HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2009 July 01.

Published in final edited form as:

Nat Methods. 2009 January ; 6(1): 67–69. doi:10.1038/nmeth.1286.

Sensitive, specific polymorphism discovery in bacteria using massively parallel sequencing

Chad Nusbaum[1], **Toshiro K. Ohsumi**[1,7], **James Gomez**[1], **John Aquadro**[1], **Thomas C. Victor**[2], **Robert M. Warren**[2], **Deborah T. Hung**[1,3], **Bruce W. Birren**[1], **Eric S. Lander**[1,4,5,6], and **David B. Jaffe**[1]

[1] Broad Institute of Massachusetts Institute of Technology and Harvard, 320 Charles Street, Cambridge, MA 02141, USA

[2] Department of Science and Technology / National Research Foundation Centre of Excellence in Biomedical Tuberculosis Research / Medical Research Council Centre for Molecular and Cellular Biology, Department of Biomedical Science, Stellenbosch University, PO Box 19063, Tygerberg 7505, Cape Town, South Africa

[3] Department of Molecular Biology and Molecular Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115, USA

[4] Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02139, USA

[5] Department of Biology, Massachusetts Institute of Technology, 31 Ames Street, 68-132, Cambridge, MA 02139, USA

[6] Department of Systems Biology, Harvard Medical School, 200 Longwood Avenue, Alpert 536, Boston, MA 02115, USA

Abstract

Our variant ascertainment algorithm VAAL uses massively parallel DNA sequence data to identify differences between bacterial genomes at high sensitivity and specificity. VAAL found ~98% of differences (including large indels) between pairs of strains from three species while calling no false positives. Further, VAAL pinpointed a single mutation between *Vibrio* genomes, identifying an antibiotic's site of action by finding the difference(s) between a drug sensitive strain and a resistant derivative.

The bacterial world that surrounds us is of enormous importance for studies ranging from human medicine to environmental ecology. Even within 'species' of bacteria, there is enormous genetic variation. The ability to routinely compare the genomes of many bacterial

Correspondence should be addressed to D.B.J. (jaffe@broad.mit.edu).

[7]Current address: Department of Molecular Biology, Massachusetts General Hospital, Richard B. Simches Research Center, 185 Cambridge Street, CPZN-7250, Boston MA 02114.

This variant ascertainment algorithm, or VAAL, uses short sequence reads of haploid bacterial genomes to first perform a local assembly of the reads, and then compare these assemblies to the reference genome. This allows VAAL the detection of all types of variants ranging from single nucleotide polymorphisms to large insertions or deletions.

strains and identify their differences is thus of tremendous value to understanding their impact on human health and biology.

In this work we address two key applications: comparison of a new strain with a previously sequenced reference strain and comparison of a laboratory-induced mutant strain to the parental strain. An example for the former would be the comparison of multiple clinical isolates of tuberculosis, which might differ in their resistance to various drugs. We expect to find genetic variation at a number of sites across the genome; the critical genetic differences underlying a phenotype becomes apparent because they are seen in multiple strains. An example for the latter would be a mutant strain resistant to a particular antibiotic. We expect only one mutation in such a comparison, which can be used to pinpoint the protein target of the antibiotic. This is valuable because there is currently no simple way to identify the sites of action of a drug.

Pioneering work on these applications used available technology, often in multi-stage processes. For example whole-genome shotgun Sanger sequencing was followed by directed resequencing1–2, or an initial microarray screen was followed by a validation microarray3. In each case the second stage was designed based on the results of the first stage. With the advent of massively parallel sequencing methods4–7, a streamlined approach is feasible, facilitating routine comparison at dramatically lower cost. Indeed, substantial sequence for a bacterial strain may now be generated for roughly $1,000, and the costs are likely to drop further.

We require new computational methods to take advantage of the new inexpensive data types and deliver highly accurate results. Methods are needed that can find the full spectrum of variation, including insertions and deletions of arbitrary size, since all may be phenotypically important. Yet a single base change can confer antibiotic resistance in bacteria, a ‘needle in the haystack’ problem that exemplifies the need for high specificity. While initial work8–13 on this problem has been promising, there has been no systematic and controlled investigation of general variation detection using new technology data. For example, insertion/deletion (indel) detection has not yet been systematically tested on real data.

We developed a new ‘variant ascertainment algorithm’ VAAL, and applied it to a series of comparisons, including between naturally occurring strains, and between antibiotic-sensitive and derived resistant strains. All the experiments were controlled, facilitating rigorous assessment of answers. [AU could you briefly explained how this control was done. Do you mean by comparison to the reference?] For each strain we generated a single lane of unpaired 36 base reads from the Illumina platform, thereby minimizing costs. The source code for VAAL and all experimental data sets are made freely available (Supplementary Resources). All applications of VAAL were carried out with the distributed version of the code and default arguments.

VAAL takes as input short reads from a ‘sample genome’ and a known sequence for a related ‘reference genome’. VAAL assembles the reads14,15, assisting with the reference genome. This ‘assisted assembly’ indicates which bases are ‘trusted’ and which are not. By

comparing the sample genome assisted assembly to the reference genome, VAAL deduces as output a list of differences between them.

There are several steps in the algorithm (Supplementary Methods). Briefly, we assign each sample read an approximate position on the reference genome and then group reads by position. Next we assemble each group, then glue the assemblies along regions where sample and reference genomes agree, yielding a single assembly of the sample genome. Next we call bases in the assembly trusted that are strongly supported by the reads. Finally, to identify polymorphisms, we compare the sample read data assembly to the reference genome and identify differences between them that lie in trusted parts of the assembly. To do this, we identify regions containing disagreements between the assembly and the reference that are flanked by regions where the two agree. These disagreements include substitutions, insertions, deletions, and more complex changes (multiple nearby differences) that we refer to as composites. All are reported as polymorphisms by VAAL. In the text, composites are treated as single polymorphisms, but VAAL also provides output in which each composite is parsed into substitutions, insertions, and deletions.

We designed VAAL to discover sequence differences between related genomes, with high sensitivity and specificity. To demonstrate this, we compared sequence data from ‘sample genomes’ to finished reference sequences for related isolates of the same species. Finished genomes were used so the true answer was known in all cases. The dataset for this work consisted of 36 base Illumina reads from three previously finished bacterial genomes (*S. aureus*, *E. coli*, *M. tuberculosis*, Supplementary Table 1).

Genomes contain repetitive sequences longer than the reads, within which polymorphisms cannot be called unambiguously. To avoid these ambiguities, VAAL defines uncallable regions of the genome in which polymorphisms cannot be called unambiguously with the given data, declaring the remainder of the genome to be callable (Supplementary Methods). The size of the callable fraction depends on the amount of duplicated sequence in the genome. It also depends on read length and the minimum overlap K (here, 28). For the data and genomes used here, the callable fraction ranges from 91.6–96.7% (Table 1). In evaluating the algorithm, we defined the set of callable polymorphisms, which reflect unambiguous differences, as those residing in the callable regions, and used these to evaluate sensitivity (Table 1).

First, we used VAAL to compare the sample reads to the finished sequences of the genomes from which they were generated. Importantly, no polymorphisms were called. This demonstrates high specificity: any reported polymorphisms would have been false positives.

Next we performed the actual between-strain comparison with VAAL, and evaluated the performance on the callable polymorphisms (Table 1).

For *E. coli* (moderate GC content, 51%), VAAL found 100% of callable polymorphisms. The called insertions included three exceeding the read length, of sizes 93, 390, and 970 bp.

For *S. aureus* (low GC content, 33%), VAAL found >99% of substitutions, 95% of indels, and 90% of composites.

For *M. tuberculosis* (high GC content, 66%), VAAL found 97% of substitutions, 83% of indels, and 95% of composites.

Briefly, VAAL showed high sensitivity in finding nearly all of the possible callable polymorphisms. In no cases were any false positives called. There were two cases where *bona fide* polymorphisms were called that lay outside the set of callable polymorphisms. The mechanism for these rare events is explained in Supplementary Methods. VAAL performed well even on complex composite events (Supplementary Methods). We also investigated the failure modes of the few polymorphisms that were missed, and how coverage influences the fraction of missed polymorphisms (Supplementary Methods).

We note that the three bacterial genomes employed have GC content ranging from 33% to 66%, a range encompassing ~80% of previously sequenced bacteria (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). Testing behavior over a range of sequence content is important because sequencing system's behaviors vary considerably as a function of GC content.

We next attempted to discover *single* base differences between a parental strain and a derived mutant strain. In this case, we sequenced both the parental and mutant strains and compared them to a known reference strain.

We compared each of five *Vibrio cholerae* isolates sensitive to the antibiotic rifampicin with derived rifampicin-resistant isolates (Supplementary Methods), anticipating that each resistant isolate would likely have acquired a single resistance mutation. To do this, for each of the five sensitive-resistant pairs, we sequenced both strains and assembled them with VAAL, assisted by a single finished reference (Supplementary Table 4). Then we used VAAL to compare the two assemblies to each other, yielding a list of differences (Table 2). For all five sensitive-resistant pairs, we found exactly one difference in the entire genome. In all cases, the difference was in the *rpoB* gene, which encodes the β subunit of RNA polymerase, rifampicin's known target (Supplementary Results).

We compared VAAL to other algorithms (Supplementary Results): MAQ and VAAL call different numbers of single-nucleotide polymorphisms (SNPs), suggesting distinct advantages; MAQ does not call indels from unpaired reads, as in this paper; a Velvet-VAAL hybrid yielded results comparable to VAAL at high coverage but inferior results at lower coverage, as expected since Velvet assembles de novo.

Sensitive and specific polymorphism discovery with single-molecule sequencing data is a major open problem. VAAL approaches it by grouping reads and then performing a local assembly, taking into account the relationships between individual reads and the reference, and between reads themselves. Thus, the algorithm can find differences (e.g. long insertions) that cannot be found by methods dependent upon alignments of single reads to the reference. Importantly, VAAL does not contain special rules for each polymorphism type (SNP, SNP cluster, indel, *etc*.) and thus can discover all differences without presuppositions.

We have demonstrated the application of massively parallel sequencing to two critical problems in bacterial genetics. First, we showed reliable detection of sequence differences

between two related bacterial strains, including long insertions and deletions, with very high sensitivity, and no observed false positives, across a wide range of genome compositions, using data from only a single lane of unpaired 36 base Illumina reads. Second, we demonstrated discovery of a single mutation conferring antibiotic resistance, without prior knowledge of its location. The first application enables rapid discovery of variations underlying medically or biologically important phenotypes. The second application enables the rapid discovery of the targets of current or newly developed antibiotics, by analyzing a series of resistant mutants. While VAAL requires relatively high coverage, this coverage is inexpensive. Costs are now sufficiently low that analysis of hundreds or thousands of bacterial strains is practical.

In its present form, VAAL works on haploid genomes, providing highly sensitive and specific polymorphic detection for bacteria. Generalization of the algorithm to large, complex, diploid genomes is the next important goal. In particular, it should soon be practical to sequence targeted subsets (such as regions or exons) from hundreds of human samples.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank M. Borowsky, S. Young, J. Maguire, B. Cockerham, M. Grabherr for useful feedback during software development, I. Shlyakhter for computational infrastructure, T. Shea for validating tuberculosis reference differences, the Broad Institute Sequencing Platform for data generation, C. Russ for scientific project management, S. Gabriel, M. Zody for helpful comments, and the National Human Genome Research Institute for support.

References

1. Alm RA, et al. Nature. 1999; 397:176–180. [PubMed: 9923682]
2. Read TD, et al. Science. 2002; 296:1976–1979. [PubMed: 12004075]
3. Albert TJ, et al. Nature Methods. 2005; 2:951–953. [PubMed: 16299480]
4. Shendure J, et al. Science. 2005; 309:1728–1732. [PubMed: 16081699]
5. Service RF. Science. 2006; 311:1544–1546. [PubMed: 16543431]
6. Harris TD, et al. Science. 2008; 320:106–109. [PubMed: 18388294]
7. Bentley DR, et al. Nature. 2008; 456:53–59. [PubMed: 18987734]
8. Andries K, et al. Science. 2005; 307:214–215. [PubMed: 15653490]
9. Velicer GJ, et al. Proc Natl Acad Sci USA. 2006; 103:8107–8112. [PubMed: 16707573]
10. Hillier LW, et al. Nature Methods. 2008; 5:183–188. [PubMed: 18204455]
11. Brockman W, et al. Genome Research. 2008; 18:763–770. [PubMed: 18212088]
12. Holt KE, et al. Nature Genetics. 2008; 40:987–993. [PubMed: 18660809]
13. Li H, Ruan J, Durbin R. Genome Research. 2008; 18:1851–1858. [PubMed: 18714091]
14. Pevzner PA, Tang H, Waterman MS. Proc Natl Acad Sci USA. 2001; 98:9748–9753. [PubMed: 11504945]
15. Butler J, et al. Genome Research. 2008; 18:810–820. [PubMed: 18340039]

Table 1

Results of polymorphism discovery using VAAL

sample organism	related strain (Genbank accession)	callable fraction of genome	callable SNPs	called SNPs	callable indels	called indels	callable composites	called composites	false positives
S. aureus USA 300	COL NC_002951.2	96.5%	687	684	63	60	116	104	0
E. coli K12 MG1655	DH10B CP000948.1	91.6%	116	116	10	10	13	13	0
M. tuberculosis F11	H37Rv AL123456.2	96.7%	757	736	65	54	19	18	0

Reads from three bacterial strains were compared to reference sequences from other strains using VAAL. Polymorphisms were grouped in three categories: SNP, indel or complex changes (composite). Events that are within 28 bases of each other are reported as a single composite event. False positives: incorrect polymorphisms reported by the algorithm. None were found.

Table 2

***Vibrio cholerae*: observed differences between resistant isolates and their sensitive parents**

	resistant isolates				
	JA-G-02	**RIF3-1**	**RIF4-1**	**RIF5-1**	**RIF6-1**
Differences with sensitive cultures	341920 (*rpoB*) A → T, D516 Asp → Val	341920 (*rpoB*) A → T, D516 Asp → Val	341949 (*rpoB*) C → T, H526 His → Tyr	341920 (*rpoB*) A → T, D516 Asp → Va	341949 (*rpoB*) C → T, H526 His → Tyr

Five rifampicin-resistant isolates were compared to the five sensitive cultures in which the mutations originated (pairs of rows in Supplementary Table 4), using an intermediate reference AE00385{2,3}.1 to facilitate the comparison. In each case VAAL found exactly one mutation, and it was on AE003852.1. One-based coordinates along with the mutated gene (*rpoB)* in which the mutations occurred are shown together with the *E. coli*-based numbering of the corresponding codon (D516 or H526), and the amino acid change that occurred.