VarCon: An R Package for Retrieving Neighboring Nucleotides of an SNV

Cancer Informatics
Volume 19: 1–3

DOI: 10.1177/1176935120976399

Johannes Ptok, Stephan Theiss and Heiner Schaal
Institute of Virology, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany.

ABSTRACT: Reporting of a single nucleotide variant (SNV) follows the Sequence Variant Nomenclature (http://varnomen.hgvs.org/), using an unambiguous numbering scheme specific for coding and noncoding DNA. However, the corresponding sequence neighborhood of a given SNV, which is required to assess its impact on splicing regulation, is not easily accessible from this nomenclature. Providing fast and easy access to this neighborhood just from a given SNV reference, the novel tool VarCon combines information of the Ensembl human reference genome and the corresponding transcript table for accurate retrieval. VarCon also displays splice site scores (HBond and MaxEnt scores) and HEXplorer profiles of an SNV neighborhood, reflecting position-dependent splice enhancing and silencing properties.

KEYWORDS: SNPs, alternative splicing, R package, sequence retrieval, HEXplorer score, HBond score

RECEIVED: September 1, 2020. **ACCEPTED:** November 1, 2020.

TYPE: Technical Advances

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Forschungskommission of the Medical Faculty, Heinrich Heine Universität Düsseldorf (2020-12) to H.S.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHORS: Stephan Theiss, Institute of Virology, Medical Faculty, Heinrich Heine University Düsseldorf, D-40225 Düsseldorf, Germany.
Email: theiss@uni-duesseldorf.de

Heiner Schaal, Institute of Virology, Medical Faculty, Heinrich Heine University Düsseldorf, D-40225 Düsseldorf, Germany.
Email: schaal@uni-duesseldorf.de

Introduction

Comparing genomic DNA sequences of individuals of the same species reveals positions where single nucleotide variations (SNVs) occur. When localized within the coding sequence of a gene, SNVs can, among others, affect which amino acids are encoded by the altered codon, potentially leading to disease. Approximately 88% of human SNVs associated with disease are, however, not located within the coding sequence of genes, but within intronic and intergenic sequence segments.[1] Nevertheless, annotations referring to the coding sequence of a specific transcript are still widely used, for example, c.8754+3G>C (BRCA2 and Ensembl transcript ID ENST00000544455), referring to the third intronic nucleotide downstream of the splice donor (SD) at the position of the 8754th coding nucleotide. Based on its position information referring to the coding sequence (c.) or alternatively to the genomic (g.) position (eg, g.1256234A>G), our tool VarCon retrieves an adjustable SNV sequence neighborhood from the reference genome. To visualize possible effects of SNVs on splice sites or splicing regulatory elements, which play an increasing role in cancer diagnostics and therapy,[2] VarCon additionally calculates HBond scores[3] of SDs and MaxEnt scores[4] of splice acceptor (SA) sites and HEXplorer scores of the retrieved sequences[9].

Implementation

VarCon is an R package which can be executed from Windows, Linux, or Mac OS. It executes a Perl script located in its directory and therefore relies on prior installation of some version of Perl (eg, Strawberry Perl). In addition, the human reference genome must be downloaded as fasta file (or zipped fasta.gz) with Ensembl chromosome names ("1" for chromosome 1) and subsequently uploaded into the R working environment, using the function "prepareReferenceFasta" to generate a large DNAStringset (file format of the R package Biostrings). To translate SNV positional information, referring to the coding sequence of a transcript, a transcript table has to be additionally uploaded to the working enviroment. The transcript table has to contain exon and coding sequence coordinates of every transcript from Ensembl. Two zipped transcript table csv-files which either refer to the genome assembly GRCh37 or GRCh38 can be downloaded from https://github.com/caggtaagtat/VarConTables.

As the transcript table with the GRCh38 genomic coordinates (currently from Ensembl version 100) will be updated with further releases, a new transcript table can be downloaded using the Ensembl Biomart interface. Any newly generated transcript table, however, must contain the same columns and column names as described in the documentation of the current transcript tables for correct integration. As, for instance, in cancer research the transcript which is used to refer to genomic positions of SNVs is often the same, a gene-to-transcript conversion table can be used for synonymous usage of certain gene names (or gene IDs) and transcript IDs (Ensembl ID). VarCon deliberately does not rely on Biomart queries using the Biomart R package, as these might be blocked by firewalls.

Due to its structure, the VarCon package can accept any genome and transcript table combination which is available on Ensembl and thus additionally permits usage for any other organism represented in the Ensembl database.[5] The combination of already existing tools like Mutalyzer,[6] SeqTailor,[7] or ensembldb[8] can lead to similar results during the variation conversion and DNA sequence extraction. However, VarCon holds additional benefits, namely, its straightforward usage even on a large-throughput scale, its independence due to the direct data entry, and its instant graphical representation of splicing regulatory elements and intrinsic splice site strength.

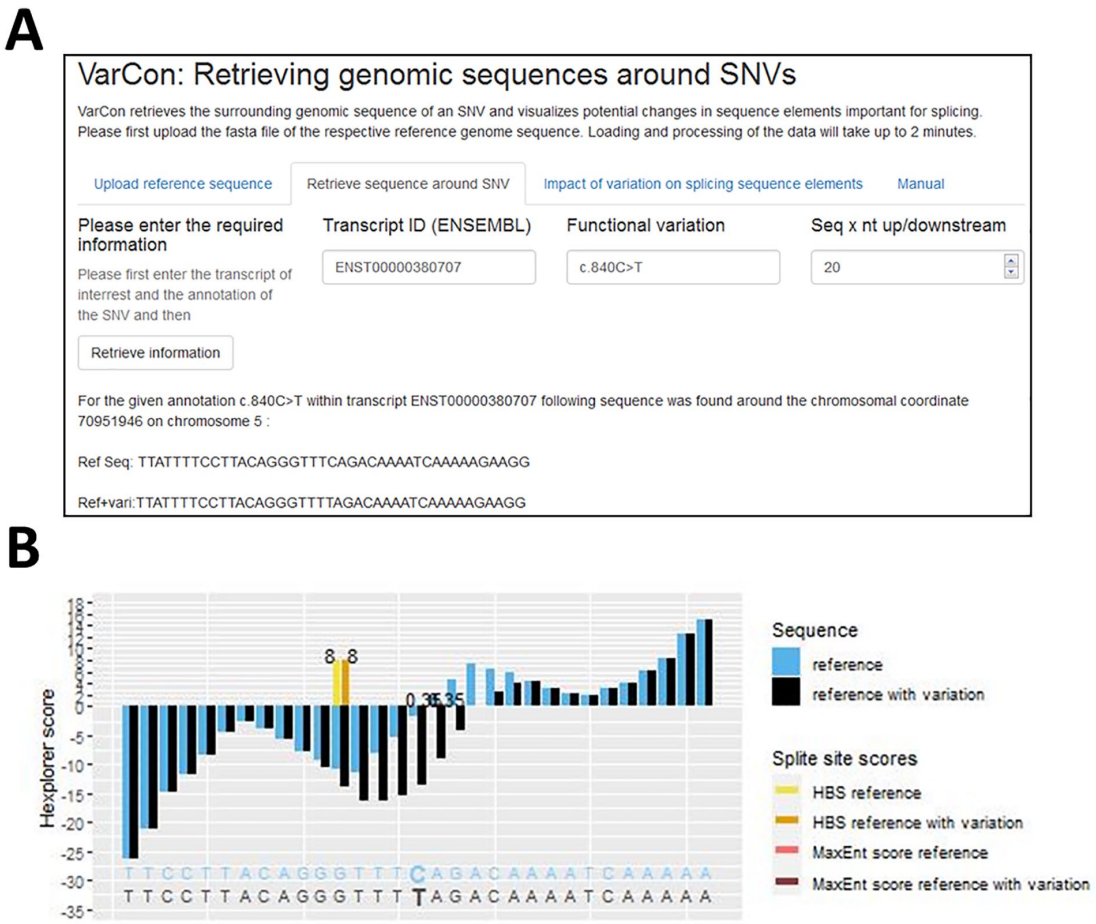

Figure 1. (A) Exemplary screenshot of VarCon GUI, querying the SNV c.840C>T in gene *SMN1* (transcript ENST00000380707). (B) HEXplorer plot of the sequence neighborhood of the same SNV. Bar plot depicting the HZ_{EI}-score for each nucleotide of the reference sequence in a ±20 nt neighborhood around the position of the variation with (black) or without (blue) the c.840C>T variation. HBond scores of donor sequences within the reference sequence are shown in yellow. HBond scores of donor sequences within the reference sequence with the variation are colored orange. GUI indicates graphical user interface; SNV, single nucleotide variant.

After upload of the human reference genome, selection of the appropriate transcript table and a potential gene-to-transcript conversion table, a transcript ID (or gene name) and an SNV (whose positional information either refers to the coding ["c."] or genomic ["g."] sequence) are requested during the execution of the main function of the package. VarCon then uses the information of the transcripts' exon coordinates to translate the SNV positional information to a genomic coordinate, if needed. Then the genomic sequence around the SNV position is retrieved from the reference genome in the direction of the open reading frame and committed to further analysis, both with and without the SNV.

For analysis of an SNV impact on splicing regulatory elements, VarCon calculates the HZ_{EI} score profile of reference and SNV sequences from the HEXplorer algorithm[9] and visualizes both in a bar plot. The HEXplorer score assesses splicing regulatory properties of genomic sequences, their capacity to recruit splicing regulatory proteins to the pre-mRNA transcript. Highly positive (negative) HZ_{EI} scores indicate sequence segments, which enhance (repress) usage of both downstream 5' splice sites and upstream 3' splice sites.

In addition, intrinsic strengths of SD and SA sites are visualized within the HZ_{EI} score plot. Splice donor strength is calculated by the HBond score, based on hydrogen bonds formed between a potential SD sequence and all 11 nucleotides of the free 5′ end of the U1 snRNA. Splice acceptor strength is calculated by the MaxEnt score, which is essentially based on the observed distribution of SA sequences within the reference genome, while also taking into account dependencies between both non-neighboring and neighboring nucleotide positions.[4]

VarCon can either be executed using integrated R package functions according to the manual on github or with a GUI (graphical user interface) application based on R package shiny with the integrated function "startVarConApp".

Example

The sequence variation c.840C>T within the seventh exon of the *SMN2* gene (Ensembl transcript ID: ENST00000380707) is associated with spinal muscular atrophy. Previous studies have shown that this sequence variation results in a change in splicing regulatory protein binding, increasing skipping of exon 7. Entering this variation and the transcript ID into VarCon (Figure 1A) leads to the following bar plot visualizing this effect with a delta HZ_{EI} of −71.76 (Figure 1B).

Acknowledgements

We would like to thank Gene Yeo for his kind approval to integrate the MaxEnt scoring algorithm into VarCon.

Author Contributions

JP developed the R code of the VarCon package and drafted the manuscript. ST and HS supervised the project and also wrote the manuscript.

Availability

VarCon is available at https://github.com/caggtaagtat/VarCon and released under the MIT License. After installation of the package, an attached shiny app can be started with the integrated function "startVarConApp".

ORCID iDs

Johannes Ptok https://orcid.org/0000-0002-0322-5649
Heiner Schaal https://orcid.org/0000-0002-1636-4365

REFERENCES

1. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 2009;106:9362-9367.
2. Dong X, Chen R. Understanding aberrant RNA splicing to facilitate cancer diagnosis and therapy. *Oncogene*. 2020;39:2231-2242.
3. Freund M, Asang C, Kammler S, et al. A novel approach to describe a U1 snRNA binding site. *Nucleic Acids Res*. 2003;31:6963-6975.
4. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*. 2004;11:377-394.
5. Birney E, Andrews TD, Bevan P, et al. An overview of Ensembl. *Genome Res*. 2004;14:925-928.
6. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum Mutat*. 2008;29:6-13.
7. Zhang P, Boisson B, Stenson PD, et al. SeqTailor: a user-friendly webserver for the extraction of DNA or protein sequences from next-generation sequencing data. *Nucleic Acids Res*. 2019;47:W623-W631.
8. Rainer J, Gatto L, Weichenberger CX. ensembldb: an R package to create and use Ensembl-based annotation resources. *Bioinformatics*. 2019;35:3151-3153.
9. Erkelenz S, Theiss S, Otte M, Widera M, Peter JO, Schaal H. Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. *Nucleic Acids Res*. 2014;42:10681-10697.