Pipeline and Varant annotation tool for identifying causal variants in inherited rare disorders

Kunal Kundu[1], Sadhna Rana[1], Ajithavalli Chellappan[1], Uma Sunderam[1], Jennifer M. Puck[2], Steven E. Brenner[3], Rajgopal Srinivasan[1#]

[1]Tata Consultancy Services Ltd, Innovation Labs, Hyderabad, India, [2] Department of Pediatrics, University of California, San Francisco, CA 94143-0519, USA
[3]University of California, Berkeley, CA 94720, USA [#]Corresponding author: email address raj@atc.tcs.com

Overview

We have developed a pipeline for the analysis of genomic variant data, having distinctive features that enabled solving numerous clinical cases related to SCID (Severe Combined Immunodeficiency disease) and related diseases.

Key features of the pipeline

- **Multiple variant callers** carefully tuned for exome data to yield high quality call set and an **extensible framework** to include additional callers.
- **Reporting of extensive quality metrics** for mapping, gene coverage and called variants.
- **Comprehensive variant annotation by Varant,** an open source tool developed by us.
- **Gene prioritization module** integrating gene annotations, protein interaction networks, pathways and text mining methods.

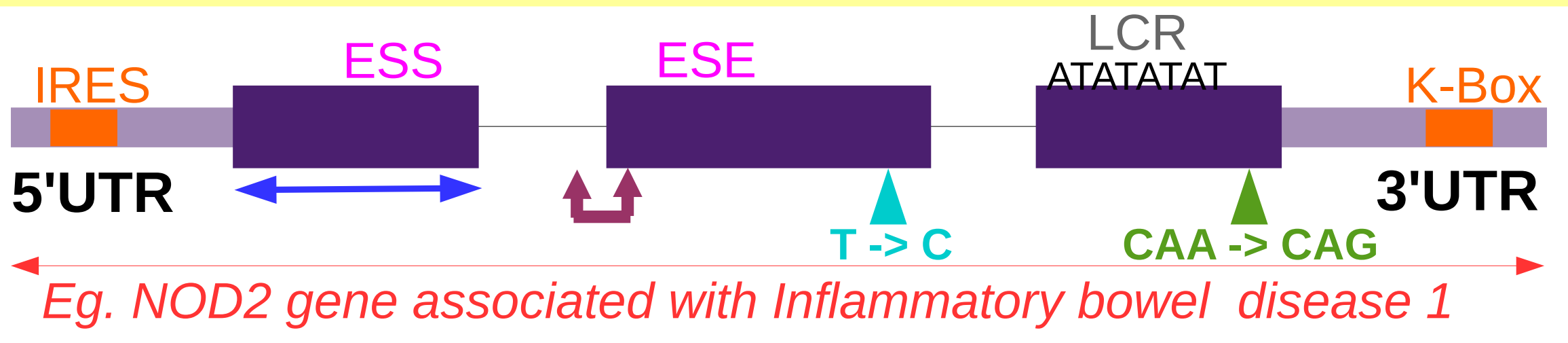
Our pipeline has identified likely causative variants cases where typical protocols would have been expected to fail.

Varant: An open source variant annotation tool

Varant provides features comparable with other tools, and ...

	Varant	Annovar	snpEff	VEP
License	AGPLv3	Commercial*	LGPLv3	Modified Apache
Language	Python	Perl	Java	Perl
Variant Type	SNP, Indel, MNP	SNP, Indel	SNP, Indel, MNP	SNP, Indel, MNP
Input Format	vcf	tsv, vcf	vcf, bed	tsv, vcf, HGVS notation
Output Format	vcf, tsv, xls	tsv	vcf, tsv	vcf, tsv, json object
Multiple Gene definition supported	●	●	●	●
Uses HGVS notation	●	●	●	●
Uses Sequence Ontology terms			●	●
Region – Intergenic, Intron, Exon, UTR	●	●	●	●
SpliceSites (Donor/Acceptor)	●	●	●	●
Mutation Type – NonSyn, StopGain etc	●	●	●	●
dbSNP, 1000Genomes(MAF), ESP(MAF)	●	●	●	●
Polyphen2 & SIFT predictions	●	●	●	●
Clinically significant variations	●	●	●	●
Protein Domain	●	●	●	●
Variant position conservation	●	●	●	
miRNA binding site	●	●	●	
CADD predictions	●	●	●	
GWAS phenotype	●	●	●	
TFBS	●	●	●	

... Varant provides key annotations, in addition to those in other tools.



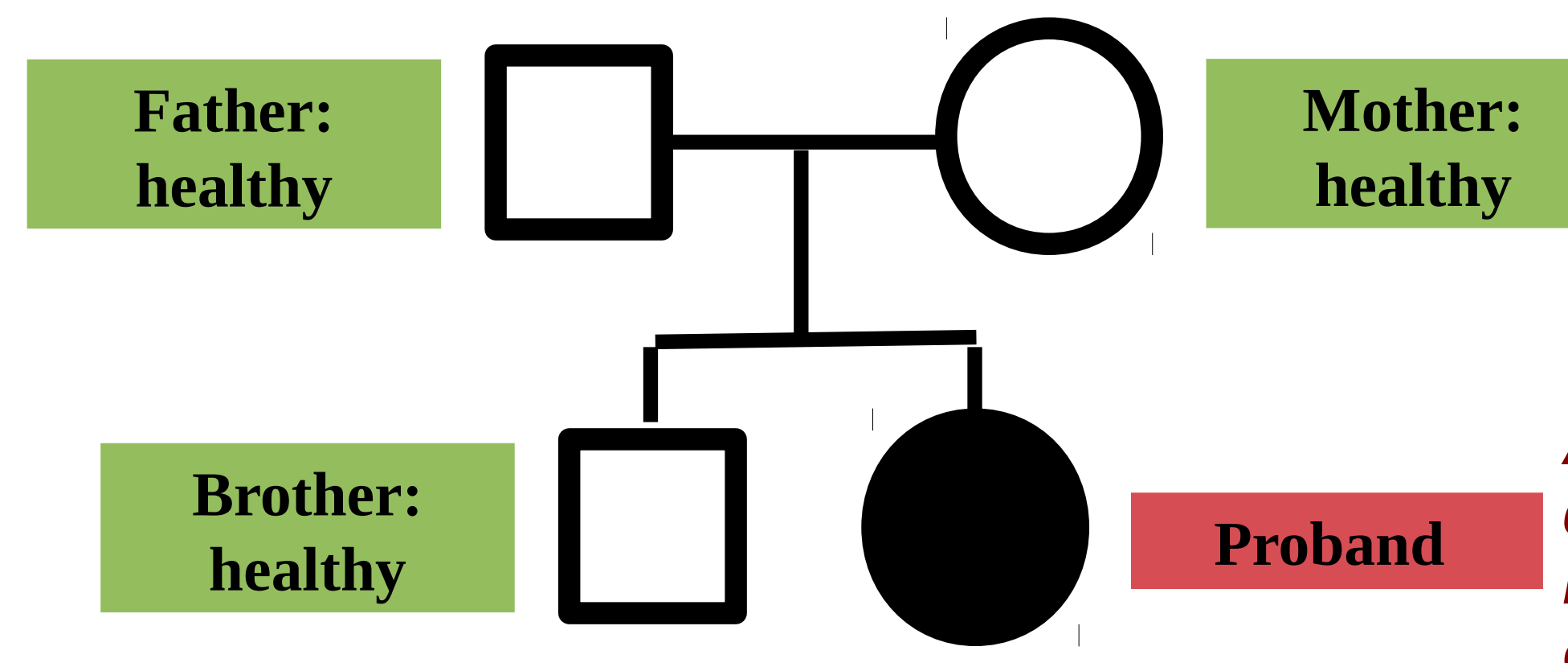
Eg. NOD2 gene associated with Inflammatory bowel disease 1

Diseases associated with gene | Capture Region | UTR Motifs
Low Complexity Region(LCR) | ESE/ESS Sites | eQTLs
Flag variants at boundary | Codon usage

*But free personal, academic, and non-profit use only

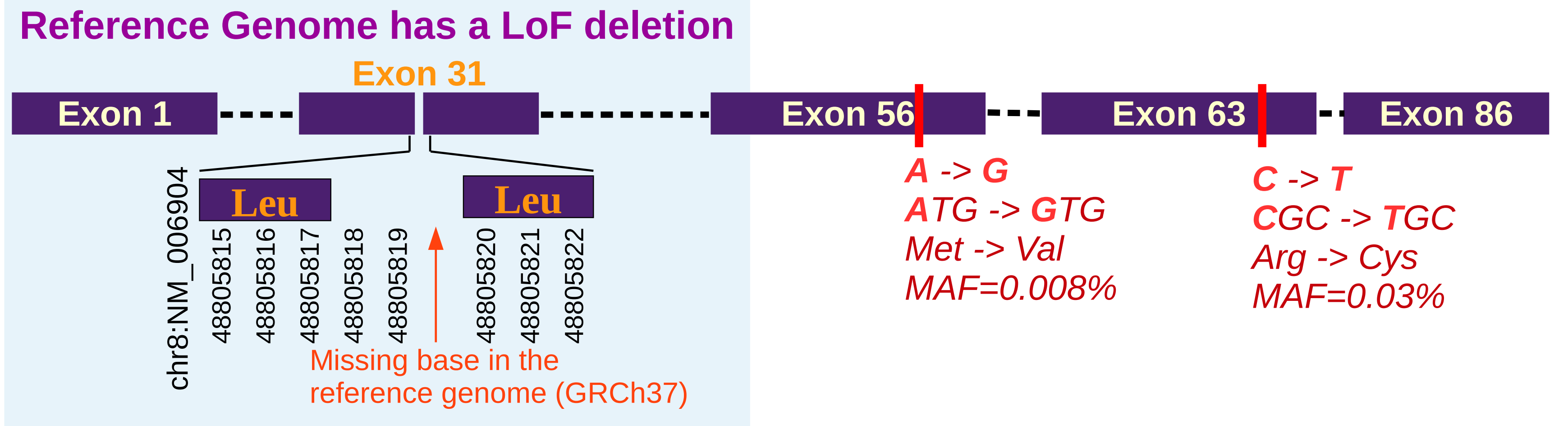
Our pipeline solved cases that would likely have been missed by others

Identified potential causative variants in the presence of inconsistencies in reference genome



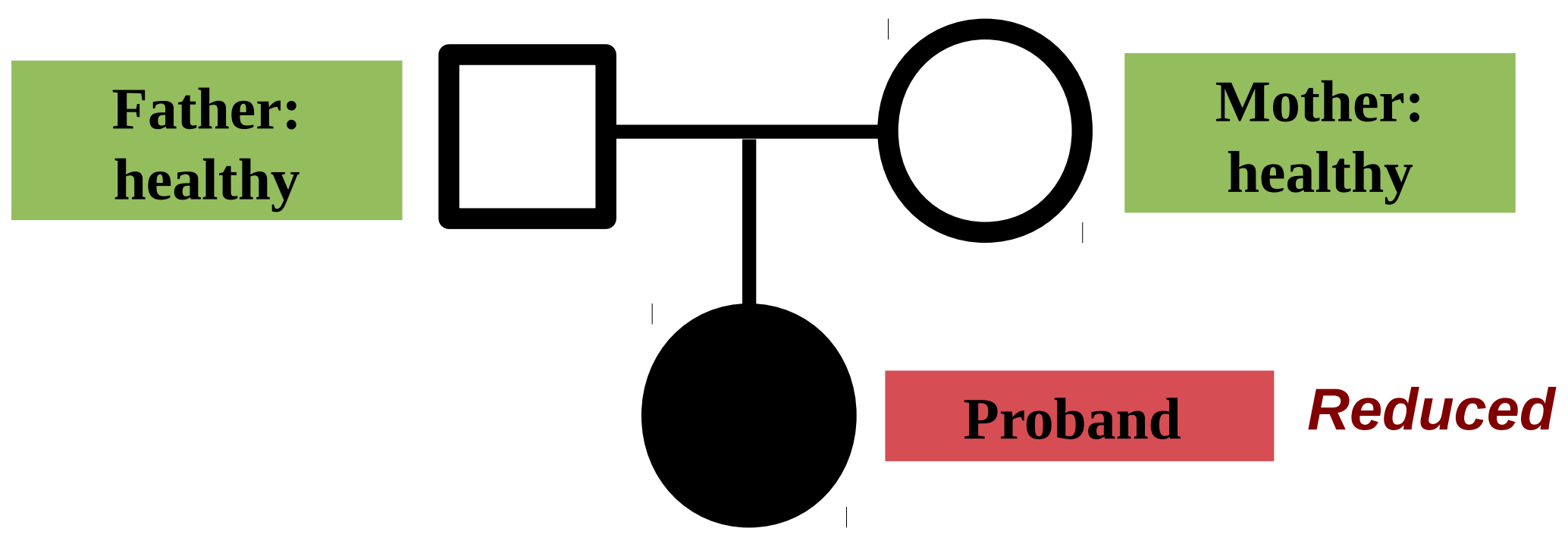
Affected with SCID, an immune disorder characterized by reduction or complete absence of T lymphocytes

Our pipeline shortlisted 2 compound heterozygous variants in the *PRKDC* gene (known to be associated with SCID) in the proband.



- Varant flagged the *PRKDC* gene with **CDS inconsistency** annotation meaning the coding sequence of *PRKDC* gene was not multiple of 3.
- The missing bases in the reference genome were found to be upstream of the prioritized variants in *PRKDC* gene.
- Manual inspection revealed that the prioritized variants in *PRKDC* gene were non-synonymous relative to the normal coding sequence.
- **Several tools like Annovar[7] do not warn about such genomic CDS anomalies, and thus these variants would have been overlooked.**

Haploinsufficiency annotation of a gene in a family with immune related disorder



Reduced T lymphocytes

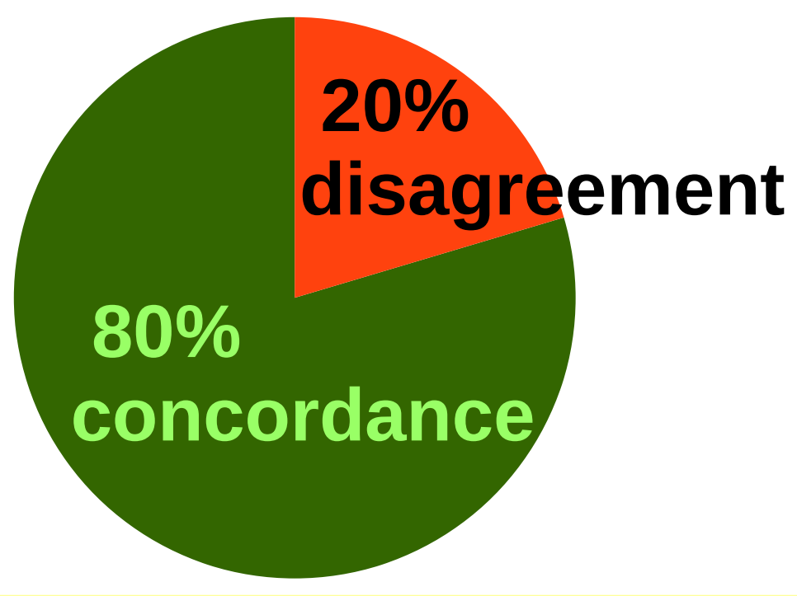
Our pipeline shortlisted a *de-novo* heterozygous variant in the *BCL11B* gene in the proband.
The gene prioritization module's haploinsufficient gene annotation (compiled by manual curation of literature report) recognized *BCL11B*[11] gene implying a single copy of mutant gene is sufficient to cause disease.

Such stand alone annotations would usually lead to variants being prioritized for review.

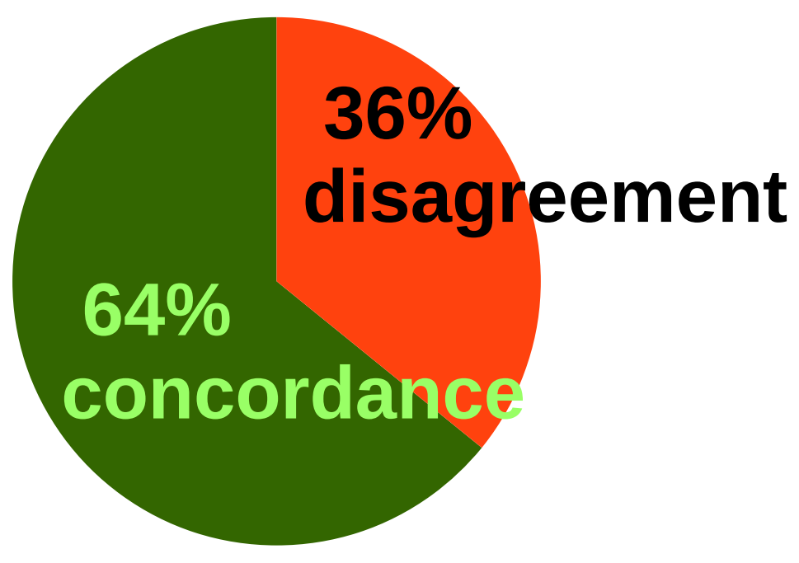
Varant has 80% concordance with other tools

To estimate the accuracy of Varant, annotations for 1.9 million variants (SNPs and Indels) present in ESP[3] vcf were extensively compared among Varant, Annovar[7], snpEff[6] and VEP[5].

Annotation comparison across all genomic region among 4 tools

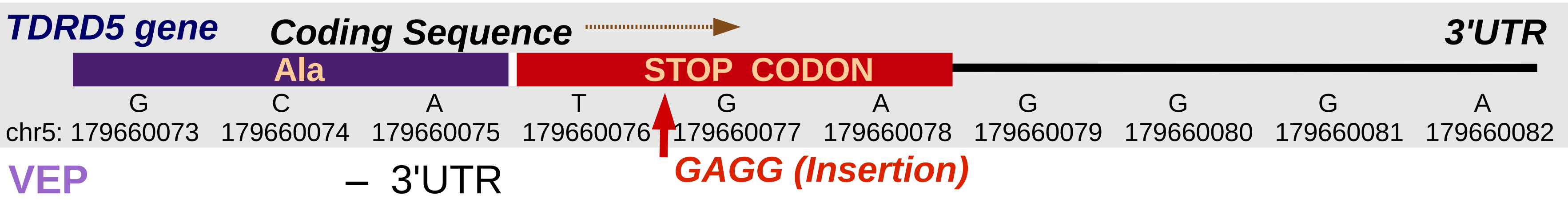


Annotation comparison for 70,347 loss of function variants(FrameShift, StopLoss, StopGain & Splicing) among 4 tools



When Varant disagrees with other methods, its predictions are superior

Insertion variant that does not alter stop codon

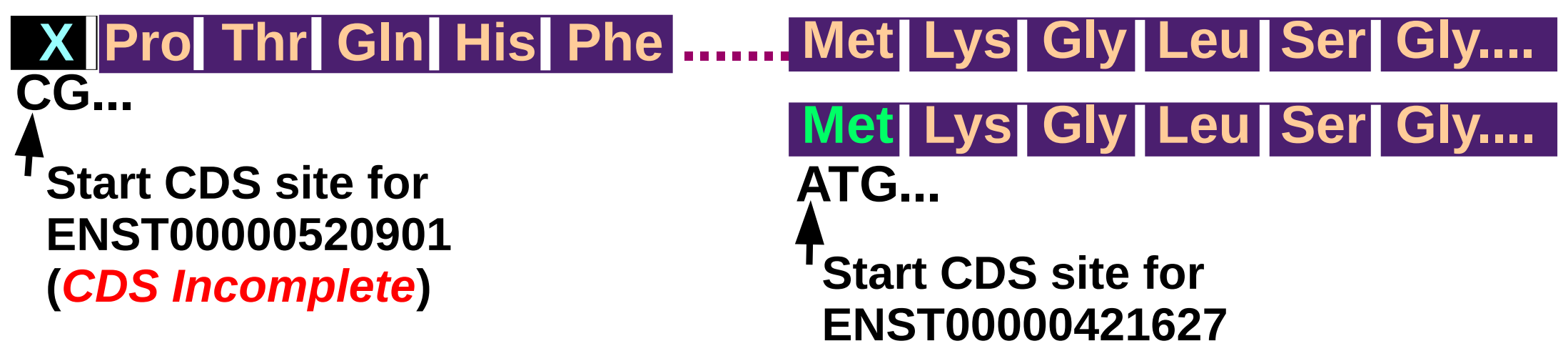

TDRD5 gene
Coding Sequence
3'UTR
Ala
STOP CODON
G C A T G A G G G A
chr5: 179660073 179660074 179660075 179660076 179660077 179660078 179660079 179660080 179660081 179660082
GAGG (Insertion)

VEP – 3'UTR
Annovar, SnpEff – FrameShift (incorrect)
Varant – **3'UTR;NoCDSChange**

Mutation type computation on incomplete transcripts

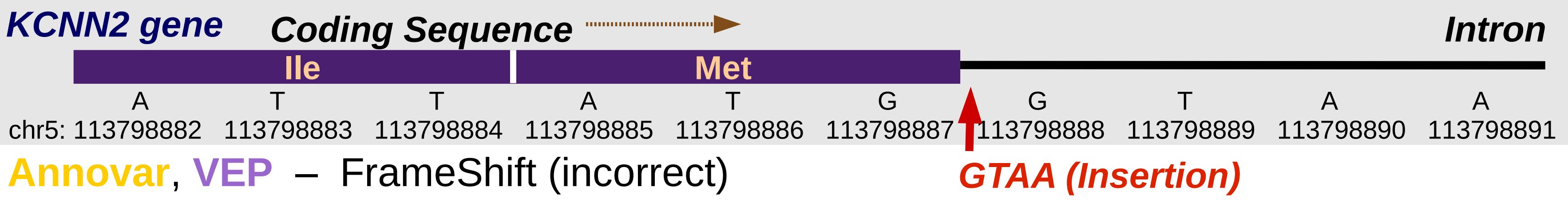
For all coding variants on transcripts whose CDS was incomplete -

	Mutation type computed	Warned about incomplete CDS
Varant		●
SnpEff	●	●
VEP	●	
Annovar		


X
Pro
Thr
Gln
His
Phe
Met
Lys
Gly
Leu
Ser
Gly....
CG...
Start CDS site for ENST00000520901 (CDS Incomplete)
Met
Lys
Gly
Leu
Ser
Gly....
ATG...
Start CDS site for ENST00000421627

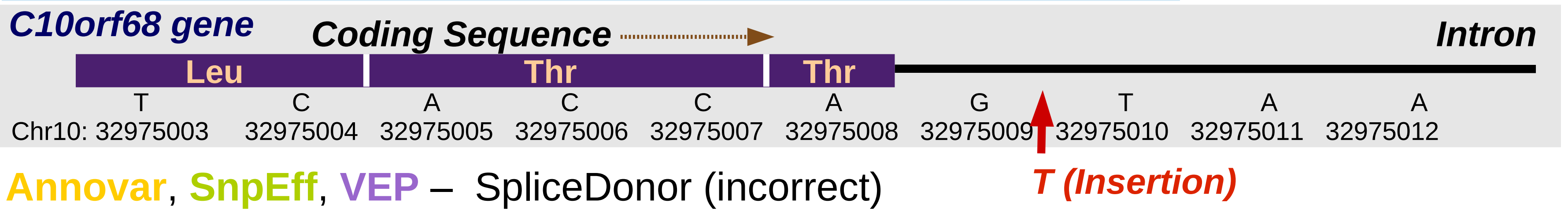
Mutation type computed on incomplete transcripts were incorrect because of change in reading frame

Insertion variant at exon intron boundary that does not alter CDS


KCNN2 gene
Coding Sequence
Intron
Ile
Met
A T T A T G G T A A
chr5: 113798882 113798883 113798884 113798885 113798886 113798887 113798888 113798889 113798890 113798891
GTAA (Insertion)

Annovar, VEP – FrameShift (incorrect)
SnpEff – SpliceDonor (incorrect)
Varant – **Exon_Intron_boundary;Close_to_SpliceDonor**

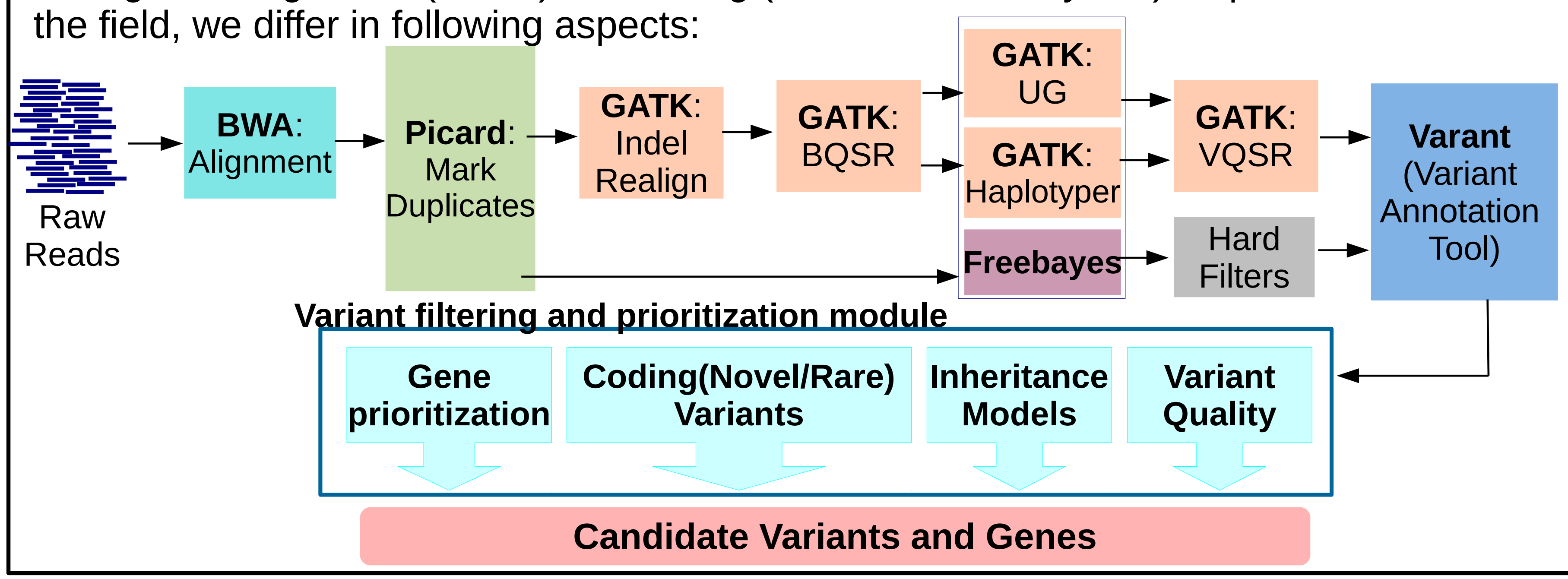
Insertion variant at splice donor site that do not disrupt splice site


C10orf68 gene
Coding Sequence
Intron
Leu
Thr
Thr
T C A C C A G T A A
Chr10: 32975003 32975004 32975005 32975006 32975007 32975008 32975009 32975010 32975011 32975012
T (Insertion)

Annovar, SnpEff, VEP – SpliceDonor (incorrect)
Varant – **Intronic; Close_to_SpliceDonor**

Exome Analysis Pipeline

Our pipeline uses series of steps to identify causal variants in rare inherited disorders. Though the alignment (BWA[8]) and calling (GATK[9] & Freebayes[10]) steps are standard in the field, we differ in following aspects:



Conclusion

- Our genome analysis pipeline generates reliable variant calls and quality variant annotation for better interpretation of human genetic variants.
- Our pipeline has identified likely causal variants in several cases where other pipelines would have been expected to fail.
- Some of the key features of our pipeline that has helped to make confident genotype-phenotype predictions includes –
  - Use of multiple callers and combined calling
  - Use of Varant which provides a broad range of annotations with equal or better precision and accuracy in comparison with other well known tools.
- Varant is freely available for use (http://compbio.berkeley.edu/proj/varant).

References

1. Boerwinkle E et al(2011). Hum Mutat. 32, 894-9. doi:10.1002/humu.21517
2. Burge CB et al. (2002) Science. 297, 1007-13. Epub 2002 Jul 11
3. Exome Variant Server, NHLBI GO Exome Sequencing Project(ESP), Seattle, WA.
4. Batzoqlou S et al. (2010). PLoS Comput Biol. 6, e1001025. doi:10.1371/journal.pcbi.1001025
5. McLaren W et al. (2010). Bioinformatics. 26, 2069-2070. doi: 10.1093/bioinformatics/btq330
6. Douglas M. Ruden et al. (2012) Fly (Austin). 6, 80–92. doi:10.4161/fly.19695
7. Hakon Hakonarson et al. (2010) Nucl. Acids Res.38, e164. doi:10.1093/nar/gkq603
8. Li H, Durbin R, (2009) Bioinformatics, 25, 1754-1760. doi:10.1093/bioinformatics/btp324
9. DePristo MA, et al. (2011) Nat Genet, 43, 491-498. doi: 10.1038/ng.806
10. Garrison E, Marth G, (2012). ArXiv e-prints: 1207.3907.
11. Dang VT, et al. (2008) Eur J Hum Genet, 16, 1350-1357. doi:10.1038/ejhg.2008.111