OPEN ACCESS

ISSN: 2996-7449

Total CR, et al. (2025): 211
American J Medi Re Heal Sci, 2025; Vol.3(2):100116
Doi: 10.71010/AJMRHS.2025-e116

Original Article

American Journal of Medical Research & Health Sciences

Systems Bioinformatics Analyses Recapitulate Inherited Anemia Loci at The Molecular Level

Chelsea R. Total[1], Molly E. Sawyer[1], Zoë E. Gaetjens[1], Azhar Rahama[1], Felix S. Vega[1], Andrew D. Burton[1], Alexis K. Mobilia[1], Julia H. Karcic, D.P.M.[1], Matthew D. Gacura, Ph.D.[1], Richard G. Ligo, Ph.D.[2], Theodore Yeshion, Ph.D.[3], and Gary Vanderlaan, Ph.D.[1*]

[1]Department of Biology, Gannon University, Erie, PA, USA
[2]Mathematics Department, Gannon University, Erie, PA, USA
[3]Forensic Investigation Center, Criminal Justice & Criminalistics, Gannon University, Erie, PA, USA

[*]**Correspondence:** Gary Vanderlaan, Ph.D, Email: vanderla002@gannon.edu

Citation: Total CR, Sawyer ME, Gaetjens ZE, Rahama A, Vega FS, et al. (2025) Systems Bioinformatics Analyses Recapitulate Inherited Anemia Loci at The Molecular Level. American J Medi Re Heal Sci: AJMRHS-116.

Received Date: 08 April, 2025; **Accepted Date:** 16 April, 2025; **Published Date:** 22 April, 2025

Abstract

Anemia represents a medical challenge due to a significant myriad of etiologies underlying its clinical presentation. Leading causes of anemia manifestations include nutritional deficiencies, infectious disease agents, and host genetics. In the latter case, numerous genes are associated with the various inherited anemias, and each genetic hematologic disorder mechanistically manifests with unique disease pathophysiology patterns accompanying key modes of anemia manifestation. Here we provide a quantifiable framework to classify nearly 200 anemia-enriched genes based on their allelic categorical distributions and permutation pattern of clinical manifestations. Our work leverages a convergence of multivariate statistical tools married to applied mathematical approaches in the form of topological data analysis to detect gene similarity interactions and connectivity networks. Such a systems bioinformatics approach permits an investigation of the relationship amongst anemia-enriched genetic loci which further grants the means for taxonomic classification to relate genotype to phenotype. In doing so, our bioinformatics pipelines are also able to reconstruct the molecular complexity of both the canonical and non-canonical anemias in a quantifiable fashion.

Keywords: anemia; hematological disorders; erythropoiesis; hemolysis; clotting abnormalities; globinopathies

1. Introduction

Anemia is a major public health priority, with a global prevalence of nearly 2 billion cases across all age groups (1–4). There are many etiologies underlying anemia manifestations in the clinic, including age, infection, inflammation, trauma, dietary intake, pregnancy, and host genetics (5,2). Enhanced risk of mortality is seen in elderly patients who present with anemia and comorbidities (5–9). The most common cause of anemia is iron deficiency, especially in pediatric cases (10–14). Nutritional anemias can also manifest due to dietary deficiency in pyridoxine (15,16), folate (17–19), or cobalamin (20,21). Infectious agents, including parvovirus (22,23), *Plasmodium* spp. (24–27), and schistosomes (28–30) as well as improper inflammatory responses (31) comprise leading causes of acquired anemias. Reproductive biology status in women contribute to global anemia rates as well, compounded by altered prenatal vitamin and mineral requirements during pregnancy terms (32–35). As might be expected for a phenotype derived from such a wide berth of etiologies, the precise definition of anemia in the clinic can be quite variegated, ranging from decreased serological hemoglobin, abnormal mean corpuscular volume (MCV), and/or atypical red blood cell (RBC) morphology (36).

The inherited anemias are dictated by germline transmission and thus are reliant upon host genetics (37). The elucidation of the precise genetic lesions underpinning the various inherited anemias is an ongoing process, with past estimates of approximately 70 distinct genetic loci collectively responsible for the inherited anemias (37). A multitude of molecular, cellular, and physiological mechanisms constitute the inherited anemias (37). Defects in RBC formation rates by attenuated erythropoiesis are well documented in several inherited anemias, including Fanconi Anemia (38–42), Diamond-Blackfan Anemia (43–46), and Congenital Dyserythropoiesis Anemia (47,48). Additionally, prior to terminal RBC differentiation stages, erythropoietic progenitors must synthesize not only hemoglobin tetramers but also execute sufficient heme biosynthesis (49–51). Defective heme anabolism constitutes many inherited anemias of the sideroblastic form (52,53). Conversely, inherited anemias can manifest as a block on normal removal rates of mature RBCs. Enhanced rates of erythrocyte destruction via hemolysis, typically by splenic macrophages in the red pulp, are seen clinically in cases involving hemolytic-uremic syndrome (54), hereditary spherocytosis (55,56), and glucose-6-phosphate dehydrogenase (G6PD) deficiency (57,58). At an erythrocyte operational level, mutations that alter either the amount or the function of the hemoglobin tetramer comprise the globinopathies: Sickle-Cell Anemia (59), α-thalassemia (60–62), and ß-thalassemia (63–66). Inherited anemias can also manifest indirectly from extensive blood loss as seen in hemophiliacs (67–69) due to misregulated clotting cascades, or in patients suffering from Von Willebrand disease due to attenuated clotting at the subendothelial layer (70). In related fashion, repeat wound formation can increase patient risk of anemia development. In patients diagnosed with epidermolysis bullosa, genetic defects in sustaining structural connectivity of the epidermis to the dermis results in detachment of the two layers, which constitutively contributes to low-grade bleeding complications, and thus greater risk for bouts of anemia (71–73).

In patients diagnosed with mitochondrial diseases, such as *POLG*-related disorders, the presentation of anemia in the clinic nearly quadruples the risk of patient mortality compared to mitochondrial disease patients that do not present with anemia manifestations (74–76).

For the inherited anemias alone, there is thus an expansively, disparate and complex set of genetic and molecular factors underlying a shared anemia manifestation in the clinic. In turn, not every genetic locus drives the inherited anemia phenotype to the same degree, and within a single locus, allelic penetrance is quite variable, ranging from null lesions to subtle hypomorphs to silent polymorphisms. Here we report on a systems biology approach to taxonomically classify the various etiologies of the numerous inherited anemias in a quantitative manner that addresses allelic variability as well as loci-associated phenotypes.

2. Materials and Methods

All custom python source code used in this article for loci scraping, data collation, bioinformatics analyses, and data visualization are freely available for examination and can be downloaded at https://github.com/VitamOrdinatio/inherited_anemias

2.1. Data acquisition of anemia loci

2.1.1. Scraping the Genetic Testing Registry (GTR) for anemia conditions

We leveraged the Genetic Testing Registry (GTR) database that is administered by the National Institute of Health (NIH) National Center for Biotechnology Information (NCBI) (77). The GTR is a public resource that tracks clinically relevant genetic testing assays (77). As of late November 2024, the GTR database contained testing information regarding 67,624 genetic assays for 26,106 genetic conditions comprising 18,705 underlying gene etiologies (77). Search queries against the GTR database can be further filtered using integration with the Online Mendelian Inheritance in Man (OMIM) and NCBI GeneReviews, both of which represent high quality, manually curated databases (77–79). We wrote custom python scripts using the python requests and beautifulsoup4 modules to sequentially scrape NCBI GTR for three query terms (i.e., anemia, willebrand, and hereditary factor) with OMIM and GeneReviews filter facets toggled on (77–79). Our scrape at this stage yielded a total of 178 unique genetic conditions with some degree of hematological disorder manifestation of the highest curation status (77–79).

2.1.2. Scraping GTR conditions for associated anemia gene etiology

To programmatically retrieve the underlying genetic loci responsible for our scraped list of 178 hematological disorders, we performed a second scrape operation against the NCBI GTR database (77). A custom python script scraped all monogenic and polygenic etiologies underlying each genetic condition. We next manually collated several conditions arriving at a list of 164 unique genes across 170 unique genetic conditions. Two entries (H19-ICR and HBB-LCR) were removed as they represented regulatory control elements rather than protein-encoding genes.

2.1.3. MitoCarta collation

MitoCarta is a database that tracks 1,136 nuclear-encoded mitochondrial gene products in the human condition (80–82). Approximately a third of the genetic loci obtained in our original anemia loci scrape were found on the MitoCarta roster. Additionally, mitochondrial DNA (mtDNA) genes from the Cambridge Reference Sequence (CRS) were also found in the original NCBI GTR scrape for anemia-enriched loci (83–85). Pearson syndrome (OMIM #557000) represents a sideroblastic anemia driven by large mtDNA deletions manifesting as lesions across the 37 essential mtDNA loci (86–88). In eukaryotes, all mtDNA replication is accomplished via the catalytic POLG subunit of the mtDNA polymerase complex. The *POLG*-related disorders include Alpers-Huttenlocher (OMIM #203700), MNGIE (OMIM #613662), SANDO/SCAE (OMIM #607459), autosomal dominant progressive external ophthalmoplegia (adPEO, OMIM #157640) and autosomal recessive PEO (arPEO, OMIM #258450) (76,75). Previous work indicated that roughly 2/3rd of all examined *POLG* patients exhibited anemia, and suffered from a nearly four-fold decrease in survivorship when compared to *POLG* patients without anemia presentations (74). We thus collated *POLG1* (aka *POLG*), *POLG2*, and all 37 mtDNA loci to our list of GTR-scraped anemia loci, resulting in a final anemia locus list composed of 199 unique genes linked by varying degrees of hematological disorder significance.

2.1.4. Scraping ClinVar for categorical allele distributions

The NCBI ClinVar database is a public repository that archives documented gene sequence variations following nomenclature standards established by the Human Genome Variation Society (HGVS) under the auspices of the Human Genome Organization (HUGO) (89–92). A dedicated ClinVar accession number is assigned to each unique gene variant along with useful metrics such as disease classification and molecular lesion type (92). Although an application programming interface (API) exists for ClinVar, we utilized custom python scripts to scrape the ClinVar site of data pertaining to our master anemia locus list (91). For any given allele, ClinVar provides an estimation of clinical disease significance (92,93). We specifically removed alleles of the uncertain significance or the conflicting classification categories (92,93). Thus, we retrieved a total of 112,534 unique alleles across 199 anemia loci, and each allele exhibited a ClinVar disease classification of either 1) benign, 2) likely benign, 3) likely pathogenic or 4) pathogenic.

2.1.5. Gene ontology (GO) and gene set enrichment analysis (GSEA)

The Human Phenotype Ontology (HPO) database tracks over 18,000 unique phenotype abnormality terms related to human disease (94). Gene ontology (GO) is a bioinformatics taxonomy approach to systematically assess how different genes and their corresponding gene products behave across different databases, including HPO (94–96). Historically, gene set enrichment analysis (GSEA) was pioneered for transcriptomic experiments that yielded read counts for each expressed locus (95,96). However, functional enrichment analysis can still be performed on simple gene lists sans read count data using the g:Profiler g:GOSt platform (97). We performed GSEA on our list of 199 anemia loci using g:GOSt and pulled down the enrichment analysis results for just the HPO terms (94,97). Of 18K possible HPO terms, a total of 619 unique HPO enrichment terms for each locus on our anemia list of 199 genes were extracted with statistically-significant adjusted p-values (i.e., p_{adj}) per standard g:GOSt settings (94,97).

2.2. Analytical pipeline for the scraped ClinVar alleles

The ClinVar dataset comprised exactly 112,534 unique alleles spanning 199 unique anemia-enriched genes with each allele assigned one of four possible ClinVar disease severity classifications: benign (B), likely benign (LB), likely pathogenic

(LP), and pathogenic (P). Our analytical pipeline for the scraped ClinVar allelic set consisted of 1) standard visualization methods, 2) multivariate statistical analyses (MSA), and 3) topological data analysis (TDA). Source code for all steps can be found on our GitHub page. All graphical plots were generated using a combination of pythonic matplotlib, seaborn, and pyCirclize modules.

2.2.1. Standard visualization of ClinVar anemia alleles

Raw allele counts or relative ClinVar categorical allele frequencies were visualized per each of 199 anemia loci. Dataframes were sorted in descending order contingent on allele counts or normalized allele frequencies using python's pandas and numpy libraries. Pythonic seaborn pairplots were generated for pairwise comparisons for any two retained ClinVar disease categories (i.e., benign, likely benign, likely pathogenic, and pathogenic).

2.2.2. Multivariate statistical analysis (MSA) of ClinVar anemia alleles

A suite of multivariate statistical tools was employed using various python libraries customized for the scraped ClinVar anemia allelic dataset.

2.2.2.1. Correlation coefficients

Correlation coefficients (r) were generated using the python's pandas library with dataframe class-defined correlation methods. Heatmaps of resulting correlation coefficient relationships for each pairwise ClinVar allelic categorical comparison was generated using python's seaborn library.

2.2.2.2. t-Distributed Stochastic Neighbor Embedding (t-SNE)

Simultaneous dimensional reduction and potential cluster identification can be achieved using the machine learning algorithm known as t-distributed stochastic neighbor embedding (t-SNE) (98–100). To implement t-SNE on the ClinVar anemia allelic dataset, we leveraged python's sklearn.manifold TSNE module with a learning rate initialized to 50, and graphically visualized the t-SNE results using various seaborn and matplotlib functions.

2.2.2.3. Principal component analysis (PCA)

Principal component analysis (PCA) is another statistical method that effectively reduces dimensionality of the original dataset by grouping highly correlated variables into principal components (101–103). For programmatic PCA implementation, we utilized several tools in the pythonic sklearn module, including StandardScaler and MinMaxScaler methods, as well as the sklearn.decomposition PCA method. Explained variance ratios, and relationships amongst each principal component (PC) following decomposition of the ClinVar anemia allelic dataset were extracted using PCA pythonic class attributes. Pairwise scatter plots between any two of four principal components were visualized using pythonic seaborn and matplotlib graphical methods.

2.2.2.4. PCA-reduced k-Means clustering

The *k*-means clustering method is an algorithmic approach to grouping data points into clusters contingent on their distance to defined centroids (104–107). For detecting such patterns across the ClinVar categorical allele counts derived from 199 anemia genes, unsupervised learning using *k*-means clustering operations was performed leveraging the python sklearn.cluster library, Kmeans. To maximize aggregate cluster generation by the *k*-means algorithm, we leveraged the first two of four principal components to reduce the dimensionality of the original ClinVar anemia allelic dataset (101–103). A programmatic pipeline was constructed using the sklearn.pipeline method known as Pipeline which permits user-friendly arguments and parameters passed to different pipe segments of the pipeline. In this case, we generated a preprocessor pipe segment that utilized a MinMaxScaler and a restriction to employ only the first two principal components, followed by a clusterer pipe segment with the following initialization settings to maximize reproducibility: init set to 'k-means++' to accelerate convergence, n_clusters set to six cluster targets, n_init set to 50 initializations which returns the results with the lowest sum of the squared error (SSE), and max_iter set to 500 to control the maximum number of iterations that are executed for each initialization of the k-means algorithm. Scatterplots of PCA-filtered *k*-means clustering of the ClinVar anemia allelic dataset were generated using pythonic seaborn and matplotlib module functions.

2.2.3. Topological data analysis (TDA) of ClinVar anemia alleles

Topological data analysis (TDA) is a non-statistical, advanced geometry method that can investigate the shape of data using a given distance metric. In the life sciences, TDA as an applied mathematical tool has been broadly utilized for research on cancer, medical imaging, molecular structures, and organismal biology (108–110). Of note, TDA resembles PCA as both are fully capable of dimensionality reduction while vastly differing in implementation. For the scraped ClinVar dataset, which comprised 112,534 anemia alleles across 199 anemia loci, each anemia allele has an exclusive disease classification attribute of either benign (B), likely benign (LB), likely pathogenic (LP), and pathogenic (P). We first generated per-locus categorical allele frequencies suitable for distance matrix calculations. Next, we employed the pythonic kmapper (aka KeplerMapper) module which is capable of TDA operations (111). The KeplerMapper workflow generally consists of projecting the data, grouping the resulting image, applying a clustering algorithm to the preimage of the groups, and building a simplicial complex that summarizes gene interactions in our case (111). We set our mapper object to utilize 7 bins with a 25% overlap while the scikit.learn DBSCAN (density-based spatial clustering of applications with noise) method was passed an epsilon set to the median value from the distance matrix and a minimum cluster size set to 3. Python's kmapper provides an interactive HTML output file that consists of nodes (aka clusters) that contain varying numbers of genes (i.e., node members). Circos plots are one way to visualize the connectivity data intrinsic to TDA, and we generated python Circos plots using the pyCirclize library (112–115).

2.3. Analytical pipeline for the enriched HPO terms encountered in the scraped anemia locus list

The HPO GSEA anemia dataset comprised exactly 619 unique HPO terms that exhibited varying degrees of intersecting loci derived from a list of 199 scraped anemia genes. Our analytical pipeline for this anemia locus-phenotype dataset leverages two overall strategies: 1) a text-mining operation coupled to wordcloud syntheses to visualize the most frequently encountered gene names and HPO phenotypic terms, and 2) abstraction to multidimensional space for subsequent connectivity mapping using TDA-digested and Circos visualization. Source code for all steps can be found on our GitHub page dedicated to this body of work. All graphical plots were generated using a combination of pythonic matplotlib, seaborn, wordcloud, and pyCirclize modules.

2.3.1. Standard visualization of the HPO / GSEA anemia loci

A truth table was first generated for the g:Profiler g:GOSt algorithm that received the list of 199 anemia genes and returned GSEA metrics from the HPO database (94,97). This truth table was essentially a programmatic conversion of the g: GOSt intersections column (i.e., the gene names that exhibited HPO term enrichment at a statistically significant level); the resulting HPO truth table consisted of 199 unique rows (unique genes) by 619 unique columns (unique phenotypes). At the intersection of each gene and HPO term, a value of zero indicates false and a value of one indicates true (i.e., the HPO term was statistically enriched for that locus). Next, we used the python wordcloud module to generate graphical visualizations for 1) the most frequently encountered gene names or 2) the most frequently encountered phenotypes across the entire truth table (i.e., 199 genes x 619 phenotypes). For gene frequency wordcloud visualization, the python random library was used to rearray gene names followed by a standard wordcloud generation with interpolation set to 'bilinear'. Graphical output of the wordcloud was performed using python's matplotlib library. Stopwords provide a masking filter, and while no stopwords were used for gene name wordclouds, stopwords that removed non-specific terms in the names of HPO terms were employed. A list variable of defined stopwords can be found by examination of python source code archived on our GitHub page.

2.3.2. Topological data analysis (TDA) of the HPO / GSEA anemia loci

The HPO / GSEA dataframe was abstracted into multidimensional space by generating distance matrices utilizing the HPO / GSEA truth table that consists of 619 unique phenotypes by 199 unique gene names. This approach effectively represents each gene as a single point in 619^{th}-dimensional space, where each axis discretely services a unique HPO enrichment term. We reduced dimensionality using pythonic kmapper to implement TDA, and investigated the various TDA components for custom illustration summaries using BioRender or pyCirclize for Circos plot generation (112–115). The settings for TDA operations on the HPO / GSEA data were identical to those employed for the ClinVar anemia alleles; the mapper object was set to use 7 bins with a 25% overlap and the scikit learn's DBSCAN was implemented as the clusterer with an epsilon parameter set to the median value from the distance matrix and the minimum cluster size set to 3 (111).

3. Results

3.1. Unique allele deposition varies across known loci underlying inherited anemias

Our custom NCBI GTR scrape pipeline yielded 199 unique genetic loci that are associated with inherited anemias. For each gene, a total of six allelic categories were additionally scraped from NCBI ClinVar: benign (B), likely benign (LB), likely pathogenic (LP), pathogenic (P), conflicting classifications (CC) and uncertain significance (US). A total of 192,296 unique alleles were obtained from these six ClinVar categories, and across the entire 199 loci, there was an average of ~160 alleles per category and gene. Total allele counts per ClinVar category varied across genes, ranging from 0 to 5,092 unique allele accessions. We removed the CC and US categories, leaving us with a total of 112,534 alleles across 4 well-defined ClinVar allele categories (B, LB, LP and P). These 112,534 alleles can be further partitioned as 12,349 benign, 57,621 likely benign, 10,944 likely pathogenic, and 31,620 pathogenic ClinVar allele classifications. More importantly, these ~110K alleles thus represent the best defined polymorphisms for the most curated anemia genes to date. A log2-transformation of total allele counts for each of these four ClinVar categories reveals elevated allele depositions for known inherited anemias, such as molecular members of the Fanconi Anemia complex as well as genes involved in Epidermolysis Bullosa (Figure 1A-1B). The four retained ClinVar allelic categories exhibited a propensity for positive correlation in pairwise fashion (i.e., r = 0.67-0.83) (Figure 1C). Further examination of each of the four ClinVar allelic categories was performed using pairplot visualization (Figure 2). Identity scatterplots reveal normal distributions of raw allele counts for each of the four ClinVar allele categories (Figure 2, diagonal). Pairwise scatterplots further illustrate the positive correlation trends for each comparison (Figure 2). When each of the 199 anemia-enriched genes are sorted in descending order based on total ClinVar allele counts derived for each of the 4 ClinVar categories (i.e., B, LB, LP, and P), seven of the top twenty loci are associated with the Fanconi Anemia complex (Figure 3A). Interestingly, not a single mtDNA gene locus was found in the top 20 list of anemia genes by allele count, but the *POLG* gene, which encodes the catalytic subunit of the mtDNA polymerase, was found on the list at position 13 (i.e., there were a total of 1600 *POLG* alleles: 125 B, 1049 LB, 180 LP and 246 P categories). Raw allele counts are informative but do not encapsulate locus essentiality nor locus stringency requirements for constancy across evolutionary time. To better understand the overall stringency for constancy possessed by each of the 199 anemia genes, we calculated relative categorical allelic frequencies (i.e., B.freq, LB.freq, LP.freq, and P.freq) per genetic locus, and resorted the normalized loci in descending order based on the sum of LP.freq and P.freq, which together represent the most capacity for pathogenicity (Figure 3B). Viewed from a relative categorical allele frequency reflecting such problematics alleles (i.e., LP and P classifiers), mtDNA genes (*MT-RNR2*, *MT-TL2*, and *MT-TS2*) comprised 3 of the top 20 loci (Figure 3B). Of note, the *POLG* locus and genes encoding the Fanconi Anemia complex were not found when sorted by pathogenicity frequency (Figure 3B). The globinopathies that are clinically defined by *HBA1*, *HBA2* or *HBB* lesions were amongst the most abundant loci by total allele counts (Figure 3A: *HBB*) or exhibited elevated levels of pathogenicity allele frequencies (Figure 3B: *HBA1* and *HBA2*). Hemophilias characterized by clotting factor abnormalities in Factor VIII and Factor IX were reflected at elevated positions in normalized pathogenicity frequency sorts but did not appear in the top 20 loci when sorted by raw allele counts (Figure 3B: *F8* and *F9*). Conversely, lesions afflicting the integrity of the extracellular matrix (ECM), or the ability to tether to the ECM exhibited exceptionally high total allele counts in our ClinVar scrape (Figure 3A: *COL3A1*, *COL4A1*, *COL7A1* and *PLEC*).

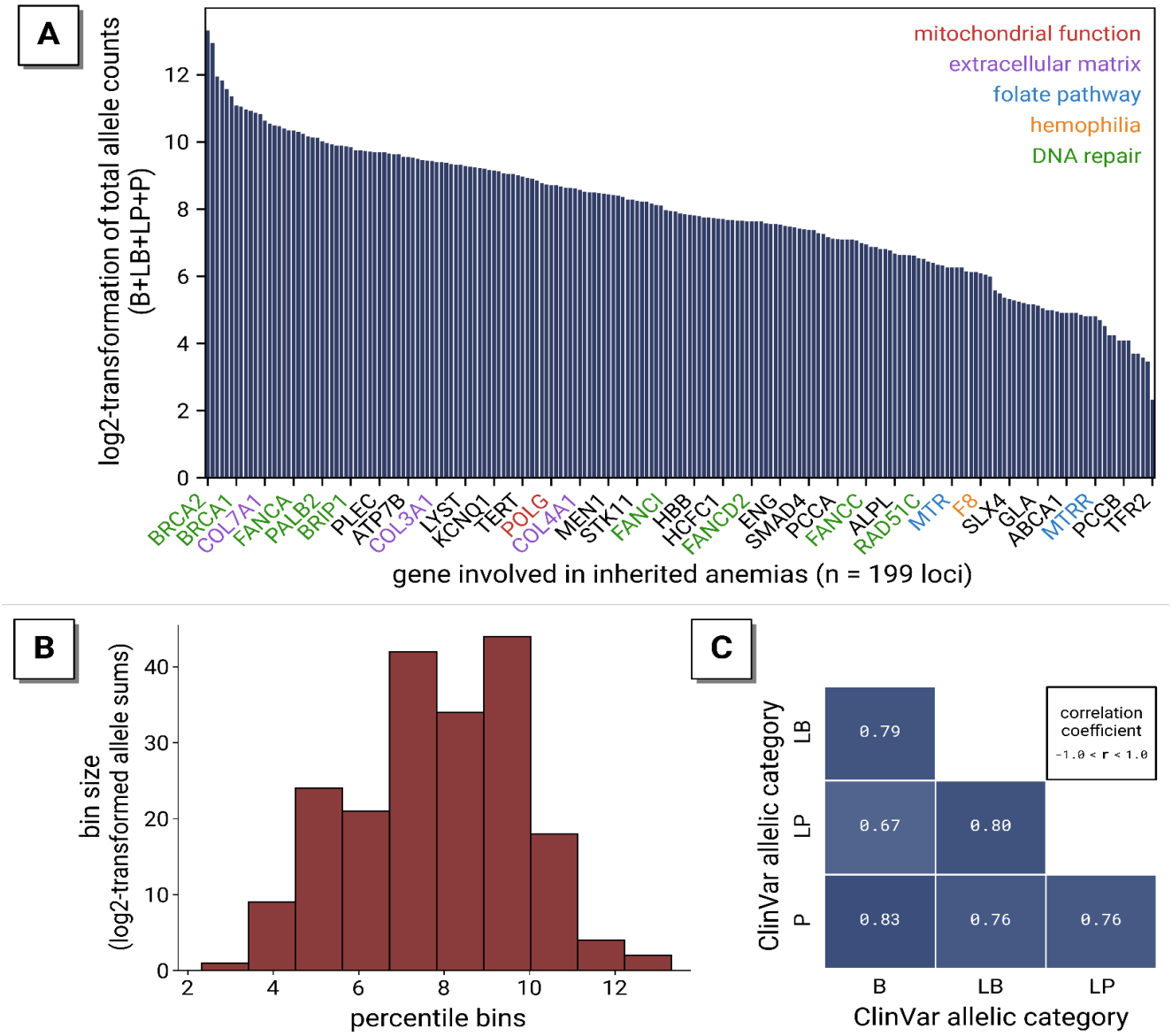

Figure 1: ClinVar allele prevalence across 199 anemia-enriched genetic loci.

Known allelic classifiers for a total of 199 anemia-enriched genes were systematically analyzed from the NCBI ClinVar database. **(A)**: Crude visualization of loci is performed by a log2-transformation of total allele counts derived from the sum of four ClinVar allelic categories: benign (B), likely benign (LB), likely pathogenic (LP) and pathogenic (P). Shown here are the most prevalent 34 loci of 199 total anemia genes. Defects in translesion DNA repair pathways manifests as Fanconi anemia and constitute some of the most defined genetic loci as measured by total unique allele counts. **(B)**: A histogram plot reveals a normal distribution for all 112,534 unique alleles across all 199 anemia loci. **(C)**: Linear regression performed on all alleles across all anemia-enriched loci reveals positive correlation coefficients (0.67 – 0.83) for each unique pairwise comparison for any two given ClinVar categories.

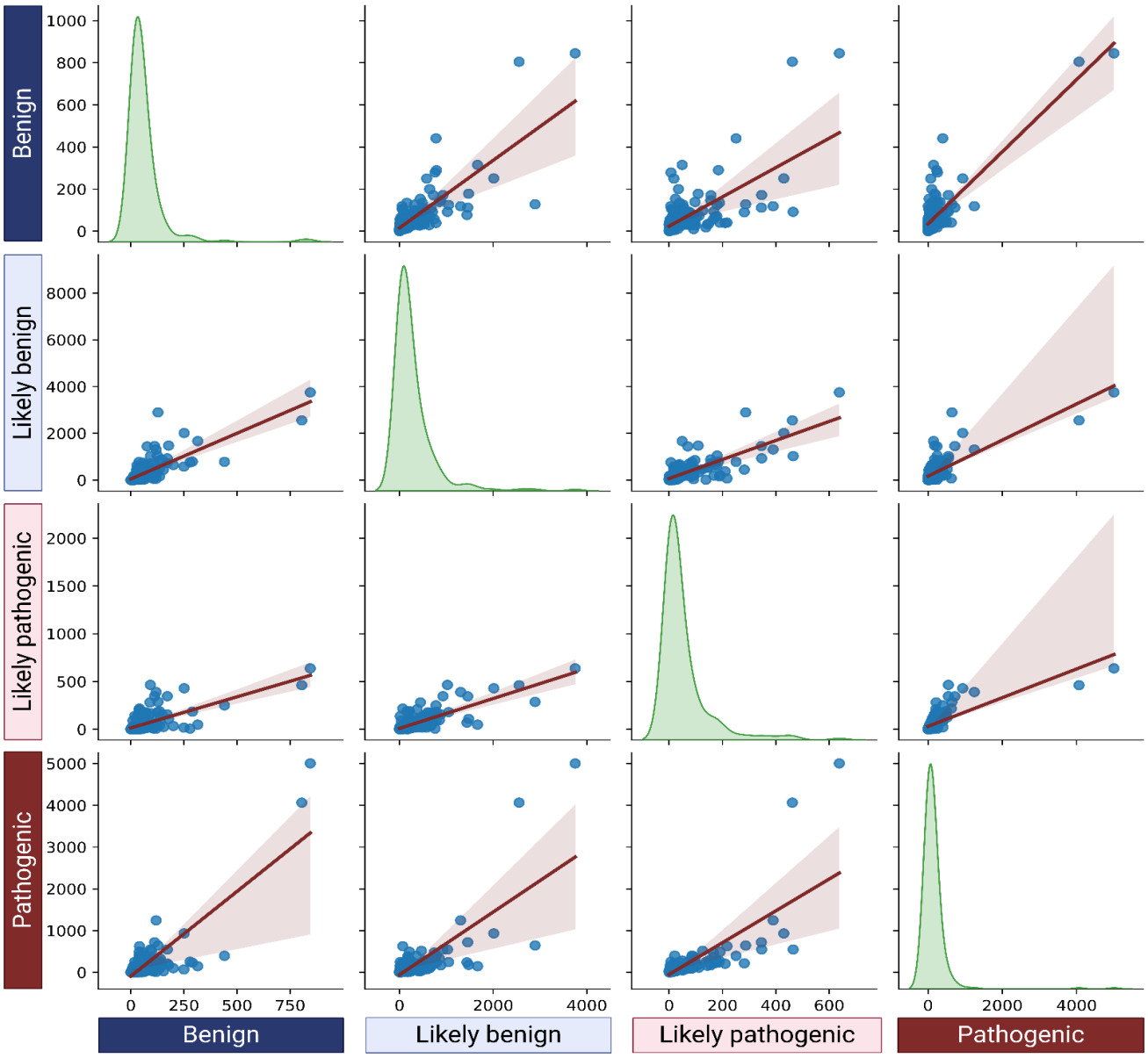

Figure 2: Pairplot visualization of ClinVar allelic categorical frequencies across 199 anemia-enriched loci.

Across all 199 anemic genes, all 112,534 alleles from each of four ClinVar allele categories (i.e., benign, likely benign, likely pathogenic and pathogenic) were visualized using seaborn pairplots. All four ClinVar allelic categories exhibited similar allelic prevalence distributions as shown in diagonal identity plots (green). Linear regressions are shown using best-fit lines (red).

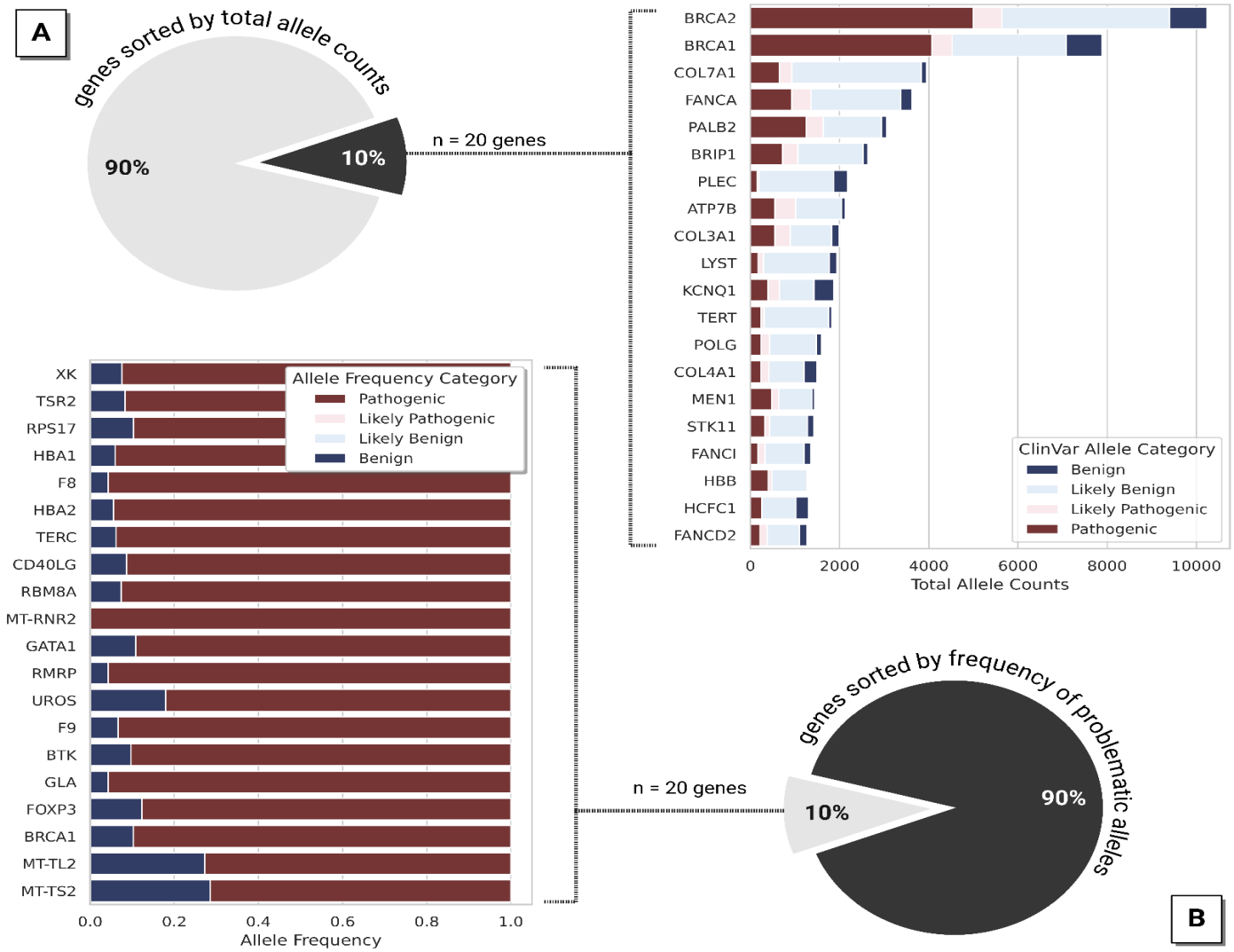

Figure 3: ClinVar allelic sorts by total allele counts or pathogenicity frequency.

(A): The top 20 genes, comprising of roughly 10% of the entire anemia loci set, were visualized by their sums of well-defined ClinVar allelic categorical assignments. Germline inherited lesions in *BRCA1*, *BRCA2*, *FANCA*, *PALB2*, *BRIP1*, and *FANCD2* result in defective translesion DNA repair which in turn potentiates DNA instability in afflicted patients. In fact, 7 of the top 20 genes by total allele counts manifest the inherited anemia known as Fanconi anemia. (B): For each anemia locus, relative allelic frequencies were calculated for the benign (B), likely benign (LB), likely pathogenic (LP) and pathogenic (P) allelic categories. The frequency sum of normalized LP and P categories was used to determine which genes constituted the most problematic allele burdens on a per-locus basis. Canonical anemias including the globinopathies and thalassemias (i.e., *HBA1* and *HBA2*), Fanconi anemia (i.e., *BRCA1*), Diamond-Blackfan anemia (i.e., *RPS17*) and clotting abnormalities (i.e., factors *F8* and *F9*) comprise the top 10% of all anemia-enriched loci. Of note, mitochondrial loci (i.e., *MT-RNR2*, *MT-TL2*, and *MT-TS2*) also exhibit similar problematic allelic frequencies.

3.2. Multivariate statistical analysis (MSA) of the alleles and loci associated with inherited anemias reveals strong support for clustering potential

Sorting gene lists based on raw allele counts or normalized allele frequencies is useful but fails to capture the complexity of the entire dataset. We next examined how each of the four ClinVar categories [i.e., benign (B), likely benign (LB), likely pathogenic (LP), and pathogenic (P)], varied across all 199 anemia loci by performing a series of multivariate statistical analyses (MSA). To ascertain a crude propensity for loci clustering, we performed t-distributed stochastic neighbor embedding (t-SNE) analyses on the 112,534 alleles derived from all 199 anemia loci while retaining each allele's disease classification status (i.e, B, LB, LP, or P). As an unsupervised means of dimensional reduction, crude clusters can be visualized across both t-SNE dimensions for all alleles of the anemia locus cohort (Figures 4A-4B). Interestingly, the t-SNE clusters seem to exhibit an inversely proportional relationship across each of the t-SNE dimensions (Figure 4A). We next performed principal component analysis (PCA) on the same dataset (i.e., 199 loci of 112,534 alleles from 4 ClinVar categories). Explained variance ratios for the interactions of each of four various principal components (PCs) illustrate that PC1 accounts for ~82.7% of the dataset and 91.3% of the variance in the dataset can be explained by PC1 and PC2 combined (Figures 4C-4D). Pairwise comparisons of each principal component using pairplot scatters further depict how each principal component interacts with one another (Figure 5). Of note, PC identity scatterplots illustrate that PC2, PC3 and PC4 are normally distributed, while PC1 is skewed (Figure 5, diagonals). Taken together, t-SNE thus illustrates cluster potential in an unsupervised fashion while PCA reveals overall

explained variance contributions. Like t-SNE methodology, k-means clustering algorithms act in an unsupervised fashion but can be benefit from PCA-mediated dimensional reduction techniques. As the first two principal components in PCA account for >90% of the explained variance in the anemia dataset (i.e., 199 genes of 112,534 alleles in 4 ClinVar categories), we first reduced the dimensionality of the anemia allelic dataset and then performed iterative k-means clustering operations (Figure 6). A total of six distinct clusters is seen with some adjoining clusters (i.e., cluster 0 and 3) while others like cluster 1 and 4 are quite distant (Figure 6).

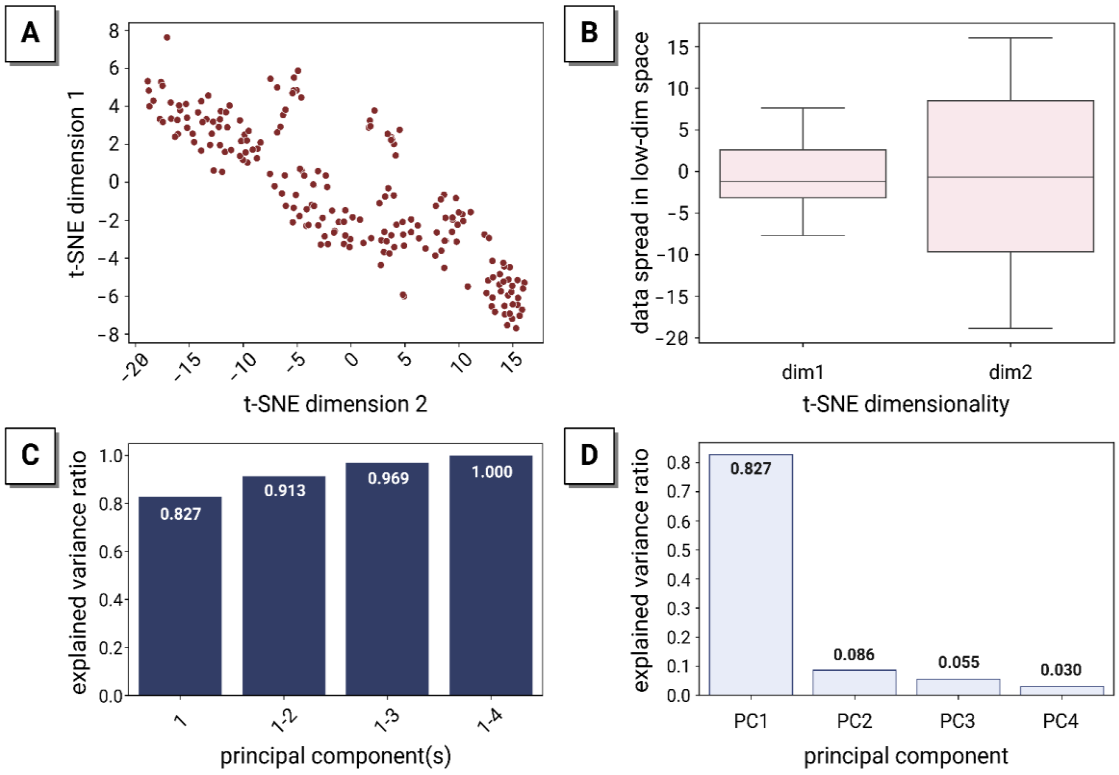

Figure 4: Principal component analysis and t-SNE visualization of ClinVar allelic categories across 199 anemia-enriched loci.

Dimensional reduction tools reveal the levels of variance found in the ~100K alleles comprising known ClinVar allelic categories across ~200 anemia loci. **(A-B)**: Dimensional reduction using t-SNE indicates high potential for cluster definitions as evidenced in scatterplots and boxplot variance assessments. **(C-D)**: Principal component analysis (PCA) reveals that PC1 and PC2 together account for >90% of the explained variance, of which PC1 alone sustains ~82.7%. Of note, PC1's ruleset indicates approximately equal weights applied to all four ClinVar categories for all alleles of anemia loci.

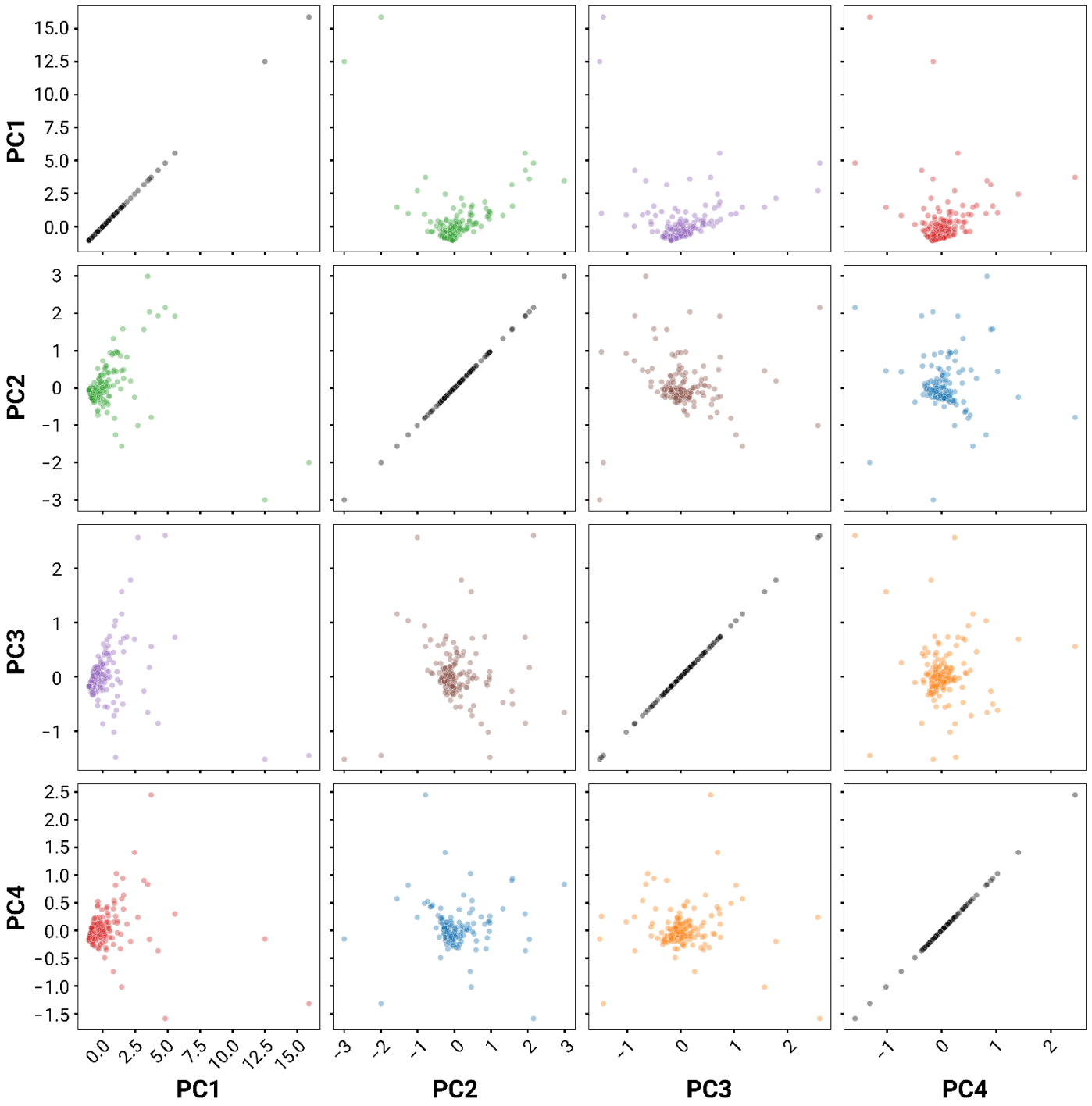

Figure 5: Pairplot visualization of principal components following PCA of ClinVar allelic categories across 199 anemia-enriched loci.

Scatter diagrams for each pairwise principal component (PC) provide a framework to visualize the relationships between each comparison, revealing how data points vary for each PC's dimensions alongside an assessment of correlation trends (i.e., positive or negative). Diagonal plots are PC identity scatterplots that reveal specific PC distribution patterns: PC1 exhibits a skewed distribution while PC2-PC4 are normally-distributed. For unique pairwise comparisons, PC1's relationship is equivocal to all other PCs. A subtle positive correlation can be seen for PC4 when compared against either PC2 or PC3.

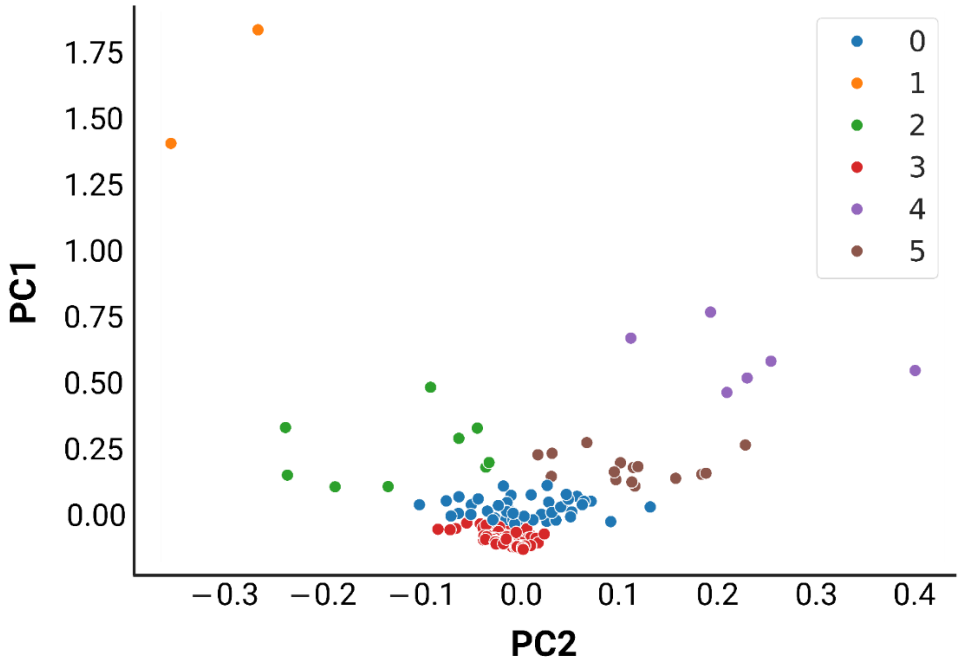

Figure 6: Visualization of ClinVar allelic categories across 199 anemia-enriched loci via PCA-filtered, k-means clustering.

Principal component analysis (PCA) reveals that PC1 and PC2 together account for ~91% of the explained variance ratio across all 4 ClinVar allele categories for each of the 112,534 alleles derived from all 199 anemia genes. An unsupervised machine learning algorithm known as k-means clustering can then be applied on the dimensionally reduced dataset by retaining only PC1 and PC2. Six clusters are seen here, with cluster 1 (orange) distally located to all other centroids. Meanwhile, cluster 3 (red) exhibits short mathematical distances to cluster 0 (blue) indicating a higher degree of similarity.

3.3. Topological data analysis (TDA) of the alleles and loci involved in inherited anemias defines distinct gene interactions

Topological data analysis (TDA) is an applied mathematics tool that relies on projecting data to multidimensional space, and then reducing dimensionality to investigate how datum points interact with one another. In biology, TDA has been utilized in various subdisciplines, including oncology, structural biology, molecular biology, and organismal biology (108–110). Here we sought to project the 112,534 unique alleles drawn from 4 ClinVar categories across 199 anemia loci into four-dimensional space where each axis serviced one of four normalized (i.e., relative) ClinVar categorical allele frequencies. In doing so, each genetic locus would exist as a single point in 4D space and with TDA-mediated dimensional reduction, we can mathematically investigate how the various 199 anemia loci are related to one another. Circos plot diagrams reveal specific gene connectivity interactions across a total of six distinct clusters (Figure 7).

The six nodes and their accompanying gene occupants exhibit a flare configuration (Figure 8). The largest cluster (i.e., node 1) contained 107 genes and the smallest cluster (i.e., node 6) comprised only 4 loci for the ClinVar anemia allele dataset (Figure 8). Averages of each nodes' ClinVar categorical allele frequencies (i.e., B.freq, LB.freq, LP.freq and P.freq) were calculated to approximate how each of the six clusters differed from one another. Node 1 for example is primarily defined by elevated levels of likely benign alleles relative to other allele categories while node 6 is driven mostly by high relative pathogenicity frequencies (Figure 8).

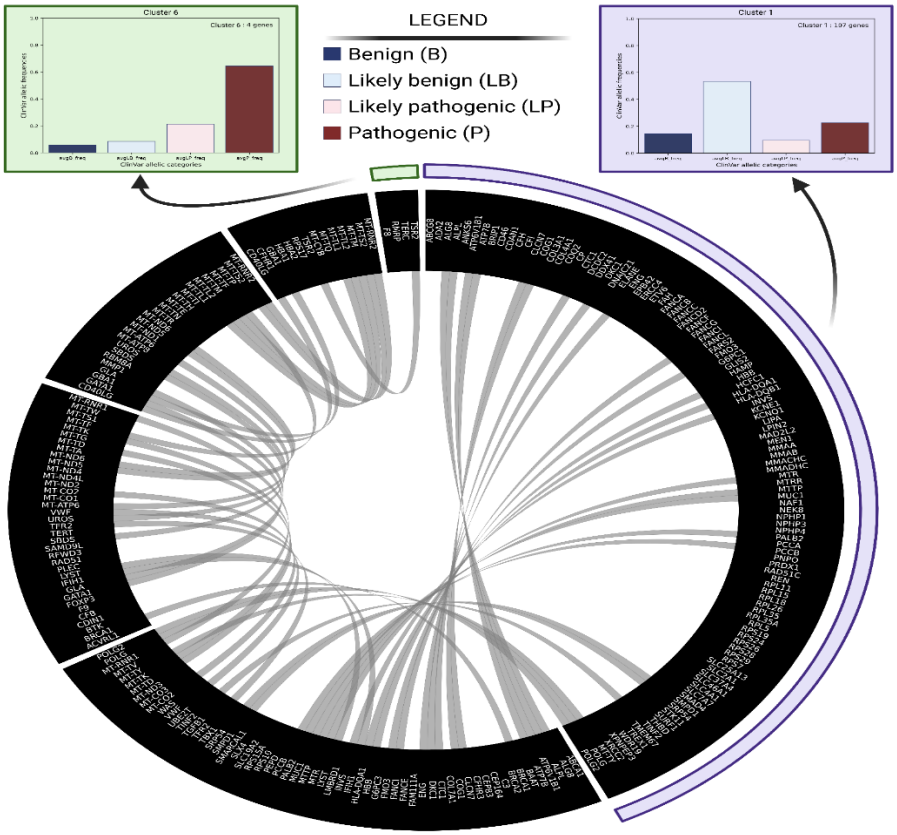

Figure 7: Circos plot of topological data analysis (TDA) of ClinVar allelic categories for 199 anemia-enriched loci.

Topological data analysis (TDA) is an advanced geometry method that permits dimensional reduction on multidimensional datasets. Circos plots are a means of visualizing all node interactions. Connections between genes that populate different clusters (aka nodes) are shown in grey. A total of six clusters are shown here in a flare arrangement. Node definitions are plotted for the most distal clusters in the flare arrangement. Cluster 1 contained the most genes and is defined by a high frequency of likely benign alleles (purple). In contrast, cluster 6 is defined primarily by pathogenic allele prevalence and harbors the fewest number of loci (green).

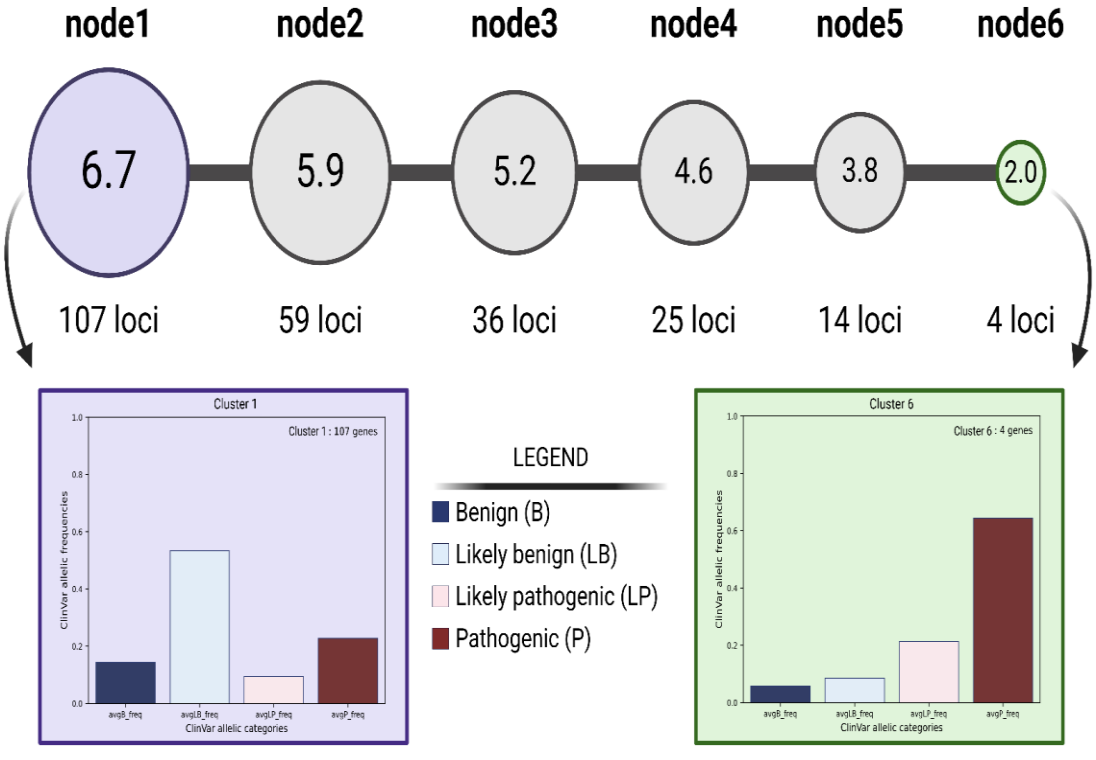

Figure 8: TDA node relationships for ClinVar allelic categories across 199 anemia-enriched loci.

A simple flare arrangement approximates the overall gene relationships with respect to ClinVar categorical allelic frequencies in multidimensional space. Six distinct nodes are connected linearly with decreasing numbers of gene membership from node 1 to node 6. On average, node 1 (purple) is identified primarily by high allelic frequencies in the NCBI ClinVar allelic category. Meanwhile, node 6 (green) is defined by elevated allelic frequency levels in pathogenic allele categories. Numbers shown within each node are the log2-transformed gene numbers for that node. The precise number of genes placed in a cluster is listed below the corresponding node.

3.4. Gene set enrichment analysis (GSEA) of anemia loci reinforces known molecular interactions

Reliance on purely total allele counts or normalized categorical allele frequencies encapsulates to a certain degree the stringency for a locus to resist change (i.e., mutation) but does not reflect a locus' known contributions towards phenotype or clinical manifestation of anemia. To bolster genotype to phenotype functional analyses, we leveraged gene set enrichment analysis (GSEA) using bioinformatics tools originally inspired by transcriptomic studies. We used the g:GOSt method of the g:Profiler suite to seek out phenotypic enrichment terms for each of our 199 anemia loci queried against the Human Phenotype Ontology (HPO) database which stores over 18,000 unique phenotypic terms (i.e., clinical manifestations) belonging to various human afflictions. Datamining the HPO database with our 199 anemia loci using the g:GOSt algorithm revealed 23,626 enrichment term intersections, of which 619 phenotypes were unique HPO terms. To visualize this dataset (i.e., 199 anemia genes by 619 phenotypes), we programmatically generated wordcloud diagrams that represented the most frequent (top 40) genes or phenotypes (Figure 9). In such wordcloud plots, relative term frequencies are approximated by relative font sizes. Interestingly, the most frequently observed gene names derived from numerous genes of the Fanconi Anemia DNA repair pathway, including *FANCA*, *FANCB*, *FANCC*, *FAND2*, *ERCC4*, *BRCA1*, *BRCA2* (Figure 9A). To a lesser extent, genes associated with Diamond-Blackfan anemia, such as *RPS10* and *RPL11*, were often frequently encountered (Figure 9A). A few mitochondrial DNA loci were also frequently encountered (i.e., *MT-CO3*, *MT-ND1*, and *MT-ND4*), which are of particular importance when considering the anemia manifestations that can occur in patients diagnosed with Pearson's syndrome (Figure 9A). Gene ontology analysis of the frequency of phenotypic terms derived from the HPO database when cross-referenced against our list of 199 anemia loci reveals expected terms centering around anemia (Figure 9B).

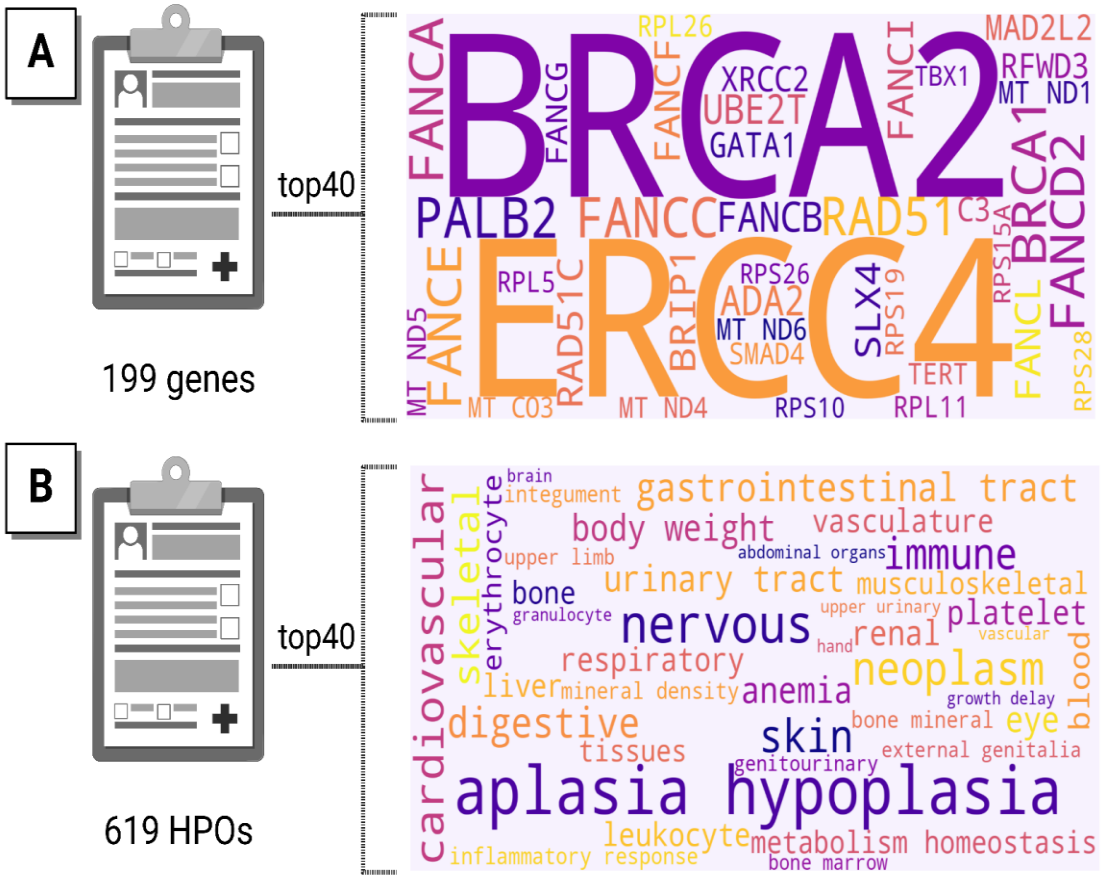

Figure 9: Wordcloud visualization of statistically significant Human Phenotype Ontology (HPO) gene set enrichment analysis (GSEA) on 199 anemia-enriched loci.

Gene ontology leverages various databases that employ tagged features for each gene. We leveraged the Human Phenotype Ontology (HPO) database to systematically examine how our list of 199 scraped genes might exhibit statistically significant enrichment of phenotypic tags. To assess the results, wordclouds of the top 40 terms were plotted. Font sizes of terms indicate either gene name frequency or HPO term frequency. **(A)**: The most frequently encountered genes with HPO enrichment are dominated by Fanconi anemia pathway loci (i.e., *ERCC4*, *BRCA2*, *FANCA*, *FANCB*, *FANCC*, etc.). Of note, the *RPS10* and *RPL11* genes that are known to drive Diamond-Blackfan anemia are also prominent. Mitochondrial loci (i.e., *MT-CO3*, *MT-ND1*, and *MT-ND4*) are also frequently represented across the HPO enrichment terms; large mtDNA deletions constitute the molecular basis for Pearson syndrome which can present with anemia manifestations. **(B)**: Of the ~18,000 unique enrichment terms available at the HPO database, 619 HPO terms exhibit statistically significant gene set enrichment values across the 199 scraped loci in this study. The top 40 most abundantly described phenotype components are illustrated here. The HPO terms aplasia, hypoplasia, and cardiovascular represent the most commonly encountered phenotypic enrichments for these 199 loci.

3.5. Canonical and non-canonical inherited anemias can be visualized phenotypically as distinct entities using TDA

To investigate the behavior of each of the 199 anemia loci with respect to a gene's precise possession of a unique permutation of statistically significant HPO terms, we projected each locus into 619th-dimensional space where each axis serviced a single phenotypic term. Leveraging TDA, we effectively reduced this multidimensional projection into just two dimensions, and while collapsing dimensionality, we captured the behavior of each gene with respect to one another. Twelve distinct clusters or nodes are observed in aggregate (Figures 10-11). Each of the twelve nodes is positioned in exactly one of three connected components: component 1 has seven nodes, component 2 contains three nodes, and component 3 possesses two nodes (Figures 10-11). Close examination of each gene name found in each node of each component reveals a molecular reconstitution of canonically studied inherited anemias, such as Diamond-Blackfan anemia and Fanconi anemia (Figure 11A-11B). For Diamond-Blackfan anemia, there were a total of 12 genes across all three nodes of component 2 that encode known large ribosomal subunits while 19 genes from two of three nodes of component 2 specify the small ribosomal subunits (Figure 11A). For Fanconi anemia defined by component 3, node 3.1 and node 3.2 encapsulated 13 and 11 genes, respectively (Figure 11B). Pearson syndrome is characterized by a sideroblastic anemia caused by large scale deletions of mtDNA genes. Similarly, *POLG*-related disorders are a type of mitochondrial disease in which the catalytic subunit for the mtDNA polymerase complex is abrogated. In patients diagnosed with *POLG* lesions, anemia manifestations elevate the risk of lethality by nearly four-fold (74). TDA of the 619th-dimensional anemia loci / HPO phenotype dataset revealed a third component comprising genes that are critical in Pearson's syndrome as well as *POLG*-related disorders (Figure 11C). All 37 essential mitochondrial genes were mapped across 6 of 7 nodes in component 1 (i.e., nodes 1.1-1.6 but not 1.7) (Figure 11C). Of particular importance, the *POLG* locus occupied the most connected node (i.e., node 1.4) (Figure 11C). Taken together, TDA digestion of HPO GSEA on the 199 anemia loci reconstitutes known molecular interactions in inherited anemias of the canonical (e.g. Diamond-Blackfan anemia and Fanconi anemia) and the non-canonical (e.g. Pearson's syndrome and *POLG*-related disorders) form.

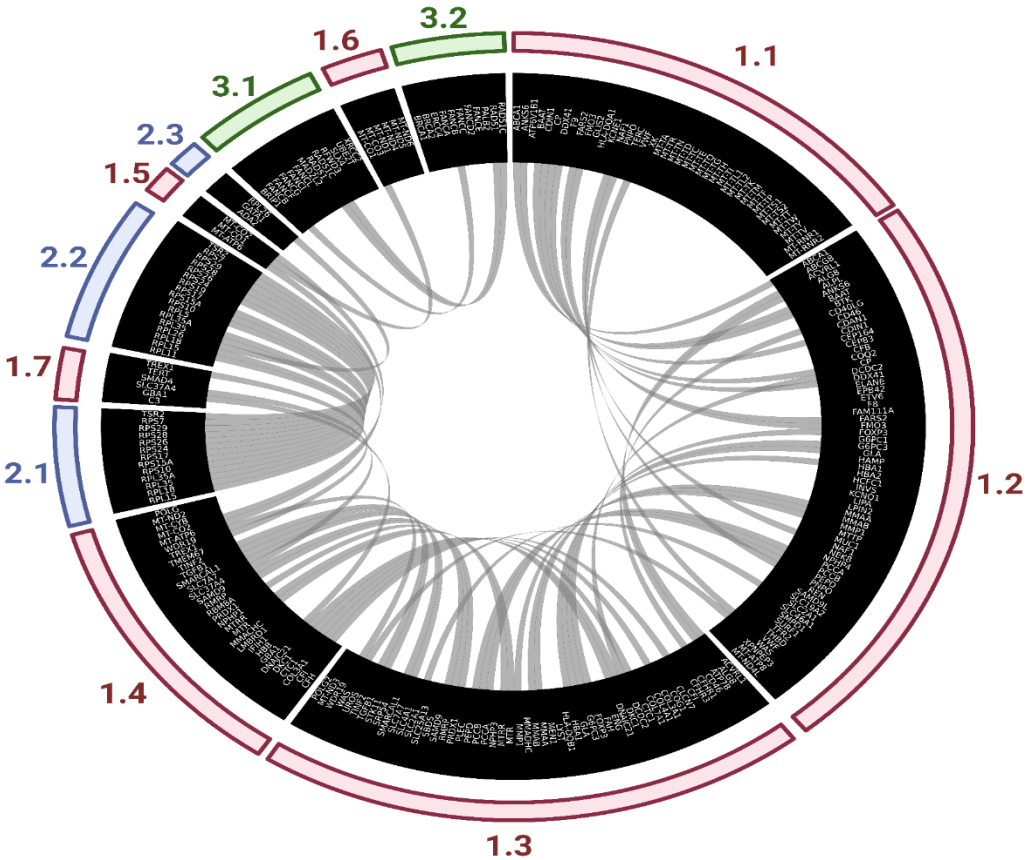

Figure 10: Circos plot of topological data analysis (TDA) performed on statistically significant Human Phenotype Ontology (HPO) gene set enrichment analysis (GSEA) terms for 199 anemia-enriched loci.

Topological data analysis is an advanced geometry method that permits dimensional reduction on multidimensional datasets. We mapped in 619th dimensional space the precise "phenotypic" position of each of genetic locus, and then reduced dimensionality by applying the TDA algorithm. A Circos visualization of TDA output illustrates how each gene relates to all other anemia loci. Of note, three distinct components with varying numbers of clusters (aka nodes) are visible. Component 1 (red) is made of 7 total nodes (i.e., 1.1 – 1.7) while component 2 (blue) consists of three nodes (i.e., 2.1-2.3) and component 3 (green) just two nodes (i.e., 3.1 and 3.2). Gene connections between different nodes are shown in grey. Component 2 (blue) is almost entirely made up of genes that are associated with Diamond-Blackfan anemia. Component 3 (green) consists solely of genes involved in Fanconi anemia. Lastly, component 1 (red) includes numerous mtDNA loci dispersed throughout the component's node architecture as well as *POLG2* and *POLG1* loci found in 1.3 and 1.4, respectively. The significance of mtDNA deletions manifests in both Pearson's syndrome as well as *POLG*-related disorders.

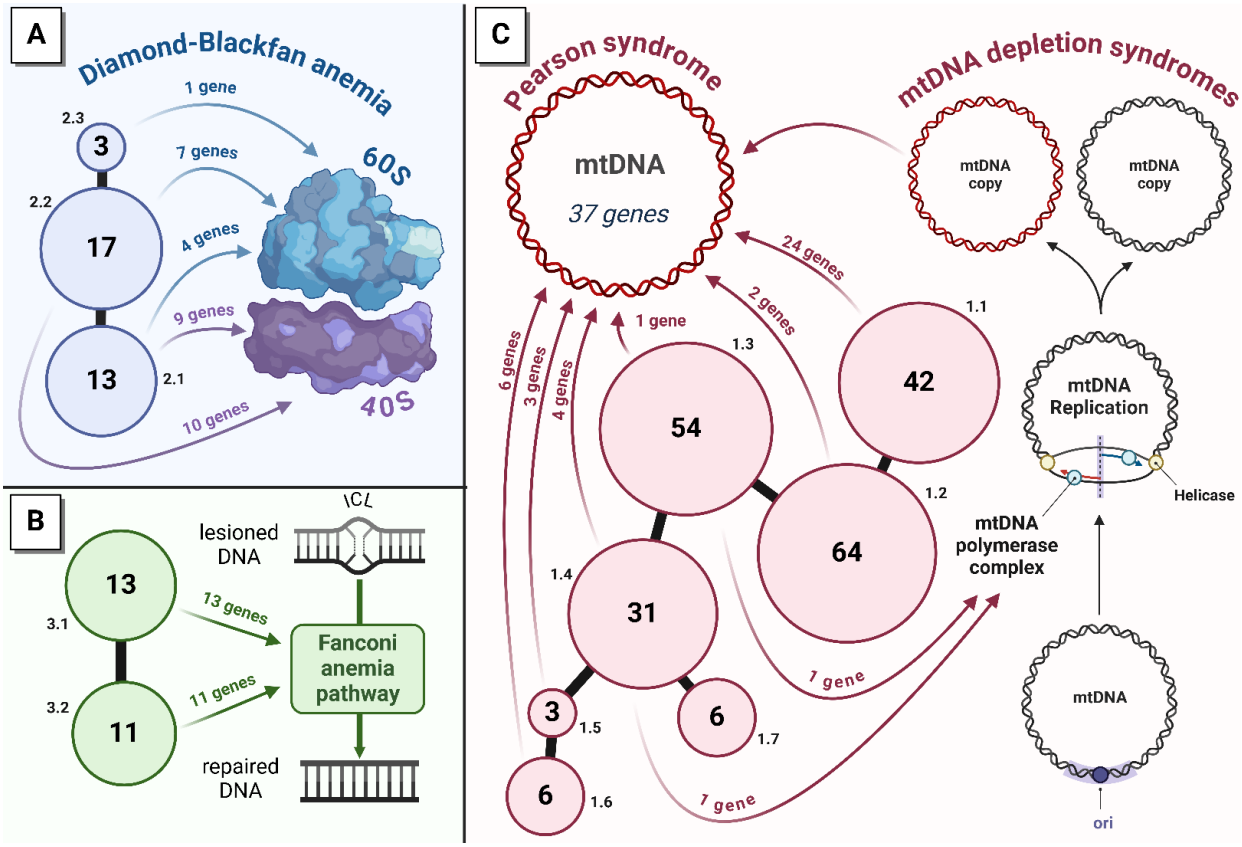

Figure 11: TDA node relationships of statistically significant Human Phenotype Ontology (HPO) gene set enrichment analysis (GSEA) terms for 199 anemia-enriched loci.

A summary figure indicating the overall prevalence of known anemia genes associated with known phenotypic patterns is shown here. The nodes of the three components resulting from the TDA mapper algorithm are graphically illustrated via filled-in circles (i.e., blue, green or red). For any given component, specific nodes are drawn as circles of varying diameters proportional to the number of genes (i.e., log2-transformed gene sums) placed within the node by the TDA algorithm. **(A)**: The Diamond-Blackfan anemia (DBA) component is shown in blue and consists of 3 nodes. Nearly all genes found in each of the three DBA nodes contribute to either large ribosomal subunit or small ribosomal subunit activity. DBA patients exhibit anemia primarily as an erythropoietic block. **(B)**: The Fanconi anemia (FA) component is shown in green and is made up of 2 distinct nodes. All genes found in each of the FA nodes have known molecular functions in DNA repair, specifically for resolving translesion DNA adducts. Deficits in these genes enhance mutagenesis rates in FA patients, and an erythropoietic block is implicated in anemia manifestations experienced by this cohort. **(C)**: The mitochondrial function component is shown in red. The mitochondrial chromosome is made of 37 essential mitochondrial genes, and 6 of 7 nodes are enriched in similar HPO term permutations within this component for all 37 mtDNA genes. A key central node (i.e., 1.4) of this component contains the *POLG1* gene which encodes the mitochondrial DNA polymerase subunit.

4. Discussion

The numerous inherited anemias distinctly manifest from a cadre of cellular, molecular and physiological perturbations (36,37) (Figure 12). Cellular production rates of erythrocytes (aka red blood cells, or RBCs) are facilitated by erythropoiesis by common myeloid progenitors (CMPs) in the bone marrow of long bones (49,51) (Figure 12A). Defective erythropoiesis can thus manifest as an inherited anemia, and an erythropoietic block is the main driving force for anemia manifestations in patients diagnosed with congenital dyserythropoiesis anemia (47,48), Fanconi anemia (38–42,116–118), Diamond-Blackfan anemia (43–46,119–122), and *POLG*-related disorders coupled to anemia (74–76) (Figure 12A). Conversely, elevated rates of cellular destruction by splenic macrophages in the red pulp results in the hemolytic anemias, as seen in hereditary spherocytosis (55,56), glucose-6-phosphate dehydrogenase

deficiency (57,58), and in various forms of hemolytic-uremic syndromes (54) (Figure 12B). The globinopathies affect either the structure and function of hemoglobin subunits (i.e., Sickle-cell anemia) or manifest due to *HBA* or *HBB* copy number loss (i.e., thalassemias) (59–62,64,65) (Figure 12C). Lastly, abnormalities in clotting action due to diminished subendothelial function (i.e., Von Willebrand disease) or attenuated clotting cascade activation (i.e., hemophilia) represent an additional means of manifesting anemia due to perturbed physiological functions (67–70,123–125) (Figure 12D). This last category can be expanded to include other extensive bleeding conditions, such as epidermolysis bullosa (72,73,126,127). Anemia, as an inherited phenotype, is thus understandably complex due to the many genetic means by which it might manifest through the germline. Our work here helps establish a taxonomic framework in which to quantify the overall similarities amongst this wide berth of inherited anemias.

To this end, we performed a systems bioinformatics analysis of the best-curated loci underlying inherited anemias. Our approach expanded anemia loci counts from prior documentation of ~70 genes to roughly 200 genes in this study (37). The novelty of our overall analysis is tethered to a two-pronged approach, leveraging both multivariate statistical analysis (MSA) and topological data analysis (TDA) on allelic and phenotypic data derived from ~200 unique scraped anemia loci. Although both MSA and TDA effectively perform dimensional reduction on multidimensional datasets, each approach is couched in an entirely different field of either statistics or applied mathematics, respectively (101,108). Nonetheless, our bioinformatics pipeline reveals that both MSA and TDA converge in agreement. More importantly, our analytical pipeline sheds light not only on the well-documented (aka canonical) anemias, such as Diamond-Blackfan anemia and Fanconi anemia, but also on the non-canonical anemias, which we define as genetic disorders that exhibit poor prognostic outcomes when accompanied by anemia manifestations. Excellent examples of what we consider non-canonical inherited anemias include the genetic disorders that afflict the mitochondria, such as Pearson's syndrome and *POLG*-related disorders (74–76,86–88). Mitochondrial diseases are typically thought of as a genetic disorder with primarily neurological manifestations, and are successfully treated as such; however, treating additionally for anemia presentation in afflicted patients with *POLG*-related disorders has the potential to quadruple patient survivorship outcomes (74). Leveraging known allele counts and allele frequencies, we show overall connectivity maps amongst genes associated with anemia manifestations across disparate genetic conditions. We also demonstrate a direct means of systematically relating phenotype to gene function to recapitulate known molecular functions. The utility of this approach should not be understated as it reveals gene similarities agnostic of known disease states. For instance, the following 10 genes from our list of 199 anemia loci are each implicated in anemia manifestations involving nutritional perturbations centered on B-complex vitamins: *HCFC1*, *LMBRD1*, *MMAA*, *MMAB*, *MMACHC*, *MMADHC*, *MTR*, *MTRR*, *PRDX1*, *SLC46A1* (128,129). Across all 10 genetic loci, effective cobalamin (vitamin B12) utilization is perturbed in some fashion except for lesions in *SLC46A1* which affect folate (vitamin B9) metabolism (128,129). A common clinical manifestation for nutritional anemias caused by inadequate intake of dietary cobalamin and/or folate is that of megaloblastic anemia (13–16,20,21). Consulting our HPO TDA analysis reveals that all 10 of these genes that are known to be involved in either folate or cobalamin utilization fall within component 1, spanning the three nodes 1.2, 1.3 and 1.4 (Figure 11). Thus, these genetic loci, which play an outsized role at the intersection of host genetics and adequate nutrients, are recapitulated in multidimensional space, and interestingly occupy adjacent clusters to those of the mitochondrial diseases, such as Pearson's syndrome (nodes 1.1-1.6) and *POLG*-related disorders (nodes 1.3-1.4) (76,86–88) (Figure 11). Perhaps this close node proximity should be unsurprising as the B-complex vitamins are essential for a wide set of biological processes, including mitochondrial electron transport chain activity during the oxidative phosphorylation (OXPHOS) stages of cellular respiration (76,130,131). In fact, momentary bursts in mitochondrial biogenesis are critical as erythropoietic progenitors approach RBC terminal maturation stages, as an expanded cellular pool of mitochondrial organelles 1) services cellular bioenergetic requirements for increased OXPHOS activity to offset elevated *HBA1*, *HBA2*, and *HBB* expression costs, and 2) facilitate four of the eight steps in heme anabolism in their concerted quest to assemble an adequate amount of functional hemoglobin tetramers (49–51,131,132). Other examples can help illustrate the utility of our systems bioinformatics analysis of inherited anemias. For instance, the disease condition known as epidermolysis bullosa manifests as ECM connectivity phenotypes resulting in detachment of the epidermis from the underlying dermal regions of the skin (133). Consequently, extensive tissue damage accompanied by localized, chronic bleeding complications can manifest as an indirect form of anemia in patients afflicted with epidermolysis bullosa (72,73,126,127). In our TDA digestion of HPO GSEA performed on anemia-enriched loci, *COL7A1* and *PLEC* genes are found in node 1.3, and mutations in either gene can result in different disease forms of epidermolysis bullosa (Figure 11). In a similar fashion, *VWF* occupies node 1.2 while clotting factors IX (*F9*) and VIII (*F8*) populate 1.1 and 1.2, respectively (Figure 11). Three genetic loci (*CFH*, *CFHR1*, and *CFHR3*) contribute to hemolytic-uremic syndrome (54,134,135), and these three loci occupied adjacent nodes along component 1 (i.e., *CFHR1* and *CFHR3* were positioned in 1.3 next to *CFH* in 1.4) (Figure 11). The proximity of loci responsible for several distinct genetic disorders (i.e., epidermolysis bullosa, hemophilia A, hemophilia B, VWF disease, and various hemolytic-uremic syndromes) provides a means to mathematically relate such otherwise seemingly disparate genetic conditions. To our knowledge, the approach employed in our work is thus the first to quantitatively measure and visualize how similar such genetic loci lie in phenotypic and categorical allelic multidimensional space with respect to the inherited anemias.

The most limiting aspect of our approach lies in our original datamining scrape operations in which we pursued only genetic loci that had existing accession entries simultaneously across three databases: NCBI GTR, OMIM, and NCBI GeneReviews. For some well-studied inherited anemias, a dedicated accession page in a single database would preclude them from our analytical pipeline. For instance, iron-refractory iron deficiency anemia (IRIDA) is an inherited anemia that manifests as a microcytic, hypochromic anemia due to lesions at the *TMPRSS6* locus (136). While IRIDA as a genetic disorder exists in the OMIM database (OMIM #206200), and likewise, *TMPRSS6* gene is OMIM-cataloged (OMIM *609862), there is a striking lack of a dedicated IRIDA accession at NCBI GeneReviews.

Therefore, *TMPRSS6* as a gene involved in IRIDA forms of microcytic anemia was missing in our list of 199 anemia loci, and thus we did not capture how *TMPRSS6* might interact with other anemia genes in our systems bioinformatics analytical pipeline. Similarly, of the eight genes involved in the heme biosynthesis pathway utilized by erythropoietic progenitors (i.e., *ALAS2*, *ALAD*, *HMBS*, *UROS*, *UROD*, *CPOX*, *PPOX*, and *FECH*), only the *UROS* locus was captured in our NCBI GTR scrape (50,51). Lesions at all eight loci typically manifest with varying degrees of sideroblastic forms of anemia and are thus important when considering a comprehensive classification scheme for anemias of genetic etiologies (50–53). Thus, a major limitation to our study is that, in pursuing the most well-curated genetic conditions that exhibit anemia manifestations, we may have inadvertently set those criteria too high, which may effectively exclude other inherited anemias such as IRIDA and sideroblastic anemias from our current systems bioinformatics analysis (52,53,136).

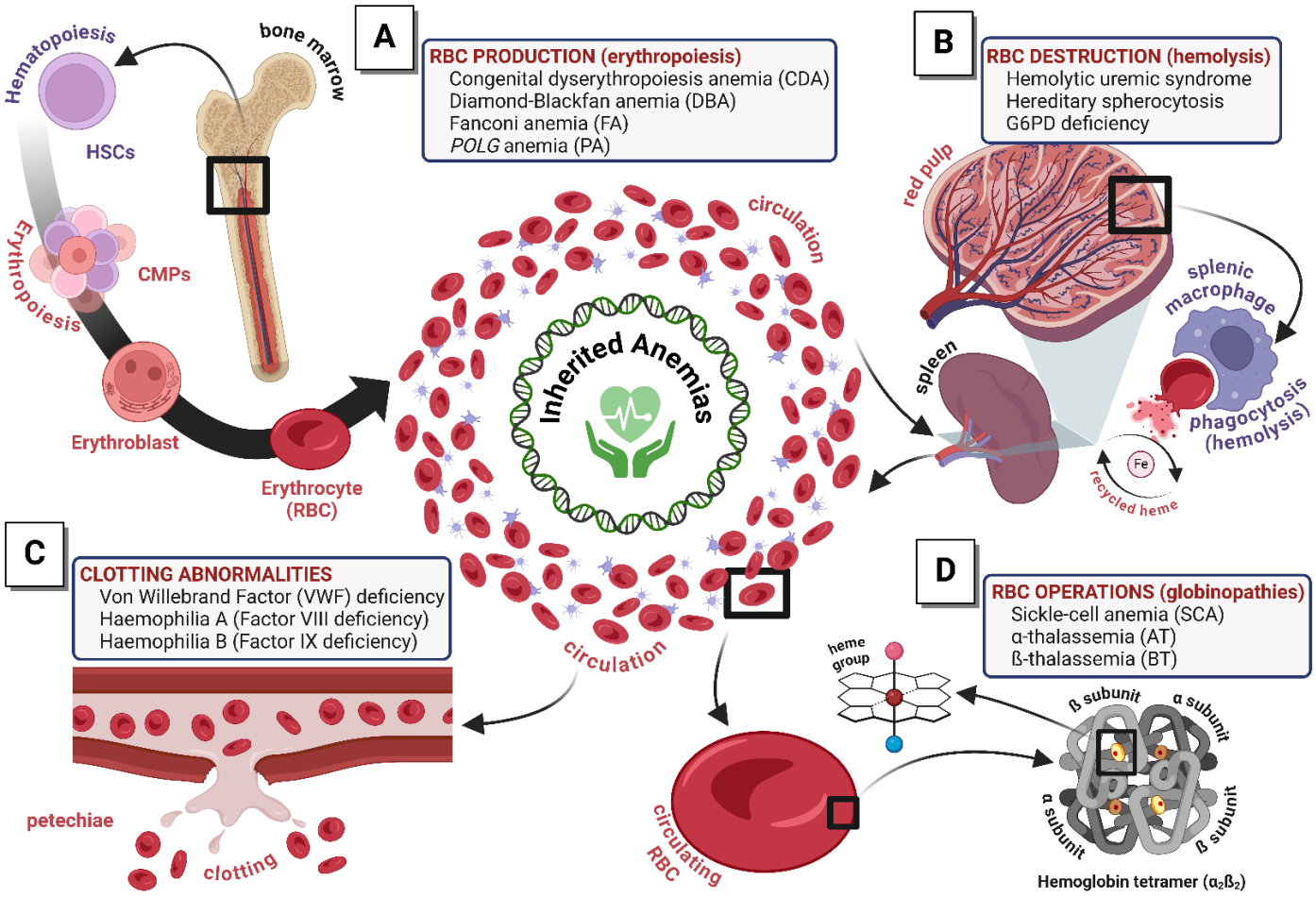

Figure 12: Summary of inherited anemias.

The genetic basis underlying the inherited anemias is well established at the cellular and molecular level. **(A)**: Abnormalities in erythropoiesis can result in reduced red blood cell (RBC) production in the bone marrow as a cellular basis for anemia. Congenital dyserythropoiesis anemia (CDA), Fanconi anemia (FA) and Diamond-Blackfan anemia (DBA) are canonical inherited anemias of this type. Non-canonical, inherited anemias like those accompanying *POLG*-related disorders also likely represent an RBC synthesis block as well. **(B)**: Elevated levels of hemolysis, primarily by splenic macrophages in the red pulp of the spleen, can also result in a cellular basis for anemia. Examples of increased RBC destruction rates are typified by hereditary spherocytosis and glucose-6-phosphate dehydrogenase (G6PD) deficiency. **(C)**: Clotting disorders, either due to defective clotting cascades in the vascular lumen (i.e., hemophilia A or hemophilia B) or defective Von Willebrand Factor action in the subendothelium, comprises a cellular-physiological basis for anemia. **(D)**: Defective hemoglobin production or operation constitutes several globinopathies, and includes Sickle-cell anemia, and the various thalassemias.

5. Conclusions

Future work should address this study's limitations, and thus include additional loci. Namely, a comprehensive system bioinformatics analysis of anemia manifestations in any genetic disorder might reveal hitherto undetected gene-phenotype patterns. With an expanded list of genes, there is a good chance that additional, deeper component and/or cluster definitions reveal themselves as related to categorical allele levels and/or the extent of phenotypic manifestation. In doing so, genetic disorders that occasionally manifest with anemia, and as such are not typically considered inherited anemias by clinicians, might receive a TDA-mediated taxonomic classification. Close integration with clinicians may thus drive better prognostic outcomes in patients cases involving inherited anemias. For example, the recognition of poor survivorship of patients diagnosed with *POLG*-related disorders whom simultaneously present with anemia manifestations in the clinic warrants

expanded Kaplan-Meier survivorship surveillance for all inherited anemia loci (74). As expected, the overall complexity underlying any given allele's penetrance (i.e., degree of loss-of-function), any particular gene's level of essentiality (i.e., stringent constancy due to a dearth of gene duplication events), or any given genotype (i.e., complete dominance versus haploinsufficiency) dramatically increases the level of understanding needed to enhance clinical outcomes. A combination of MSA and TDA approaches can thus help elucidate key interactions buried underneath such layers (i.e., allele-level, gene-level, and genotype-level) that contribute towards disease complexity in a combinatorial way. An interesting future topic building on this article's current taxonomic framework might be to analyze known treatments for the inherited anemias in multidimensional space, and after using MSA and TDA to collapse dimensionality, to check if TDA treatment clusters resemble TDA allelic and/or TDA phenotypic clusters. Such work might provide a quantitative means of relating not only genotype to phenotype, but also to treatment mode and to survivorship surveillance, in the overall pursuit of enhancing patient outcomes.

Acknowledgments
The authors would like to acknowledge the highly collaborative environment fostered by disparate departments at Gannon University. Such an environment provides an ample opportunity for engaged, interdisciplinary research projects.

Funding
There are no sources of funding to declare.

Author contributions
Conceptualization, C.R.T., R.G.L., T.Y., G.V.; methodology, M.E.S., F.S.V., A.D.B., M.D.G., R.G.L., T.Y., G.V.; software, M.E.S., F.S.V., A.D.B., M.D.G., R.G.L., G.V.; validation, C.R.T., J.H.K., M.D.G., R.G.L., T.Y., G.V.; formal analysis, M.D.G., R.G.L., G.V.; investigation, M.E.S., F.S.V., A.D.B., J.H.K., T.Y., G.V.; resources, C.R.T., M.D.G., R.G.L., T.Y., G.V.; data curation, G.V.; writing—original draft preparation, G.V.; writing—review and editing, C.R.T., M.E.S., Z.E.G., A.R., F.S.V., A.D.B., A.K.M., J.H.K., M.D.G., R.G.L., T.Y., G.V.; visualization, M.D.G., R.G.L., G.V.; supervision, R.G.L., T.Y., G.V.; project administration, G.V.; funding acquisition, n/a. All authors have read and agreed to the published version of the manuscript.

Conflict of interest
The author(s) declare no conflict of interest.

Data availability statement
All python scripts employed in this body of work are freely available and can be found on our GitHub site: https://github.com/VitamOrdinatio/inherited_anemias. Great care has been taken to heavily document the codebase for legibility and comprehension using verbose documentation in the source code. GitHub subfolders housing 15 different python scripts for 15 different bioinformatics tasks are available for examination, download, verification and/or experimental replication. In total, we provide over 4,200 lines of python code across all 15 GitHub subfolders. Additionally, each python script contains a header block defining requisite virtual environment construction using pip venv installation techniques in BASH.

Sample availability
The author(s) declare that no physical samples were used in this study.

Supplementary materials
Although no supplementary material is cited throughout this study, readers are encouraged to visit this paper's dedicated GitHub repository for each pipeline step's cadre of input and output files.

References

1. Garcia-Casal MN, Dary O, Jefferds ME, Pasricha S. Diagnosing anemia: Challenges selecting methods, addressing underlying causes, and implementing actions at the public health level. Annals of the New York Academy of Sciences. 2023 Jun;1524(1):37–50.
2. Gardner WM, Razo C, McHugh TA, Hagins H, Vilchis-Tella VM, Hennessy C, et al. Prevalence, years lived with disability, and trends in anaemia burden by severity and cause, 1990–2021: findings from the Global Burden of Disease Study 2021. The Lancet Haematology. 2023 Sep;10(9):e713–34.
3. Hess SY, Owais A, Jefferds MED, Young MF, Cahill A, Rogers LM. Accelerating action to reduce anemia: Review of causes and risk factors and related data needs. Annals of the New York Academy of Sciences. 2023 May;1523(1):11–23.
4. Mildon A, Lopez De Romaña D, Jefferds MED, Rogers LM, Golan JM, Arabi M. Integrating and coordinating programs for the management of anemia across the life course. Annals of the New York Academy of Sciences. 2023 Jul;1525(1):160–72.
5. Martinsson A, Andersson C, Andell P, Koul S, Engström G, Smith JG. Anemia in the general population: prevalence, clinical correlates and prognostic impact. Eur J Epidemiol. 2014 Jul;29(7):489–98.
6. Penninx BWJH, Pahor M, Woodman RC, Guralnik JM. Anemia in Old Age Is Associated With Increased Mortality and Hospitalization. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences. 2006 May 1;61(5):474–9.
7. Li Z, Zhou T, Li Y, Chen P, Chen L. Anemia increases the mortality risk in patients with stroke: A meta-analysis of cohort studies. Sci Rep. 2016 May 23;6(1):26636.
8. Karoopongse E, Srinonprasert V, Chalermsri C, Aekplakorn W. Prevalence of anemia and association with mortality in community-dwelling elderly in Thailand. Sci Rep. 2022 Apr 30;12(1):7084.
9. Luiz MM, Schneider IJC, Kuriki HU, Fattori A, Corrêa VP, Steptoe A, et al. The combined effect of anemia and dynapenia on mortality risk in older adults: 10-Year evidence from the ELSA cohort study. Archives of Gerontology and Geriatrics. 2022 Sep;102:104739.
10. Haschke F, Javaid N. Nutritional Anemias. Acta Paediatrica. 1991 Apr;80(s374):38–44.
11. Brabin BJ, Premji Z, Verhoeff F. An Analysis of Anemia and Child Mortality. The Journal of Nutrition. 2001 Feb;131(2):636S-648S.
12. Ozdemir N. Iron deficiency anemia from diagnosis to treatment in children. Turk Arch Ped. 2015 Apr 10;50(1):11–9.
13. Gallagher PG. Anemia in the pediatric patient. Blood. 2022 Aug 11;140(6):571–93.

14. Martinez-Torres V, Torres N, Davis JA, Corrales-Medina FF. Anemia and Associated Risk Factors in Pediatric Patients. PHMT. 2023 Sep;14:267–80.
15. Stover PJ, Field MS. Vitamin B-6. Advances in Nutrition. 2015 Jan;6(1):132–3.
16. Wilson MP, Plecko B, Mills PB, Clayton PT. Disorders affecting vitamin B6 metabolism. J of Inher Metab Disea. 2019 Jul;42(4):629–46.
17. Fenech M. Folate (vitamin B9) and vitamin B12 and their function in the maintenance of nuclear and mitochondrial genome integrity. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2012 May;733(1–2):21–33.
18. Santoyo-Sánchez A, Aponte-Castillo JA, Parra-Peña RI, Ramos-Peñafiel CO. Dietary recommendations in patients with deficiency anaemia. Revista Médica Del Hospital General De México. 2015 Jul;78(3):144–50.
19. Socha DS, DeSouza SI, Flagg A, Sekeres M, Rogers HJ. Severe megaloblastic anemia: Vitamin deficiency and other causes. CCJM. 2020 Mar;87(3):153–64.
20. Oh R, Brown DL. Vitamin B12 deficiency. Am Fam Physician. 2003 Mar 1;67(5):979–86.
21. Green R, Allen LH, Bjørke-Monsen AL, Brito A, Guéant JL, Miller JW, et al. Vitamin B12 deficiency. Nat Rev Dis Primers. 2017 Jun 29;3(1):17040.
22. Person DonaldA. Gloves and socks syndrome. The Lancet. 1996 Apr;347(9008):1125–6.
23. Chisaka H, Morita E, Yaegashi N, Sugamura K. Parvovirus B19 and the pathogenesis of anaemia. Reviews in Medical Virology. 2003 Nov;13(6):347–59.
24. Burkhardt J, Anemana SD, Gellert S, Cramer JP, Ehrhardt S, Laryea S, et al. Manifestation And Outcome Of Severe Malaria In Children In Northern Ghana. The American Journal of Tropical Medicine and Hygiene. 2004 Aug 1;71(2):167–72.
25. Reyburn H. Association of Transmission Intensity and Age With Clinical Manifestations and Case Fatality of Severe Plasmodium falciparum Malaria. JAMA. 2005 Mar 23;293(12):1461.
26. Tripathy R, Parida S, Das L, Mishra DP, Tripathy D, Das MC, et al. Clinical Manifestations and Predictors of Severe Malaria in Indian Children. Pediatrics. 2007 Sep 1;120(3):e454–60.
27. Perkins DJ, Were T, Davenport GC, Kempaiah P, Hittner JB, Ong'echa JM. Severe Malarial Anemia: Innate Immunity and Pathogenesis. Int J Biol Sci. 2011;7(9):1427–42.
28. Friedman JF, Kanzaria HK, McGarvey ST. Human schistosomiasis and anemia: the relationship and potential mechanisms. Trends in Parasitology. 2005 Aug;21(8):386–92.
29. Adam I, ALhabardi NA, Al-Wutayd O, Khamis AH. Prevalence of schistosomiasis and its association with anemia among pregnant women: a systematic review and meta-analysis. Parasites Vectors. 2021 Mar 2;14(1):133.
30. Gebrehana DA, Molla GE, Endalew W, Teshome DF, Mekonnen FA, Angaw DA. Prevalence of schistosomiasis and its association with anemia in Ethiopia, 2024: a systematic review and meta-analysis. BMC Infect Dis. 2024 Sep 27;24(1):1040.
31. Županić-Krmek D, Sučić M, Bekić D. Anemia Of Chronic Disease: Illness Or Adaptive Mechanism. Acta Clin Croat. 2014;53(3):348–54.
32. Sifakis S, Pharmakides G. Anemia in Pregnancy. Annals of the New York Academy of Sciences. 2000 Apr;900(1):125–36.
33. Goonewardene M, Shehata M, Hamad A. Anaemia in pregnancy. Best Practice & Research Clinical Obstetrics & Gynaecology. 2012 Feb;26(1):3–24.
34. Noronha JA, Khasawneh EA, Seshan V, Ramasubramaniam S, Raman S. Anemia in Pregnancy—Consequences and Challenges: A Review of Literature. Journal of South Asian Federation of Obstetrics and Gynaecology. 2012 Apr;4(1):64–70.
35. Karami M, Chaleshgar M, Salari N, Akbari H, Mohammadi M. Global Prevalence of Anemia in Pregnant Women: A Comprehensive Systematic Review and Meta-Analysis. Matern Child Health J. 2022 Jul;26(7):1473–87.
36. Vieth JT, Lane DR. Anemia. Emergency Medicine Clinics of North America. 2014 Aug;32(3):613–28.
37. Russo R, Marra R, Rosato BE, Iolascon A, Andolfo I. Genetics and Genomics Approaches for Diagnosis and Research Into Hereditary Anemias. Front Physiol. 2020 Dec 22;11:613559.
38. Kottemann MC, Smogorzewska A. Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. Nature. 2013 Jan 17;493(7432):356–63.
39. Brosh RM, Bellani M, Liu Y, Seidman MM. Fanconi Anemia: A DNA repair disorder characterized by accelerated decline of the hematopoietic stem cell compartment and other features of aging. Ageing Research Reviews. 2017 Jan;33:67–75.
40. García-de-Teresa B, Rodríguez A, Frias S. Chromosome Instability in Fanconi Anemia: From Breaks to Phenotypic Consequences. Genes. 2020 Dec 21;11(12):1528.
41. Kolinjivadi AM, Crismani W, Ngeow J. Emerging functions of Fanconi anemia genes in replication fork protection pathways. Human Molecular Genetics. 2020 Oct 20;29(R2):R158–64.
42. Rageul J, Kim H. Fanconi anemia and the underlying causes of genomic instability. Environ and Mol Mutagen. 2020 Aug;61(7):693–708.
43. Flygare J, Karlsson S. Diamond-Blackfan anemia: erythropoiesis lost in translation. Blood. 2007 Apr 15;109(8):3152–4.
44. Lipton JM, Ellis SR. Diamond-Blackfan Anemia: Diagnosis, Treatment, and Molecular Pathogenesis. Hematology/Oncology Clinics of North America. 2009 Apr;23(2):261–82.
45. Engidaye G, Melku M, Enawgaw B. Diamond Blackfan Anemia: genetics, pathogenesis, diagnosis and treatment. The Journal of the International Federation of Clinical Chemistry and Laboratory Medicine. 2019 Mar 1;30(1):67–81.
46. Da Costa L, Leblanc T, Mohandas N. Diamond-Blackfan anemia. Blood. 2020 Sep 10;136(11):1262–73.
47. Iolascon A, Heimpel H, Wahlin A, Tamary H. Congenital dyserythropoietic anemias: molecular insights and diagnostic approach. Blood. 2013 Sep 26;122(13):2162–6.
48. Iolascon A, Andolfo I, Russo R. Congenital dyserythropoietic anemias. Blood. 2020 Sep 10;136(11):1274–83.
49. Liu X, Zhang Y, Ni M, Cao H, Signer RAJ, Li D, et al. Regulation of mitochondrial biogenesis in erythropoiesis by mTORC1-mediated protein translation. Nat Cell Biol. 2017 Jun 1;19(6):626–38.

50. Ricci A, Di Betto G, Bergamini E, Buzzetti E, Corradini E, Ventura P. Iron Metabolism in the Disorders of Heme Biosynthesis. Metabolites. 2022 Aug 31;12(9):819.
51. Menon V, Slavinsky M, Hermine O, Ghaffari S. Mitochondrial regulation of erythropoiesis in homeostasis and disease. Br J Haematol. 2024 Aug;205(2):429–39.
52. Ducamp S, Fleming MD. The molecular genetics of sideroblastic anemia. Blood. 2019 Jan 3;133(1):59–69.
53. Abu-Zeinah G, DeSancho MT. Understanding Sideroblastic Anemia: An Overview of Genetics, Epidemiology, Pathophysiology and Current Therapeutic Options. JBM. 2020 Sep;Volume 11:305–18.
54. Morel-Maroger L, Kanfer A, Solez K, Sraer JD, Richet G. Prognostic importance of vascular lesions in acute renal failure with microangiopathic hemolytic anemia (hemolytic-uremic syndrome): Clinicopathologic study in 20 adults. Kidney International. 1979 May;15(5):548–58.
55. Bolton-Maggs PHB, Stevens RF, Dodd NJ, Lamont G, Tittensor P, King M -J., et al. Guidelines for the diagnosis and management of hereditary spherocytosis. Br J Haematol. 2004 Aug;126(4):455–74.
56. Polizzi A, Dicembre LP, Failla C, Matola TD, Moretti M, Ranieri SC, et al. Overview on Hereditary Spherocytosis Diagnosis. Int J Lab Hematology. 2024 Oct 28;ijlh.14376.
57. Luzzatto L, Ally M, Notaro R. Glucose-6-phosphate dehydrogenase deficiency. Blood. 2020 Sep 10;136(11):1225–40.
58. Garcia AA, Koperniku A, Ferreira JCB, Mochly-Rosen D. Treatment strategies for glucose-6-phosphate dehydrogenase deficiency: past and future perspectives. Trends in Pharmacological Sciences. 2021 Oct;42(10):829–44.
59. Kavanagh PL, Fasipe TA, Wun T. Sickle Cell Disease: A Review. JAMA. 2022 Jul 5;328(1):57.
60. Vijian D, Wan Ab Rahman WS, Ponnuraj KT, Zulkafli Z, Mohd Noor NH. Molecular Detection of alpha Thalassemia: A Review of Prevalent Techniques. MMJ [Internet]. 2021 [cited 2024 Dec 11]; Available from: https://medeniyetmedicaljournal.org/jvi.aspx?pdir=medeniyet&plng=eng&un=MEDJ-14603
61. Lal A, Vichinsky E. The Clinical Phenotypes of Alpha Thalassemia. Hematology/Oncology Clinics of North America. 2023 Apr;37(2):327–39.
62. Tamary H, Dgany O. Alpha-Thalassemia. In: Adam MP, Feldman J, Mirzaa GM, Pagon RA, Wallace SE, Amemiya A, editors. GeneReviews [Internet]. Seattle, Washington, USA: University of Washington-Seattle; 2024. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1435/
63. Kattamis A, Forni GL, Aydinok Y, Viprakasit V. Changing patterns in the epidemiology of β-thalassemia. European J of Haematology. 2020 Dec;105(6):692–703.
64. Ali S, Mumtaz S, Shakir HA, Khan M, Tahir HM, Mumtaz S, et al. Current status of beta-thalassemia and its treatment strategies. Molec Gen & Gen Med. 2021 Dec;9(12):e1788.
65. Langer A. Beta-Thalassemia. In: Adam MP, Feldman J, Mirzaa GM, Pagon RA, Wallace SE, Amemiya A, editors. GeneReviews [Internet]. Seattle, Washington, USA: University of Washington-Seattle; 2024. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1426/
66. Rao E, Kumar Chandraker S, Misha Singh M, Kumar R. Global distribution of β-thalassemia mutations: An update. Gene. 2024 Feb;896: 148022.
67. Karim MA, Jamal CY. A Review on Hemophilia in Children. Bangladesh J Child Health. 2013 Jun 18;37(1):27–40.
68. Zimmerman B, Valentino LA. Hemophilia: In Review. Pediatrics In Review. 2013 Jul 1;34(7):289–95.
69. Alblaihed L, Dubbs SB, Koyfman A, Long B. High risk and low prevalence diseases: Hemophilia emergencies. The American Journal of Emergency Medicine. 2022 Jun;56: 21–7.
70. Mojzisch A, Brehm MA. The Manifold Cellular Functions of von Willebrand Factor. Cells. 2021 Sep 8;10(9):2351.
71. Rousselle P, Laigle C, Rousselet G. The basement membrane in epidermal polarity, stemness, and regeneration. American Journal of Physiology-Cell Physiology. 2022 Dec 1;323(6):C1807–22.
72. So J, Teng J. Epidermolysis Bullosa Simplex. In: Adam MP, Feldman J, Mirzaa GM, Pagon RA, Wallace SE, Amemiya A, editors. GeneReviews [Internet]. Seattle, Washington: University of Washington-Seattle; 2022. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1369/
73. Youssefian L, Vahidnezhad H, Uitto J. Kindler Syndrome. In: Adam MP, Feldman J, Mirzaa GM, Pagon RA, Wallace SE, Amemiya A, editors. GeneReviews [Internet]. Seattle, Washington: University of Washington-Seattle; 2022. Available from: https://www.ncbi.nlm.nih.gov/books/NBK349072/
74. Hikmat O, Charalampos T, Klingenberg C, Rasmussen M, Tallaksen CME, Brodtkorb E, et al. The presence of anaemia negatively influences survival in patients with POLG disease. J of Inher Metab Disea. 2017 Nov;40(6):861–6.
75. Rahman S, Copeland WC. POLG-related disorders and their neurological manifestations. Nat Rev Neurol. 2019 Jan;15(1):40–52.
76. Stumpf JD, Saneto RP, Copeland WC. Clinical and Molecular Features of POLG-Related Mitochondrial Disease. Cold Spring Harbor Perspectives in Biology. 2013 Apr 1;5(4):a011395–a011395.
77. Rubinstein WS, Maglott DR, Lee JM, Kattman BL, Malheiro AJ, Ovetsky M, et al. The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. Nucleic Acids Research. 2012 Nov 26;41(D1):D925–35.
78. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype–gene relationships. Nucleic Acids Research. 2019 Jan 8;47(D1):D1038–43.
79. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Research. 2022 Jan 7;50(D1):D20–6.
80. Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong SE, et al. A Mitochondrial Protein Compendium Elucidates Complex I Disease Biology. Cell. 2008 Jul;134(1):112–23.
81. Calvo SE, Clauser KR, Mootha VK. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic Acids Res. 2016 Jan 4;44(D1):D1251–7.
82. Rath S, Sharma R, Gupta R, Ast T, Chan C, Durham TJ, et al. MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. Nucleic Acids Research. 2021 Jan 8;49(D1):D1541–7.

83. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. Nature. 1981 Apr 9;290(5806):457–65.
84. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet. 1999 Oct;23(2):147–147.
85. Bandelt HJ, Kloss-Brandstätter A, Richards MB, Yao YG, Logan I. The case for the continuing use of the revised Cambridge Reference Sequence (rCRS) and the standardization of notation in human mitochondrial DNA studies. J Hum Genet. 2014 Feb;59(2):66–77.
86. Pearson HA, Lobel JS, Kocoshis SA, Naiman JL, Windmiller J, Lammi AT, et al. A new syndrome of refractory sideroblastic anemia with vacuolization of marrow precursors and exocrine pancreatic dysfunction. The Journal of Pediatrics. 1979 Dec;95(6):976–84.
87. Rötig A, Bourgeron T, Chretien D, Rustin P, Munnich A. Spectrum of mitochondrial DNA rearrangements in the Pearson marrow-pancreas syndrome. Hum Mol Genet. 1995;4(8):1327–30.
88. Lee Y, Kim T, Lee M, So S, Karagozlu MZ, Seo GH, et al. De Novo Development of mtDNA Deletion Due to Decreased POLG and SSBP1 Expression in Humans. Genes. 2021 Feb 17;12(2):284.
89. Den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Human Mutation. 2016 Jun;37(6):564–9.
90. Harrison SM, Riggs ER, Maglott DR, Lee JM, Azzariti DR, Niehaus A, et al. Using ClinVar as a Resource to Support Variant Interpretation. CP Human Genetics [Internet]. 2016 Apr [cited 2024 Dec 1];89(1). Available from: https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/0471142905.hg0816s89
91. Zhang X, Minikel EV, O'Donnell-Luria AH, MacArthur DG, Ware JS, Weisburd B. ClinVar data parsing. Wellcome Open Res. 2017 May 23;2:33.
92. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. Nucleic Acids Research. 2020 Jan 8;48(D1):D835–44.
93. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research. 2018 Jan 4;46(D1):D1062–7.
94. Gargano MA, Matentzoglu N, Coleman B, Addo-Lartey EB, Anagnostopoulos AV, Anderton J, et al. The Human Phenotype Ontology in 2024: phenotypes around the world. Nucleic Acids Research. 2024 Jan 5;52(D1):D1333–46.
95. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25–9.
96. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L, Mi H. PANTHER: Making genome-scale phylogenetics accessible to all. Protein Science. 2022 Jan;31(1):8–22.
97. Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Research. 2023 Jul 5;51(W1):W207–12.
98. Hinton G, Roweis S. Stochastic Neighbor Embedding. In: Becker S, Thrun S, Obermayer K, editors. Advances in Neural Information Processing Systems 15 [Internet]. Vancouver, British Columbia, Canada: MIT Press; 2002. p. 1–8. Available from: https://papers.nips.cc/paper_files/paper/2002/hash/6150ccc6069bea6b5716254057a194ef-Abstract.html
99. van der Maaten L, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008;9(86):2579–605.
100. Li W, Cerise JE, Yang Y, Han H. Application of t-SNE to human genetic data. J Bioinform Comput Biol. 2017 Aug;15(04):1750017.
101. Pearson K. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1901 Nov;2(11):559–72.
102. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 1933;24(6):417–41.
103. Hotelling H. Relations between two sets of variates. Biometrika. 1936 Dec 1;28(3–4):321–77.
104. Steinhaus H. Sur la division des corps matériels en parties. Bulletin de l'Académie Polonaise des Sciences. 1956;4(12):801–4.
105. Forgy EW. Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classifications. In Tallahassee, Florida, USA: Wiley; 1965. Available from: http://www.jstor.org/stable/2528559
106. Macqueen J. Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J, editors. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics [Internet]. California, USA: University of California Press; 1967. p. 281–97. Available from: https://digitalassets.lib.berkeley.edu/math/ucb/text/math_s5_v1_article-17.pdf
107. Lloyd S. Least squares quantization in PCM. IEEE Trans Inform Theory. 1982 Mar;28(2):129–37.
108. Amézquita EJ, Quigley MY, Ophelders T, Munch E, Chitwood DH. The shape of things to come: Topological data analysis and biology, from molecules to organisms. Developmental Dynamics. 2020 Jul;249(7):816–33.
109. Loughrey CF, Fitzpatrick P, Orr N, Jurek-Loughrey A. The topology of data: opportunities for cancer research. Wren J, editor. Bioinformatics. 2021 Oct 11;37(19):3091–8.
110. Singh Y, Farrelly CM, Hathaway QA, Leiner T, Jagtap J, Carlsson GE, et al. Topological data analysis in medical imaging: current state of the art. Insights Imaging. 2023 Apr 1;14(1):58.
111. Van Veen H, Saul N, Eargle D, Mangham S. Kepler Mapper: A flexible Python implementation of the Mapper algorithm. JOSS. 2019 Oct 17;4(42):1315.
112. Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res. 2009 Sep;19(9):1639–45.
113. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. Bioinformatics. 2014 Oct 1;30(19):2811–2.
114. Shimoyama. pyCirclize: circular visualization in Python [Internet]. 2023 [cited 2024 Apr 1]. Available from: https://github.com/moshi4/pyCirclize

115. Kolbinger FR, Veldhuizen GP, Zhu J, Truhn D, Kather JN. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. Commun Med [Internet]. 2024 Apr 11 [cited 2024 Sep 17];4(1). Available from: https://www.nature.com/articles/s43856-024-00492-0
116. Sondalle SB, Longerich S, Ogawa LM, Sung P, Baserga SJ. Fanconi anemia protein FANCI functions in ribosome biogenesis. Proc Natl Acad Sci USA. 2019 Feb 12;116(7):2561–70.
117. Gueiderikh A, Maczkowiak-Chartois F, Rouvet G, Souquère-Besse S, Apcher S, Diaz JJ, et al. Fanconi anemia A protein participates in nucleolar homeostasis maintenance and ribosome biogenesis. Sci Adv. 2021 Jan;7(1):eabb5414.
118. Gueiderikh A, Maczkowiak-Chartois F, Rosselli F. A new frontier in Fanconi anemia: From DNA repair to ribosome biogenesis. Blood Reviews. 2022 Mar;52:100904.
119. Boria I, Garelli E, Gazda HT, Aspesi A, Quarello P, Pavesi E, et al. The ribosomal basis of diamond-blackfan anemia: mutation and database update. Hum Mutat. 2010 Dec;31(12):1269–79.
120. Ellis SR, Gleizes PE. Diamond Blackfan Anemia: Ribosomal Proteins Going Rogue. Seminars in Hematology. 2011 Apr;48(2):89–96.
121. Kapralova K, Jahoda O, Koralkova P, Gursky J, Lanikova L, Pospisilova D, et al. Oxidative DNA Damage, Inflammatory Signature, and Altered Erythrocytes Properties in Diamond-Blackfan Anemia. IJMS. 2020 Dec 17;21(24):9652.
122. Piantanida N, La Vecchia M, Sculco M, Talmon M, Palattella G, Kurita R, et al. Deficiency of ribosomal protein S26, which is mutated in a subset of patients with Diamond Blackfan anemia, impairs erythroid differentiation. Front Genet. 2022 Dec 12;13: 1045236.
123. Huang S, Ninivaggi M, Chayoua W, De Laat B. VWF, Platelets and the Antiphospholipid Syndrome. IJMS. 2021 Apr 18;22(8):4200.
124. El-Mansi S, Nightingale TD. Emerging mechanisms to modulate VWF release from endothelial cells. The International Journal of Biochemistry & Cell Biology. 2021 Feb;131: 105900.
125. Manz XD, Bogaard HJ, Aman J. Regulation of VWF (Von Willebrand Factor) in Inflammatory Thrombosis. ATVB. 2022 Nov;42(11):1307–20.
126. Pfendner E, Lucky A. Dystrophic Epidermolysis Bullosa. In: Adam MP, Feldman J, Mirzaa GM, Pagon RA, Wallace SE, Amemiya A, editors. GeneReviews [Internet]. Seattle, Washington: University of Washington-Seattle; 2018. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1369/
127. Pfendner E, Lucky A. Junctional Epidermolysis Bullosa. In: Adam MP, Feldman J, Mirzaa GM, Pagon RA, Wallace SE, Amemiya A, editors. GeneReviews [Internet]. Seattle, Washington: University of Washington-Seattle; 2018. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1369/
128. Fowler B, Froese DS, Watkins D. Disorders of Cobalamin and Folate Transport and Metabolism. In: Saudubray JM, Baumgartner MR, García-Cazorla Á, Walter J, editors. Inborn Metabolic Diseases: Diagnosis and Treatment [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2022. p. 511–29. Available from: https://doi.org/10.1007/978-3-662-63123-2_28
129. Watkins D, Venditti CP, Rosenblatt DS. Chapter 51 - Vitamins: cobalamin and folate. In: Rosenberg RN, Pascual JM, editors. Rosenberg's Molecular and Genetic Basis of Neurological and Psychiatric Disease [Internet]. 6th ed. Academic Press; 2020. p. 687–97. Available from: https://www.sciencedirect.com/science/article/pii/B9780128139554000519
130. Russell OM, Gorman GS, Lightowlers RN, Turnbull DM. Mitochondrial Diseases: Hope for the Future. Cell. 2020 Apr;181(1):168–88.
131. Vercellino I, Sazanov LA. The assembly, regulation and function of the mitochondrial respiratory chain. Nat Rev Mol Cell Biol. 2022 Feb;23(2):141–61.
132. Jain V, Yang WH, Wu J, Roback JD, Gregory SG, Chi JT. Single Cell RNA-Seq Analysis of Human Red Cells. Front Physiol. 2022 Apr 20;13: 828700.
133. Bardhan A, Bruckner-Tuderman L, Chapple ILC, Fine JD, Harper N, Has C, et al. Epidermolysis bullosa. Nat Rev Dis Primers. 2020 Sep 24;6(1):78.
134. Zhang K, Lu Y, Harley K, Tran MH. Atypical Hemolytic Uremic Syndrome: A Brief Review. Hematology Reports. 2017 Jun 1;9(2):7053.
135. Muff-Luett M, Nester C. The Genetics of Ultra-Rare Renal Disease. J Pediatr Genet. 2016 Feb 23;05(01):033–42.
136. De Falco L, Sanchez M, Silvestri L, Kannengiesser C, Muckenthaler MU, Iolascon A, et al. Iron refractory iron deficiency anemia. Haematologica. 2013 Jun 1;98(6):845–53.

Copyright:* © *2025 *Vanderlaan G. This Open Access Article is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*