

Chromosome 17 *TP53* gene variant calling and protein structure prediction

Xinyi Zhang CID: 02020037

Introduction

Single Nucleotide Polymorphisms (SNPs) are common genetic variant in humans. It is a genomic substitution at a single base presented in at least 1% of the population. Identifying SNPs is vital in determining how humans develop diseases and respond to drugs and chemicals. *TP53* (Tumour Protein 53) is one of the most commonly mutated genes in human cancer (Palomar-Siles et al., 2022) that can be caused by SNPs. The *TP53* gene codes for the p53 protein, which is a tumor suppressor. *TP53* is located on human chromosome 17 7,661,779-7,687,538. The gene is essential in regulating apoptosis, cell cycle, and metabolism (Mello & Attardi, 2018). However, SNPs cause it to be associated with cancer such as Li-Fraumeni syndrome (Guha & Malkin, 2017), esophageal cancer (Niyaz et al., 2020), adrenocortical carcinoma (Ribeiro et al., 2001), and many other cancers. Thus, it is vital to figure out the potential variation within the *TP53* gene, aiding localization of the cancer-causing SNPs.

Methods

To analyze SNPs, the process involves aligning Illumina sequencing reads with a reference sequence using the BWA algorithm (Li and Durbin, 2009). Unwanted duplicates are then removed using GATK (Van der Auwera et al., 2020). Differences between the reference and the sequence reads are identified by SAMtools (Li et al., 2009) and GATK. A final raw VCF file is then generated. GATK and Base quality score recalibration (BQSR) is used to clean up and filter the data further. SnpEff program is used to predict the effect of SNPs and annotate their significance (Cingolani et al., 2012), while bcftools can further annotate the VCF file to indicate known SNPs from the database and add rs numbers for SNPs (Danecek et al., 2021). By indicating the position of the *TP53* gene (7,661,779-7,687,538), a final processed VCF file is produced. In order to further investigate the known variants with rs numbers annotated by bcftools, BioMart is used to identify their clinical significance and find the UniProt ID for the transcripts that contain the variants (Durinck et al., 2009). Further analysis of the variants is done on the Ensembl (Cunningham et al., 2022), MissenseDB-3D (Khanna et al., 2021), and PhyreRisk (Ofoegbu et al., 2019).

Results and Discussion

The SNPs identified within the *TP53* gene region are rs1416898259, rs587782596, and rs786201057. Their basic features are listed in table 1.

Table 1: SNP Variants identified in *TP53* gene and their features

rs number	Chromosome	Position	Type of variant	Has clinical significance or not
rs1416898259	Chr 17	7675724	Intron variant	No
rs587782596	Chr 17	7675071	Missense variant	Possibly

				pathogenic
rs786201057	Chr 17	7675995	Missense variant	Possibly pathogenic

rs1416898259 has an ancestral allele of Adenine (A) with a less than 0.0008% chance of being substituted by Guanine (G) in African/African American descent. BioMart suggests that it does not exhibit any clinical significance (Cunningham et al., 2022). PhyreRisk indicates that it acts as an intron variant affecting alternative splicing (Ofoegbu et al., 2019). Therefore, further study on this intron variants is needed to look for evidence for disrupted splicing in p53 transcript.

There is less than 0.0008% chance that rs587782596 missense variant substitutes ancestral base G with A or Thymine (T) in Non-Finnish European. While the missense variant rs786201057 also has less than 0.0008% chance of ancestral base substitution from G to A or T or Cysteine (C) and is often happen in African/African American. The REVEL tool estimates that both rs587782596 (0.891 for G/A; 0.778 for G/T) and rs786201057 (0.925 for G/A; 0.963 for G/T or G/C) are likely to be disease-causing. MetaLR tool also indicates a damaging pathogenic effect for rs587782596 (0.989 for G/A; 0.983 for G/T) and rs786201057 (0.989 for G/A; 0.991 for G/C; 0.992 for G/T).

Both SNPs show evidence of developing a type of hereditary cancer-predisposing syndrome called Li-Fraumeni syndrome which is an inherited cancer predisposition syndrome that makes the family members inclined to develop various cancers including breast cancer, osteosarcomas, and brain tumors (Bougeard et al., 2015). rs587782596 has evidence showing that it causes G into A (Raymond et al., 2013) or T substitution (Kato et al., 2003), while rs786201057 has clear evidence for alteration from G to A (Bougeard et al., 2015). Some specific kinds of carcinomas are also triggered by rs786201057 SNP, including acute myeloid leukemia, lung adenocarcinoma, and pancreatic adenocarcinoma with G replaces by A/T/C (Chang et al., 2016). Type 2 diabetes is also proven to be caused by rs786201057 (Pipal et al., 2022).

The wild-type protein structure of the transcript predicted by the MissenseDB-3D (Khanna et al., 2021) is shown in Figure 1 (a). In the mutant structure, arginine (Arg) at UniProt position 181 is changed into cysteine (Cys) and causes missense mutation. This mutation is included in the isoform 1 of the transcript, which is ENST000002693059. The mutant protein structure is shown in Figure 1 (b). According to PhyreRisk, the new protein structure is caused by the missense variant rs587782596 (Ofoegbu et al., 2019). However, MissenseDB-3D does not classify the new structure as damaging.

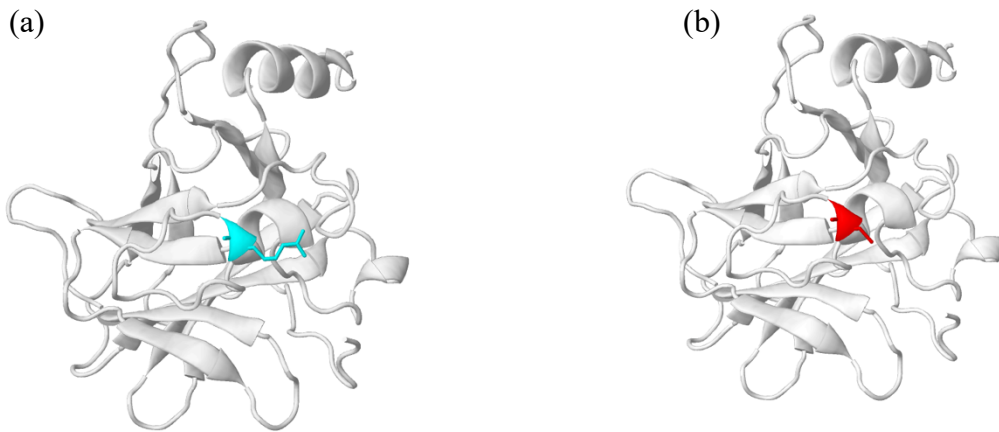


Figure 1: Structure of the protein missense variant rs587782596 produced. (a): wild type structure, with Arg (cyan) at position 181 (b): mutant structure, replace arginine with Cys (red) at position 181.

This suggests that the inclusion of the rs587782596 variant in the transcript leads to a structural change in the p53 protein, but it is not considered destructive. However, it's important to notice that the altered p53 protein may not interact effectively with other proteins like MDM2 (Fischer et al., 2016) and BRCA1 (Niyaz et al., 2020), potentially leading to the development of severe cancers or Li-Fraumeni syndrome. Also, p53 is a DNA-binding protein that can activate genes in the nucleus. Arg 181 maintains the p53 intra-dimer interaction and impact the DNA-binding stability of p53 (Timofeev & Stiewe, 2021). Hence, if Arg 181 changed into Cys 181, p53 property of DNA-binding is affected. Further research on the mutant p53 protein in this case should be done.

The protein structure that is produced by the variant rs786201057 is shown in figure 2. According to PhyreRisk, threonine (Thr) 125 can be replaced by Arg or lysine (Lys) or methionine (Met) when the ancestral base changes (Ofoegbu et al., 2019).

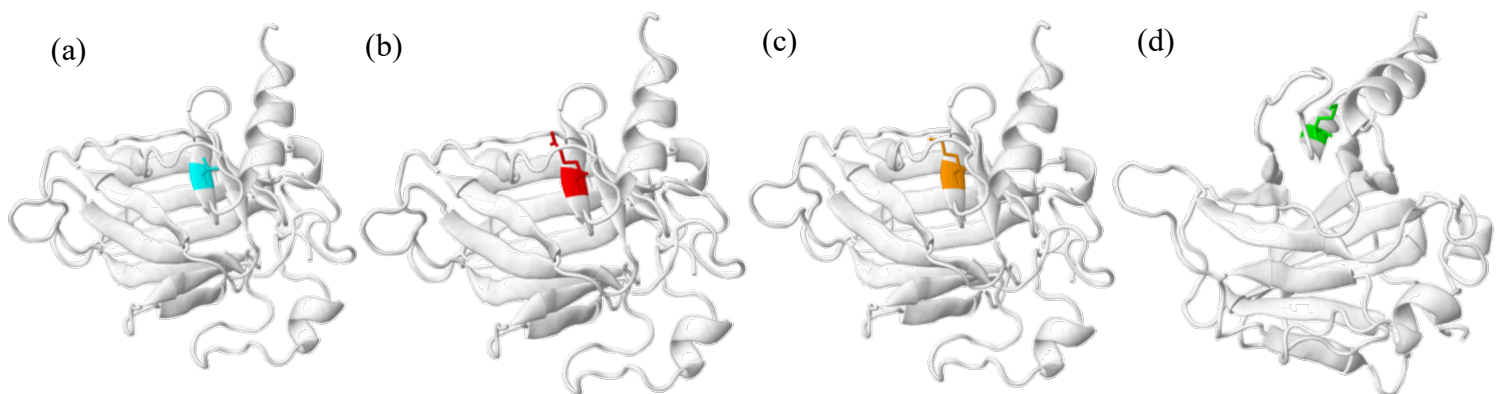


Figure 2: Structure of the protein missense variant rs786201057 produced. (a): wild type structure, with Thr (cyan) at position 125 (b)-(d): mutant structures, replaced Thr 125 with: (b) Arg (red), (c) Lys (orange), (d) Met (green).

Predicted by MissenseDB-3D, all three substitutions cause structural damage to protein structure (Khanna et al., 2021). Arginine 125 causes buried to exposed switch of variant residue since Arg has RSA of 14.5% while Thr has RSA of 0.7%. Furthermore, the MolProbity clash score increases from 14.35 to 34.90 after substitution. When replaced by Lys, the uncharged residue converts into charged residue and the buried H-bond is broken. Met, on the other hand, also causes buried H-bond breakage. Unlike rs587782596, the missense variant rs786201057 brings negative changes to the p53 protein, potentially explaining its association with various cancers and diseases reported on Ensembl.

However, further analysis is needed to establish the exact relationship of these variants with diseases. The evidence supporting the association of rs786201057 with various cancers and diabetes is limited to two papers on Biomart, reducing its persuasiveness. Additionally, the probabilities of variant occurrence are unreliable as they are based on a subset of the population. While Ensembl provides current knowledge, the actual impact of these variants needs to be determined through experiments such as gene editing to assess their clinical significance. Furthermore, what variant calling, and protein predictions fail to mention is the information about protein-protein interactions. For example, the intron SNP rs1416898259 affects alternative splicing, which can hinder translation and alteration of open reading frame. rs587782596 does not cause p53 protein damage but computational analysis does not give information about the p53 protein interaction with other proteins.

In summary, TP53 variants, including intron and missense variants, may play a role in diseases such as Li-Fraumeni syndrome, cancers, and type 2 diabetes. They can induce changes in protein structure, although not necessarily harmful. However, further analysis and experimental investigation to understand their functional implications beyond superficial bioinformatics analysis are required.

References

- Bougeard, G., Renaux-Petel, M., Flaman, J.-M., Charbonnier, C., Fermey, P., Belotti, M., Gauthier-Villars, M., Stoppa-Lyonnet, D., Consolino, E., Brugières, L., Caron, O., Benusiglio, P.R., Bressac-de Paillerets, B., Bonadona, V., Bonaïti-Pellié, C., Tinat, J., Baert-Desurmont, S. & Frebourg, T. (2015) Revisiting Li-Fraumeni Syndrome From TP53 Mutation Carriers. *Journal of Clinical Oncology*. 33 (21), 2345–2352. doi:10.1200/JCO.2014.59.5728.
- Chang, M.T., Asthana, S., Gao, S.P., Lee, B.H., Chapman, J.S., Kandoth, C., Gao, J., Socci, N.D., Solit, D.B., Olshen, A.B., Schultz, N. & Taylor, B.S. (2016) Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature biotechnology*. 34 (2), 155–163. doi:10.1038/nbt.3391.
- Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. & Ruden, D.M. (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*. 6 (2), 80–92. doi:10.4161/fly.19695.

Cunningham, F., Allen, J.E., Allen, J., Alvarez-Jarreta, J., Amode, M.R., et al. (2022) Ensembl 2022. *Nucleic Acids Research*. 50 (D1), D988–D995. doi:10.1093/nar/gkab1049.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M. & Li, H. (2021) Twelve years of SAMtools and BCFtools. *GigaScience*. 10 (2), giab008. doi:10.1093/gigascience/giab008.

Durinck, S., Spellman, P.T., Birney, E. & Huber, W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*. 4 (8), 1184–1191. doi:10.1038/nprot.2009.97.

Fischer, M., Grossmann, P., Padi, M. & DeCaprio, J.A. (2016) Integration of TP53, DREAM, MMB-FOXM1 and RB-E2F target gene analyses identifies cell cycle gene regulatory networks. *Nucleic Acids Research*. 44 (13), 6070–6086. doi:10.1093/nar/gkw523.

Guha, T. & Malkin, D. (2017) Inherited TP53 Mutations and the Li–Fraumeni Syndrome. *Cold Spring Harbor Perspectives in Medicine*. 7 (4), a026187. doi:10.1101/cshperspect.a026187.

Kato, S., Han, S.-Y., Liu, W., Otsuka, K., Shibata, H., Kanamaru, R. & Ishioka, C. (2003) Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proceedings of the National Academy of Sciences of the United States of America*. 100 (14), 8424–8429. doi:10.1073/pnas.1431692100.

Khanna, T., Hanna, G., Sternberg, M.J.E. & David, A. (2021) Missense3D-DB web catalogue: an atom-based analysis and repository of 4M human protein-coding genetic variants. *Human Genetics*. 140 (5), 805–812. doi:10.1007/s00439-020-02246-z.

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25:1754-60. [PMID: 19451168]

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25 (16), 2078–2079. doi:10.1093/bioinformatics/btp352.

Mello, S.S. & Attardi, L.D. (2018) Deciphering p53 signaling in tumor suppression. *Current Opinion in Cell Biology*. 51, 65–72. doi:10.1016/j.ceb.2017.11.005.

Niyaz, M., Ainiwaer, J., Abudurehman, A., Zhang, L., Sheyhidin, I., Turhong, A., Cai, R., Hou, Z. & Awut, E. (2020) Association between TP53 gene deletion and protein expression in esophageal squamous cell carcinoma and its prognostic significance. *Oncology Letters*. 20 (2), 1855–1865. doi:10.3892/ol.2020.11709.

Ofoegbu, T.C., David, A., Kelley, L.A., Mezulis, S., Islam, S.A., Mersmann, S.F., Strömich, L., Vakser, I.A., Houlston, R.S. & Sternberg, M.J.E. (2019) PhyreRisk: A Dynamic Web Application to Bridge Genomics, Proteomics and 3D Structural Data to Guide Interpretation of Human Genetic Variants. *Journal of Molecular Biology*. 431 (13), 2460–2466. doi:10.1016/j.jmb.2019.04.043.

Palomar-Siles, M., Heldin, A., Zhang, M., Strandgren, C., Yurevych, V., van Dinter, J.T., Engels, S.A.G., Hofman, D.A., Öhlin, S., Meineke, B., Bykov, V.J.N., van Heesch,

- S. & Wiman, K.G. (2022) Translational readthrough of nonsense mutant TP53 by mRNA incorporation of 5-Fluorouridine. *Cell Death & Disease*. 13 (11), 1–17. doi:10.1038/s41419-022-05431-2.
- Pipal, K.V., Mamtani, M., Patel, A.A., Jaiswal, S.G., Jaisinghani, M.T. & Kulkarni, H. (2022) Susceptibility Loci for Type 2 Diabetes in the Ethnically Endogamous Indian Sindhi Population: A Pooled Blood Genome-Wide Association Study. *Genes*. 13 (8), 1298. doi:10.3390/genes13081298.
- Raymond, V.M., Else, T., Everett, J.N., Long, J.M., Gruber, S.B. & Hammer, G.D. (2013) Prevalence of germline TP53 mutations in a prospective series of unselected patients with adrenocortical carcinoma. *The Journal of clinical endocrinology and metabolism*. 98 (1), E119-25. doi:10.1210/jc.2012-2198.
- Ribeiro, R.C., Sandrini, F., Figueiredo, B., Zambetti, G.P., Michalkiewicz, E., Lafferty, A.R., DeLacerda, L., Rabin, M., Cadwell, C., Sampaio, G., Cat, I., Stratakis, C.A. & Sandrini, R. (2001) An inherited p53 mutation that contributes in a tissue-specific manner to pediatric adrenal cortical carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*. 98 (16), 9330–9335. doi:10.1073/pnas.161479898.
- Timofeev, O. & Stiewe, T. (2021) Rely on Each Other: DNA Binding Cooperativity Shapes p53 Functions in Tumor Suppression and Cancer Therapy. *Cancers*. 13 (10), 2422. doi:10.3390/cancers13102422.
- Van der Auwera, G. A. et al. (2020) *Genomics in the cloud : using Docker, GATK, and WDL in Terra*. First edition. Sebastopol, CA: O'Reilly Media.