Idris Rasheed
Course: Bioinformatics
Instructor: Professor Ryan Koehler

Bioinformatics Analysis of TP53 Gene and its Relationship with Li-Fraumeni syndrome

Introduction

Purpose of this Paper
Given the role of the TP53 gene in the development of Li-Fraumeni syndrome (LFS), the purpose of this paper is to explore the genetic analysis of TP53 by looking at its gene expression and germline mutation.

Overview of Li-Fraumeni syndrome
Li-Fraumeni syndrome (LFS) is a rare hereditary cancer disorder that predisposes people to a wide variety of early-onset cancers such as soft tissue and bone sarcomas, breast cancer, brain tumors, and adrenocortical carcinoma (ACC) [1,11]. Consequently, LFS is also known as the sarcoma, breast, leukemia, and adrenal gland (SBLA) syndrome [1]. Breast cancer and sarcoma make up ~60% of the tumors associated with LFS. In 1969, at the National Cancer Institute, Dr. Frederick Li and Dr. Joseph Fraumeni, Jr. discovered LFS. While studying pediatric and familial cancers, they learned of four families with children and young adults suffering from multiple early-onset cancers. In 1982, British researchers were the first to publish a report on LFS after studying two families with a history of various forms of cancer in youths [11].

Causes of Li-Fraumeni syndrome
Researchers learned that the main culprit of LFS is a mutated TP53 gene that disrupts the p53 protein's job. TP53 is known as the "guardian of the genome" because of its role in cancer tumor suppression. The TP53 gene's primary function is to regulate DNA repair and cell division by instructing p53 protein to perform tumor suppression by preventing cells from dividing uncontrollably. Once the gene becomes mutated, p53 protein cannot reduce the speed of cell division, and malignant tumorigenesis occurs[1,11].

Genetic Analysis of TP53

Overview of TP53
As aforementioned, TP53 is known as the "guardian of the genome" because of its role in cancer tumor suppression on chromosome 17. Specifically, TP53 gives instructions for creating the p53 protein, which suppresses tumors by regulating cell division. The p53 protein located in the nucleus is responsible for repairing a damaged cell and causing a cell to self-destruct when the damaged cell cannot be repaired [1,5]. When a mutation occurs in TP53, not only can the p53 protein not perform its job, the alternative splicing of TP53 also leads to alternate promoters, which in turn causes transcript variants and isoforms. Outside of LFS, germline TP53 mutations contribute to 15% - 20% of all inherited cancers [5].

Germline Mutation
Since LFS is a phenotype related to TP53 germline mutations in cancer patients, we searched through the International Agency for Research on Cancer (IARC) for the types of mutation variants and the affected alleles. The majority of germline TP53 mutations are missense mutations, with silent mutations being the least. In other words, an alteration to one amino acid in a protein has a more substantial impact on germline mutation than a change in a single nucleotide base that does not alter the amino acid or amino acid functionality in TP53 [1].

Idris Rasheed
Course: Bioinformatics
Instructor: Professor Ryan Koehler

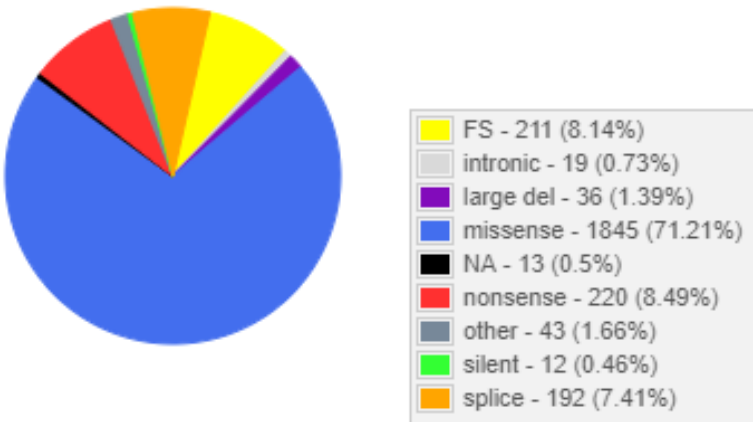


Figure 2 [1]

Single Nucleotide Variant Search

We now know missense mutations are the most common germline mutations in TP53, and we need to find single nucleotide variants (SNVs) related to LFS. Through the National Center for Biotechnology Information (NCBI) SNP database, we searched for missense SNVs associated with the TP53 gene and a pathogenic clinical significance for LFS by editing the filters [10]. After establishing the filters, we have nine missense SNVs: rs11540652, rs28934574, rs28934576, rs121913343, rs397514495, rs397516434, rs587782144, rs730882008, and rs121913343.

Idris Rasheed
Course: Bioinformatics
Instructor: Professor Ryan Koehler

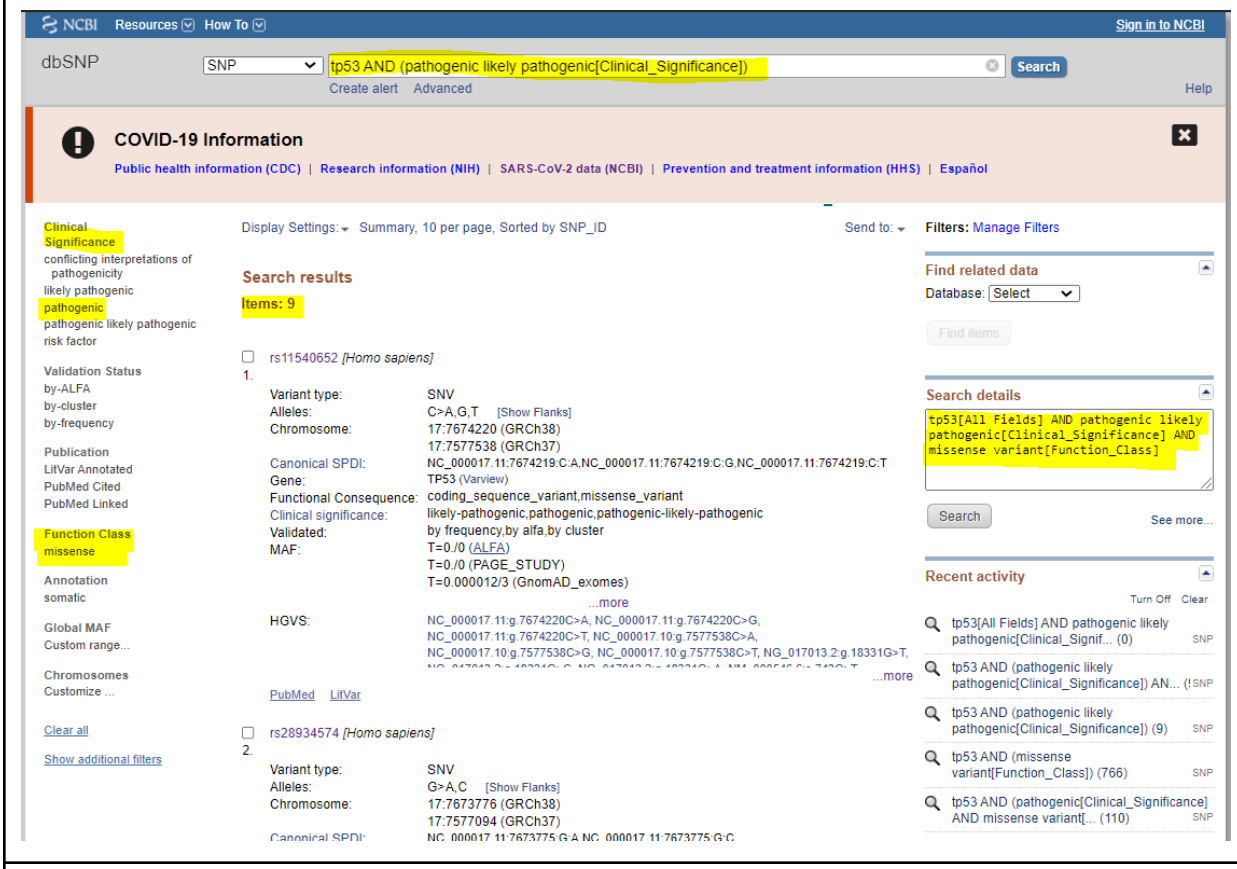


Figure 4 [10]

From there, we calculated their Combined Annotation Dependent Deletion (CADD) score on https://cadd.gs.washington.edu/. The CADD tool annotates and scores variants to analyze the correlation of allelic diversity and the pathogenicity of both coding and non-coding variants. We copied the chromosome position from the NCBI's Reference SNP (rs) Report and pasted it into CADD's SNV Lookup. For each SNV, we chose the reference and alternate allele with the highest Phred quality and CADD score. The Phred quality score measures the accuracy of the quality of the identification of the nucleobases. The higher the Phred score, the more accurate the quality of the allele is [2]. The CADD annotates SNVs and measures how harmful the variant is. As CADD decreases, the more deleterious the SNV is with the cutoff, usually within 10 - 20 [6].

SNV rs11540652 has the highest Phred score at 27.8 and a CADD score below 10. Based on its Phred score, rs11540652 has a high-quality nucleotide identification with an accuracy of approximately 99.9%. Its CADD score is far below 10 - indicating it is a harmful variant of LFS. The other SNVs, except for rs28934574 and rs730882008, have very low-quality scores, but every SNV is a deleterious variant.

Idris Rasheed
Course: Bioinformatics
Instructor: Professor Ryan Koehler

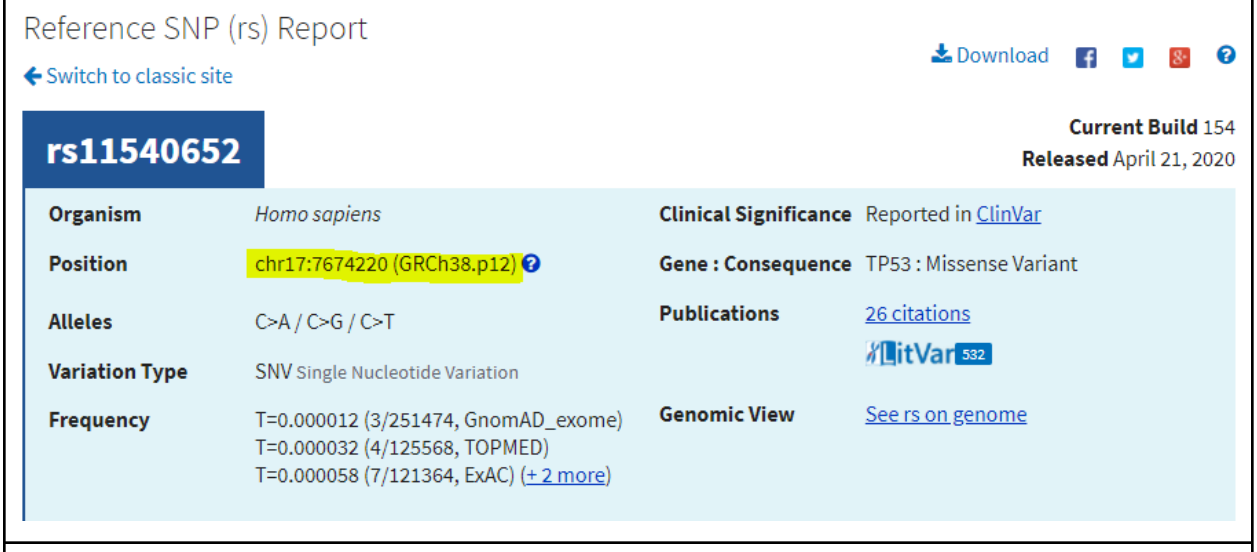


Figure 5 [6]

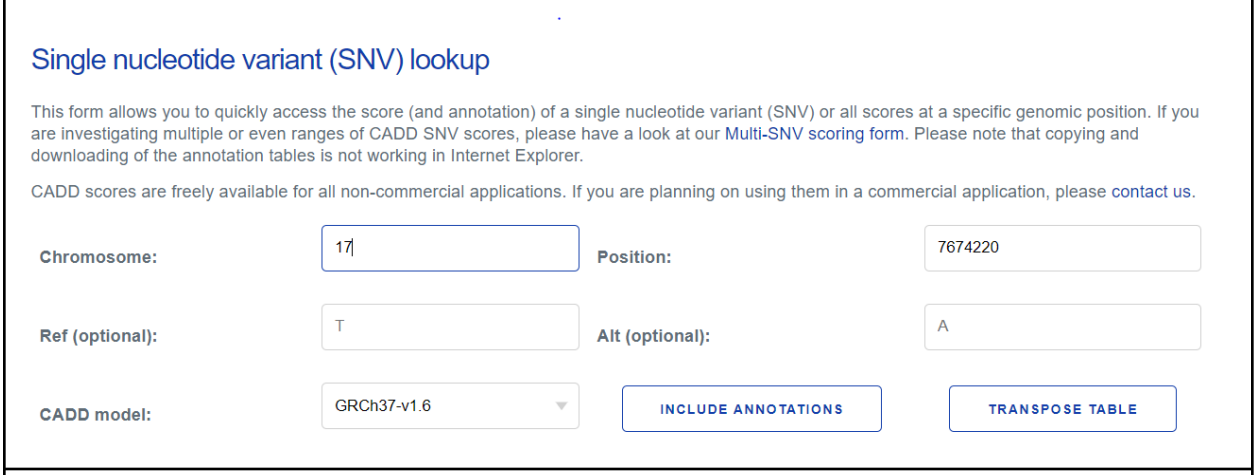


Figure 6 [6]

Index	SNV ID	Position	Ref Allele	Alt Allele	PHRED Score	Accuracy	CADD Score
1	rs11540652	chr17:7674220	T	C	27.8	0.998340413	4.009439
2	rs28934574	chr17:7673776	A	G	9.29	0.882239403	0.488046
3	rs28934576	chr17:7673802	T	A	2.229	0.4014506	-0.027558
4	rs121913343	chr17:7673803	G	A	4.015	0.603265471	0.085259
5	rs397514495	chr17:7675070	C	T	0.172	0.038830458	-0.407615
6	rs397516434	chr17:7670669	T	C	6.012	0.749504459	0.211657
7	rs587782144	chr17:7675139	A	G	2.616	0.452479985	-0.000488
8	rs730882008	chr17:7673775	C	T	7.568	0.824934729	0.327899
9	rs121913343	chr17:7673803	G	A	4.015	0.603265471	0.085259

Table 1

Idris Rasheed
Course: Bioinformatics
Instructor: Professor Ryan Koehler

Germline Mutations of Exon and Introns

The X-axis is the index for the introns and exons for TP53, and the Y-axis is the germline mutation frequency. The exon is the segment of a DNA or RNA molecule that contains information coding for a protein or peptide sequence. The introns are the segment of a DNA or RNA molecule that does not code for proteins. For example, you have Exon-Intron-Exon-Intron-Exon; the exons flank the introns. As demonstrated in the chart below, germline mutation occurs mainly on exon 5, exon 7, and exon 8. The germline TP53 gene mutations rarely appear from introns, and this is supported by Figure 3 [1].

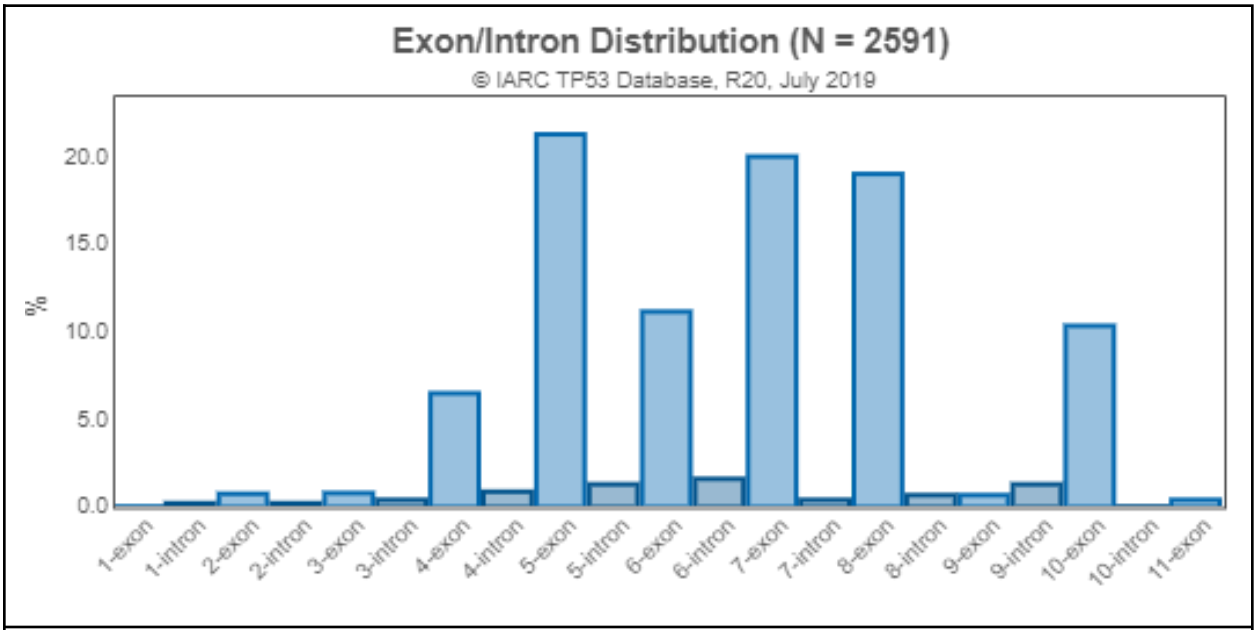


Figure 7 [1]

The X-axis is the codon number, and the Y-axis is the germline mutation frequency (%). A codon is a trinucleotide sequence of DNA or RNA that is assigned to a specific amino acid. As demonstrated in the table, codon 248 experiences the highest percentage of germline mutation at its location. Additionally, codon 273 has the second-highest proportion of mutation and codon 337 has the third-highest frequency. Germline mutations rarely appear on the lower end and higher end of the codon distribution [1].

Idris Rasheed
Course: Bioinformatics
Instructor: Professor Ryan Koehler

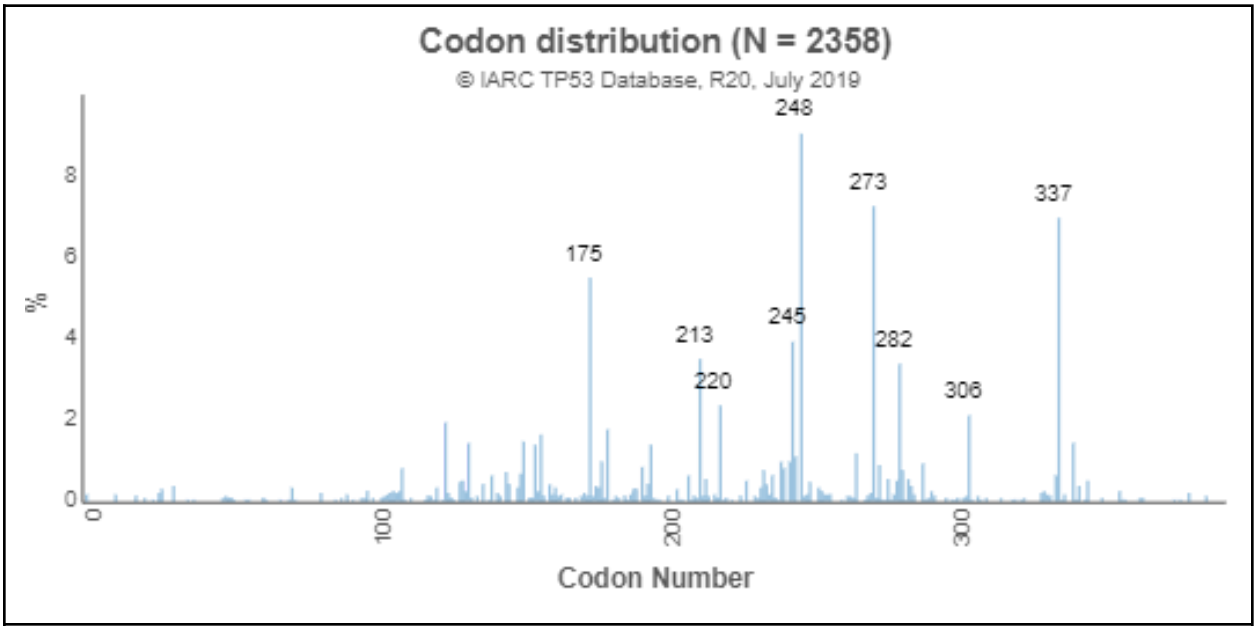
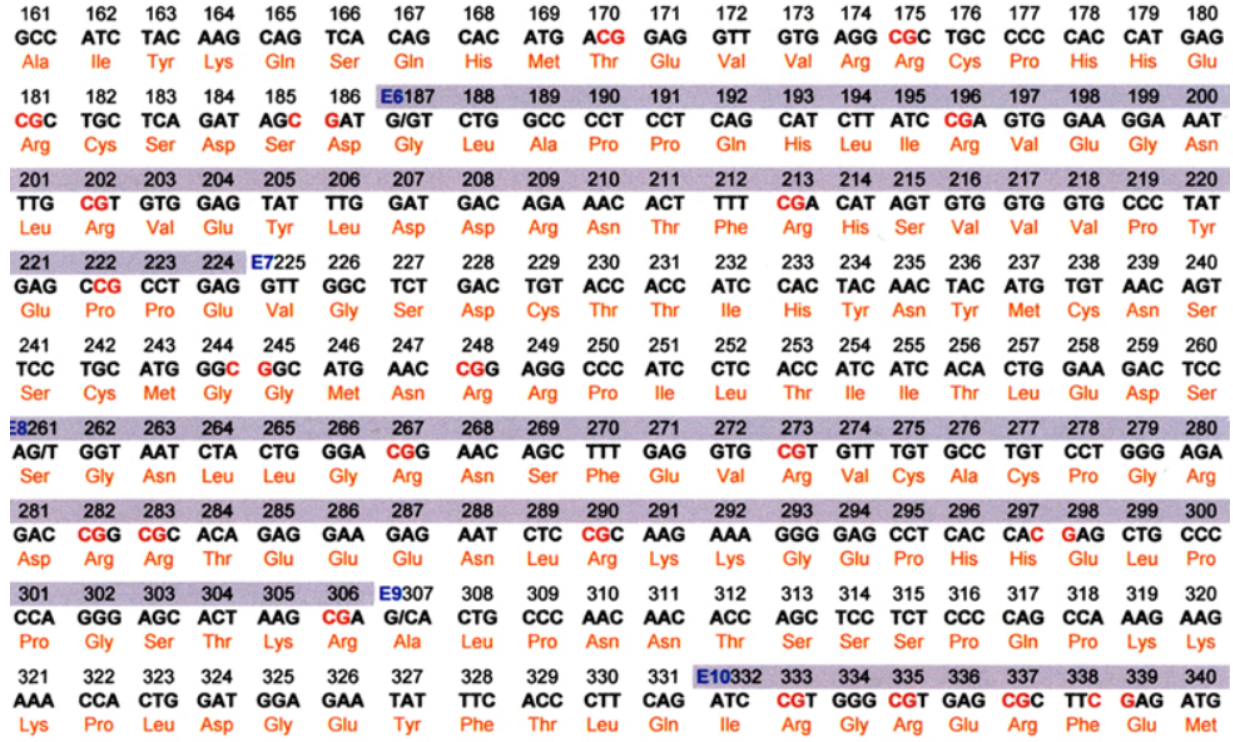


Figure 8 [1]

The image below shows the TP53 three-letter amino acids by codon. The blue represents the exon boundaries, and the red represents the CpG sites [1]. CpG sites are areas on a genome where a cytosine nucleotide is followed by a guanine nucleotide on the strand of nucleic acid. They are essential because DNA methylation occurs in their regions which plays a crucial role in gene regulation [3]. We see codons 248, 273, and 337 have CpG sites and are arginine amino acids.

Idris Rasheed
Course: Bioinformatics
Instructor: Professor Ryan Koehler



161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180
GCC	ATC	TAC	AAG	CAG	TCA	CAG	CAC	ATG	ACG	GAG	GTT	GTG	AGG	CGC	TGC	CCC	CAC	CAT	GAG
Ala	Ile	Tyr	Lys	Gln	Ser	Gln	His	Met	Thr	Glu	Val	Val	Arg	Arg	Cys	Pro	His	His	Glu
181	182	183	184	185	186	E6187	188	189	190	191	192	193	194	195	196	197	198	199	200
CGC	TGC	TCA	GAT	AGC	GAT	G/GT	CTG	GCC	CCT	CCT	CAG	CAT	CTT	ATC	CGA	GTG	GAA	GGA	AAT
Arg	Cys	Ser	Asp	Ser	Asp	Gly	Leu	Ala	Pro	Pro	Gln	His	Leu	Ile	Arg	Val	Glu	Gly	Asn
201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220
TTG	CGT	GTG	GAG	TAT	TTG	GAT	GAC	AGA	AAC	ACT	TTT	CGA	CAT	AGT	GTG	GTG	GTG	CCC	TAT
Leu	Arg	Val	Glu	Tyr	Leu	Asp	Asp	Arg	Asn	Thr	Phe	Arg	His	Ser	Val	Val	Val	Pro	Tyr
221	222	223	224	E7225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240
GAG	CCG	CCT	GAG	GTT	GGC	TCT	GAC	TGT	ACC	ACC	ATC	CAC	TAC	AAC	TAC	ATG	TGT	AAC	AGT
Glu	Pro	Pro	Glu	Val	Gly	Ser	Asp	Cys	Thr	Thr	Ile	His	Tyr	Asn	Tyr	Met	Cys	Asn	Ser
241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260
TCC	TGC	ATG	GGC	GGC	ATG	AAC	CGG	AGG	CCC	ATC	CTC	ACC	ATC	ATC	ACA	CTG	GAA	GAC	TCC
Ser	Cys	Met	Gly	Gly	Met	Asn	Arg	Arg	Pro	Ile	Leu	Thr	Ile	Ile	Thr	Leu	Glu	Asp	Ser
E8261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280
AG/T	GGT	AAT	CTA	CTG	GGA	CGG	AAC	AGC	TTT	GAG	GTG	CGT	GTT	TGT	GCC	TGT	CCT	GGG	AGA
Ser	Gly	Asn	Leu	Leu	Gly	Arg	Asn	Ser	Phe	Glu	Val	Arg	Val	Cys	Ala	Cys	Pro	Gly	Arg
281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300
GAC	CGG	CGC	ACA	GAG	GAA	GAG	AAT	CTC	CGC	AAG	AAA	GGG	GAG	CCT	CAC	CAC	GAG	CTG	CCC
Asp	Arg	Arg	Thr	Glu	Glu	Glu	Asn	Leu	Arg	Lys	Lys	Gly	Glu	Pro	His	His	Glu	Leu	Pro
301	302	303	304	305	306	E9307	308	309	310	311	312	313	314	315	316	317	318	319	320
CCA	GGG	AGC	ACT	AAG	CGA	G/CA	CTG	CCC	AAC	AAC	ACC	AGC	TCC	TCT	CCC	CAG	CCA	AAG	AAG
Pro	Gly	Ser	Thr	Lys	Arg	Ala	Leu	Pro	Asn	Asn	Thr	Ser	Ser	Ser	Pro	Gln	Pro	Lys	Lys
321	322	323	324	325	326	327	328	329	330	331	E10332	333	334	335	336	337	338	339	340
AAA	CCA	CTG	GAT	GGA	GAA	TAT	TTC	ACC	CTT	CAG	ATC	CGT	GGG	CGT	GAG	CGC	TTC	GAG	ATG
Lys	Pro	Leu	Asp	Gly	Glu	Tyr	Phe	Thr	Leu	Gln	Ile	Arg	Gly	Arg	Glu	Arg	Phe	Glu	Met

TP53 Coding Sequence Figure 9 [1]

Idris Rasheed
Course: Bioinformatics
Instructor: Professor Ryan Koehler

Gene Expression

We entered the TP53 gene in the UCSC Genome Browser's database and built box plots of 54 tissues from 17,382 samples and 948 donors. We did this to see which tissue will demonstrate the highest median Transcripts Per Million (TPM) in the TP53 gene. Cells - EBV-transformed lymphocytes have the highest median TPM with 72.95 TPM. The gene expression boxplots for the TP53 gene indicate many outliers on the upper threshold compared to the lower threshold. The brain tissues tend to have low TPM and outliers on the upper end [4].

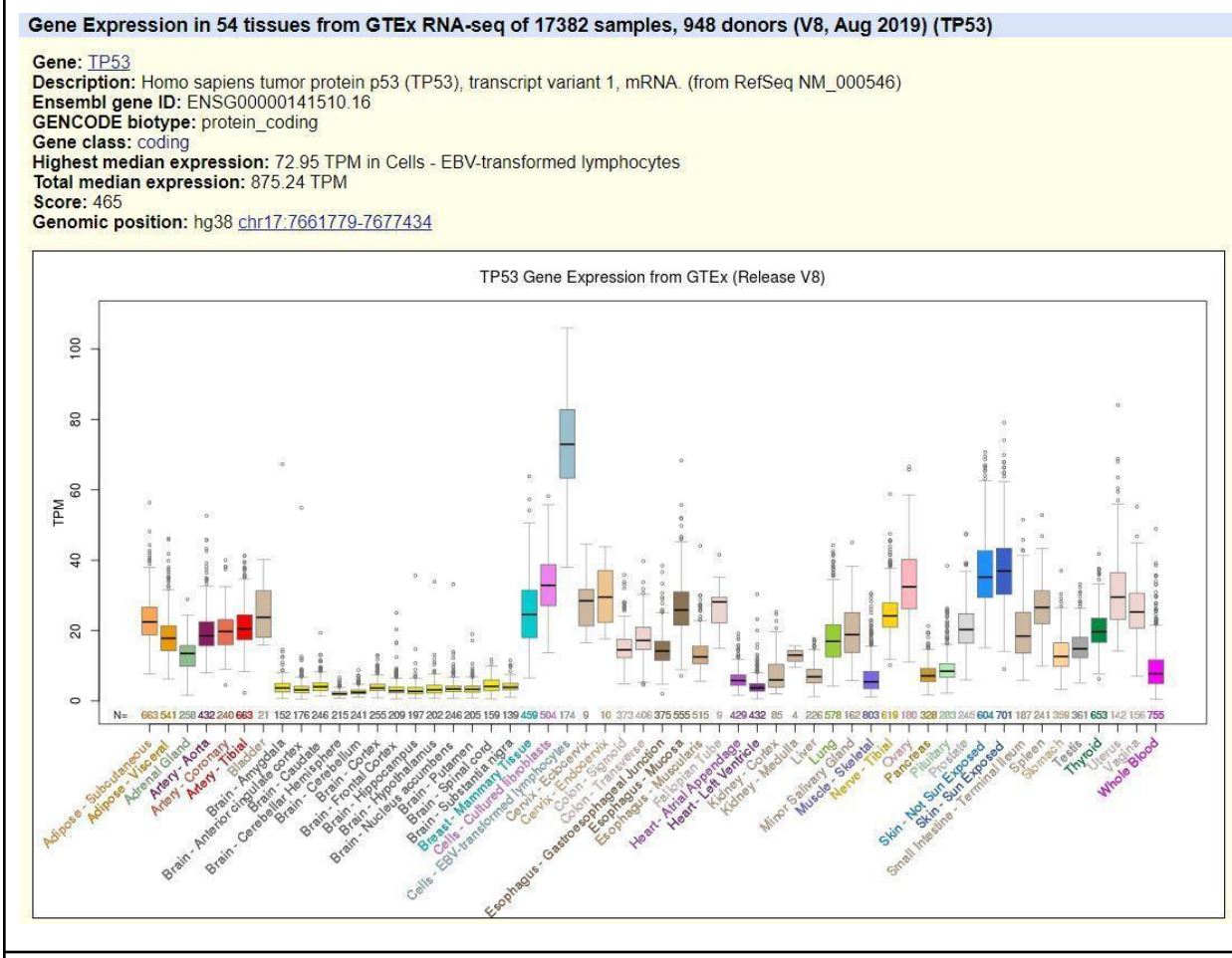


Figure 1 [4]

TP53 Gene Impact on Li-Fraumeni Syndrome

As noted before, LFS is a rare hereditary cancer disorder that predisposes people to a wide variety of early-onset cancers such as soft tissue and bone sarcomas, breast cancer, brain tumors, and adrenocortical carcinoma (ACC). Researchers now know that the main culprit of LFS is a mutated TP53 gene that disrupts the p53 protein's job of cellular division regulation. In what was explored before, understanding the TP53 gene provides insight into cancer causes in LFS.

Tumors associated with Li-Fraumeni syndrome (LFS)

Idris Rasheed
Course: Bioinformatics
Instructor: Professor Ryan Koehler

Approximately 82.21% of LFS-specific cancerous tumors are:

- **Breast Tumors:** The most common cancerous tumors for female LFS patients comprise 31.46% of tumors associated with the germline TP53 mutation. Breast cancer for LFS individuals occurs around age 33, usually before a woman enters menopause. In one clinical test, there was a 54% incidence rate for breast cancer in females by age 70.
- **Soft tissue sarcomas:** The most prevalent in children with LFS and appear in children under five years old. They make up 17%-27% of all the cancers from LFS patients, and males with soft sarcomas account for 22% of LFS cases, and females account for 15%.
- **Brain tumors**: Account for 11.15% of TP53 germline mutations and up to 14% of LFS cancers. In LFS patients, glioblastomas and astrocytomas are the most common brain tumors, and ependymomas, choroid plexus carcinomas, and supratentorial primitive neuroectodermal are the least common tumors. By the age of 70, 6% of women and 19% of men with LFS will get a brain cancer diagnosis.
- **ACC:** Occurs in 6%-13% of LFS patients and is typically detected before the age of five for children and before the age of 40 for adults. Specifically, the southern Brazilian *TP53* founder variant is related to the number of high-risk ACC cases in children. ACC makes up 55% of cancer cases in children with the southern Brazilian *TP53* founder variant and 23% of adults.
- **Bone Sarcomas:** Account for 9.3% of TP53 germline mutations and up to 16% of LFS cancers. LFS patients are typically diagnosed with bone cancer before age 30, and bone sarcomas affect 5% of women and 11% of men with LFS by age 70 [1,7].

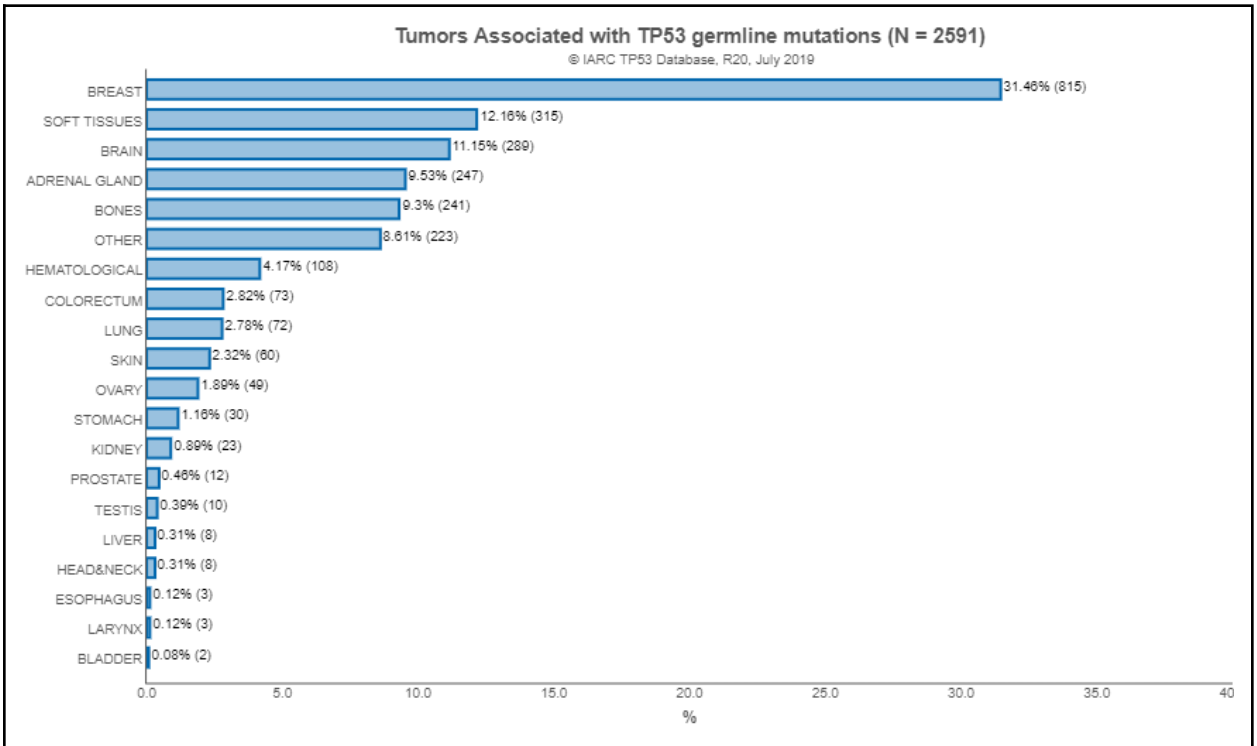


Figure 10 [1]

Conclusion

The TP53 gene on chromosome 17 is "the guardian of the genome" by ensuring stable tumor suppression, DNA repair, and cell division through the creation of p53 protein. Mutated TP53 changes a single amino acid in p53, causing the protein to lose control of cell division and unable to trigger programmed cell death in cells with mutated or damaged DNA. The mutation is linked to the formation of cancer tumors, commonly breast and brain tumors, soft tissue and bone sarcomas, and ACC. The inability to suppress tumors led researchers to connect LFS and to p53 germline mutations in a cell. LFS patients inherit one copy of a mutated TP53 gene from each cell, and in some cases, inherited from a parent. The parent has a 50% chance of passing an LFS-causative pathogenic variant to the offspring [1,5]. 75% of patients have this germline TP53 mutation. For patients without a parental history of LFS, the disease can occur from a TP53 pathogenic variant. Cancers associated with the TP53 gene have low survivability rates, and chemotherapy treatments are less effective than other cancers—also, patients who have cancer due to this gene experience high relapse rates than other cancer patients [7,8,9].

Idris Rasheed
Course: Bioinformatics
Instructor: Professor Ryan Koehler

References

1. Bouaoun L, Sonkin D, Ardin M, Hollstein M, Byrnes G, Zavadil J, Olivier M. IARC TP53 database. (2016, September). Retrieved May 29, 2021, from https://p53.iarc.fr/
2. G. (Ed.). (2021, February 1). Phred-scaled quality scores. Retrieved May 05, 2021, from https://gatk.broadinstitute.org/hc/en-us/articles/360035531872-Phred-scaled-quality-scores
3. Jang, H., Shin, W., Lee, J., & Do, J. (2017, May 23). CpG and NON-CPG methylation IN epigenetic gene regulation and brain function. Retrieved May 30, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5485512/
4. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. UCSC Genome Browser. (2006, June 12). Retrieved May 29, 2021, from https://genome.ucsc.edu/
5. Nichols, K. E., & Malkin, D. (2015, June 22). Genotype Versus Phenotype: The Yin and Yang of Germline TP53 Mutations in Li-Fraumeni Syndrome. Retrieved from https://ascopubs.org/doi/10.1200/JCO.2015.61.5757
6. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. (2021, February 22). Retrieved May 30, 2021, from https://cadd.gs.washington.edu/
7. Robles, A., Jen, J., & Harris, C. (2016, September 1). Clinical outcomes of tp53 mutations in cancers. Retrieved May 29, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5008065/
8. Schneider, K. (2019, November 21). Li-Fraumeni syndrome. Retrieved May 30, 2021, from https://www.ncbi.nlm.nih.gov/books/NBK1311/
9. Sorrell, A., Espenschied, C., Culver, J., & Weitzel, J. (2013, February 17). Tumor protein p53 (TP53) testing and LI-FRAUMENI syndrome: Current status of clinical applications and future directions. Retrieved April 18, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3627545/
10. Gene. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – Retrieved May 30, 2021, from https://www.ncbi.nlm.nih.gov/snp/
11. What is lfs? (2021, April 15). Retrieved April 18, 2021, from https://www.lfsassociation.org/what-is-lfs/