A Bioinformatics Overview of Publicly Available Fanconi Anemia Mutation Data

Lazo B. Ali[1]

Supervised By: Dr. Joseph Rebehmed[1]

[1]Department of Computer Science and Mathematics
School of Arts and Sciences
Lebanese American University

Abstract

Fanconi anemia (FA) is a rare genetic disorder characterized by genomic instability, congenital abnormalities, and cancer predisposition, arising from mutations in a network of >22 DNA repair genes. While major variant repositories such as ClinVar and the Leiden Open Variation Database (LOVD) curate FA-associated mutations, inconsistencies in transcript usage, formatting, and classification standards have impeded cross-platform analyses. Here, we present a harmonized meta-dataset of 91,720 GRCh37-aligned FA variants derived from both sources. Using a reproducible pipeline, we converted all entries into a standardized VCF format, resolved redundant and malformed entries, and functionally annotated variants with ANNOVAR and dbNSFP v4.2a. We identified 15,096 overlapping mutations between databases, yet over 20% of LOVD records lacked ClinVar cross-references, and >8,000 nonsynonymous SNVs exhibited conflicting pathogenicity assertions. Structural mapping revealed non-random distribution of variants, with hotspot localization and mutation-intolerant regions in essential FA genes. Surprisingly, several synonymous SNVs were consistently labeled pathogenic, likely due to effects on splicing or RNA stability. We further show that frequently reported variants—especially exonic stopgains—dominate pathogenic calls across key genes such as FANCI. Our analysis highlights the pitfalls of overreliance on in silico predictors and underscores the need for curated, multi-omic data integration. This harmonized FA mutation compendium provides a transparent framework for genotype–phenotype interpretation and offers a model for reconciling specialized and clinical archives across rare genetic diseases.

1. Introduction

1.1 Clinical and Genetic Features of Fanconi Anemia

Fanconi anemia (FA) is a rare inherited bone marrow failure syndrome and cancer-predisposing condition, occurring in roughly 1 per 100,000–250,000 births worldwide (Bhandari et al., 2025a). Clinically, FA is characterized by progressive pancytopenia (aplastic anemia) often manifesting in childhood, a spectrum of congenital abnormalities, and markedly increased cancer susceptibility (Bhandari et al., 2025b; Fiesco-Roa et al., 2019). Common physical anomalies involve multiple organ systems – for example, skeletal malformations (especially radial ray defects of the thumb/forearm), renal and cardiac defects, growth retardation, and skin pigmentation changes – often overlapping with features of the VACTERL-H association and PHENOS syndrome that are frequently observed in FA patients (Fiesco-Roa et al., 2019). FA is genetically heterogeneous: pathogenic variants in more than 22 different genes (FANCA, FANCB, FANCC, etc.) have been identified as causes, defining complementary groups A, B, C and so on. These FANC genes collectively encode the proteins of the FA/BRCA DNA repair pathway, a crucial genome maintenance mechanism (Fiesco-Roa et al., 2019).

Inheritance is usually autosomal recessive (with the exception of the X-linked FANCB), and most patients carry biallelic loss-of-function variants in a given FA gene. A clinical diagnosis of FA is classically confirmed by a positive chromosome breakage test – patient cells exhibit hypersensitivity to DNA cross-linking agents (e.g. diepoxybutane or mitomycin C) with numerous chromosomal breaks and radial formations in vitro (Ameziane et al., 2012; Chowdhry et al., 2014) (See Figure 1).

Molecular genetic testing is then used to pinpoint the affected gene and specific mutations, which is essential for family counseling and emerging therapeutic decisions (e.g. stem cell transplant donor selection, preimplantation genetic diagnosis). Nevertheless, the phenotypic manifestation of FA exhibits significant heterogeneity, and the physician's interpretation of such manifestation substantially influences the differential diagnosis of FA. Notably, patients with minor congenital anomalies tend to receive delayed diagnoses compared to those with major congenital anomalies. Additionally, individuals lacking hematological symptoms encounter diagnostic delays, even in the presence of notable congenital abnormalities (Auerbach, 2009).

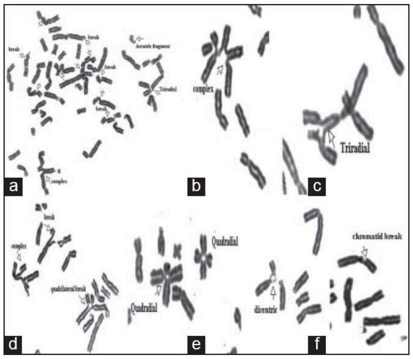

Figure 1. Mitomycin-C-induced chromosomal aberrations in Fanconi anemia cells.

1.2 DNA Repair Mechanism and Pathogenic Mutations in FA

At a cellular level, FA is fundamentally a disorder of DNA repair. The FA pathway's normal role is to maintain genomic stability by orchestrating the repair of DNA interstrand crosslinks (ICLs) – a particularly toxic type of DNA damage that prevents strand separation during replication (Muniandy et al., 2010). Upon ICL detection (recognized by FANCM and associated sensor proteins), the FA "core complex" of proteins (including FANCA, FANCB, FANCC, FANCE, FANCF, FANCG, FANCL, FANCM, etc.) is recruited to the site of damage. This core complex functions as an E3 ubiquitin ligase that monoubiquitinates the FANCD2/I proteins (together forming the ID2 complex) .

Monoubiquitinated FANCD2–FANCI then localizes to chromatin at stalled replication forks and coordinates downstream repair processes by interacting with other DNA repair factors. These downstream effectors include nucleases and homologous recombination proteins (many with their own gene names in the FA complementation list, such as FANCD1 which is BRCA2, FANCN which is PALB2, and others) that collectively resolve the crosslink and restore DNA integrity (Jacquemont & Taniguchi, 2007).

Functional implications of FA-associated mutations are severe: a biallelic loss of function at any step of this pathway can abrogate ICL repair. Cells lacking a functional FA pathway accumulate DNA breaks, chromosome fusions, and gross chromosomal rearrangements, especially after exposure to endogenous aldehydes or exogenous cross-linking agents. This genomic instability drives the clinical phenotype – bone marrow progenitor cells undergo attrition (leading to aplastic anemia) and patients are highly prone to malignancies such as acute myeloid leukemia and squamous cell carcinomas. Structural implications of the mutations are evident at both protein and chromosomal levels (Deans & West, 2011).

Many FA gene mutations produce truncated or misfolded proteins that cannot assemble into the large FA repair complexes, or fail to ubiquitinate the ID2 complex, thereby interrupting the cascade of repair (Mamrak et al., 2017). The downstream result is the hallmark chromosomal fragility of FA cells, often visualized as chromatid breaks and radial chromosome structures in metaphase spreads, as shown in Figure 1. This unique cellular phenotype (reflected in the positive DEB/MMC stress test) is a direct consequence of the defective FA/BRCA repair mechanism and remains a defining diagnostic and research feature of the syndrome.

1.3 Pathways, Networks, and Genotype-Phenotype Correlations

Advances in bioinformatics and systems biology have begun to shed light on the complex molecular landscape of FA, complementing traditional genetic studies. The FA/BRCA DNA repair pathway itself has been well delineated in pathway databases like KEGG and Reactome, which map the sequence of protein interactions from ICL recognition to resolution. These curated pathway analyses underscore how tightly integrated the FA proteins are with other genome maintenance processes. For instance, the monoubiquitination of FANCD2/FANCI by the FA core complex links the FA pathway to the broader homologous recombination and DNA damage response networks.

Beyond this canonical pathway mapping, protein–protein interaction (PPI) network studies have identified broader interactomes of FA proteins. Using tools such as the STRING database, researchers have constructed FA-centered interaction networks to find key hub proteins and co-regulated pathways that might contribute to the disease phenotype.

For example, a proteomic study by Hou et al., 2020, of bone marrow cells in FA patients integrated with STRING PPI analysis revealed a subnetwork of deregulated proteins and highlighted a set of 12 hub proteins potentially driving the progression from bone marrow failure to leukemia . Such network analyses suggest that FA mutations may perturb not only DNA repair but also intersecting pathways – for instance, cell cycle regulation, immune response, and oxidative stress pathways – through direct or indirect protein interactions. Concurrently, pathway enrichment analyses (using Gene Ontology and KEGG) on omics data from FA cells have illuminated which biological processes are most affected. Notably, unbiased proteome and transcriptome profiling have shown that FA cells exhibit differential expression of proteins involved in cytoskeletal organization, immune signaling, and metabolic processes. In one study, down-regulated proteins in an FA patient's cells were enriched in pathways related to actin cytoskeleton regulation and glucose metabolism, while up-regulated proteins were enriched in stress response pathways.

These findings hint that the consequences of FA gene mutations extend into diverse cellular functions, explaining some of the phenotypic variability (such as developmental abnormalities and cellular hypersensitivity) observed in patients. Crucially, bioinformatics platforms have also enabled large-scale genotype–phenotype correlation studies in FA. The Fanconi Anemia Mutation Database (FAMD) – a comprehensive repository of reported FA mutations – has facilitated meta-analyses of case data from around the world. By aggregating hundreds of variants across the FA genes, FAMD and related efforts allow researchers to identify patterns linking specific genotypes to clinical outcomes. For example, a recent review of over 1100 FA cases found that the presence of certain mutations correlates with distinct phenotypic patterns: patients with null (truncating) mutations in FANCD2 or in the downstream "ID2 complex" genes tended to have a higher frequency of congenital anomalies (e.g. renal malformations, microcephaly, short stature), whereas those with hypomorphic variants had somewhat milder physical presentations. Likewise, variants in different complementation groups can lead to different cancer risks; for instance, biallelic mutations in FANCD1 (BRCA2) or FANCN (PALB2) are known to cause early-onset and more severe malignancies, highlighting a genotype–phenotype gradient even within FA (Fiesco-Roa et al., 2019).

Such insights have only been possible through the integration of genotype data on a global scale. The FAMD, established in 1998, continues to compile novel FA mutations and now utilizes LOVD3.0 to systematically capture variant pathogenicity, population origin, and associated clinical information. By querying this database, researchers can draw connections between a patient's specific mutation profile and their clinical course, informing both prognosis and potential personalized therapies.

While the Fanconi Anemia Mutation Database, hosted on the Leiden Open Variant Database (FAMD/LOVD) (Fokkema et al., 2021) has long served as a key repository for curated FA variants, the U.S. National Center for Biotechnology Information's ClinVar (Landrum et al., 2014) archive now encompasses an even broader collection of submissions—aggregating thousands of FA-related variants from diagnostic laboratories, research groups, and clinical testing services worldwide (ClinVar). *Despite this wealth of data, to date there has been no comprehensive study reconciling and harmonizing FA mutations across FAMD/LOVD and ClinVar.* Such an integrative meta‑analysis—linking variant annotations, pathogenicity classifications, etc… across both platforms—remains a critical next step toward improving global genotype–phenotype correlations and enabling more consistent, evidence‑based interpretation of FA gene variants.

The current study aims to fill this critical gap by systematically integrating and comparing FA-related variants from both the Rockefeller University's Fanconi Anemia Mutation Database (FAMD/LOVD) and the NCBI ClinVar archive. Through standardized normalization to the GRCh37 and GRCh38 reference genome, conversion into a unified variant call format (VCF), and comprehensive functional annotation of GRCh37-mapped variants using ANNOVAR, we provide a harmonized dataset that enables direct comparison of variant content, classification discrepancies, and annotation completeness. By quantifying overlaps, identifying unreferenced but concordant mutations, and highlighting divergent pathogenicity assessments, this work lays the foundation for a more unified and transparent interpretation framework. Finally, we look at most reported pathogenic mutations across key players in the FA pathway to highlight how pathway knowledge aids clinicians in assessing pathogenicity. Ultimately, our analysis offers insights into how legacy and modern databases can complement each other to support more accurate diagnostics, research, and clinical decision-making in Fanconi anemia.

2. Results and Discussion

2.1 Data Overview

"Database Inconsistency Hinders Analysis"

A major obstacle in reconciling Fanconi anemia variant data between the Rockefeller LOVD (FAMD) and ClinVar is the lack of consistent, up-to-date cross-referencing with standardized reference sequences. Many mutations in the LOVD are annotated using outdated transcript versions. For instance, while the most current transcript for FANCA is NM_000135.4, the LOVD reports mutations on the earlier NM_000135.3. This discrepancy complicates direct comparison with ClinVar entries, which generally follow updated transcript annotations.

Moreover, the LOVD does not provide variants in widely adopted formats such as VCF, whereas ClinVar supports structured VCF downloads. Thus, a rigorous time consuming reconciliation process is needed before inter-database analysis.

After full reconciliation (see Methods), a total number of 91720 variants were retained, of these 10 were duplicated within the *same* database, and 15096 of them were mentioned in both databases. Interestingly, despite this significant overlap the vast majority of mutations reported on the LOVD (19904) were not cross referenced with ClinVar, despite a `ClinVar ID` column being present. ClinVar did not cross-reference the LOVD at all. Additionally, despite both databases claiming to abide by the ACMG/AMP guidelines (Richards et al., 2015), the classification of mutations in the LOVD did not follow standards. Thus, manual relabelling of consequences was necessary (See Methods for exact method). The distribution of labels before and after manual relabelling can be seen in Figure 2.

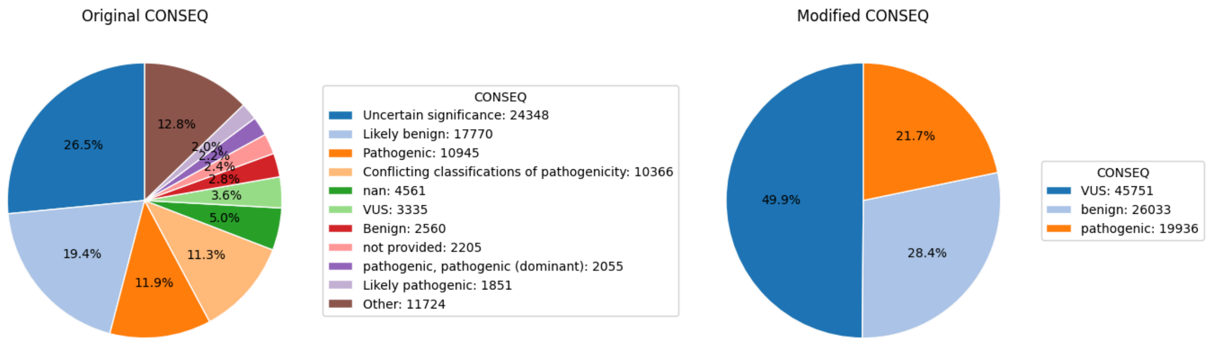

Figure 2. Distribution of mutation consequences, Original vs Modified.

Going back to the overlap between databases. Of the 15096 overlapping records, only 12579 of them had concordant pathogenicity labels. 2508 of common variants had semi-conflicting labels, where one database reported it as either pathogenic or benign and the other reported it as a variant of unknown significance (VUS) after relabelling. The remaining 9 records contained totally conflicting pathogenicity labels and were therefore dropped from downstream analysis. A summary of this overlap can be seen in Figure 3. It is worth noting that all of these totally conflicting mutations were in the BRCA1, BRCA2, and FANCA genes. Genes among those with the highest number of records.

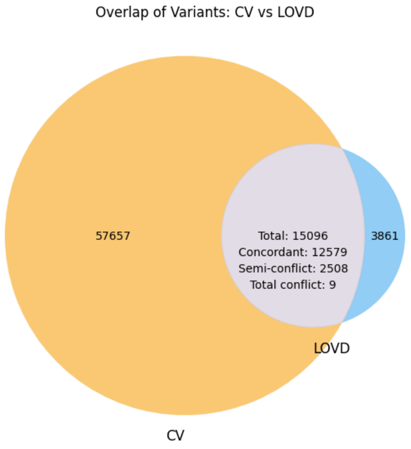

Figure 3. Venn diagram of overlapping records between databases.

Visually, it is not clear whether the proportion of pathogenic variants per gene scales with the total number of variants reported, as seen in Figure 4. For example, UBE2T, FANCG, FANCL and RAD51C each harbor a higher percentage of pathogenic mutations than many genes with far larger variant catalogs. This observation is contradicted by correlation analysis: both Pearson's *r:* 0.647 and Spearman's *ρ:* 0.485 between total variant count and percent pathogenic are statistically significant, confirming a non-random, albeit weak, relationship between reporting depth and pathogenicity rate.

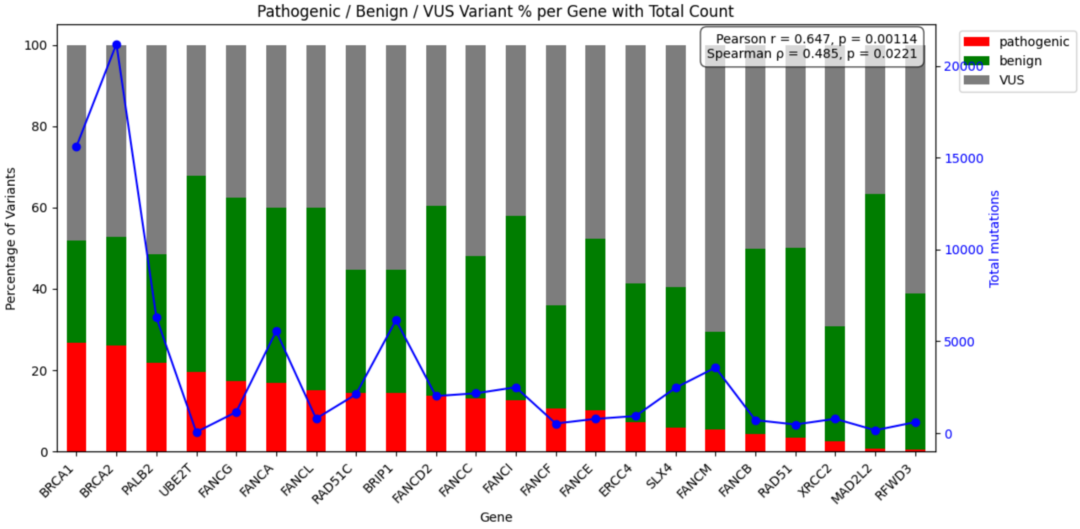

Figure 4. Relationship between unique variants per gene and pathogenicity rate. Showing weak but statistically significant correlation.

2.2 Annotation Results

"Mutations are Localized Away from Certain Regions"

After merging the data and refactoring conflicting classifications, each variant was annotated with ANNOVAR (see methods). Of those variants with a functional annotation, the vast majority of them were predictably exonic (Figure 5.a). The rest of the functional annotations are led by intronic mutations. Most strikingly, despite our rigorous labelling, the vast majority of splicing mutations and those classified as exonic;splicing are pathogenic. Clearly highlighting that splicing mutations are the least tolerated.

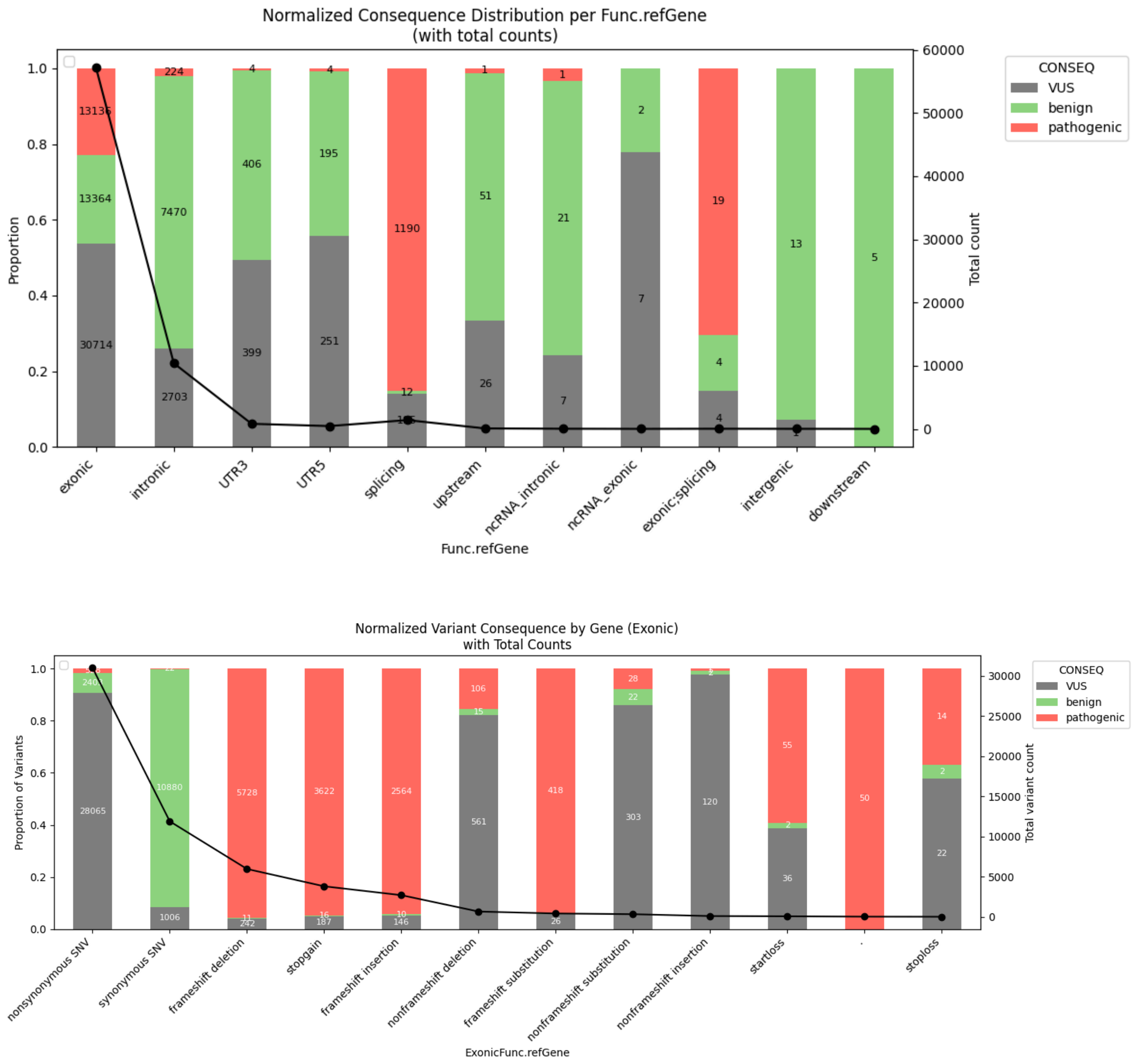

Figure 5. (a) Distribution of Mutations by Functional Type and Pathogenicity (b) Distribution of Exonic Mutations by Protein Consequences.

Of those Exonic mutations, the vast majority are nonsynonymous SNVs of which are mostly classified as VUS (Figure 5.b). This is presumably partly due to our labelling and partly due to the central role of the FA pathway.

First, while only about 500 of the nonsynonymous SNV mutations are labelled as Pathogenic in our final dataset, at least 8900 of them were labelled with “Conflicting classifications of pathogenicity” in the original dataset. Thus, given that a mutation causes a nonsynonymous amino acid substitution, it is not conclusive to whether that mutation will be pathogenic or not. This sheds important light on the fact that Genomic level analysis and protein-change consequences are not enough to conclude the impact of a mutation in FA associated genes.

On the other hand, this might be due to the central role of the FA pathway. It is plausible to assume that unambiguously pathogenic nonsynonymous SNVs are not tolerated early on in embryonic development, hence later sequencing and mutation reporting is not possible. Of nonsynonymous mutations, those that do allow for development up to a point where the patient’s DNA is sequenced, they are not necessarily pathogenic, hence the conflicting classification.

To investigate this further, all nonsynonymous SNVs along with their pathogenicity were plotted on the exons, as seen in Figure 6. As Figure 6 shows and we know from Figure 5, the vast majority of these nonsynonymous SNVs are either of unknown consequence or benign, highlighting the fact that these loci tolerate mutations to some extent. While Figure 6 only shows the plots for 2 genes, most genes displayed mutation localization away from certain exons/regions. The lack of reported pathogenic mutations in these regions further strengthen the hypothesis that these regions do not tolerate any mutations early on in embryonic development and thus subsequent sequencing of these variants was not possible. This necessitates further investigation into the structural and functional role of these loci and how it relates to the function of the respective genes.

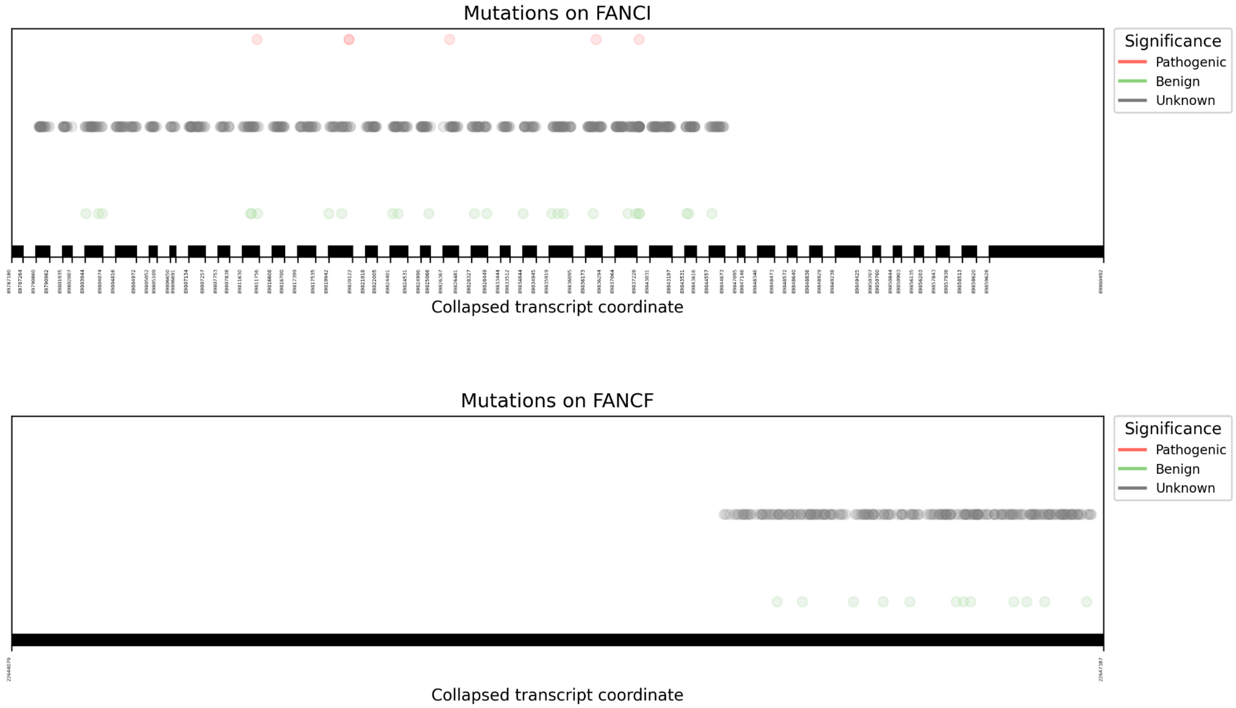

Figure 6. Localization of Synonymous SNVs Across 2 genes.

Also notably, the vast majority of exonic mutations that result in frameshifts or stop gains are pathogenic, this point will be further highlighted in section 2.4. While the story with non frameshift deletions and substitutions are the same as nonsynonymous SNVs. Finally, turning to the domains of pathogenic mutations that fell onto InterPro domains after annotation, the most mutated domains seem to be, as expected, related to DNA repair and protein-protein interactions, however local position within domain seems to be important, as not all mutations within these domains are pathogenic. A summary can be seen in Figure 7.

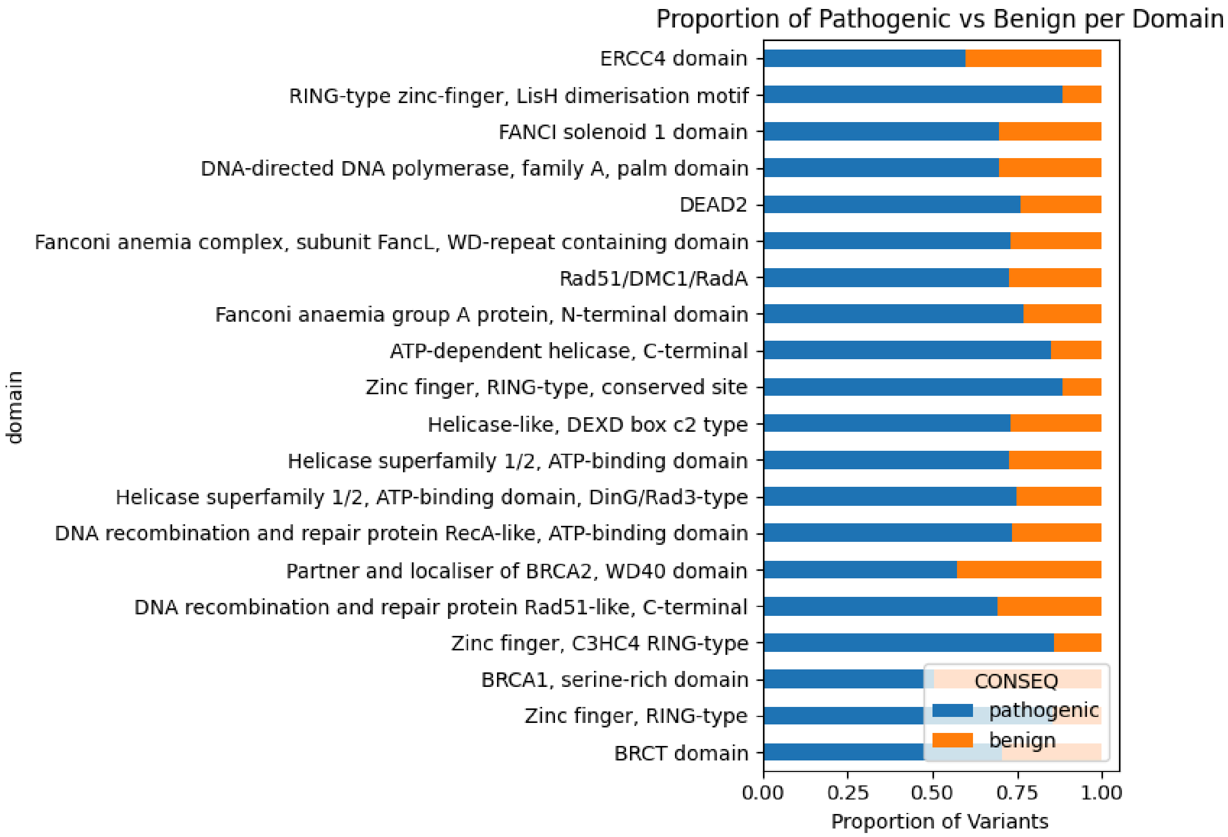

Figure 7. Top 10 InterPro domains that contain the most pathogenic mutations.

2.3 Investigating the Presence of Pathogenic Synonymous SNVs

"Genome Level Analyses May not be Enough"

As we have seen with nonsynonymous SNVs, it appears that an amino acid change is not sufficient to conclude pathogenicity. Another feature of the dataset that highlights the insufficiency of genomic level analyses for conclusive classification is seen in the synonymous SNV bar in Figure 4.b. While, predictably, the vast majority of synonymous SNVs, which are genomic changes that do not lead to protein level changes, are benign, 24 of them are classified as pathogenic despite our stringent labelling. In the original data, 754 synonymous SNV exonic mutations were labelled as pathogenic one way or another. Table 1 shows that these mutations are restricted to only a few genes.

Gene	BRCA2	BRCA1	FANCA	PALB2	FANCC	BRIP1	FANCD2
Count	6	5	5	4	2	1	1

Table 1. Genes that contain pathogenic yet synonymous SNVs.

The pathogenicity of these mutations is worth further investigation. We hypothesize that these mutations are pathogenic due to their structural effect on the transcript level. They either cause a mRNA splicing disruption, or cause the mRNA to have an unstable secondary structure. Additionally, while due to codon degeneracy the transcripts lead to the same amino acids, different codons can code for the same amino acid under different kinetics. One final possibility is that these sites overlap with microRNA regulatory motifs. Further investigation is indeed necessary.

2.4 Investigating Frequency of Mutation Reports

“Most Common Mutations are Stopgain”

Thus far we have talked about unique mutations, i.e those which represent a unique nucleotide change at a unique position. We will now turn our attention to how many times a specific mutation was reported. Overall, the vast majority of mutations were reported only a few times as seen in Figure 8.

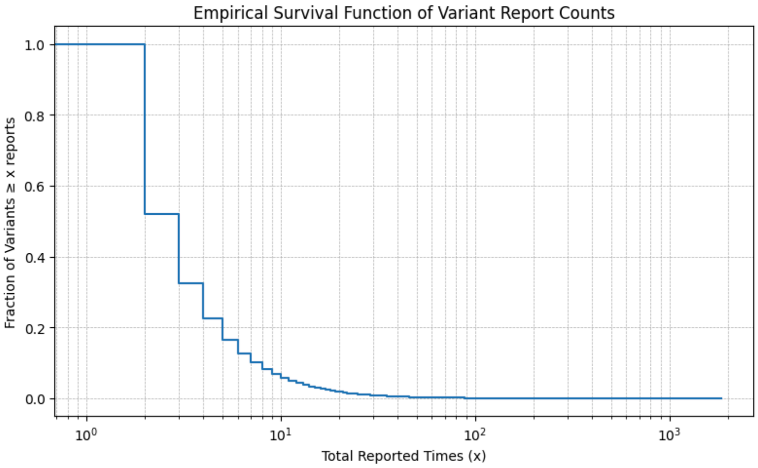

Figure 8. ECDF of Variant Report Counts showing that at least 80% of mutations were reported less than 10 times.

While the trend reinforces the idea that FA mutations are rare, it is worth investigating the top 10 most reported mutations for each gene, as they contain the most likely mutations a clinician may encounter. Overall, the type of pathogenic mutations most reported in the two databases is summarized in Figure 9. As we can observe, the majority are exonic and of those exonic the majority are stopgain mutations.

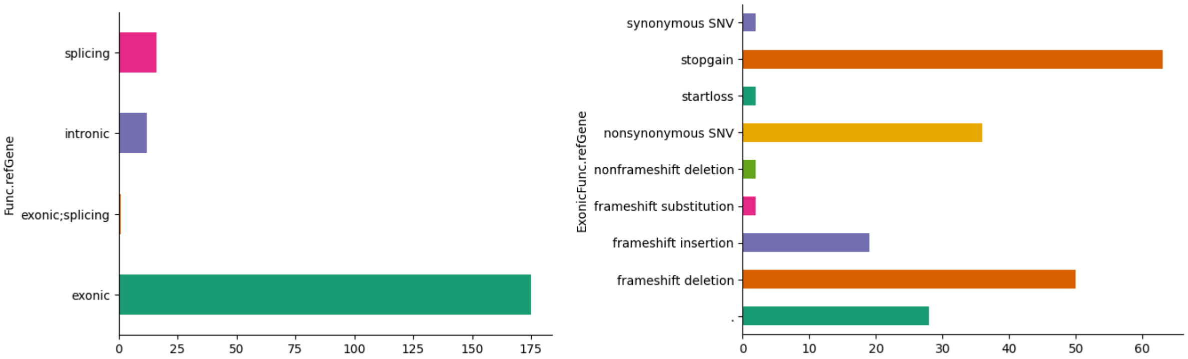

Figure 9. Mutation types for top 10 reported pathogenic mutations per gene.

If we take for example the top reported mutation on FANCI, a major player in the FA pathway, we see the mutation reported in Table 2.

ID	Number of Reports	Total Reported	% of Total	Type	Typeof Exonic	AA Change
CV.FANCI00002	15	3835	0.39%	exonic	stopgain	p.R1285X

Table 2. Top reported mutation FANCI is a stopgain.

Structurally, this mutation results in the early termination of translation and the resulting protein will lack the C-terminal domain highlighted in Figure 10. The literature suggests that the C-terminus region of FANCI contains a leucine zipper motif necessary for protein-protein and protein-DNA interactions (Yuan et al., 2009). The lack of this domain thus hinders the ability of monoubiquitinated FANCD/I to further proceed in the pathway. It remains to be investigated why stop loss mutations are more tolerated than nonsynonymous SNVs and make it past development.

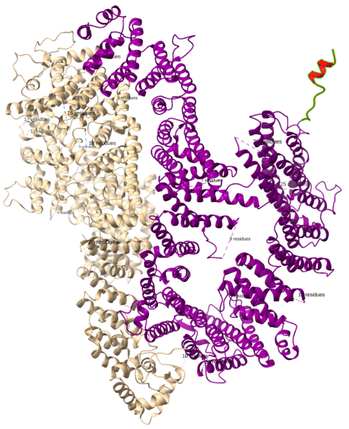

Figure 10. Top reported FANCI (Purple) mutation results in the improper translation of the C-terminal domain highlighted (Red). PDB: 3S4W (Joo et al., 2011)

2.5 Functional Impact Scores and Mutation Consequence

“What is a Pathogenic Mutation?”

The ACMG/AMP guidelines (Richards et al., 2015) establish a five‑tier framework—Pathogenic, Likely Pathogenic, Uncertain Significance, Likely Benign and Benign—for Mendelian variant interpretation, based on 28 evidence codes spanning population frequency, computational predictions, functional data, segregation, and other categories; each code carries a defined strength (very strong, strong, moderate or supporting) and must be combined according to specified rules to reach a final classification.

Both ClinVar and the LOVD fully adopt the ACMG/AMP five categories for germline variant assertions, recording each submitter’s interpretation. On ClinVar conflicts are reported only when submitters’ calls differ among the ACMG/AMP categories, and ClinVar allows expert panels (e.g., ClinGen VCEPs) to issue consensus classifications that carry top confidence.

Examining Figure 11 in light of these standards shows that variants ClinVar or LOVD label “pathogenic” almost invariably carry extreme in silico warnings: very high CADD-phred scores (which rank a variant’s predicted deleteriousness on a PHRED‑like scale derived from genome-wide SNV distribution) , high REVEL ensemble scores (0–1, integrating 13 individual predictors to estimate missense pathogenicity), low SIFT scores (<0.05 indicates predicted deleterious impact based on evolutionary conservation), and high PolyPhen-2 scores (0–1, reflecting predicted disruption of protein structure/function).

These computational metrics which align tightly with clinical assertions are supposed to serve only as supporting evidence under ACMG/AMP (PP3/BP4) rather than as standalone proof of pathogenicity. However, such a tight fit requires further investigation and reexamination of submission records, which are available on ClinVar.

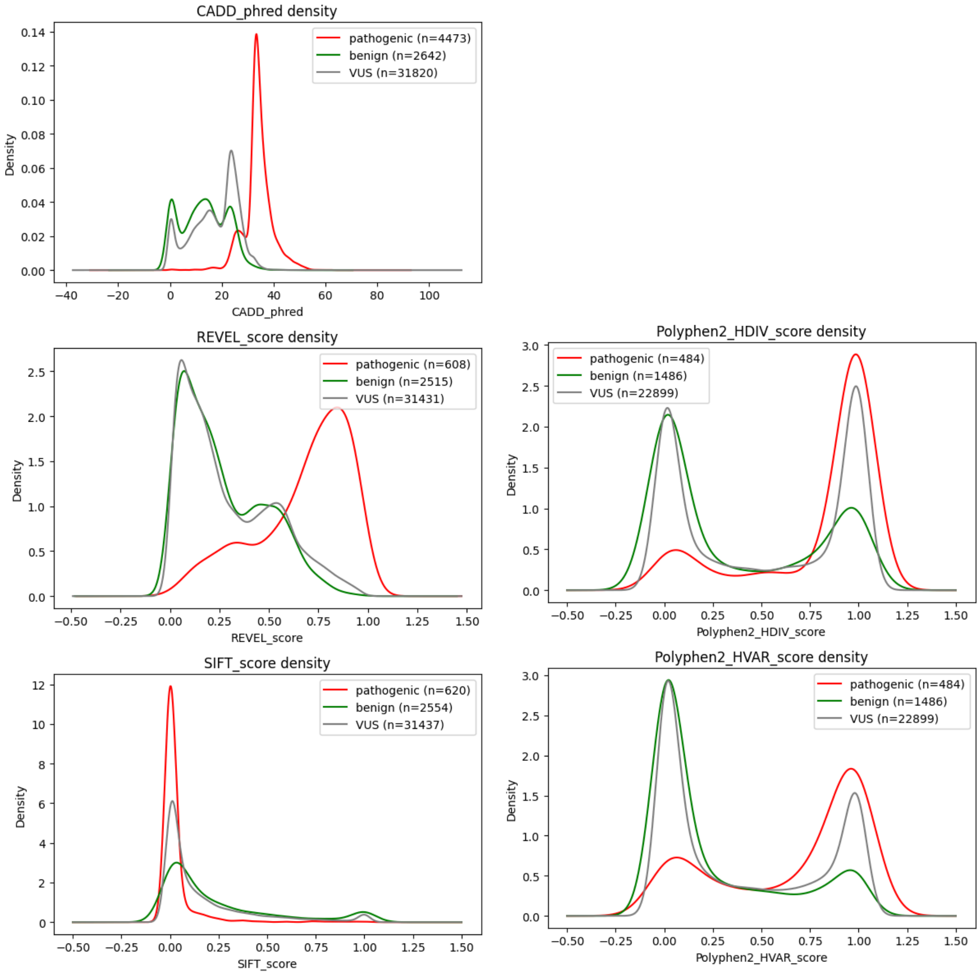

Figure 11. Distribution of mutation classifications over in-silico prediction scores.

3. Conclusion

In this study, we present a unified and meticulously curated dataset of Fanconi Anemia (FA)-associated mutations by integrating two major resources: the LOVD (FAMD) and NCBI ClinVar databases. Through systematic reconciliation, we revealed substantial discrepancies in transcript versions, file formats, and cross-referencing. Although both databases nominally adhere to ACMG/AMP guidelines for five-tier classification, our integration showed that over 20% of LOVD entries lacked corresponding ClinVar records, and many variants carried conflicting or "unknown" assertions when merged. These inconsistencies underscore the urgent need for harmonized reference sequences, standardized VCF exports, and automated cross-linking between disease-specific and broader clinical archives.

Our pipeline resolved these inconsistencies, enabling in-depth analysis of 91,720 GRCh37-aligned variants. Functional annotation and in silico benchmarking confirmed that high CADD, REVEL, and PolyPhen-2 scores—and low SIFT scores—closely track with pathogenic labels. While this concordance illustrates the predictive strength of computational models, it also raises concern: such metrics may unduly influence clinical interpretation in the absence of orthogonal validation. The surprising presence of synonymous SNVs labeled as pathogenic highlights critical gaps in current understanding of FA mutations. These variants likely exert their effects via mRNA splicing, structural instability, codon usage kinetics, or disruption of regulatory motifs—mechanisms not captured by sequence-level tools alone.

Our broader analyses uncovered non-random distributions of mutations across genes and within exons, with striking regions of apparent developmental intolerance. We also observed that recurrent stopgain mutations dominate the most frequently reported pathogenic variants, particularly in core pathway genes like FANCI. These observations suggest that genomic consequence alone is insufficient for pathogenicity inference—functional context and developmental impact must be considered.

Looking forward, similar integration efforts in other Mendelian diseases—such as BRCA-associated breast cancer or cystic fibrosis—have begun to bridge specialist databases and ClinVar, yielding richer genotype–phenotype maps and enabling cross-disease comparisons of variant tolerance and hotspot architecture. Extending our approach, future work should incorporate multi-omics readouts (transcriptomic, proteomic, chromatin), high-throughput functional assays, and rigorous re-evaluation of submission metadata. Only through such cross-disciplinary, data-driven scrutiny can we refine pathogenicity frameworks, improve diagnostic consistency, and accelerate the translation of genomic insights into patient care.

This harmonized framework not only advances variant interpretation in FA, but also serves as a model for integrative curation in other rare genetic disorders—supporting the shift toward precision diagnostics and evidence-based classification in clinical genomics.

4. Methods

4.1 Data Acquisition

Variant data for the 22 known Fanconi Anemia (FA) genes was obtained from two primary sources:

- ClinVar: A comprehensive variant_summary.tsv file aligned to GRCh37 was downloaded via NCBI FTP.
- LOVD: Variants were retrieved via web scraping of the Leiden Open Variation Database (LOVD3.0), using a custom script that paginates through the HTML tables of each gene's variant entries.

4.2 Preprocessing and Normalization

All LOVD-scraped files were harmonized to a consistent schema using fixScrape.py. Each gene's data was then parsed using parse.py, which extracts genomic HGVS coordinates (g.) and assigns standardized variant IDs (e.g., LOVD.FANCA00001). Genomic contigs were inferred from a pre-defined dictionary of GRCh37 contig accessions. The extracted HGVS descriptions were converted into Variant Call Format (VCF) using hgvsToVCF.py, which locally normalizes and parses each variant using the Biocommons hgvs library with a pre-indexed GRCh37 FASTA. ClinVar data for each gene was filtered from variant_summary.tsv, rearranged into a custom tabular VCF using main.sh, and augmented with synthetic IDs using addClinVarID.py.

4.3 Variant Conversion and Cleanup

Both LOVD and ClinVar variants were converted into unified VCF format using tsvToVcf.py. To ensure base-level validity:

- Ambiguous IUPAC bases and malformed entries were removed using removeIUPACambg.py.
- Chromosome fields were updated to GRCh37 contig names using replaceCHROM.py.

A combined dataset was produced by concatenating all cleaned VCFs, followed by deduplication and normalization using bcftools norm in normVCF.sh. Only biallelic SNPs and valid entries (POS ≥ 1) were retained.

4.4 Annotation and Labeling

The merged and cleaned dataset was annotated using ANNOVAR with the following databases:

- refGene for gene-based annotations.
- dbNSFP v4.2a for deleteriousness scores including CADD-phred, REVEL, SIFT, and PolyPhen2-HDIV/HVAR.

Custom labeling of the consequence field (CONSEQ) was performed using label.py, which mapped inconsistent clinical interpretations to a three-class system: pathogenic, benign, and VUS (uncertain significance).

4.5 Downstream Analysis and Visualization

Final annotated tables were analyzed in Python (capstone.ipynb) using pandas, matplotlib, and numpy.

4.6 Reproducibility

All processing steps are wrapped in the reproduce.sh script, which automates data download, normalization, annotation, merging, and preparation for analysis. The script supports re-running the entire pipeline from scratch, provided the reference genome and ClinVar variant summary are available. Downstream analysis and figure generation can be found in scripts/dataAnalysis.ipynb

5. References

Ameziane, N., Sie, D., Dentro, S., Ariyurek, Y., Kerkhoven, L., Joenje, H., Dorsman, J. C., Ylstra, B., Gille, J. J. P., Sistermans, E. A., & de Winter, J. P. (2012). Diagnosis of Fanconi Anemia: Mutation Analysis by Next-Generation Sequencing. *Anemia*, *2012*, e132856. https://doi.org/10.1155/2012/132856

Auerbach, A. D. (2009). Fanconi anemia and its diagnosis. *Mutation Research*, *668*(1–2), 4–10. https://doi.org/10.1016/j.mrfmmm.2009.01.013

Bhandari, J., Thada, P. K., Killeen, R. B., & Puckett, Y. (2025a). Fanconi Anemia. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK559133/

Bhandari, J., Thada, P. K., Killeen, R. B., & Puckett, Y. (2025b). Fanconi Anemia. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK559133/

Chowdhry, M., Makroo, R. N., Srivastava, P., Kumar, M., Sharma, S., Bhadauria, P., & Mahajan, A. (2014). Clinicohematological correlation and chromosomal breakage analysis in suspected Fanconi anemia patients of India. *Indian Journal of Medical and Paediatric Oncology : Official Journal of Indian Society of Medical & Paediatric Oncology*, *35*(1), 21–25. https://doi.org/10.4103/0971-5851.133706

Deans, A. J., & West, S. C. (2011). DNA interstrand crosslink repair and cancer. *Nature Reviews Cancer*, *11*(7), 467–480. https://doi.org/10.1038/nrc3088

Fiesco-Roa, M. O., Giri, N., McReynolds, L. J., Best, A. F., & Alter, B. P. (2019). Genotype-Phenotype Associations in Fanconi Anemia: A Literature Review. *Blood Reviews*, *37*, 100589. https://doi.org/10.1016/j.blre.2019.100589

Fokkema, I. F. A. C., Kroon, M., López Hernández, J. A., Asscheman, D., Lugtenburg, I., Hoogenboom, J., & den Dunnen, J. T. (2021). The LOVD3 platform: Efficient genome-wide sharing of genetic variants. *European Journal of Human Genetics*, *29*(12), 1796–1803. https://doi.org/10.1038/s41431-021-00959-x

Hou, H., Li, D., Gao, J., Gao, L., Lu, Q., Hu, Y., Wu, S., Chu, X., Yao, Y., Wan, L., Ling, J., Pan, J., Xu, G., & Hu, S. (2020). Proteomic profiling and bioinformatics analysis identify key regulators during the process from fanconi anemia to acute myeloid leukemia. *American Journal of*

Translational Research, *12*(4), 1415–1427.

Jacquemont, C., & Taniguchi, T. (2007). The Fanconi anemia pathway and ubiquitin. *BMC Biochemistry*, *8*(1), S10. https://doi.org/10.1186/1471-2091-8-S1-S10

Joo, W., Xu, G., Persky, N. S., Smogorzewska, A., Rudge, D. G., Buzovetsky, O., Elledge, S. J., & Pavletich, N. P. (2011). Structure of the FANCI-FANCD2 complex: Insights into the Fanconi anemia DNA repair pathway. *Science (New York, N.Y.)*, *333*(6040), 312–316. https://doi.org/10.1126/science.1205805

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, *42*(Database issue), D980-985. https://doi.org/10.1093/nar/gkt1113

Mamrak, N. E., Shimamura, A., & Howlett, N. G. (2017). Recent discoveries in the molecular pathogenesis of the inherited bone marrow failure syndrome Fanconi anemia. *Blood Reviews*, *31*(3), 93–99. https://doi.org/10.1016/j.blre.2016.10.002

Muniandy, P. A., Liu ,Jia, Majumdar ,Alokes, Liu ,Su-ting, & and Seidman, M. M. (2010). DNA interstrand crosslink repair in mammalian cells: Step by step. *Critical Reviews in Biochemistry and Molecular Biology*, *45*(1), 23–49. https://doi.org/10.3109/10409230903501819

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., & ACMG Laboratory Quality Assurance Committee. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *17*(5), 405–424. https://doi.org/10.1038/gim.2015.30

Yuan, F., El Hokayem, J., Zhou, W., & Zhang, Y. (2009). FANCI Protein Binds to DNA and Interacts with FANCD2 to Recognize Branched Structures. *The Journal of Biological Chemistry*, *284*(36), 24443–24452. https://doi.org/10.1074/jbc.M109.016006

Acknowledgements

I would like to thank my Family, the Lebanese American University, and Dr Joseph for their support throughout the beautiful journey I had learning bioinformatics at LAU. This is only the beginning, I still have a lot to learn – like not leaving a project for the last week. Mistakes are not failures, but lessons; with every breath I take, I’m grateful that I’m a wiser man than I was before the breath.