Context analysis of SNP containing sites in mammalian genomes

Nataly S. Safronova

Novosibirsk State University, Novosibirsk, Russia, taschasafronova@mail.ru

Irina Abnizova

Welcome Trust Sanger Centre, Cambridge, United Kingdom ia1@sanger.ac.uk

Rene te Boekhorst

School of Computer Science, University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB United Kingdom R.TeBoekhorst@herts.ac.uk

Vladimir N. Babenko, Irina V. Medvedeva, Yuriy L. Orlov

The Institute of Cytology and Genetics, Novosibirsk, Russia {bob,brukaro,orlov}@bionet.nsc.ru

Due to technological breakthrough in the next-generation sequencing (NGS) technologies volume of available genomic data exponentially grows each year, including information on natural variability in mammalian genomes. Fast growth of NGS data demands development of new methods of analysis, processing and pre-processing of raw sequence reads. Applications of sequencing analysis include raw data (sequencing reads) in transcripts and personal genome studies, detection of individual genetic polymorphisms ("1000 genomes" project). The study of genomic context of single nucleotide polymorphisms (SNPs) represented in the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/) is of greater interest. Association of SNP position in human genome with mononucleotide repeats was shown earlier. We updated context analysis of SNP containing sites using novel software based on text complexity estimates and extended applications of SNP analysis from human "1000 genomes" data to rat and mouse genomes.

Competition of sequencing technologies, such as Illumina Solexa, SOLiD, PacBio leads to problem of data formats incompatibility, and more important, to non-reliable conclusion based on wrong DNA reads mapping to reference genome. Pre-processing and data filtration is important for effective detection of single nucleotide polymorphisms, detection of transcription factor binding sites. It was earlier shown that low complexity sequence regions, poly-tracks and simple repeats are related to systematic errors in genome sequence reads mapping and interpretation of sequencing results. The sequence complexity measures could be roughly dived to entropy estimates, linguistic complexity and algorithmic estimates including Lempel-Ziv compression method (Orlov et al., 2006). The measures of data quality are especially important for variant calling: in the particular case of SNP calling, a great number of false-positive SNPs may be obtained. We found earlier that not only the probability of sequencing errors (i.e. the quality value) is important to distinguish an FP-SNP but also the conditional probability of "correcting" this error (the "second best call" probability, conditional on that of the first call) (Abnizova et al., 2012). Surprisingly, around 80% of mismatches can be "corrected" with this second call. We have developed several measures to distinguish between sequence errors and candidate SNPs, based on a base call's nucleotide context and its mismatch type. The project aim is to realize and apply algorithms of DNA sequence complexity estimates to analysis of sequencing in human genome and in model organisms.

We studied context dependencies in broader scale in human, mouse and rat genomes using several complexity measures. Nucleotide text complexity is important mathematical features to explore fully the contextual dependencies in the sequences, is the complexity of the text (Orlov et al., 2006). A wide range of complexity measures estimates different features of the nucleotide text: linguistic complexity relates to oligonucleotide vocabularies, complexity estimation by Lempel-Ziv compression relates to structure of repeats in the text, Shannon entropy counts variation of nucleotides. These algorithms which were previously used in “Complexity” software developed at the Institute of Cytology and Genetics in Novosibirsk (http://wwwmgs.bionet.nsc.ru/mgs/programs/lzcomposer/) have been re-implemented in a computer program with supplements weight complexity measures and measures the rotation of the monomers.

We analyzed the nucleotide sequences containing SNPs in the human, mouse and rat genomes by in-house program (in C++) calculating the averaged text complexity profiles. We analyzed more than 2.7 million SNP containing sites (+/-20 nt) in the human genome presented at the UCSC Genome Browser tables and in the “1000 genomes” project (http://www.1000genomes.org/data). The presence of low complexity sites in the flanking regions around SNPs in the human genome was statistically shown. The same effect was confirmed for sample of SNPs in mouse and in rat genomes. Effect of mononucleotide repeats adjacent to a SNP position (Siddle et al., 2011) was confirmed on new data including model mammalian genomes. An association between the distribution of SNPs and the presence of microsatellites is supported by several lines of evidence. Microsatellites exhibit high levels of length polymorphism such that heterozygous individuals can be viewed as carrying microdeletions, potentially enhancing the local mutation rate. Stresses associated with the unusual base-stacking of purine–pyrimidine repeats or other structural properties could also be responsible for the mutational biases observed in regions flanking microsatellites.

We’d like discuss problems arising for SNP detection by modern NGS technologies (Abnizova et al., 2012). It was shown earlier, that short mononucleotide repeats can cause errors in DNA reads during sequencing, at least for previous Illumina sequencing technologies. Database curation and validation based on population studies should protect SNP detection from possible bias. As we found, changes in complexity profiles are related to basic properties of DNA texts.

Overall, low complexity profiles keep more information extending just measures of mononucleotide patches. The irregularities of mutation hot-spots in genome have been shown earlier on a limited data. The molecular mechanism of the observed effect of lowering the text complexity on flanks of SNP genome position can be explained by the increased frequency of double-helix DNA breaks in flanking positions.

The research has been supported by ICG SB RAS budget project VI.61.1.2 and RFBR 14-04-01906, 15-54-10060.

1. I. Abnizova et al. (2012) Analysis of context-dependent errors for illumina sequencing, *J Bioinform Comput Biol*., **10**(2):1241005.
2. Y.L. Orlov, R. Te Boekhorst, I.I. Abnizova (2006) Statistical measures of the structure of genomic sequences: entropy, complexity, and position information, *J Bioinform Comput Biol*., **4**, 523-36.
3. K. J. Siddle et al. (2011) Bases adjacent to mononucleotide repeats show an increased single nucleotide polymorphism frequency in the human genome, *Bioinformatics,* **27**(7), 895-8.