Single nucleotide variation context in human genome

Vera Enes, João Manuel Rodrigues and Vera Afreixo

April 6, 2018

Single nucleotide variation context in human genome

Vera Enes
vera.enes@ua.pt
Department of Mathematics, University of Aveiro.

João M.O.S. Rodrigues
jmr@ua.pt
IEETA-Institute of Electronic Engineering and Informatics of Aveiro
Department of Electronics, Telecommunications and Informatics, University of Aveiro.

Vera Afreixo
vera@ua.pt
iBiMED-Institute of Biomedicine
IEETA-Institute of Electronic Engineering and Informatics of Aveiro
Department of Mathematics, University of Aveiro.

Abstract

We use the data made available by the 1000 Genomes Project to investigate variation context in the human genome population. We observe that word frequencies in the vicinity of single nucleotide variation (SNV) sites are associated with the type of variation.

1 Introduction

Over a decade ago, the initial sequencing and analysis of what is still considered the reference human genome revealed that less than 2% of the sequence codes for proteins [5]. However, recent evidence suggests that at least 80% of the genome is transcribed or, at least, biochemically active at some point [4]. This evidence highlights the need to understand the biological function of this vast region of the genome, as well as, the evolutionary constraints acting over it. It also highlights the inadequacy of investigating evolutionary constraints by solely considering mammalian conservation criteria, and the need to develop new methodologies to investigate these constraints within a species [13].

The 1000 Genomes Project was a pioneering effort in population sequencing [7]. Their sequence variants with respect to the GRCh37 reference human genome assembly, including single nucleotide variation (SNV), small insertions and deletions (indels), and larger structural variants, are available in the variant call format (VCF, [1, 2, 6]). The single nucleotide variation (SNV) is a genetic variation in a single nucleotide that occurs at a specific position in the genome. There are 6 types of SNVs, which can further be classified as transitions, $C \leftrightarrow T$ and $A \leftrightarrow G$, or as transversions, $A \leftrightarrow C$, $G \leftrightarrow T$, $A \leftrightarrow T$ and $C \leftrightarrow G$.

The genome variations could be a random phenomenon or could be an evolutive/adpative phenomenon. There are some natural questions to ask: Is the variation occurrence position independent from the sequence neighborhood context? Did the neighborhood context presents specific motifs to each type of variation? Are the motifs position in long or short range from the variation position?

Previous work points to a non-random nature of variation occurrences. Variations sites were studied in the mouse genome, and it was concluded that there is a nucleotide bias in the neighborhood of the variation positions, and the association effect decreases with distance from the variation position [11, 14]. In [10], the neighborhoods of 15,110 single nucleotide variations in the bovine genome were analysed, and the authors verified an association between C_pG content and some types of variation. Using the 1000 Genomes Project data, [3] discusses the association between the oligonucleotide neighbourhood variation context (emphasising the heptanucleotide context) and the single nucleotide variations in the human genome.

Here, with the data made available by the 1000 Genomes Project, we present an approach based on the frequency of words in the neighborhood of each annotated variation, to identify new context variation patterns.

2 Material and Methods

We used the GRCh37 reference human genome assembly [9], and version 3 (March 16, 2012) of the Phase 1 integrated variant call set, based on both low coverage and exome whole genome sequencing data from 1,092 individuals [8]. For this study, only the 22 human autosome pairs were considered.

VCF files contain a header followed by variant call records, one per line. Figure 1 shows an extract of the chromosome 1 VCF file from the 1000 Genomes data. (For the sake of legibility, some details were ommited.)

```
##fileformat=VCFv4.1
...
##reference=GRCh37
#CHROM POS    ID  REF ALT  QUAL FILTER INFO FORMAT   HG00096     HG00097
1      10583 ... G   A    100  PASS   ...  GT:DS:GL 0|0:...:... 0|0:...:...
1      10611 ... C   G    100  PASS   ...  GT:DS:GL 0|0:...:... 0|1:...:...
...
1      46402 ... C   CTGT 31   PASS   ...  GT:DS:GL 0|0:...:... 0|0:...:...
```

Figure 1: Excerpt of the chromossome 1 VCF file from the 1000 Genomes phase 1 data. Ommited fields and lines were replaced by ellipsis.

Each record contains several fields of information for a single variation site. The CHROM and POS fields identify the site of variation relative to the reference genome (GRCh37, in this case). The kind of variation is encoded in the REF and ALT fields, which specify the reference allele and alternative allele observed in individual samples. The FORMAT field specifies the encoding used for the remaining fields, each of which contains annotations on a specific individual sample. For example, on the first record, both the HG00096 and the HG00097 samples have `0|0` on the genotype (`GT`) subfield. This means that at this site (position 10583 on chr1), both of these individuals are homozygous with the reference allele, that is, both are $G|G$. On the second record (for position 10611 on chr1), we see that HG00097 is heterozygous `0|1`, meaning it has genotype $C|G$. Other individuals on the same record may be heterozygous `1|0`, meaning $G|C$, homozygous with the reference allele `0|0`, meaning $C|C$, or homozygous with the alternative allele `1|1`, meaning $G|G$. Other fields and subfields in the records include further information, such as the quality or confidence level of the variant calls, but this was not used in this work.

We wrote a short C program, optimized for the specific VCF format used in the 1000 Genomes data, to preprocess the VCF files. This preprocessing consisted of: (1) discarding unwanted fields; (2) selecting only SNV records (rejecting indels and structural variants); and (3) counting samples with each of the `GT` types, `0|0`, `0|1`, `1|0` or `1|1`. This produced much smaller files, with just a few columns, as shown in Figure 2.

```
#CHROM POS    ID  REF ALT C0|0: C0|1: C1|0: C1|1:
1      10583 ... G   A   783   304   0     5
1      10611 ... C   G   1051  37    4     0
1      13302 ... C   T   849   192   45    6
```

Figure 2: Excerpt of the output of the preprocessing stage for the chromossome 1 VCF file. The last four columns show the number of individual samples of each genotype.

The preprocessed output files were then imported into the R software environment [12] for further data normalization and statistical processing. Normalization involved classifying the (REF, ALT) pair into one of the six SNV types, merging the heterozygous counts, and possibly swapping the homozygous counts of each record. The records were then grouped according to the SNV type. Then, the DNA segments in the immediate vicinity to the left and to the right of each SNV site were recovered from the reference genome. An example of the result is shown in Table 1.

Finally, we selected the words of length k located d nucleotides to the right and to the left of each SNV site (see example in Table 1), and produced contingency tables across all of the genome, for each SNV type. This was repeated for words of length $k = 1, 2, 3$, and displacements $d = \pm 1, \pm 2, \ldots, \pm 50$.

CHR	POS	SNV	$A\|A$	$A\|G$	$G\|G$	Left flank	Right flank
1	10583	$A \leftrightarrow G$	5	304	783	*CCCTCGCGGT*	*CT***CT***CCGGGT*
1	54421	$A \leftrightarrow G$	881	202	9	*TAATTGCTTT*	*TC***AC***TCATAT*
1	54490	$A \leftrightarrow G$	13	149	930	*ATACTCTACC*	*GG***CT***TCTGGA*
1	55330	$A \leftrightarrow G$	0	1	1091	*TACTATTTAC*	*CT***TC***AGTAAA*

Table 1: Variation data after preprocessing and normalization. Records are shown only for SNVs of type $A \leftrightarrow G$. The last columns show the left- and right-flanking sequences around the SNV site. Words of length $k = 2$ located $d = +2$ nucleotides to the right of the SNV are highlighted.

The patterns of relative word frequency around the SNV sites are described through the differences to the corresponding average word frequencies in the full data. The association between SNV sites and the type of variation are evaluated with the standard statistical tools: chi-square test and ϕ measure.

3 Results and Discussion

Figure 3 shows nucleotide ($k = 1$) frequency patterns around the SNV sites, grouped by transversions and transitions. The genome sites under study with highest association effect with nucleotide frequencies are in the imediate vicinity ($\pm1, \ldots, \pm4$) around the SNV sites. The bias around transitions is much larger than around transversions. Around transversion sites, complementary nucleotides display symmetrical frequency patterns. This is not visible around transition sites.

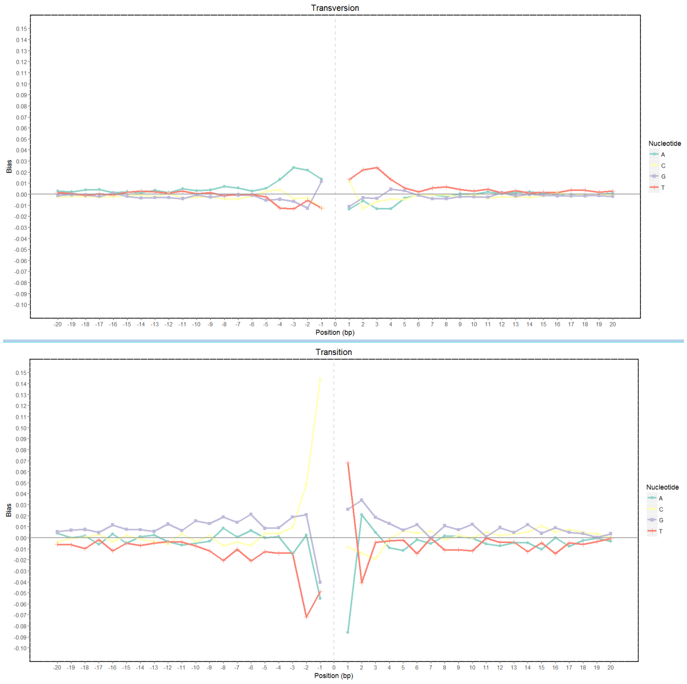

Figure 3: Single nucleotide frequency patterns around transversions (top), and around transitions (bottom).

Each SNV type has specific occurrence context. For example, at the +1-site to the right of a $G \leftrightarrow T$ variation, GT is the most favored dinucleotide, and distinct patterns are observed on the left and right flanks (see Fig. 4, top). Around $C \leftrightarrow G$ transversions, the frequency patterns of reverse-complementary dinucleotides seem to be symmetrical (see Fig. 4, bottom).

Funding

This work was supported by Portuguese funds through the iBiMED - Institute of Biomedicine, IEETA - Institute of Electronics and Informatics Engineering of Aveiro and the Portuguese Foundation for Science and Technology ("FCT–Fundaçao para a Ciência e a Tecnologia"), within projects: UID/BIM/04501/2013 and PEst-OE/EEI/UI0127/2014.

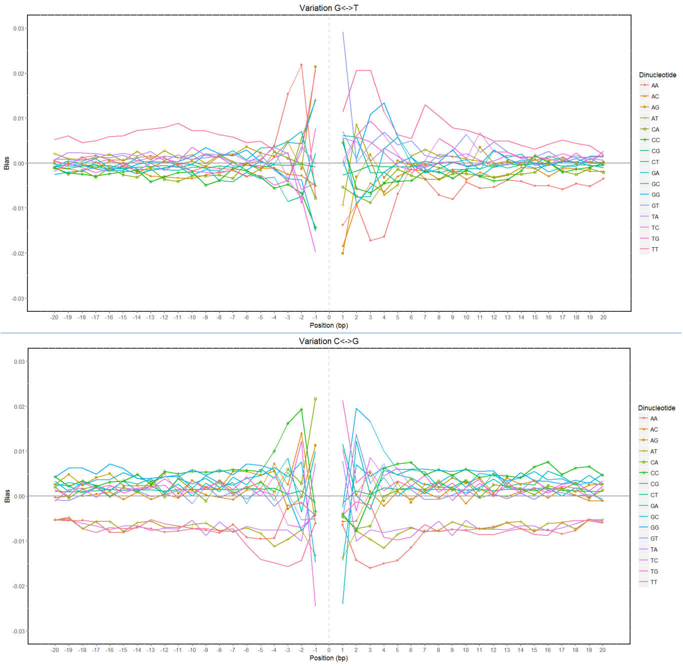

Figure 4: Dinucleotide frequency patterns around $G \leftrightarrow T$ variations (top), and around $C \leftrightarrow G$ variations (bottom).

References

[1] Consortium Project 1000Genomes. An integrated map of genetic variation from 1092 human genomes. *Nature*, 491:56–65, 2012.

[2] The 1000 Genomes Project Consortium (2010). A map of human genome variation from populationscale sequencing. *Nature*, 467: 1061–1073, 2010.

[3] Varun Aggarwala and Benjamin F Voight. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics*, 48, 2016.

[4] The ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57–74, 2012.

[5] The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[6] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011. doi: 10.1093/bioinformatics/btr330.

[7] The 1000 genomes project. 2016. `http://www.1000genomes.org`.

[8] The 1000 genomes project data release: Integrated variant call set for phase 1 version 3. 2016. `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521`.

[9] GRCh37 Reference human genome assembly. 2016. `ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.37.3/`.

[10] Zhihua Jiang, Wu Xiao-Lin, Ming Zhang, and et al. The complementary neighborhood patterns and methylation-to-mutation likelihood structures of 15.110 single-nucleotide polymorphisms in the bovine genome. *Genetics*, 180(1):639–647, 2008.

[11] Zackery E. Plyler, Aubrey E. Hill, Christopher W. McAtee, and et al. SNP formation bias in the murine genome provides evidence for parallel evolution. *Genome Biology and Evolution*, 7(9):2506–2519, 2015.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

[13] L. Ward and M. Kellis. Evidence of abundant purifying selection

in humans for recently acquired regulatory functions. *Science*, 337: 1675–1678, 2012.

[14] F. Zhang and Z. Zhao. The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs. *Genomics*, 84(5):785–795, 2004.