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

Supplemental Material

1 Evaluation

1.1 Definition of a mis-assembly

We evaluate the correctness of VALET on a simulated and synthetic metagenomic dataset. We ran VALET on the assemblies and compared the errors found with the reference-based mis-assemblies detected by QUAST [**?**]. QUAST uses the Plantagoras 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. If any part of a region flagged by VALET overlaps with a mis-assembled region reported by QUAST, we consider it a true positive (mis-assembly correctly identified by our method).

1.2 VALET achieves high sensitivity on a simulated metagenomic community

We examine the sensitivity of VALET on a toy simulated metagenomic community consisting of four bacteria at varying abundances: *Bacteroides vulgatus* (80x), *Bacillus cereus* (60x), *Actinomyces odontolyticus* (40x), and *Acinetobacter baumannii* (20x). Approximately four million reads are simulated using WGSIM [**?**] with default parameters. The dataset was assembled using IDBA-UD [**?**], MetaVelvet [**?**], SPAdes [**?**], and SOAPdenovo2 [**?**].

Across all assemblers, VALET detects greater than 80% of mis-assemblies detected by QUAST (Supplemental Table 1).IDBA-UD has the greatest N50 after breaking the assembly at regions marked by QUAST (206.7 Kbp), followed by SPAdes (128.1 Kbp), MetaVelvet (29.5 Kbp), and SOAPdenovo2 (10.8 Kbp). These rankings match those provided by VALET (Supplemental Figure 1).

						Mis-assembly signatures			Suspicious regions			
Assembler	Len (Mbp)	Ctgs	N50 (Kbp)	NA50	Errs	Num	Valid	Sens	Num	Valid	Sens	Mismatches per Kbp
IDBA-UD	16.5	200	208.8	206.7	36	804	36	100.00%	25	8	22.20%	23.95
MetaVelvet	16.3	1,117	29.5	29.5	21	2,802	19	90.50%	4	2	9.50%	35.52
SPAdes	16.4	330	130.9	128.1	37	1,117	31	83.80%	17	4	10.80%	22.43
Soapdenovo2	12.3	2,161	10.8	10.8	1	4,983	1	100.00%	2	0	0%	13.37

Table 1: VALET results for simulated mock community consisting of four bacteria at varying abundances: *Bacteroides vulgatus* (80x), *Bacillus cereus* (60x), *Actinomyces odontolyticus* (40x), and *Acinetobacter baumannii* (20x). Reads were assembled using the four provided assemblers. General assembly statistics include length in Mbp (Len), number of contigs (Ctgs), N50 contig size (N50), N50 of contigs after broken at mis-assemblies (NA50), number of errors detected by QUAST (Errs), number of flagged regions by VALET (Num), number of flagged regions that overlap an error found by QUAST (Valid), sensitivity (Sens), and mismatches per Kbp.

						Mis-assembly signatures			Suspicious regions			
Assembler	Len (Mbp)	Ctgs	N50 (Kbp)	NA50 (Kbp)	Errs	Num	Valid	Sens	Num	Valid	Sens	Mismatches per Kbp
MEGAHIT	192.3	19,145	38.9	33.5	770	30,377	268	34.80%	2,239	100	13.00%	92.24
Omega	194	10,284	44.1	11.9	56,917	1,425,127	55,108	96.10%	17,758	13,935	96.80%	98.55

Table 2: VALET results for assemblies of the Shakya et al. [?] dataset. General assembly statistics include length in Mbp (Len), number of contigs (Ctgs), N50 contig size (N50), N50 of contigs after broken at mis-assemblies (NA50), number of errors detected by QUAST (Errs), number of flagged regions by VALET (Num), number of flagged regions that overlap an error found by QUAST (Valid), sensitivity (Sens), and mismatches per Kbp.

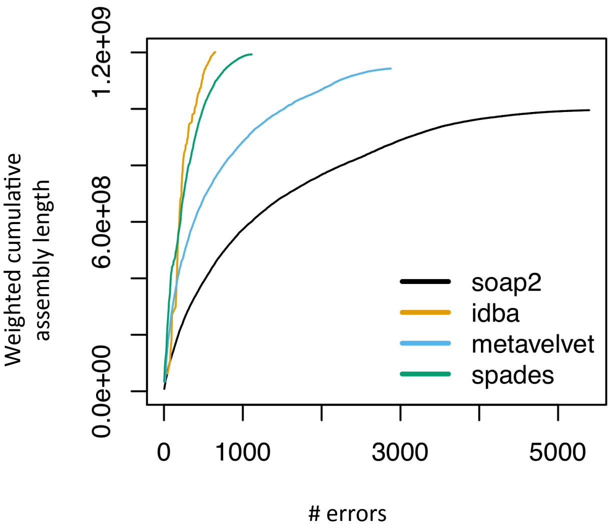

Figure 1: FRC plot provided by VALET of a simulated mock community.

1.3 VALET accurately evaluates assemblies of a 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 [**?**]. The dataset consists of 53.4 million reads (101 bp in length). Due to the greater size and complexity of this dataset compared to the previous simulation, we assemble the dataset using two recent, fast metagenomic assemblers: Omega [**?**] and MEGAHIT[**?**]. We run VALET on the assemblies and compare the errors found with those reference-based mis-assemblies detected by QUAST (Table 1.2).

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.

1.4 Mis-assemblies not validated by QUAST

We investigate the high false positive rate by examining a small number of regions flagged by VALET, but not marked by QUAST within the MEGAHIT assembly. One contig, roughly 25 Kbp in size, had a 5 Kbp region at the start of the contig marked as high coverage (Figure 2). This region was roughly 4x the median coverage of the remaining contig. Using NCBI's BLAST [**?**] 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.

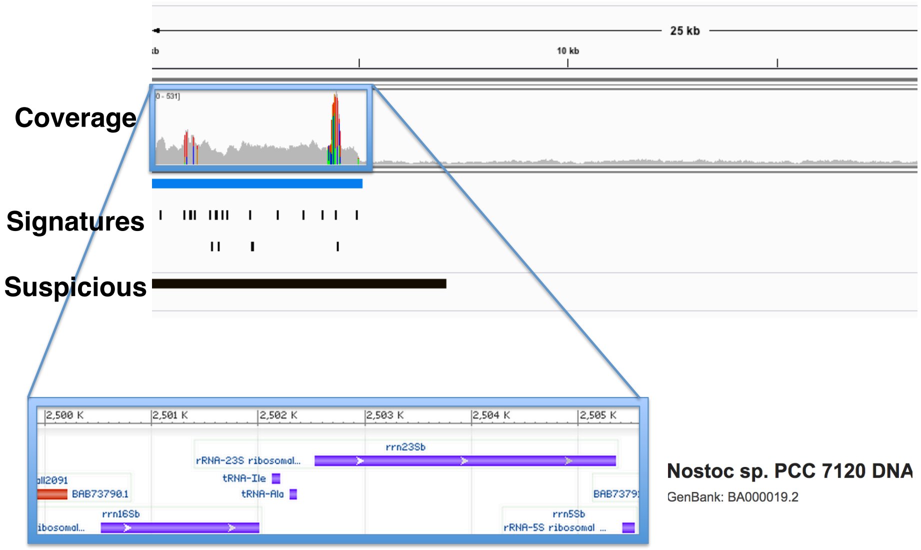

Figure 2: A closer examination of a region flagged by VALET, but no misassembly reported by QUAST. This region contained 16S, 23S, and 5S rRNA genes and was found at four locations in the *Nostoc sp. PCC 7120 genome*.