ssahaSNP – A Polymorphism Detection Tool on a Whole Genome Scale

Zemin Ning[1], Mario Caccamo[1], James C. Mullikin[2]
[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus,
Hixton, Cambridge CB10 1SA UK
[2]Genome Technology Branch/NHGRI, NIH, 5625 Fishers Lane, Bethesda, MD 20852, USA
zn1@sanger.ac.uk

Abstract

We present a software package which can detect homozygous SNPs and indels on a eukaryotic genome scale from millions of shotgun reads. Matching seeds of a few kmer words are found to locate the position of the read on the genome. Full sequence alignment is performed to detect base variations. Quality values of both variation bases and neighbouring bases are checked to exclude possible sequence base errors. To analyze polymorphism level in the genome, we used the package to detect indels from 20 million WGS reads against the draft WGS assembly. From the dataset, we detected a total number of 663,660 indels, giving an estimated average indel density at about one indel every 2.48 kilobases. Distribution of indels length and variation of indel mapped times are also analyzed.

1. Introduction

Fast and accurate detection of genomic polymorphism is increasingly under demand for various applications, such as cancer related diseases, clinical diagnostics, and even on heterozygosity analysis in genome sequencing. In the past few years, a number of systems have been reported with different methods in mining genomic polymorphisms and variations, such as visual comparison of sequence traces in local BAC regions, reduced representation shotgun (RRS), and more recently whole genome alignment by placing a randomly shotgun read to the genome. However, the efforts coordinated by The SNP Consortium (TSC) were mainly focused on single nucleotide polymorphisms (SNP). A study on insertions/deletions (indels) was recently reported [1] and there is a need for systematic studies on structured polymorphisms. In this paper, we outline a package ssahaSNP which detects both SNPs and indels with a fast speed without any sequence repeat masking.

2. SSAHA2 and ssahaSNP

SSAHA2 is a package combined SSAHA [2] with phrap/cross_match developed by Phil Green at the University of Washington [3]. Matching seeds, a few exactly matched kmer words are detected from the database by the SSAHA algorithm. SSAHA achieves its fast search speed by converting sequences information into a "hash table" data structure, which can then be searched very rapidly for matches. When the location information of matching seeds is obtained, we then cut off both query and subject sequences and pass the two sequences to cross_match for full alignment. Extra sequences with a given edge length are used for both query and subject to extend the alignment length. In terms of software implementation, cross_match has been imbedded into the SSAHA system and alignment functions are used as libraries.

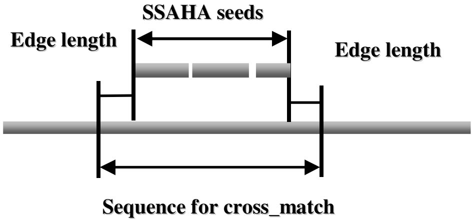

Figure 1. Extra sequences with a given edge length are used to extend the length of alignment.

As a fast tool capable of efficient processing of large data sets, ssahaSNP was used in the SNP detection by the international SNP consortium [4]. In the early version of ssahaSNP, there was an alignment module in the package. However, the alignment quality is not good enough in handling middle sized or long indels. Also SNP calling was carried out for every piece of alignment on the genome and it then relies on the parsing code to exclude those SNP candidates which are mapped multiple times. In the new version of ssahaSNP, we use ssaha2 as the alignment tool to place genomic reads on finished or draft assembly sequences. Highly repetitive elements are filtered out by ignoring those kmer words with high occurrence numbers. For those less repetitive or non-repetitive reads, we place them uniquely on the reference genome sequence and find the best alignment according to the pair-wise alignment score if there are multiple seeded regions.

From the best alignment, SNP candidates are screened, taking into account the quality value of the base with variation as well as the quality values in the neighbouring bases, using neighbourhood quality standard (NQS) [5]. There is no widely accepted quality standard for insertions/deletions. In ssahaSNP, we still use NQS for single base deletions to the reference sequence. For other indels with a length more than one base, we don't check quality values. To ensure the indels are detected with high confidence, a conservative method is adopted and we only report the cases in which exactly the same indel is mapped by two or more shotgun reads.

3. Results

A whole genome shotgun (WGS) assembly is generated as part of the zebrafish genome project at the Sanger Institute. In the initial phase of WGS reads production, the DNA samples were extracted from more than 1000 5 day old embryos (http://www.sanger.ac.uk/project/D_rerio). Multiple DNA sources lead to a very high polymorphism level in the data set and consequently leave tremendous technical challenges in sequence assembling. On the other hand, this polymorphic dataset also offers opportunities on genomic variation studies.

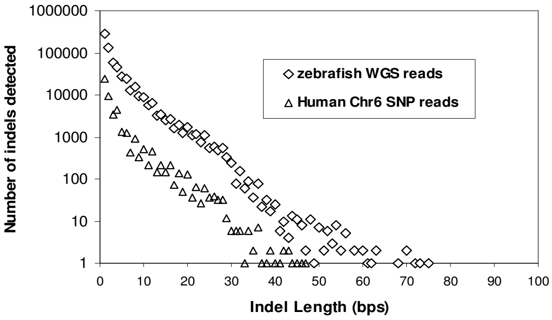

Figure 2. Distribution of the indel length.

We used the ssahaSNP package to detect indels from 20 million WGS reads against the draft WGS assembly as well as the finished clone contigs of 800 Mbps. The total number of detected indels is 663,660. Given the genome size of 1.65 GB, this indicates that the average indel density is at about one indel every 2.48 kilobases. Distribution of indel length over the number of detected indels is shown in Figure 1, where the data of Human Chromosome 6 is superimposed for comparison. In the human dataset, we processed 1.61 millions reads whose DNA samples were from three cell lines. With a total number of 47,692 detected indels, this gives the average indel density at about one indel per 3.58 kilobases. It is seen from the figure that for both datasets the majority of the indels are short, with a length N50 = 2 (half of the indels are less or equal to 2 bps).

In the zebrafish dataset, the shotgun coverage is about 7X and for human Chr6 the coverage is about 6X. Even with multiple haplotypes, it is likely the same indels would be mapped more than two times by the shotgun reads. Figure 2 shows variations of indel number against the times of indel mapped. For the zebrafish dataset, there is a long tail in the figure that extends beyond 100. This is because the WGS assembly is only a draft and many repetitive or long duplicated regions are still not represented in the assembly. For the reads belonging to these gapped regions, the correct location does not exist, thus the code maps reads to these regions to levels that are much higher than the average shotgun coverage. For the finished Chr6 sequence, the situation is much better, with only a very small percentage mapping more often than expected.

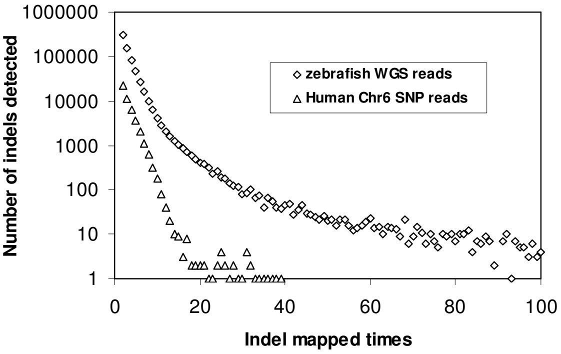

Figure 3. Histogram of the number with the times indels are mapped.

4. References

[1] Bennett, E.A. et al 2004. Natural Genetic Variation Caused by Transposable Elements in Humans. Genetics 168:933-951.
[2] Ning, Z., Cox, A.J. and Mullikin, J.C. 2001. SSAHA: A Fast Search Method for Large DNA Databases. Genome Research 11:1725-1729.
[3] http://www.phrap.com/
[4] The International SNP Map Working Group 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409(6822):928-933.
[5] Mullikin, J.C. et al 2000. An SNP map of human chromosome 22. Nature 407(6803):516-520.

Acknowledgement

The project is funded by the Wellcome Trust.