Genomics Proteomics Bioinformatics 11 (2013) 77–85

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Interpretation, Stratification and Evidence for Sequence Variants Affecting mRNA Splicing in Complete Human Genome Sequences

Ben C. Shirley [1], Eliseos J. Mucaki [2], Tyson Whitehead [3], Paul I. Costea [4], Pelin Akan [4], Peter K. Rogan [1,2,5,*]

[1] *Department of Computer Science, Middlesex College, The University of Western Ontario, London, ON N6A 5B7, Canada*
[2] *Department of Biochemistry, Schulich School of Medicine and Dentistry, The University of Western Ontario, London, ON N6A 5C1, Canada*
[3] *SHARCNET, London, ON N6A 5B7, Canada*
[4] *Royal Institute of Technology, Science for Life Laboratory, Solna 17165, Sweden*
[5] *Cytognomix Inc., London, ON N6G 4X8, Canada*

Received 6 December 2012; revised 16 January 2013; accepted 21 January 2013
Available online 14 March 2013

KEYWORDS

Mutation;
mRNA splicing;
Information theory;
Next-generation sequencing;
Genome interpretation

Abstract Information theory-based methods have been shown to be sensitive and specific for predicting and quantifying the effects of non-coding mutations in Mendelian diseases. We present the Shannon pipeline software for genome-scale mutation analysis and provide evidence that the software predicts variants affecting mRNA splicing. Individual information contents (in bits) of reference and variant splice sites are compared and significant differences are annotated and prioritized. The software has been implemented for CLC-Bio Genomics platform. Annotation indicates the context of novel mutations as well as common and rare SNPs with splicing effects. Potential natural and cryptic mRNA splicing variants are identified, and null mutations are distinguished from leaky mutations. Mutations and rare SNPs were predicted in genomes of three cancer cell lines (U2OS, U251 and A431), which were supported by expression analyses. After filtering, tractable numbers of potentially deleterious variants are predicted by the software, suitable for further laboratory investigation. In these cell lines, novel functional variants comprised 6–17 inactivating mutations, 1–5 leaky mutations and 6–13 cryptic splicing mutations. Predicted effects were validated by RNA-seq analysis of the three aforementioned cancer cell lines, and expression microarray analysis of SNPs in HapMap cell lines.

* Corresponding author.
E-mail: progan@uwo.ca (Rogan PK).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

Production and hosting by Elsevier

Introduction

The volume of human next-generation sequencing (NGS) data requiring bioinformatic analysis has necessitated development of high-performance software for genome scale assembly and analysis [1]. Genomic variations found in these analyses,

http://dx.doi.org/10.1016/j.gpb.2013.01.008

particularly single nucleotide polymorphisms (SNPs), have traditionally been interpreted in terms of amino acid modifications in coding regions. Clinically-significant non-coding variants are a relatively unexplored source of pathogenic mutations and lack a general, high-throughput method to interpret their effects. We present genome-scale software to quantify the effect of mutations in the common classes of splice donor (U1) or acceptor (U2)-type sites in a high-throughput manner. Mutations predicted with this method will be useful for pinpointing potentially deleterious variants suitable for further laboratory investigation.

Clinical studies have deemed the vast majority of known variants in patients with Mendelian disorders to be of uncertain pathogenic significance (VUS) [2,3]. *Cis* mutations can affect protein translation, mRNA processing and initiation of transcription. *In silico* methods have been developed for the first two of these cases (*e.g.*, [4,5]), but have only been routinely applied for protein coding changes in genome-scale applications (*e.g.*, [6]). Many NGS studies classify mutations at only the highly conserved dinucleotides within each splice junction (*e.g.*, [7]). Although more sensitive methods have been developed which assess other conserved sequence elements [8–12], none have been scaled for the large numbers of variants generated by NGS and nor have they been validated for these data. Exonic variants in close proximity to splice junctions but outside of this window may be classified as synonymous, missense or nonsense substitutions, yet still have profound effects on splicing, which may be the predominant contributor to the phenotype. Unless multiple affected patients are reported with the same mutation, the mutations are transmitted through pedigrees, and functional assays verify their effects, these variants in patients are generally be classified as VUS. mRNA splicing mutations are common in Mendelian diseases [13,14], and it is likely that they contribute to many complex disorders. Clearly, genome-scale predictive methods that filter out benign or small changes in mRNA splicing due to sequence variation will be essential for mutation discovery in exomes, complete genomes and high-density targeted deep sequencing projects. Examination of individual variants in the laboratory with functional assays is both expensive and inefficient as many variants are not likely to be deleterious, or differ significantly in their pathogenicity.

The Automated Splice Site Analysis (ASSA) [5] server evaluates single mutations that change splice site strength with information theory-based models [15]. The average information, R_{sequence}, of a set of binding sites recognized by the same protein (such as U1 or U2) describes the conservation of these sequences. Sequences are ranked according to their individual information content (R_{i} in bits) [15–17]. Individual information content is a portable, universal measure which allows direct comparison of binding sites across the genome or transcriptome, regardless of the sequence or protein recognizer. Functional binding sites have $R_{\text{i}} > 0$, corresponding to $\Delta G < 0$ kcal/mol [18]. Strong binding sites have $R_{\text{i}} \gg R_{\text{sequence}}$, while weak sites have $R_{\text{i}} \ll R_{\text{sequence}}$. Any sequence variation may change its protein binding affinity, which is reflected by a change in the computed R_{i} of that binding site. A 1-bit change in information content (ΔR_{i}) corresponds to a $\geqslant$2-fold change in binding affinity ($100/2^{\Delta R_{\text{i}}}$). The ASSA server has been widely used and its sensitivity and specificity have previously been extensively validated in hundreds of studies of individual mutations (http://tinyurl.com/splice-server-citations). However, it requires approximately 30 s to examine a single variant and is therefore not suitable for comprehensive analysis of whole-genome sequencing data. The Shannon pipeline was developed using the same mathematical approach and information weight matrices as ASSA to carry out batch information theory-based analysis of thousands of mutations from the *BRCA1* and *BRCA2* genes in Breast Cancer Information Core Database [19]. In the present study, the software has been adapted to perform a single matrix algebraic calculation across a genome with an efficient state machine that significantly increased computational speed over ASSA. Here we describe this software tool and analyze predicted mutations with RNA-seq data from genomes of 3 cancer cell lines.

Results

Performance of the Shannon pipeline software

We implemented an efficient algorithm for high-throughput detection and interpretation of mRNA splicing mutations based on information theory-based position weight matrices of a genome-wide set of curated splice donor and acceptor sites [20]. The present study focuses on software performance, interpretation of contextual changes identified from genomic annotations and supported by genome-scale RNA-seq data. The strategy underlying the Shannon splicing pipeline is to evaluate many sequence changes by information analysis quickly; this is followed by implementation of a set of heuristics based on these results combined with genome annotations to distinguish normal splice sites from those with diminished binding and cryptic sites with competitive binding affinities.

To assess performance, all point mutations detected in the complete genomes of the three cancer cell lines were analyzed using the pipeline. Variants in the cell lines U2OS (osteosarcoma-derived), A431 (epidermoid squamous carcinoma-derived) and U251 (glioblastoma-derived) were examined and filtered to create tractable sets of variants. Predicted splice-altering mutations not found in dbSNP135 (a list of ~54 million known nucleotide polymorphisms) and those with less than 1% average heterozygosity are reported (Tables S1–3).

The software processes single nucleotide variants (SNVs) to identify and annotate putative splicing mutations with sufficient speed to analyze single or multiple genomes within a few hours. Analysis of all single nucleotide substitutions detected in the genome of the U2OS cell line – 211,049 variants – is completed in 1 h 12 min on an I7-based CPU in either Linux or Mac OSX. The speed of a genome analysis is dependent on the number of chromosomes represented in the input data. The state machine facilitates the analysis of all variants on a single chromosome with the highest efficiency because genomic data for each chromosome must be read and parsed. A complete analysis of 300 variants on a single small chromosome (*e.g.*, chromosome 22) can be completed in 5 min. Variants distributed throughout all chromosomes require at least 1 h to process. The Shannon pipeline should be executed on a machine with sufficient RAM to store the entire human genome ($\geqslant$4 Gb). When all chromosomes are represented, increasing the number of mutations results in an approximately linear increase in actual computation time, after accounting for the overhead required for memory management of genome sequences and annotations. For example, 2 h 35 min is required