Genotyping microsatellites in next-gen sequencing data

Harriet Dashnow[12], Joseph Romano[3], Thomas Abeel[3] and Alicia Oshlack[2]

1. Life Science Computation Centre, Victorian Life Sciences Computation Initiative, Carlton, VIC, Australia
2. Murdoch Childrens Research Institute, Parkville, VIC, Australia
3. Broad Institute of MIT and Harvard, Cambridge, MA, USA

Microsatellites

Microsatellites are short (2-6bp) DNA sequences repeated in tandem (Figure 1). Approximately 3% of the human genome consists of microsatellites. Most microsatellites are simple, consisting of a single type of repeat unit. Compound microsatellites consist of two or more adjacent simple microsatellites. Microsatellites have been implicated in a range of functions such as DNA replication and repair, chromatin organisation and regulation of gene expression. In bacteria, microsatellites have long been used in population studies, but have not been more extensively investigated in next-gen data.

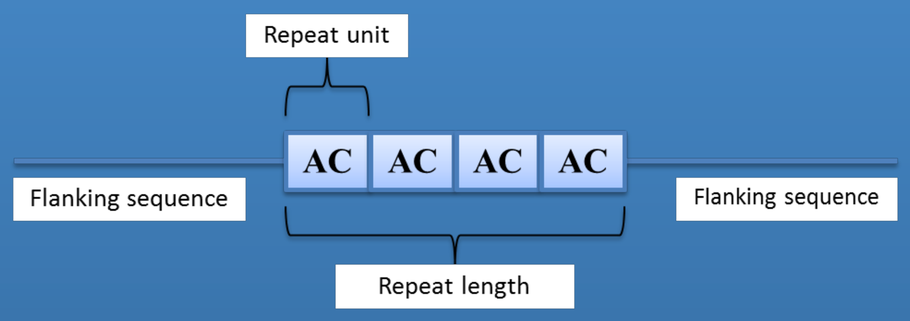
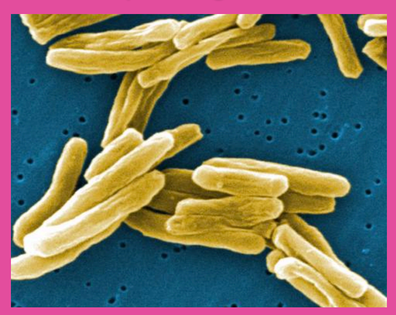


Figure 1: Microsatellites consist of a variable number of repeat units.

Mycobacterium tuberculosis

Mycobacterium tuberculosis is a bacterial pathogen, the causal agent of tuberculosis (a.k.a. consumption). There are currently ~12 million people affected worldwide and considerable concern about multiple drug resistant strains.



VNTR (variable numbers of tandem repeats) typing has been used in *M. tuberculosis* to track outbreaks and study evolution. Traditionally, 24 standard loci have been typed using PCR and gel electrophoresis. This method is expensive, time consuming and lacks resolution.

Data available:

- A closed reference genome (genbank: CP003248.2) which consists of 1 chromosome (4.41 Mb) and is approximately 66% GC
- An initial set of 160 diverse strains: 100bp PE Illumina genome sequences ~100X coverage. ~3000 strains in the full collection.

Finding Microsatellites in the *M. tuberculosis* Genome

We predicted the positions of microsatellite loci in the *M. tuberculosis* genome using Tandem Repeats Finder[1] (TRF) . TRF was run with low stringency for detection. These predictions were further filtered using the following criteria:

1. Repeat unit must be 1-6bp
2. At least 3 full copies of the repeat unit in tandem
3. At least 90% purity compared to a "perfect" microsatellite of that length (purity defined as 1 - hamming distance/sequence length)
4. No indels relative to the reference (apart from additional repeat units)

In some cases a TRF prediction did not match the criteria, but a substring of the sequence did. In these cases the longest substring to match the criteria was taken to be the microsatellite locus.

Microsatellite Loci

Of the 7056 microsatellites predicted by TRF, 6676 remained after the filtering process. The repeat motifs tend to be GC rich (Figure 2). Most of the microsatellites have 3bp repeat units (Figure 3A), consistent with in-frame expansions. All predicted microsatellites are less than 40bp in length, with most being less than 20bp (Figure 3B). This means that they are amenable to being genotyped using 100bp reads.

Figure 2: Word cloud of most common predicted repeat unit motifs.

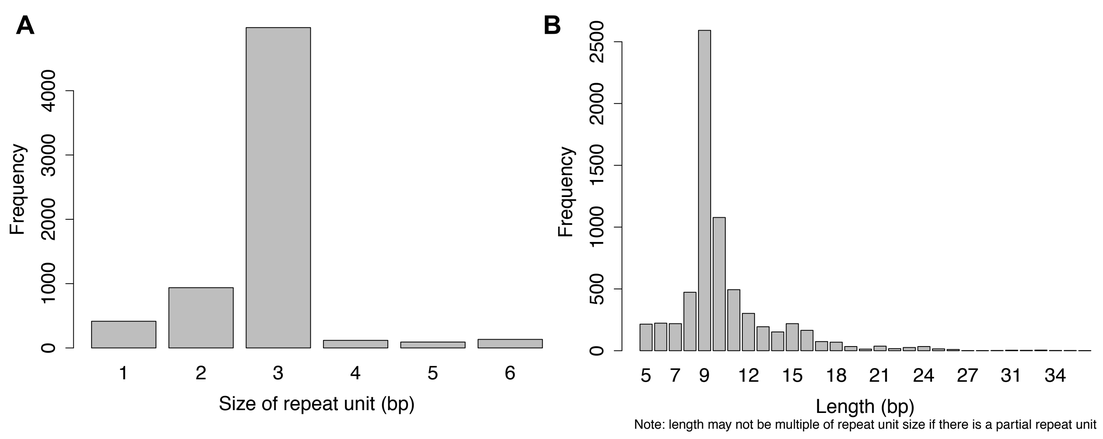


Figure 3: A Repeat unit lengths and **B** total lengths of predicted microsatellite sequences.

Genotyping Microsatellites in *M. tuberculosis*

Microsatellites were genotyped at all 6676 predicted loci in 160 strains.

- Reads were mapped to the reference genome using Bowtie2[2].
- For each microsatellite locus, all reads spanning that locus were identified.
- The number of repeat units was detected using a dynamic regular expression string matching algorithm.
- Since *M. tuberculosis* is haploid, the genotype was taken to be the most common repeat length at each locus.
- To reduce false positives, genotypes were only called for loci with at least five spanning reads.

Microsatellite variation

A microsatellite was said to be variable if the genotype differed between any two of the 160 strains. Of the 6676 loci genotyped, 1102 loci were observed to have variable microsatellites. Because genotypes are only called at loci covered by 5 or more reads, many loci had genotypes missing in some strains. To reduce the false positive rate, only loci with one or fewer missing genotypes were considered further. 453 loci matched this criterion.

Most strains had less than 20 loci that differed from the reference genome (Figure 4).

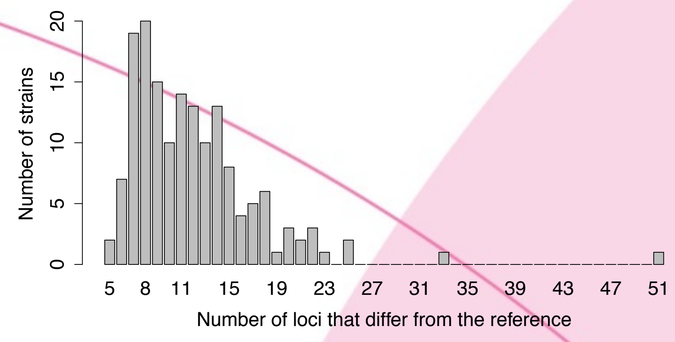


Figure 4: Number of loci per strain that vary from the reference

The variable loci were subdivided by repeat unit size. None of the 5bp loci were variable. Only four each of the 4bp and 6bp loci were variable. For the 1bp loci there is approximately equal variation across allele lengths. For the 2bp and 3bp loci (for which we have the most data points) we observe a trend towards more variability for longer loci (e.g. Figure 5). This is consistent with the observation that DNA polymerase is more likely to slip on longer microsatellites[3].

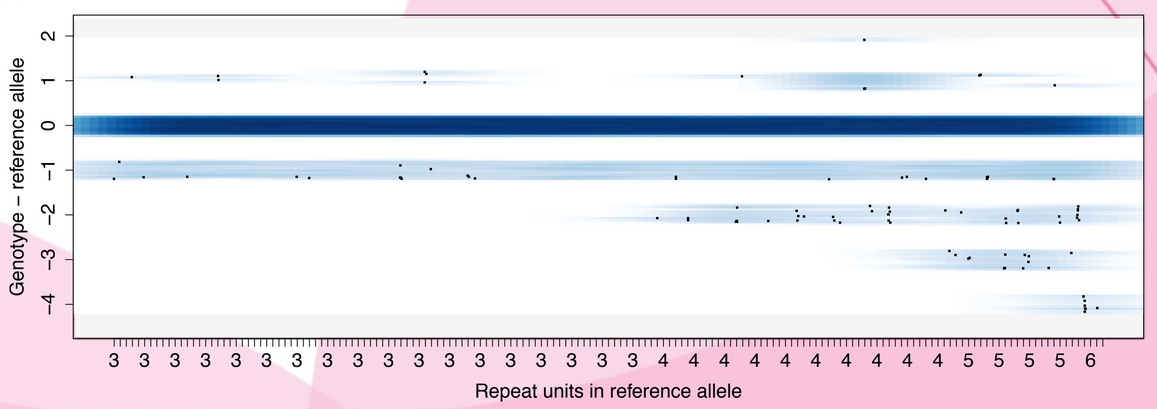


Figure 5: Variation relative to the reference allele in microsatellite loci with 3bp repeat units. Each tick on the x axis is a locus. The y axis shows genotype calls relative to the reference allele.

Future Directions

Although a number of steps have been performed to reduce the number of false positive microsatellite genotypes, there is room for improvement. One strategy is to perform local realignment of reads around the microsatellite loci to ensure indels are called consistently. Another is to develop a score for microsatellite genotypes, taking into consideration mapping qualities.

The analysis of more strains may reveal variation at additional loci and give us the power to detect trends in microsatellite variation.

References:

1. Benson, G. Tandem repeats finder: a program to analyse DNA sequences.. *Nucleic Acids Res* **27**, 573-80 (1999).
2. B Langmead, SL Salzberg. Fast gapped-read alignment with Bowtie 2.. *Nat Methods* **9**, 357-9 (2012).
3. Shinde, D. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Res.* **31,** 974–980 (2003).

Murdoch Childrens Research Institute — Healthier Kids. Healthier Future.

THE UNIVERSITY OF MELBOURNE

The Royal Children's Hospital Melbourne

vlsci LIFE SCIENCES COMPUTATION