Dashnow et al. BMC Bioinformatics 2015, 16(Suppl 2):A5
http://www.biomedcentral.com/1471-2105/16/S2/A5

BMC Bioinformatics

MEETING ABSTRACT Open Access

Genotyping microsatellites in next-generation sequencing data

Harriet Dashnow[1,2,3*], Susan Tan[4], Debjani Das[4], Simon Easteal[4], Alicia Oshlack[2,3]

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

Background

Microsatellites are short (2-6bp) DNA sequences repeated in tandem, which make up approximately 3% of the human genome [1]. These loci are prone to frequent mutations and high polymorphism with the estimated mutation rates of 10^{-2} - 10^{-6} events per locus per generation, orders of magnitude higher than other parts of the genome [2]. Dozens of neurological and developmental disorders have been attributed to microsatellite expansions [3]. Microsatellites have also been implicated in a range of functions such as DNA replication and repair, chromatin organisation and regulation of gene expression [4].

Traditionally, microsatellite variation has been measured using capillary gel electrophoresis [5]. In addition to being time-consuming, and expensive, this method fails to reveal the full complexity at these loci because it does not directly sequence the fragment but only measure the number of bases in the repeat.

Next-generation sequencing has the potential to address these problems. However, determining microsatellite lengths using next-generation sequencing data is difficult. In particular, polymerase slippage during PCR amplification introduces stutter noise. A small number of software tools have been written to genotype simple microsatellites in next-generation sequencing data [6-8], however they fail to address the issues of SNPs and compound repeats, and in some cases provide only approximate genotypes.

We have begun to develop a microsatellite genotyping algorithm that addresses these issues, providing high accuracy as well as more detailed analysis of microsatellite loci. We have validated it using high depth amplicon sequencing data of microsatellites near the *AVPR1A* gene.

Results

We found high concordance between our algorithm and repeat lengths obtained by electrophoresis, manual inspection and Mendelian inheritance (Table 1). By subsampling the reads, we found that our model is accurate to within one repeat unit down to coverages that we would expect in standard exome sequencing (Figure 1). Additionally, we detected polymorphic single nucleotide changes within some microsatellites.

Conclusions

The algorithm was approximately 95% correct at calling the exact same genotype on high depth sequencing data. When it did call a genotype incorrectly, the genotype was only one repeat unit different. The algorithm can perform at approximately 90% accuracy to within one repeat unit with as few as 20 informative reads and reaches almost 100% accuracy to within one repeat unit with 100 or more informative reads.

Future work will include expanding the algorithm to genotype compound microsatellites and further validation and comparison with other algorithms will be performed on whole genome data sets.

Table 1 Concordance of microsatellite variance calls three validation methods: electrophoresis, manual inspection and Mendelian inheritance

Validation method	Concordant #	Concordant %
Electrophoresis	9/9	100%
Manual inspection	17/18	~95%
Mendelian inheritance	18/18	100%

[1]Life Science Computation Centre, Victorian Life Sciences Computation Initiative, Carlton, VIC, Australia
Full list of author information is available at the end of the article

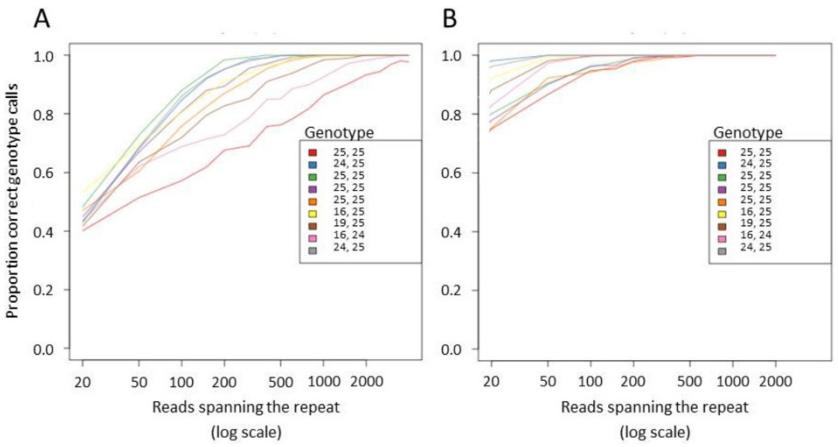

Figure 1 Genotyping accuracy at the $(AC)_n$ promoter locus as a function of the number of reads spanning the microsatellite. 20 to 3000 reads were sampled with replacement from those spanning the microsatellite. This was done 1000 times for each depth. A shows the portion of genotypes that were exactly correct, B shows the proportion of genotypes that were correct to within one repeat unit.

Authors' details
[1]Life Science Computation Centre, Victorian Life Sciences Computation Initiative, Carlton, VIC, Australia. [2]The University of Melbourne, Parkville, VIC, Australia. [3]Murdoch Childrens Research Institute, Parkville, VIC, Australia. [4]John Curtin School of Medical Research - Australian National University, Canberra, ACT, Australia.

Published: 28 January 2015

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409(6822)**:860-921.
2. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ: **Variable tandem repeats accelerate evolution of coding and regulatory sequences.** *Annual review of genetics* 2010, **44**:445-477.
3. Gatchel JR, Zoghbi HY: **Diseases of unstable repeat expansion: mechanisms and common principles.** *Nature Reviews Genetics* 2005, **6(10)**:743-755.
4. Li YC, Korol AB, Fahima T, Beiles A, Nevo E: **Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review.** *Molecular ecology* 2002, **11(12)**:2453-2465.
5. Guichoux E, Lagache L, Wagner S, Chaumeil P, LÉGer P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F, *et al*: **Current trends in microsatellite genotyping.** *Molecular Ecology Resources* 2011, **11(4)**:591-611.
6. Gymrek M, Golan D, Rosset S, Erlich Y: **lobSTR: A short tandem repeat profiler for personal genomes.** *Genome Research* 2012.
7. Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D: **Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles.** *Nucleic acids research* 2012, gks981.
8. Cao MD, Tasker E, Willadsen K, Imelfort M, Vishwanathan S, Sureshkumar S, Balasubramanian S, Bodén M: **Inferring short tandem repeat variation from paired-end short reads.** *Nucleic acids research* 2014, **42(3)**:e16-e16.

doi:10.1186/1471-2105-16-S2-A5
Cite this article as: Dashnow *et al.*: Genotyping microsatellites in next-generation sequencing data. *BMC Bioinformatics* 2015 **16**(Suppl 2):A5.