

MEETING ABSTRACT

Open Access

Normalizing alternate representations of large sequence variants across multiple bacterial genomes

Alex Salazar^{1,2}, Ashlee Earl¹, Christopher Desjardins¹, Thomas Abeel^{1,3*}

From Tenth International Society for Computational Biology (ISCB) Student Council Symposium 2014 Boston, MA, USA. 11 July 2014

Background and description

Variant-focused comparative genomics enables researchers to study the evolution of distinct genetic characteristics in bacterial populations, while avoiding the difficulties of whole-genome assembly and alignment. A major challenge in using this method is that many variant detecting tools are largely limited to predicting single nucleotide variants (SNVs) and small indels. This is a challenge because bacterial organisms do not only possess SNVs but also harbor much larger sequence variants (LSVs), such as large indels and substitutions (>25 nt), when compared to a reference genome. LSVs have been shown to play a role in shaping important biological aspects such as virulence and drug resistance as well as reporting on population structure [1-3]. Recent variant callers, such as Pilon <http://www.broadinstitute.org/software/pilon/>, can identify LSVs with single nucleotide accuracy in microbial genomes. However, one remaining challenge is that identical LSVs can be represented non-identically by a single variant detecting tool; this generally results from similarity in the flanking sequence of the variant and variability of the read quality and alignment information in that region across the different strains. As a result, alternate representations of large variants make it difficult to perform downstream analyses - such as association studies - that depend on consistent representations of variants.

We present Emu, an algorithm that resolves alternate representations of LSVs by comparing variant calls across genomes.

Results

To evaluate Emu's ability to resolve alternate representations of LSVs, we introduced 179 simulated LSVs into the H37Rv genome—a carefully curated and finished reference genome for *Mycobacterium tuberculosis* (Mtb). We then used Pilon to identify variants in a set of 146 clinical samples of Mtb that were collected in China using the modified H37Rv genome as a reference [4]. We identified a total of 10,001 unique variant representations. The average number of non-identical representations of each simulated LSV was 56 (in the range of 1 to 145). We then applied Emu to identify the non-identical representations across the genomes of the 146 clinical samples and canonicalize them to a single form. Emu reduced the total number of non-identical representations to 676 LSVs bringing the average number of non-identical representations at each LSV to 4, with 15 LSVs reduced to a single representation and no LSV having more than 25 representations.

We then investigated how Emu's ability to resolve alternate representations might impact association analyses, e.g., associating LSVs with population structure. We ran Pilon again on the set of 161 clinical samples from China, but used the unmodified H37Rv genome. Pilon identified a total of 20,512 distinct LSVs when compared to the unmodified H37Rv genome. By applying Emu, the number of distinct LSVs decreased by almost 50% to 10,936 LSVs. Emu also increased the power of association tests on the LSVs. While we initially identified a total number of 69 LSVs that were significantly associated ($p < 0.01$) with membership to a specific clade, after processing with Emu that number increased to 94.

Conclusion

Emu enables comprehensive analysis of LSVs in bacterial genomes by reducing the cross-sample noise that

¹Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
Full list of author information is available at the end of the article

results from per-sample variant calls. By normalizing our variant calls with Emu, we increased our power to utilize LSVs association tests. Pilon and Emu are open source tools that can also be applied to identify and normalize variants in other organisms.

Authors' details

¹Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

²University of California, Santa Cruz, California, USA. ³Delft Bioinformatics Lab, Delft University of Technology, Delft, Netherlands.

Published: 28 January 2015

References

1. Alland D, Lacher DW, Hazbón MH, Motiwala AS, Qi W, Fleischmann RD, Whittam TS: **Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of *Mycobacterium tuberculosis* and the utility of LSPs in phylogenetic analysis.** *J Clin Microbiol* 2007, **45**:39-46.
2. Maurelli AT, Fernández RE, Bloch CA, Rode CK, Fasano A: **"Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*.** *Proc Natl Acad Sci USA* 1998, **95**:3943-3948.
3. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, Niyogi SK, Kim EJ, Ramamurthy T, Chun J, Wood JLN, Clemens JD, Czerkinsky C, Nair GB, Holmgren J, Parkhill J, Dougan G: **Evidence for several waves of global transmission in the seventh cholera pandemic.** *Nature* 2011, **477**:462-5.
4. Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, Liu Z, Li C, Galwey N, Deng J, Zhou Y, Zhu Y, Gao Y, Wang T, Wang S, Huang Y, Wang M, Zhong Q, Zhou L, Chen T, Zhou J, Yang R, Zhu G, Hang H, Zhang J, Li F, Wan K, Wang J, Zhang X-E, Bi L: **Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance.** *Nat Genet* 2013, , September: 1-8.

doi:10.1186/1471-2105-16-S2-A8

Cite this article as: Salazar et al.: Normalizing alternate representations of large sequence variants across multiple bacterial genomes. *BMC Bioinformatics* 2015 **16**(Suppl 2):A8.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

