6,027
MPK	67.55	63.41	58.32	56.82
MQ Errs	53	20	0	0
Flg	401	165	0	0
Flg Overlap	41	16	0	0
Vld	33	17	0	0
Vld Overlap	25	11	0	0

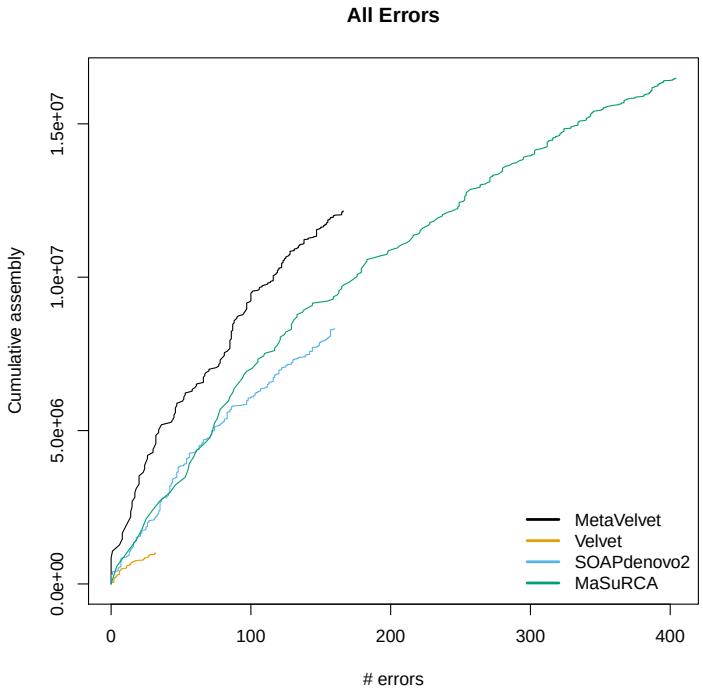

Figure 3: FRC plot provided by VALET of a simulated mock community.

Table 2: VALET results for assemblies of the Shakya et al. [Shakya *et al.*, 2013] dataset. Reads were assembled using the four provided assemblers. General assembly statistics include length in Mbp (Len), number of contigs (Ctgs), N50 contig size in kbp (N50), N50 of contigs after broken at mis-assemblies in kbp (NA50) and mismatches per were calculated for contigs >5 kbp. The number of errors detected by MetaQUAST (MQ Errs), number of flagged regions by VALET (Flg), number of flagged regions that overlap an error found by MetaQUAST (Flg Overlap), number of valid misassemblies or flagged regions with multiple misassembly signatures (Vld), number of valid revions that overlap an error found by MetaQUAST (Vld Overlap).

Metric	MetaVelvet	SOAPdenovo2	SPADES	Velvet
Len	134,157,560	150,990,751	173,364,620	133,773,517
Ctgs	5,707	5,782	5,367	6,042
N50	41,337	50,979	74,297	36,654
NA50	38,279	45,408	62,668	33,908
MPK	45.9	41.89	92.22	47.26
MQ Errs	332	472	1174	327
Flg	1111	1288	1551	1136
Flg Overlap	7	18	45	7
Vld	3	3	2	3
Vld Overlap	0	0	0	0

novo2 [Luo *et al.*, 2012] by MetAMOS [Treangen *et al.*, 2013]. We ran VALET on the assemblies and compare the errors found with the reference-based misassemblies detected by MetaQUAST (Table 2). These rankings match those provided by VALET (Figure 3).

While the MEGAHIT and Omega assemblies are close in total size (192.3 Mbp vs. 194 Mbp, respectively), MEGAHIT has nearly twice as many contigs as Omega (19,145 vs. 10,284). QUAST detects far more mis-assemblies in the Omega assembly compared to MEGAHIT (56,917 vs. 770, respectively). VALET detects 34.80% and 96.10% of these mis-assemblies found by QUAST in the MEGAHIT and Omega assemblies, respectively. While Omega has a higher N50 than MEGAHIT (44.1 Kbp vs. 38.9 Kbp), after breaking at mis-assemblies, the N50 drops well below MEGAHIT's (11.9 Kbp vs. 33.5 Kbp), illustrating why the N50 metric is not always a good indicator of assembly quality. VALET is able to accurately assess the quality of the two assemblies without using the reference genomes.

4.3 Mis-assemblies not validated by MetaQUAST

We investigated the high number of misassemblies identified by VALET and not MetaQUST by manually visualizing the misassembly regions for one the XYZ dataset.

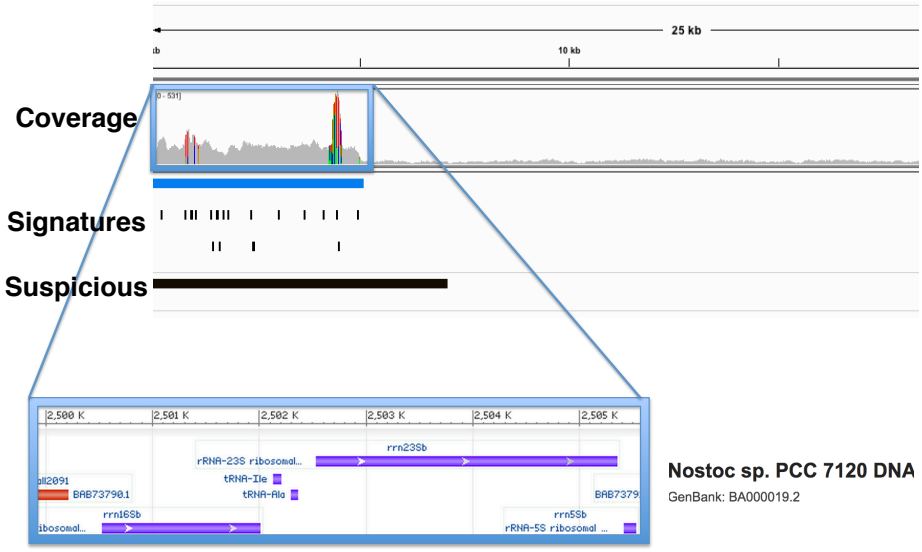

Figure 4: A closer examination of a region flagged by VALET, but no mis-assembly reported by QUAST. This region contained 16S, 23S, and 5S rRNA genes and was found at four locations in the *Nostoc sp. PCC 7120 genome*.

One contig, roughly 25 Kbp in size, had a 5 Kbp region at the start of the contig marked as high coverage (Figure 4). This region was roughly 4x the median coverage of the remaining contig. Using NCBI's BLAST [Altschul *et al.*, 1997] and reference database, the region aligned to the organism *Nostoc* sp. PCC 7120. Upon closer inspection, this region contained 16S, 23S, and 5S rRNA genes and was found at *four* locations in *Nostoc* sp. PCC 7120. This region was only found once in the assembly, so all the sequences from the repeats aligned to this region, inflating the coverage. This noticeable and consistent increase in coverage caused VALET to mark it as mis-assembled. Unsurprisingly, QUAST did not mark this as a mis-assembly because the actual sequence within this region matched the reference.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**(17), 3389–3402.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., *et al.* (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, **19**(5), 455–477.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Cox, A. J., Kruglyak, S., and Saunders, C. T. (2015). Manta: Rapid detection of structural variants and indels for clinical sequencing applications. *bioRxiv*, page 024232.

Consortium, H. M. P. *et al.* (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, **486**(7402), 207–214.

Lander, E. S. and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**(3), 231–239.

Li, H. (2013). wgsim-read simulator for next generation sequencing.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., *et al.* (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, **1**(1), 18.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2015). Metaquast: evaluation of metagenome assemblies. *Bioinformatics*, page btv697.

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*, **40**(20), e155–e155.

Shakya, M., Quince, C., Campbell, J. H., Yang, Z. K., Schadt, C. W., and Podar, M. (2013). Comparative metagenomic and rrna microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental microbiology*, **15**(6), 1882–1899.

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., Darling, A. E., Phillippy, A. M., and Pop, M. (2013). Metamos: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol*, **14**(1), R2.

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, **18**(5), 821–829.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The masurca genome assembler. *Bioinformatics*, **29**(21), 2669–2677.