SeqTailor: a user-friendly webserver for the extraction of
DNA or protein sequences from next-generation sequencing data

Peng Zhang $^{1,8}$, Bertrand Boisson $^{1,2,3}$, Jean-Laurent Casanova $^{1,2,3,4,5}$,
Laurent Abel $^{1,2,3}$, and Yuval Itan $^{6,7}$

1 St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY 10065, USA
2 Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM UMR1163, Paris 75015, France
3 Paris Descartes University, Imagine Institute, Paris 75015, France
4 Howard Hughes Medical Institute, New York, NY 10065, USA
5 Pediatric Immunology-Hematology Unit, Necker Hospital for Sick Children, Paris 75015, France
6 The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
7 Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

$^8$ Corresponding author:
Peng Zhang (pzhang@rockefeller.edu)
ABSTRACT

Human whole-genome sequencing generally reveals about 4,000,000 genetic variants, including 20,000 coding variants, in each individual studied. These data are mostly stored as VCF-format files. Although many variant analysis methods accept VCF files as input, many other tools require DNA or protein sequences, particularly for splicing prediction, sequence alignment, phylogenetic analysis, and structure prediction. However, there is currently no existing online tool for extracting DNA or protein sequences for genomic variants from VCF files with user-defined parameters in a user-friendly, efficient, and standardized manner. We developed the SeqTailor webserver to bridge this gap. It can be used for the rapid extraction of (1) DNA sequences around genetic variants, with customizable window sizes, from the hg19 or hg38 human reference genomes; and (2) protein sequences encoded by the DNA sequences around genetic variants, with built-in SnpEff annotation and customizable window sizes, from human canonical transcripts. The SeqTailor webserver streamlines the sequence extraction process, and accelerates the analysis of genetic variant data with software requiring DNA or protein sequences. SeqTailor will facilitate the study of human genomic variation, by increasing the feasibility of sequence-based analysis and prediction.

The SeqTailor webserver is freely available from http://shiva.rockefeller.edu/SeqTailor/.
INTRODUCTION

Human whole-genome sequencing (WGS) and whole-exome sequencing (WES) generate large amounts of genetic data, generally revealing about 4,000,000 variants (WGS) and 20,000 coding variants (WES), in each individual studied. These variant data are usually recorded in variant call format (VCF), which includes structured fields for chromosome (CHROM), position in reference genome sequence (POS), reference allele (REF), alternative allele (ALT) and others (e.g. QUAL, FILTER, INFO, etc.). The major human genetic variation and disease databases (e.g. gnomAD/ExAC (Lek et al. 2016), 1,000 Genomes Project (Genomes Project et al. 2015), BRAVO (TOPMed 2016), HGMD (Stenson et al. 2012), and ClinVar (Landrum et al. 2014)), and most research laboratories and publications, present and manage genetic variants in VCF. For further variant annotations, predictions and visualizations, most of the leading tools use VCF as input (e.g. ANNOVAR (Wang et al. 2010), SnpEff (Cingolani et al. 2012), VEP (McLaren et al. 2016), CADD (Kircher et al. 2014), and PopViz (Zhang et al. 2018)).

However, as the DNA sequence around the variant site presents richer information than the variant alone, many popular genomic software suites require DNA sequences as input, particularly for predictions relating to splicing (e.g. Human Splicing Finder (Desmet et al. 2009), NetGene2 (Brunak et al. 1991), Spliceman (Lim and Fairbrother 2012), RegRNA (Chang et al. 2013)), sequence alignment with reference genomes (e.g. BLAST (Altschul et al. 1997), BLAT (Kent 2002)), alignments of multiple sequences supplied by the user (e.g. MUSCLE (Edgar 2004), MAFFT (Katoh et al. 2017)), and phylogenetic analyses (e.g. PAML (Yang 2007), IQ-TREE (Trifinopoulos et al. 2016)). In addition, many protein bioinformatics tools use amino-acid sequences for protein domain identification and functional annotation (e.g. Pfam (Finn et al. 2016),
InterPro (Finn et al. 2017), PolyPhen-2 (Adzhubei et al. 2010)), protein structure prediction and homology modeling (SWISS-MODEL (Waterhouse et al. 2018), HMMER (Potter et al. 2018)), and protein feature calculation (e.g. POSSUM (Wang et al. 2017), PROFEAT (Zhang et al. 2017)). The sequence alignment tools cited above can also be applied to protein sequences. Hence, a knowledge of the reference protein sequence and of the protein sequences altered by genomic variants would make it possible to evaluate their potential effects on protein domains, structures, and functions.

Users wishing to analyze VCF data with software that requires DNA or protein sequence must therefore extract the corresponding sequence from VCF files. The major tools currently available for this purpose are BEDTools (Quinlan and Hall 2010), GATK (McKenna et al. 2010), and IGV software (Thorvaldsdottir et al. 2013). However, BEDTools and GATK are script-based methods that require programming expertise, and IGV is impractical for high-throughput applications (a detailed feature comparison is presented in the Discussion section). We therefore developed the SeqTailor webserver, which offers a user-friendly, efficient, and standardized approach for streamlining DNA and protein sequence extraction from the genetic variant data in VCF.
RESULTS

We developed the SeqTailor webserver to provide a user-friendly and efficient approach for the rapid extraction of DNA and protein sequences for genetic variants (single nucleotide variants (SNVs) and indels) in VCF, with user-defined window sizes. It accepts input from either the webpage textbox or a user-uploaded file, and it outputs the extracted sequences in the browser and in a downloadable FASTA file. SeqTailor bridges the gap between genomic variants and the tools requiring DNA or protein sequences, as illustrated in Figure 1. The SeqTailor interface is shown in Figure 2, and a schematic diagram is provided in Figure 3.

Figure 1: The SeqTailor webserver bridges the gap between genomic variants and sequence-based tools.
Figure 2: SeqTailor webserver interface: (A) the homepage, (B) the page for DNA sequence extraction, and (C) the page for protein sequence extraction.
Figure 3: SeqTailor webserver scheme to demonstrate the framework of the application.
For DNA sequence extraction, SeqTailor supports use of the hg19/GRCh37 and hg38/GRCh38 human reference genomes from the UCSC genome database (Karolchik et al. 2004), 1/0-based genomic coordinate indexing, and the choice of the both/forward/reverse strand(s). It accepts both genomic variants in VCF and genomic ranges in BED. For input in the form of genomic variants in VCF, it outputs the wild-type reference (ref) and mutated alternative (alt) DNA sequences, with user-defined up/downstream window sizes (in bp), on the selected strand(s). For input in the form of genomic ranges in BED, it outputs the DNA sequences within the intervals on the selected strand(s). SeqTailor provides the DNA sequences exactly from the START position to the END position, whereas the UCSC Genome Browser always provides an overlap between the input genomic intervals and certain genomic regions (exons, introns, UTRs, etc.), and outputs the desired DNA sequence intervals segregated into different regions.

For protein sequence extraction, SeqTailor first conducts built-in SnpEff variant annotation (Cingolani et al. 2012) for the input variants, retains the missense, stop-gained, and synonymous variants, and extracts the annotated gene symbols, transcript IDs, and amino-acid changes. Only the annotations on canonical transcripts are retained, by matching against the canonical transcripts collected from the UCSC Genome Browser and the GENCODE database (Harrow et al. 2012). SeqTailor provides options for extracting the entire protein sequence, or the sequence within a window of customized size (in aa). Finally, it outputs the canonical and altered protein sequences. As the extraction of protein sequences from genomic ranges is already well-handled by UCSC Genome Browser, we did not develop this function in SeqTailor.
Four case studies are presented in the **Supplementary Document**, including two examples of DNA sequence extraction for a SNV and an indel from the *MSH2* and *BRCA2* genes, respectively, followed by splicing prediction; one example of protein sequence extraction for a stop-gained SNV in the *CFTR* gene; and one example of protein sequence extraction for a SNV in the *BRAF* gene, followed by protein domain search and functional prediction. These examples demonstrate the practical power of SeqTailor to bridge the gap between genome variant data and sequence-based tools for downstream analysis and prediction. SeqTailor will make it efficient to investigate genomic variant data further and will render sequence-based software more accessible.

SeqTailor is designed for the rapid computational extraction of DNA and protein sequences from genetic variant data. For DNA sequence extraction, it takes about 6, 12, 13 and 14 seconds to extract the sequences of 10, 100, 1,000, and 10,000 genomic variants, respectively, from VCF files. For protein sequence extraction, all the input variants are first passed through the build-in SnpEff annotator, and only the missense, stop-gained, and synonymous variants annotated on canonical transcripts are retained for the extraction of protein sequences. It takes approximately 50, 60, 90, 100 seconds to complete the variant annotation/filtering and the protein sequence extraction for 10, 100, 1,000, and 10,000 variants, respectively. The SeqTailor runtime performance is shown in **Figure 4**.
Figure 4: Runtime performance of SeqTailor for extracting DNA or protein sequences from varying sizes of input VCF data (10, 100, 1,000, and 10,000 genomic variants)
DISCUSSION

Two major sequence extraction methods currently available, BEDTools and GATK, provide a script-based option for DNA sequence extraction, and require significant bioinformatics expertise, and they do not support the extraction of protein sequences around genomic variants. Alternatively, IGV software can be used to locate the position of the variant in the reference genome, with manually copying and pasting to extract the desired DNA sequences or the corresponding protein sequences, one at a time, which would be very time-consuming, error-prone, and impractical in high-throughput data applications. The UCSC Genome Browser provides a web interface for converting genomic ranges in BED/position format to DNA/protein sequences, but users need to go through several steps and define a range of parameters (annotation track, gene table, field table, retrieval region, etc.) to get the result. Importantly, the UCSC Genome Browser does not support sequence extraction from genomic variant data in VCF. The popular GALAXY platform (Afgan et al. 2018) and biopython library (Cock et al. 2009) do not offer options for DNA and protein sequence extraction. A comparison of these approaches with SeqTailor is presented in Table 1.
SeqTailor webserver is the first online tool for the extraction of DNA and protein sequences from genomic variants in VCF. It presents an efficient and customizable approach for this purpose, requiring only succinct input parameters and no knowledge of script programming. It will be of benefit to scientists studying human genetics, genomics, and molecular biology, enabling them to perform further analyses of their variants with DNA or protein sequence-based tools more rapidly and efficiently than is currently possible.
METHODS

Webserver

The SeqTailor webserver is run by Apache HTTP (version 2.2.15), on a Red Hat Enterprise Linux Server (version 6.9), with 8 Intel CPU processors @ 2.4GHz and 48GB RAM. The website interface is design and presented in HTML, PHP, CSS, and JavaScript. The data are stored in MySQL tables and plain text files. The sequence extraction programs are coded in Python.

Reference Genome

The human reference genome sequences (GRCh37/hg19 assembly, and GRCh38/hg38 assembly) are obtained from UCSC Genome Browser, and stored by genome and by chromosome, to accelerate data processing. Once SeqTailor receives the variants for DNA sequence extraction, it firstly groups the variants by chromosome, and then extracts all the DNA sequences from the same chromosome only once, thus enhancing the efficiency of sequence extraction.

Canonical Transcript

The canonical transcript IDs were retrieved from UCSC Genome Browser, and the protein-coding transcript translated sequences were downloaded from the GENCODE database. Merging these two resources of data, we obtained a collection of amino-acid sequences for 17,675 human canonical transcripts.

Built-in Variant Annotation

SnpEff is built in for variant annotation and effect prediction within SeqTailor, with the installed GRCh37.75 and GRCh38.85 databases. Its speed is maximized by a multi-thread configuration,
annotating the canonical transcript only, and excluding the upstream, downstream, intergenic, and intron regions. The input variants are then filtered by annotated effect, for retention of the missense, stop-gained, and synonymous variants only. The remaining transcripts are matched against the canonical transcript sequences collected from UCSC genome browser and GENCODE. The annotated amino-acid changes are then reflected in the canonical protein sequences.

**Runtime Evaluation**

The runtime evaluation was performed by applying four input data sizes (10, 100, 1,000, and 10,000 genetic variants). We generated 50 VCF files for each input size, by randomly selected variants from the gnomAD database (Lek et al. 2016), and retaining only the first five columns of information (CHROM, POS, ID, REF, and ALT). We therefore used a total of 200 VCF files with different numbers of variants to evaluate the runtime performance of SeqTailor for DNA and protein sequence extraction.
ACKNOWLEDGEMENT

We thank Y. Nemirovskaya and D. Papandrea for administrative support, F. Rapaport and B. Bigio for technical discussions, and Z. Yang for the artwork of website design. This study was supported by the Rockefeller University, and the Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai.

DISCLOSURE DECLARATION

The authors have no conflict of interest to declare.
REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7: 248-249.

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Gruning BA et al. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 46: W537-W544.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

Brunak S, Engelbrecht J, Knudsen S. 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. J Mol Biol 220: 49-65.

Chang TH, Huang HY, Hsu JB, Weng SL, Horng JT, Huang HD. 2013. An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. BMC Bioinformatics 14 Suppl 2: S4.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6: 80-92.

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wileczynski B et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422-1423.

Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37: e67.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M et al. 2017. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res 45: D190-D199.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44: D279-285.

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al. 2015. A global reference for human genetic variation. Nature 526: 68-74.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22: 1760-1774.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32: D493-496.

Katoh K, Rozewicki J, Yamada KD. 2017. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform doi:10.1093/bib/bbx108.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. Genome Res 12: 656-664.

Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46: 310-315.
Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**: D980-985.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285-291.

Lim KH, Fairbrother WG. 2012. Spliceman--a computational web server that predicts sequence variations in pre-mRNA splicing. *Bioinformatics* **28**: 1031-1032.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thor mann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* **17**: 122.

Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. 2018. HMMER web server: 2018 update. *Nucleic Acids Res* **46**: W200-W204.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN. 2012. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Curr Protoc Bioinformatics Chapter 1*: Unit1 13.

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**: 178-192.

TOPMed. 2016. Bravo.

Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res* **44**: W232-235.

Wang J, Yang B, Revote J, Leier A, Marquez-Lago TT, Webb G, Song J, Chou KC, Lithgow T. 2017. POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. *Bioinformatics* **33**: 2756-2758.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**: e164.

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriel l G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L et al. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **46**: W296-W303.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586-1591.

Zhang P, Bigio B, Rapaport F, Zhang S, Casanova J, Abel L, Boisson B, Itan Y. 2018. PopViz: a webserver for visualizing minor allele frequencies and damage prediction scores of human genetic variations. *Bioinformatics*.

Zhang P, Tao L, Zeng X, Qin C, Chen SY, Zhu F, Yang SY, Li ZR, Chen WP, Chen YZ. 2017. PROFEAT Update: A Protein Features Web Server with Added Facility to Compute Network Descriptors for Studying Omics-Derived Networks. *J Mol Biol* **429**: 416-425.