bioRxiv preprint doi: https://doi.org/10.1101/173047; this version posted August 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

GENOME ANALYSIS

TraPS-VarI: a python module for the identification of STAT3 modulating germline receptor variants

Daniel Kogan[1] and Vijay Kumar Ulaganathan[2,*]

[1]Technische Universität München, Arcisstraße 21, München 80333, Germany

[2]Max Planck Institute of Biochemistry, Department of Molecular Biology, Am Klopfersptiz 18, Martinsried 82152, Germany

To whom correspondence should be addressed.

Abstract

Motivation: Human individuals differ because of variations in the DNA sequences of all the 46 chromosomes. Information on genetic variations altering the membrane-proximal binding sites for signal transducer of transcription 3 (STAT3) is valuable for understanding the genetic basis of cancer prognosis and disease progression (Ulaganathan et al, 2015). In this regard, non-synonymous coding region mutations resulting in the alteration of protein sequence in the juxtamembrane region of the type I membrane proteins are biologically and clinically relevant. The knowledge of such rare cell line- and individual-specific germline receptor variants is crucial for the investigation of cell-line specific biological mechanisms and genotype-centric therapeutic approaches.

Results: Here we present TraPS-VarI (**Tra**nsmembrane **P**rotein **S**equence **V**ariant **I**dentifier), a python module to rapidly identify human germline receptor variants modulating STAT3 binding sites by using the genetic variation datasets in the variant call format 4.0. For the found protein variants the module also checks for the availability of associated therapeutic agents and ongoing clinical trial studies.

Availability: The Source code and binaries are freely available for download at https://gitlab.com/VJ-Ulaganathan/TraPS-VarI and the documentation can be found at http://traps-vari.readthedocs.io/.

Contact: ulaganat@biochem.mpg.de & ulaganat@icloud.com

Supplementary information: Supplementary data enclosed with the manuscript file.

1 Introduction

Approximately 30% of all open reading frames (ORFs) in the mammalian genome encode for membrane proteins (Krogh et al, 2001; Stevens & Arkin, 2000; Wallin & Heijne, 1998; Yildirim et al, 2007). And a large majority of all currently available therapeutic agents target membrane proteins[4]. Furthermore, it is becoming increasingly evident that variations affecting the amino acid sequence of membrane proteins might play a substantial role in disease susceptibility, disease progression and can determine therapeutic outcomes (Csaszar & Abel, 2001; Hargreaves et al, 2015; Molnar et al, 2016). Furthermore, surmounting data on genetic association of membrane protein variants with complex diseases such as autoimmune diseases, metabolic disorders, cardiovascular diseases, and cancer demands new tools for systematic approaches. Although there is no shortage of bioinformatics software catering the need of biological interpretation of genome datasets, the significance of functionally relevant variations in membrane proteins has largely remained under appreciated. Recently, we uncovered the occurrence of biologically relevant membrane-proximal STAT3 binding sites in type I membrane proteins that are capable of modulating the amplitude of STAT3 signaling within tissues and additionally altering the growth inhibition responses to certain pharmacological inhibitors (Ulaganathan et al, 2015; Ulaganathan & Ullrich, 2016). The knowledge on the frequencies of variations that either create, delete or expose such membrane-proximal STAT3 binding motifs in the general population and patient cohorts can thus significantly help clinicians to determine an effective therapeutic regimen and additionally help researchers identify and validate disease linked receptor variants. Towards this goal, we present TraPS-VarI, a python tool to rapidly identify functionally meaningful germline receptor variants by using the genome-wide genetic variation datasets.

2 Methods

TraPS-VarI traces genomic variants to their effects on membrane proteins using a mapping path running through nodes that include the latest human genome builds (GRCh37 and GRCh38) (Human Genome Sequencing, 2004) (MacDonald et al, 2014), Human Ensembl (Yates et al, 2016), dbSNP (Sherry et al, 2001) and Uniprot protein (2017). After determining the optimum path, the algorithm converts stepwise from genomic variant to protein substitution (**Fig. 1a**). Once the mutations on the chromosome loci are successfully converted to protein sequence alteration, the script scans the protein sequence from the end of transmem-

bioRxiv preprint doi: https://doi.org/10.1101/173047; this version posted August 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

brane segment in the C-term direction until it reaches the end of 40 amino acid distances. The effect of protein substitution on the presence, absence or alteration of any membrane proximal STAT3 binding motif namely "YxxQ" (where x is any amino acid) in the 40 amino acid long membrane proximal cytoplasmic domains are reported with their position followed by the motif in parenthesis (**Fig. 1b**). Annotations are as follows, " PRESENT ", when a motif is found but not altered by the mutation, " ABSENT ", when no motif exists, " CREATED ", when no motif is found but the mutation creates one and " DESTROYED ", when a motif is found but destroyed by the altered allele.

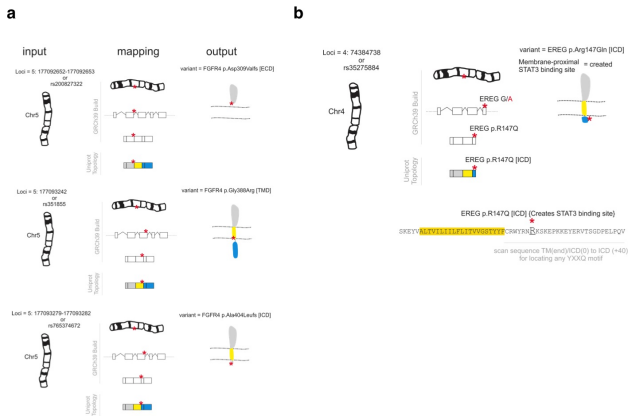


Fig. 1. Mapping of genetic mutation altering membrane proximal STAT3 binding sites. (a) Illustration of mapping of the entries namely 'chromosome' and 'position' from the .vcf file to amino acid location in UniProt protein and (b) identification of membrane proximal STAT3 binding motif YXXQ.

3 Results

TraPS-VarI processes the vcf file line by line. It takes the position, matches this against coding regions in the RefSeq database. It then matches the CDS to its appropriate UniProt entry, modifies the CDS according to the mutation and retranslates the resulting CDS. The effect of the mutation is derived from the difference between those two entries. In it's current version it only matches against UniProts main entries and not their isoforms (support for this is planned). It also looks up the position and mutation in the dbSNP dataset; if the mutation is contained there it adds the dbSNPid to the entry. It also checks the found refseq and uniprot ids against the therapeutic target database (ttd) (Zhu et al, 2012) and DrugBank database (Law et al, 2014). TraPS-Vari outputs the results in 8 columns namely, "protein id", "protein position", "predicted protein change", "mutation type", "STAT3 site", "protein domain", "ttd" and "drugbank".

3.1 Installation

TraPS-VarI will add itself as a module to python.

3.1.1 Requirements

(1) Python 3.4 or newer

(2) MySQLdb module for python (MySQL-python).

(3) Access to a MySQL database (InnoDB engine with spatial index support - v5.7 or higher).

3.1.2 Install command

Extract the package to the desired install location and run the installation script with:

python install .py

And follow the instructions.

3.1.2 Usage

python TraPSVarI.py {−p=<CHR:POS> −m=<REF/ALT>| −f=<filename> [options]} [−assembly=<assembly version >]

• -p=chr:pos to look up a single mutation (i.e. A/T) at the position chr:pos (only with -m)

• -m=REF/ALT mutation to look up (only in conjunction with -p)

• -f= to run the script on a file in vcf format

• -fout file to save the result to (default is input file traps vari output)

• -filter use to omit all lines from the result that do not contain a transmembrane protein mutation

• -assembly=37/38 (which assembly to use, dbSNP is only supported for 38)

The resulting file will have the protein position, protein allele change, mutation effect and the motif changes, hits in the ttd database and the DrugBank database appended as columns. Additionally, snippets that are readily integrable in the workflow of TraPS-VarI module are available under https://github.com/VJ-Ulaganathan.

Acknowledgements

The authors thank Prof. Dr. Axel Ullrich for supporting this work and Viet Nuyen for technical assistance.

Conflict of Interest: The authors declare no conflict of interests.

References

(2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45:** D158-D169

Csaszar A, Abel T (2001) Receptor polymorphisms and diseases. *European Journal of Pharmacology* **414:** 9-22

Hargreaves CE, Rose-Zerilli MJJ, Machado LR, Iriyama C, Hollox EJ, Cragg MS, Strefford JC (2015) Fcγ receptors: genetic variation, function, and disease. *Immunological Reviews* **268:** 6-24

Human Genome Sequencing C (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931-945

Krogh A, Larsson Br, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1. *Journal of Molecular Biology* **305:** 567-580

Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research* **42:** D1091-D1097

bioRxiv preprint doi: https://doi.org/10.1101/173047; this version posted August 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research* **42:** D986-D992

Moln√°r Jn, Szak√°cs G, Tusn√°dy GbE (2016) Characterization of Disease-Associated Mutations in Human Transmembrane Proteins. *PLOS ONE* **11:** e0151760

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29:** 308-311

Stevens TJ, Arkin IT (2000) Do more complex organisms have a greater proportion of membrane proteins in their genomes? *Proteins: Structure, Function, and Bioinformatics* **39:** 417-420

Ulaganathan VK, Sperl B, Rapp UR, Ullrich A (2015) Germline variant FGFR4 p.G388R exposes a membrane-proximal STAT3 binding site. *Nature* **528:** 570-574

Ulaganathan VK, Ullrich A (2016) Membrane-proximal binding of STAT3 revealed by cancer-associated receptor variants. *Molecular & Cellular Oncology* **3:** e1145176

Wallin E, Heijne GV (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Science* **7:** 1029-1038

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Gir√≥n CGa, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Keenan S, Lavidas I, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Nuhn M, Parker A, Patricio M, Pignatelli M, Rahtz M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Birney E, Harrow J, Muffato M, Perry E, Ruffier M, Spudich G, Trevanion SJ, Cunningham F, Aken BL, Zerbino DR, Flicek P (2016) Ensembl 2016. *Nucleic Acids Research* **44:** D710-D716

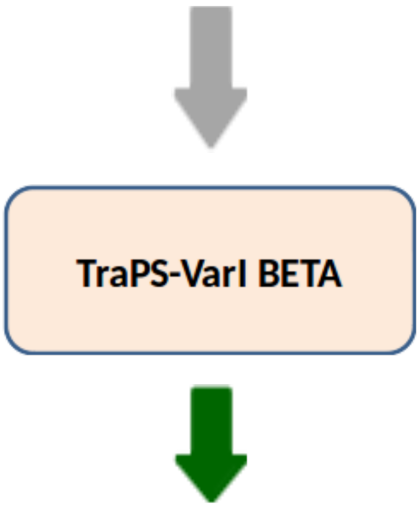
Yildirim MA, Goh K-I, Cusick ME, Barabasi A-L, Vidal M (2007) Drug[mdash]target network. *Nat Biotech* **25:** 1119-1126

Zhu F, Shi Z, Qin C, Tao L, Liu X, Xu F, Zhang L, Song Y, Liu X, Zhang J, Han B, Zhang P, Chen Y (2012) Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Research* **40:** D1128-D1136

bioRxiv preprint doi: https://doi.org/10.1101/173047; this version posted August 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Input file: .VCF genotype data

Chr	Position	ID	Ref Allele	Alt Allele
13	28034133	rs773968505	C	A
15	41507219	rs770047590	G	A
1	121096043	rs587642032	A	G



Output file-1: .traps_vari_output

Protein IDI	Protein Acession	Protein Position	Predicted Protein Change	Mutation Type	STAT3 Site	Protein Domain	Therapeutic Target Database	DrugBank
FLT3_HUMAN	P36888	596	E->*	STOP CREATED	(566, 'YKKQ'), '[PRESENT]'), ((572, 'YESQ'), '[PRESENT]')	CYTOPLASMIC	TTDNC00410[investigational],TTDS00090[approved]	DB00398 [investigational, approved] (Sorafenib), DB01268 [investigational, approved] (Sunitinib), DB05014 [investigational] (XL999), DB05213 [investigational] (AC220), DB05216 [investigational] (MP470), DB05465 [investigational] (MLN-18), DB06080 [investigational] (ABT-869), DB08901 [approved] (Ponatinib), DB09079 [approved] (Nintedanib)
LTK_HUMAN	P29376	473	R->*	STOP CREATED	(484, 'YYCQ'), [DESTROYED]	CYTOPLASMIC	-	-
FCGRB_HUMAN	Q92637	245	S->Y	SNP	(245, 'YSLQ'), [[CREATED]	CYTOPLASMIC	-	DB00028 [investigational,approved] (b'Immune Globulin Human')

Supplementary_Figure. 1 Schematic diagram describing the workflow of TraPS-VarI python module