A periodic pattern of SNPs in the human genome

Bo Eskerod Madsen (1), Palle Villesen (1), Carsten Wiuf (1,2)

1. Bioinformatics Research Center (BiRC) University of Aarhus, Hoegh-Guldbergs Gade 10, Building 1090, DK-8000 Aarhus C
2. Molecular Diagnostic Laboratory, Aarhus University Hospital, Brendstrupgaardsvej 90, DK-8200 Aarhus N

Nature Precedings : doi:10.1038/npre.2007.435.1 : Posted 12 Jul 2007

Introduction

The most frequent type of variation in the Human genome is single nucleotide polymorphisms (SNPs). Most SNPs are expected to be the result of independent single mutations, and they are therefore expected to be randomly distributed throughout the genome.
In contrast to this, we report that the distance between SNPs inside *periodic* DNA follow a periodic pattern (see Figure 1 for a definition of SNP distance and periodic DNA).

Figure 2A. The periodic pattern is caused by pairs of identical SNPs. It is seen that pairs of identical SNPs follow a 1, 2, 4, 6, 8 pattern, whereas pairs of different SNPs are almost uniformly distributed for d>1.

Figure 2B. The pattern is virtually captured by periodic DNA. The pattern is strongest in periodic DNA and nearly disappears outside periodic DNA. Also, there is a strong general overrepresentation of SNP pairs in periodic DNA when compared to the rest of the genome.

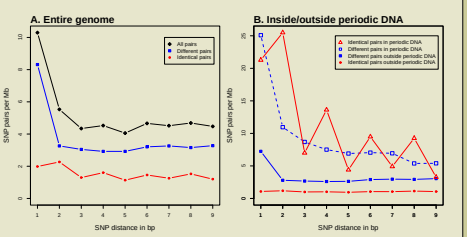
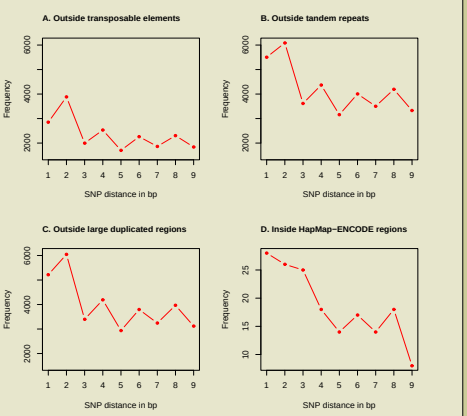


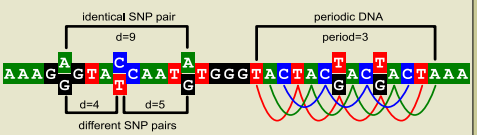
Figure 3. The periodic pattern is not restricted to known repetitive elements. Identical SNP pairs follow the periodic distribution. The pattern is found outside various well described repetitive elements like transposable elements, tandem repeats and large duplicated regions. Also the patterns seems to be present in the ENCODE regions (but note the low amount of data).



Main results

- By surveying all validated SNPs in the human genome we have found that SNPs positioned 1, 2, 4, 6 or 8 bp apart are more frequent than SNPs 3, 5, 7 or 9 bp apart (Figure 2A)
- The periodic pattern is caused by pairs of identical SNPs (Figure 2A)
- The periodic pattern is not restricted to known repetitive elements (Figure 3)
- The pattern is virtually captured by periodic DNA (Figure 2B).
- The pattern in periodic DNA is abundant throughout the genome, except in exons (Figure 4).
- Periodic DNA have 1.8 times higher SNP density than the rest of the genome.
- In the HapMap populations, genotyping failed in 21.4% of the cases for SNPs inside periodic DNA, but only in 12.3% for SNPs outside periodic DNA (P-value<10^-13).

Figure 1. SNP pairs and periodic DNA. Paris of SNPs can be divided into identical pairs (same alleles) and different pairs (different alleles).



Materials & Methods

SNP data
To avoid false patterns due to study specific biases, we used SNPs from all projects that reported to dbSNP build 125, and SNPs from HapMap phases I+II. We only selected unambiguously mapped SNPs, where the flanking sequences surrounding a SNP had exactly one hit to the human genome. To avoid SNPs with potential alignment problems on the local scale (<10bp, e.g. due to indels), we only selected SNPs that were perfectly mapped at on local scale, i.e. where the alignment of the flanking sequences and the reference genome were exactly 1 bp apart.

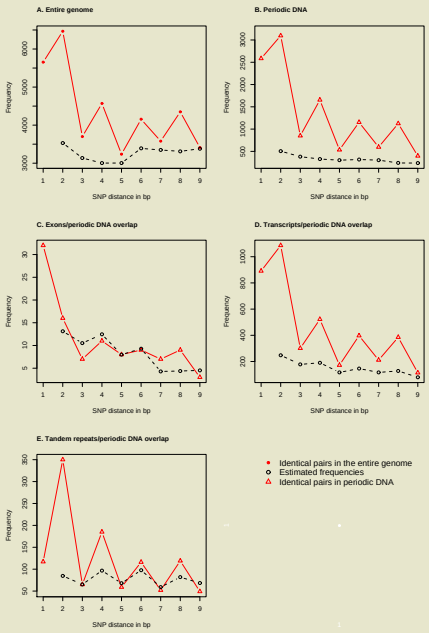
Periodic DNA
A sequence is defined as periodic DNA, with period p, if it fulfils the following three criteria:
1. The minimum length is 9 bp.
2. The pattern (e.g. AT in ATATATATAT) is repeated at least three times.
3. There are at most $p/4$ bp that do not match a periodic pattern of period p in the sequence.

Genomic elements
All information on genomic elements were downloaded from the UCSC Table Browser.

Conclusions

- There is an excess of SNPs in periodic DNA compared to non-periodic DNA
- SNPs in periodic DNA are distributed according to a 2, 4, 6, 8 pattern
- Care should be taken in analysis of SNPs from periodic DNA since SNPs in periodic DNA have a higher genotyping error rate than SNPs outside periodic DNA. The latter may have important consequences for SNP and association studies.
- Not all SNPs in the human genome are created by independent single nucleotide mutations

Figure 4. The pattern in periodic DNA is abundant throughout the genome, except in exons. (A) The entire genome (B) Periodic DNA (4.3% of the entire genome). (C) The overlap between exons and periodic DNA (0.06% of the entire genome). (D) The overlap between transcripts and periodic DNA (1.56% of the entire genome). (E) The overlap between tandem repeats and periodic DNA (0.12% of the entire genome).



For Further Information

A paper on the results is just accepted for publication in Genome Research, and will be published in September.
Please contact Bo Eskerod Madsen at e-mail: eskerod@birc.au.dk for further information.