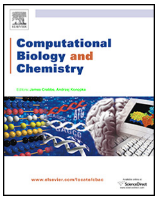
Computational Biology and Chemistry xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem



Research Article

Characterizing regions in the human genome unmappable by next-generation-sequencing at the read length of 1000 bases

Wentian Li*, Jan Freudenberg

The Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institute for Medical Research, North Shore LIJ Health System, 350 Community Drive, Manhasset, NY 11030, USA

ARTICLE INFO

Article history:
Accepted 11 July 2014
Available online xxx

ABSTRACT

Repetitive and redundant regions of a genome are particularly problematic for mapping sequencing reads. In the present paper, we compile a list of the unmappable regions in the human genome based on the following definition: hypothetical reads with length 1 kb which cannot be uniquely mapped with zero-mismatch alignment for the described regions, considering both the forward and reverse strand. The respective collection of unmappable regions covers 0.77% of the sequence of human autosomes and 8.25% of the sex chromosomes in the reference genome GRCh37/hg19 (overall 1.23%). Not surprisingly, our unmappable regions overlap greatly with segmental duplication, transposable elements, and structural variants. About 99.8% of bases in our unmappable regions are part of either segmental duplication or transposable elements and 98.3% overlap structural variant annotations. Notably, some of these regions overlap units with important biological functions, including 4% of protein-coding genes. In contrast, these regions have zero intersection with the ultraconserved elements, very low overlap with microRNAs, tRNAs, pseudogenes, CpG islands, tandem repeats, microsatellites, sensitive non-coding regions, and the mapping blacklist regions from the ENCODE project.

1. Introduction

High-throughput next-generation-sequencing (NGS or HTS) is an extension of shotgun sequencing in a highly parallel fashion. In NGS, the genome is fragmented into small pieces and portion or whole of the piece, called reads, are sequenced. The collection of all sequenced reads is assembled either by alignment to a known reference genome (*reference assembly*), or by connecting reads without any help from a reference genome (*de novo assembly*). Based on the early Sanger sequencing efforts of the human genome (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001) and subsequent work, a high quality reference sequence exists for the human genome. Therefore, NGS is almost always carried out by a reference assembly.

If two regions of the human genome with size L are identical (considering both the direct and the reverse complement strands), which could be caused either by sequence duplications or insertion of specific sequences (e.g. transposable elements), a read with length $k \leq L$ may not have a unique alignment destination. Such a read can then be considered unmappable (can also be called non-unique, redundant, repetitive, or ambiguous). If a read with length k can be uniquely mapped to the reference genome, then it can be considered mappable. This issue of the mappability of a genome at read length k is related to the description of the reference genome concerning regional uniqueness at the size of k.

As there have been several publications on the topic of mappability by NGS reads (Rozowsky et al., 2009; Cahill et al., 2010; Wang et al., 2010; Taub et al., 2010; Chung et al., 2011; Koehler et al., 2011; Lee and Schatz, 2012; Treangen and Salzberg, 2012; Derrien et al., 2012; Storvall et al., 2013; Chu et al., 2013; Ramachandran et al., 2013; Cabanski et al., 2013; Aldwairi et al., 2013; Dao et al., 2013), it is necessary to clarify the scope and definition of the term "mappability" used in this paper. We do not attempt to discuss the measure of mappability, which may strongly depend on the technology used and various factors. For example, all regions in the human genome would be unmappable if the reads length is too short; some regions are unmappable by current-standard NGS, but mappable with longer Sanger sequencing reads (Huddleston et al., 2014); and all regions can be mappable, if the genome is sequenced by reading out base-by-base from a single molecule in real time (SMRT sequencing) (Braslavsky et al., 2003; Astier et al., 2006; Eid et al., 2009; Clarke et al., 2009; Schadt et al., 2010; Loomis et al., 2013; Koren et al., 2013).

* Corresponding author.
E-mail address: wtli2012@gmail.com (W. Li).

http://dx.doi.org/10.1016/j.compbiolchem.2014.08.015
1476-9271/© 2014 Elsevier Ltd. All rights reserved.

The term mappability as it is used in the current paper refers to NGS short-read based mappability (SMRT sequencing is not discussed here); it refers to mappability of a read to a reference genome (not assemblability in the absence of a reference genome (Kinsford et al., 2010; Bradnam et al., 2013)); we ignore the failure to map a read due to the gaps or unsequenced regions in the reference genome (the N's in the reference genome) (Genovese et al., 2013); we refer to a specific length scale (e.g. read length $k = 1000$) as read length will dramatically change the nature of mappability; and it is assumed that reads are aligned to a genomic region only if there is an exact match (zero-mismatch alignment). The latter assumption is to avoid the computational difficulties in aligning the whole genome to itself with mismatches (Derrien et al., 2012).

With these sets of constraints plus the fact that we are not examining the effect of various practical issues in NGS such as genetic variants (both SNP and structural variants) in an individual genome and known correlation between paired-end reads, we can identify the unmappable regions in the genome by studying the reference genome itself. The limitation of zero-mismatch will introduce some bias in our results when the non-zero-mismatch is unavoidable: such as the situations with polymorphism and sequencing error. However, we would expect that such a bias is negligible. We only cover the unmappability on the sequence analysis aspect of the human reference genome, not unmappability due to technical difficulties in the sequencing experiment.

Previously, we have identified all non-unique k-mers up to $k = 1000$ in the GRCh37/hg19 release of the human reference genome (Li et al., 2014). We have observed that (1) percentage of non-unique k-mer counts ("tokens") decreases with read length k as a piece-wise power-law function, which is slower than an exponential or a linear decrease; (2) at a fixed k, the histogram of k-mer "types" as a function of the number of times the k-mer appearing in the genome ("copy number", "frequency" f) is mostly a negative-exponent power-law function, with the scaling exponent larger (sharper drop) for larger k's; (3) the worst-case non-unique 1000-mers (with copy number $f \geq 10$) are located in regions with either segmental duplication or LINE transposable elements (Li et al., 2014).

This paper extends the result concerning locations of non-unique or unmappable k-mers: we align all non-unique 1000-mers ($f \geq 2$) back to the reference genome. The number of these tokens exceeds 11 millions, much larger than those with $f \geq 10$ (~6000). Aligning 11 million 1000-mer tokens to the reference genome required extensive computations, which could not be done in our previous paper (Li et al., 2014). We also carry out a comprehensive intersection analysis between these unmappable regions with other genomic annotations.

Unmappable regions are the difficult-to-map portion of the genome by NGS short reads, and this difficulty is intrinsically related to the concept of complexity (Campbell, 1988; Li, 1991, 1997). On the other hand, the redundancy may lead to lower entropy, biased base/oligomer/long-k-mer composition, all features of a low level of complexity (Wootton and Federhen, 1993; Pizzi and Frontali, 2001). This implies that high-complexity in terms of read-mapping may correspond to low-complexity in terms of sequence feature.

Our interest in unmappable regions in NGS is far from a theoretical curiosity. Information on unmappable regions in the human genome is important in reliably mapping genetic variants and their association with the human diseases (Sudmant et al., 2010). The 1000 genomes project (The 1000 Genomes Project Consortium, 2012), for example, only analyze the "accessible genome" portion (94% of the whole genome) where the short reads can be reliably aligned. It is clearly desirable to know what is being missed when the unmappable regions are not studied.

2. Methods and data

Sequence of the human reference genome: We use the GRCh37/hg19 release of the human reference genome, available from the UCSC Genome Browser (http://genome.ucsc.edu/). The GRCh38 was released by Genome Reference Consortium (consists of Sanger Institute, Genome Institute at Washington University, EMBL, and NCBI) in December 2013, but is only sparsely annotated by UCSC. The availability of the newest assembly GRCh38 data was also too recent for our then ongoing analysis. The unsequenced bases (N's) are used to partition the chromosomes into subsequences, and 1000-mers across two different subsequences are not allowed. The coordinates of two pseudoautosomal regions (PARs) on X and Y chromosomes are taken from the description file to be: chrY:10001-2649520 and chrY: 59034050-59363566, and chrX:60001-2699520 and chrX: 154931044-155260560.

Counting k-mers – the DSK program: Although counting k-mers with small k values can be easily done by a Perl script, for $k = 1000$, the high memory requirement for storing all possible k-mer types on a regular sized computer presents a serious obstacle. We use a public domain program *DSK* (http://minia.genouest.org/dsk/ version 1.5031 from March 26, 2013) which trades memory with hard disk space (Rizk et al., 2013), thus counting 1000-mer in human genome is possible if the computer has enough hard disk space. The counting of 1000-mers is done by moving a 1000-base window, one position at the time. Thus we are counting overlapping 1000-mers.

Aligning 1000-mers to the human reference genome – the BLAT program: We use the BLAT program from UCSC (Kent, 2002) to align the redundant 1000-mers to the reference genome. Roughly more than ten 1000-mer types can be aligned per minute per core on our Linux computer. Though not used in this paper, we notice the existence of some alternative alignment programs (SSAHA2 (Ning et al., 2001), BWA (Li and Durbin, 2010), mrsFAST (Hach et al., 2010)).

Genomic annotation: The location of genes is obtained from the UCSC Genome Browser, including: RefSeq genes (http://www.ncbi.nlm.nih.gov/refseq/) (Pruitt et al., 2014), GENCODE genes v17 (http://www.gencodegenes.org/) (Harrow et al., 2012), and Vega genes (http://vega.sanger.ac.uk/) (Harrow et al., 2014). Other gene-related features include Yale pseudo60 (http://www.pseudogene.org/) (Zheng et al., 2007), sno/miRNA (Lestrade and Weber, 2006; Kozomara and Griffiths-Jones, 2014), tRNA, and CpG islands (Cohen et al., 2011).

Two lists with regions problematic for read-mapping, both produced by the ENCODE project (https://genome.ucsc.edu/ENCODE/) (The ENCODE Project Consortium, 2004), are examined. The Duke excluded regions are identified as those regions where mapped sequence tags are filtered out, thus are problematic for sequence tag detection. The ENCODE DAC (Data Analysis Consortium) Blacklisted Regions are manually curated for those regions that are problematic in sequence alignment (Kundaje, A comprehensive collection of signal artifact blacklist regions in the human genome, http://www.broadinstitute.org/~anshul/projects/encode/rawdata/blacklists/hg19-blacklist-README.pdf).

The segmental duplication region is retrieved from the UCSC Genome Browser genomicSuperDups track (Bailey et al., 2001, 2002). The transposable element location information is obtained from the Dfam database (http://dfam.janelia.org/) (Wheeler et al., 2013). The tandem repeats are obtained from Dfam in which the TRF program (http://tandem.bu.edu/trf/trf.html) (Benson, 1999) is run. The microsatellites from the Genome Browser were also downloaded. Though both are derived from the TRF program, the Genome Browser version of the microsatellites are not identical to the Dfam version.