Computational Biology and Bioinformatics
2025, Vol. 13, No. 2, pp. 72–85
https://doi.org/10.11648/j.cbb.20251302.13

Research Article

Structure-guided Genome-wide Association Analysis of ALK Variants with GWAS Data Using R

Tanmay Bandbe[1], **Vineeta Johri**[2], **Uma Kumari**[1, *]

[1]Bioinformatics Project and Research Institute, Noida, India
[2]NIMS Institute of Medical Science and Technology, Jaipur, India

Abstract

Anaplastic lymphoma kinase (ALK) has been linked to several hematological malignancies; however, its comprehensive genetic variability and potential disease associations are not fully understood. In this study, a structure-guided genome-wide association analysis (GWAS) of ALK variants was performed using publicly available summary statistics and R-based analytical pipelines. The GWAS datasets were acquired, filtered, and ranked based on sample size to ensure sufficient statistical power. A focused analysis on two distinct datasets, which were selected based on sample size and phenotypic diversity: one representing lymphoma-related genetic traits from the UK Biobank, and another capturing ALK-associated proteomic variation. Rigorous quality control and comprehensive data visualization were performed using a set of diagnostic and analytical plots, including volcano plots, QQ plots, histograms, size effects, and a correlation matrix heatmap of numerical variables. Regional Manhattan plots highlighted distinct, highly significant associations at the ALK locus in both datasets, enabling the identification of independent lead variants. Interpretation of the QQ plots and histograms confirmed adequate control for population stratification and minimal inflation of test statistics. Integration of insights from the effect size distribution and SE versus Beta plots provided a clear assessment of the precision and reliability of estimated genetic effects. By mapping genetic variants onto the ALK protein structure, single-nucleotide polymorphisms (SNPs) with potential functional relevance and evaluating their associations with disease phenotypes across populations were prioritized. This strategy facilitates the identification of variants likely to influence protein structure and function, thereby enhancing the interpretability of GWAS findings in a protein-centric context. This approach demonstrates the power of integrating structural bioinformatics with statistical genetics to reveal novel genotype-phenotype relationships, offering valuable insights for precision medicine and targeted ALK-directed therapies. Overall, this integrative methodology establishes a reproducible framework for detailed regional GWAS analyses, successfully pinpointing strong ALK locus associations and identifying candidate variants for subsequent functional validation relevant to the phenotypes, and assessing their potential role in therapeutic investigation for hematological malignancies.

Keywords

GWAS, ALK Gene Variants, Multiomics Data Integration, BioR Database, Bioinformatics Integration, Variant Prioritization

*Corresponding author: uma27910@gmail.com (Uma Kumari)

Received: 5 November 2025; **Accepted:** 18 November 2025; **Published:** 11 December 2025

1. Introduction

The Anaplastic lymphoma kinase (ALK) encodes a receptor tyrosine kinase that plays important roles in neural development and, when altered, can act as an oncogenic driver in multiple cancers. Somatic and germline changes in ALK are linked to several cancers and developmental disorders, and the spectrum of functionally relevant ALK variation and its structural consequences remains incompletely resolved. Activating point mutations and gene fusions that disrupt ALK's kinase domain stimulate oncogenic signalling in neuroblastoma, lung cancer, and various other tumour types, making ALK both a paradigmatic cancer gene and a promising target for translational research [1, 2]. Studies in lung cancer and blood malignancies illustrate that different ALK alterations (fusion partners, splice variants, or missense changes) can produce distinct biochemical behaviors and therapy responses, underscoring the need to move beyond association signals to mechanistic interpretation [3]. ALK is best known for fusion events (EML4–ALK in lung cancer), but somatic point mutations and rare germline variants can also impact ALK function and treatment response. Because ALK alterations are clinically actionable, several small-molecule ALK inhibitors markedly improve outcomes in ALK-positive cancers, and understanding the full spectrum of ALK variation (including common germline polymorphisms that might modulate risk, expression, or drug response) has clear translational value [4, 5].

Genome-wide association studies (GWAS) test hundreds of thousands to millions of common genetic variants across the genome to find loci statistically associated with traits or disease risk, and have become a cornerstone of human genetics. By scanning large cohorts, GWAS have revealed thousands of robust genotype–phenotype links and produced a rich trove of summary statistics that can be reused for follow-up functional work, but GWAS alone do not directly reveal which specific nucleotide changes alter protein function or disease biology [6]. Rather than identifying single causal changes, GWAS typically returns statistical signals, typically single-nucleotide polymorphisms (SNPs) that tag regions of the genome where the true causal variant(s) lie. While GWAS has been enormously productive, translating an association signal into a molecular mechanism remains a major challenge because most GWAS hits are non-coding, occur in linkage disequilibrium (LD) blocks, and can implicate multiple nearby genes or regulatory elements. Recent methodological work has therefore emphasized careful fine-mapping, integration of functional annotation, and experimental validation to move from signal to function [7, 8].

The integration of GWAS with expression Quantitative Trait Loci (eQTLs) has become a standard approach for functional interpretation in the "post-GWAS era". eQTLs link genetic variants to changes in gene expression, helping to identify the true causal gene at a GWAS locus. Recent studies, while often in non-ALK contexts, illustrate the power of combining GWAS and eQTL data to pinpoint genes for follow-up functional studies [9]. Work in the last few years has emphasized the uneven overlap between GWAS hits and cis-eQTLs, highlighting tissue specificity, limited power in available eQTL maps, and regulatory mechanisms beyond steady-state expression (splicing, chromatin contacts). For ALK, tissue context (neural vs. hematopoietic) and cell-type resolution are essential: colocalization with brain- or immune-cell eQTLs will strengthen claims that a given germline locus influences ALK biology [10]. Although ALK-centric GWAS papers are still sparse, several recent computational efforts provide templates for how to proceed. Advances include finer visualization and regional plotting tools (making inspection of LD, recombination, and local association patterns straightforward), pipelines that integrate structural variant catalogues into GWAS follow-up, and locus-centric meta-analysis frameworks that leverage cross-cohort harmonization [11]. Here, R provides a flexible and reproducible environment for implementing the full GWAS workflow and integrates structure-based scores from external tools or in-house calculations. Additionally, reproducible R packages and Bioconductor resources facilitate data wrangling, visualization, and reporting, while connectors to structural bioinformatics tools enable residue-level mapping of variants onto three-dimensional kinase models. Combining these elements supports a transparent path from statistical association to structural hypothesis, which is especially valuable for clinically actionable genes such as ALK [12].

In this study, we applied a structure-guided genome-wide association analysis targeting ALK, using publicly available GWAS summary statistics and R-based analytical workflows. We selected phenotypically distinct datasets to ensure a broad exploration of ALK's genetic landscape across different disease contexts. Further, we performed rigorous quality control, statistical testing, and visualization, followed by regional association mapping around the ALK locus. Importantly, we superimposed associated SNPs onto models of the ALK protein, enabling us to prioritize variants not just by p-value but by their potential structural and functional effects. This integrative approach capitalizes on recent advances in computational structural biology and statistical genetics to help interpret GWAS hits in a protein-centric manner. Such a framework paves the way for identifying candidate ALK variants that may influence disease risk, guide functional follow-up, and ultimately inform precision medicine strategies, including targeted ALK therapies. Our work demonstrates a reproducible, scalable pipeline in R for detailed regional GWAS that can be adapted for other genes of interest in precision medicine, especially when combined with structural bioinformatics, and highlights novel ALK locus associations in hematological malignancies.

2. Materials and Methods

2.1. GWAS Data Acquisition and Preprocessing

To explore genetic variation in the ALK gene region, we conducted a structure-guided genome-wide association analysis using publicly available GWAS summary statistics. GWAS datasets were accessed through the OpenGWAS API using the 'ieugwasr' R package, along with supporting packages including 'tidyverse', 'dplyr', 'ggplot2', 'ggrepel', 'susieR', 'coloc', and 'ACAT'. We first authenticated our session via the OpenGWAS JSON Web Token (JWT), and access was verified using the 'ieugwasr:: get_opengwas_jwt()' function. The Ensembl ID for ALK (ENSG00000171094) was employed to extract variants located within ± 50 kb of the gene using the 'variants_gene()' function. Retrieved variants included information on rsID, chromosome, and genomic position.

2.2. Selection of GWAS Datasets

We retrieved metadata for all GWAS datasets using the 'gwasinfo()' function. To concentrate on hematological malignancies, we filtered studies using a keyword-based search including terms such as "leukemia," "lymphoma," "myeloma," and related descriptors. The resulting filtered datasets were subsequently ranked by sample size to prioritize studies with greater statistical power.

2.3. Extraction of Association Statistics

For the selected GWAS study, we obtained association statistics for the ALK gene region using the 'associations()' function, encompassing variants within ± 250 kb of the gene. The data were standardized to include the following columns: rsID, chromosome, position, effect size (β), standard error (SE), $p-$value, effect allele, and effect allele frequency (EAF).

2.4. Data Visualization

We performed a series of exploratory visualizations to investigate the genetic architecture of the ALK region. These included regional Manhattan plots to highlight top associations and genome-wide significant SNPs, as well as volcano plots of effect sizes (β) against $-\log10(p-\text{values})$ to identify variants with both strong effects and statistical significance. Quantile-Quantile (QQ) plots were used to evaluate deviations from the null distribution, while histograms of $p-$values and effect sizes summarized their overall distributions. Additionally, correlation heatmaps were generated for numeric variables, including β, standard errors (SE), and $p-$values. All plots were created using the 'package and annotated with top hits using the 'ggrepel' package.

2.5. Independent SNP Selection

We subsequently conducted Linkage Disequilibrium (LD) clumping using 'Ieugwasr:: ld_clump()' function to pinpoint independent variants within the ALK locus. The resulting clumped variants were retained for downstream functional and colocalization analyses.

2.6. Statistical and Colocalization Analyses (ACAT)

We further examined independent variants by performing fine-mapping with the 'susieR' (Sum of Single Effects) package and conducting colocalization analyses with other traits using the 'coloc' package. In addition, gene-level association testing was carried out by combining variants $p-values$ through the 'ACAT' (Aggregated Cauchy Association Test) method.

3. Results

3.1. GWAS Data Acquisition and Preprocessing

The Summary-level GWAS data were obtained from five independent studies curated within the OpenGWAS and UK Biobank repositories, each corresponding to distinct phenotypic traits. The datasets were systematically harmonized to ensure uniform variant identifiers (rsIDs), chromosomal position alignment, and consistent reference/effect allele orientation. Rigorous quality control steps were implemented to exclude variants with missing $p-values$, non-biallelic loci, or low-confidence markers. All genomic positions were standardized to the GRCh37 (hg19) reference build.

3.2. Selection of GWAS Datasets

Five GWAS datasets encompassing the ALK genomic region were selected [(ebi-a-GCST90018878 (Study ID 1), ieu-b-4957 (Study ID 2), ukb-d-C_LYMPHOMA (Study ID 3), ebi-a-GCST 90018658 (Study ID 4), and prot-a-235 (Study ID 5)] based on the availability of comprehensive summary-level statistics and their relevance to cancer and immune-related phenotypes, To identified which studies merited deeper analysis, we compared the overall quality and completeness of their summary-level outputs. Among these, Study IDs 3 and 5 were prioritized for detailed investigation due to their dense SNP coverage across the ALK locus, minimal missing data in their summary statistics, and well-behaved p-value distributions with low genomic inflation. Most importantly, each of these studies reported robust association outcomes; Study ID 3 quantified disease-related association strength $(-log_{10}(p))$ for lymphoma susceptibility phenotypes, while Study ID 5 measured protein-level variation, capturing how genetic differ-

ences around ALK influence measurable protein phenotypes.

3.3. Extraction of Association Statistics

Association summary statistics, including β, SE, EAF), and $p - values$, were extracted for all SNPs located within ± 250 kb of the ALK locus (chromosome 2: 29.0–30.0 Mb). Across the analyzed datasets, the $-log_{10}(p)\ values$ spanned approximately 2.0 to > 8.0. The Study ID 3 exhibited a prominent association peak near 29.88 Mb (rs72861594; $-log_{10}(p) \sim 6.0$), whereas Study ID 5 demonstrated the strongest overall signal at 29.50 Mb (rs 62130614 ; $-log_{10}(p) \sim 8.0$). Collectively, these findings reinforce ALK as a recurrent and robust susceptibility locus across multiple, phenotypically distinct GWAS datasets.

3.4. Data Visualization

Extensive visual analyses were conducted using Manhattan, regional Manhattan, and volcano plots, complemented by QQ plots, histograms, SE–β plots, and correlation heatmaps. The QQ and histogram plots confirmed adequate correction for population stratification and the absence of genomic inflation. The Regional Manhattan plots revealed discrete, highly significant association peaks across the ALK locus, particularly clustered around 29.25 Mb and $29.50 - 29.78$ Mb. The volcano plots highlighted phenotype-specific effect patterns; Study ID 5 exhibited predominantly risk-enhancing (positive β) variants, whereas Study ID 3 demonstrated symmetric distributions indicative of smaller and more balanced genetic effects.

3.4.1. Volcano Plot

The Volcano Plots illustrate variability in both the magnitude and direction of genetic effects across different studies, likely due to differences in the specific trait being studied (as indicated by the varying study IDs) and sample populations. This comparative analysis highlights that while the ALK region consistently shows statistically significant associations across multiple studies, the biological relevance (effect size) is highly dependent on the specific phenotype being investigated.

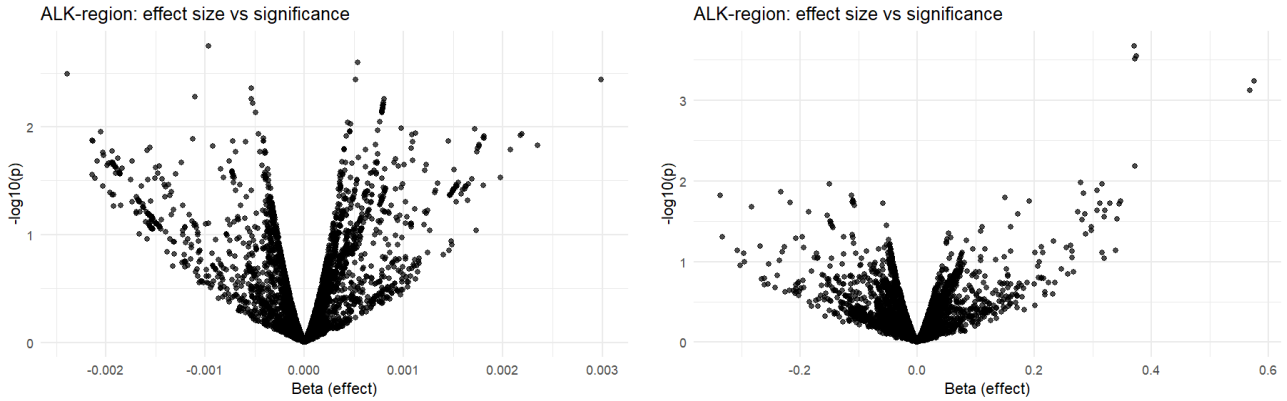

Figure 1. *Volcano Plots for Study ID 3 (left) and Study ID 5 (right)*

The Study ID 3 plot shows a classic, narrow "volcano" shape and is highly symmetrical with an extremely narrow Beta range (from $\sim -0.003\ to\ 0.003$). This suggests the genetic effects in this study are very subtle. Significant variants are observed on both the positive and negative effect size wings. The highest point on the Y-axis reaches $-log_{10}(p) \sim 2.5$. The highly significant variants correspond to the largest observed effect sizes in this narrow range. The ALK region shows associations with the trait, characterized by extremely small effect sizes but robust statistical significance for some variants. Moreover, the Study ID 5 plot is mostly symmetrical around the Beta axis, with a modest effect size range (from ~ -0.2 to 0.6). The most significant points are clustered on the positive effect size side. The highest peak is at $-log_{10}(p) > 3.5$, corresponding to an effect size of Beta ~ 0.4. Another cluster of highly significant points is found at a larger effect size of Beta ~ 0.6 with $-log_{10}(p) \sim 3.2$. The highest negative effect size (Beta ~ -0.2) is associated with low to moderate significance ($-log_{10}(p) < 1.0$). Significant associations in this study are primarily driven by variants with a positive (risk) effect, with the highest significance found at moderate positive effect sizes. In essence, associations in Study ID 3 are statistically real but biologically subtle, while Study ID 5 shows stronger, more directional effects, suggesting risk-enhancing ALK variants for that trait.

3.4.2. QQ Plot

The QQ plots for the ALK region show a similar pattern in both studies, characterized by a good fit to the null model at lower significance levels, followed by a clear upward deflection, indicating true biological association.

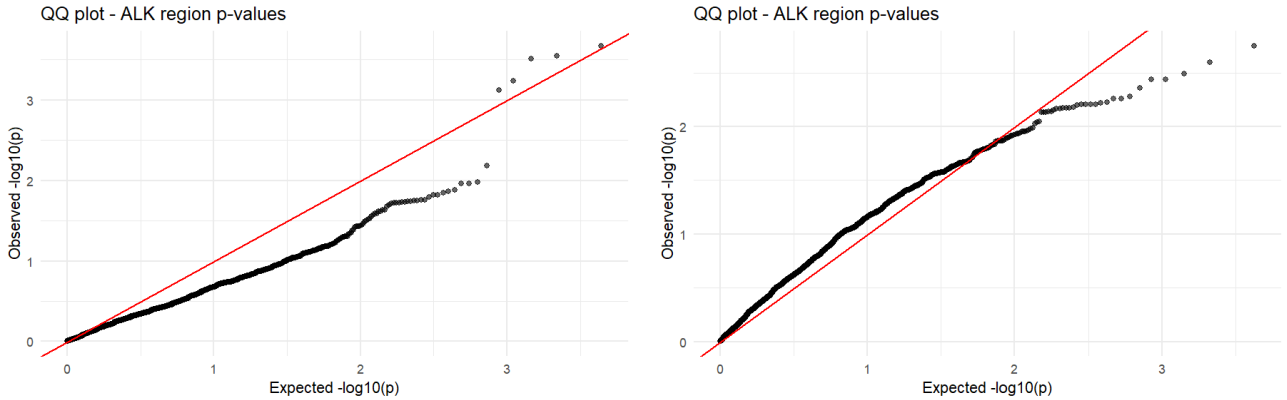

Figure 2. *QQ Plots for Study ID* 3 *(left) and Study ID* 5 *(right)*

The initial part of the Study ID 3 plot follows the null line closely up to Expected $-log_{10}(p) \sim 2.0$. A distinct, sharp upward deflection is visible starting at Expected $-log_{10}(p) \sim 2.0$. The highest observed $p-value$ reaches Observed $-log_{10}(p) \sim 2.7$ when the expected value is ~ 3.25. The plot confirms the presence of genuine genetic associations in the ALK region for the Lymphoma trait, with strong statistical confidence in the tail end of the distribution. Further, the Study ID 5 plot fit to the null model is excellent up to Expected $-log_{10}(p) \sim 2.0$. A very steep and pronounced upward deflection starts around Expected $-log_{10}(p) \sim 2.0$. The top observed $p-value$ reaches Observed $-log_{10}(p) > 3.5$, representing the strongest deviation among the five studies. The strong upward tail confirms a highly robust and significant genetic signal in the ALK region for this study's trait, indicating the greatest statistical evidence of association. Overall, both datasets contain genuine association signals, not artifacts of population structure or inflation.

3.4.3. Histogram of P-values Plot

These plots use a log scale for the $p-value$ axis, which compresses the common, non-significant $p-values$ (near 1.0) and stretches the significant $p-values$ (near 0.0), making the distribution more interpretable.

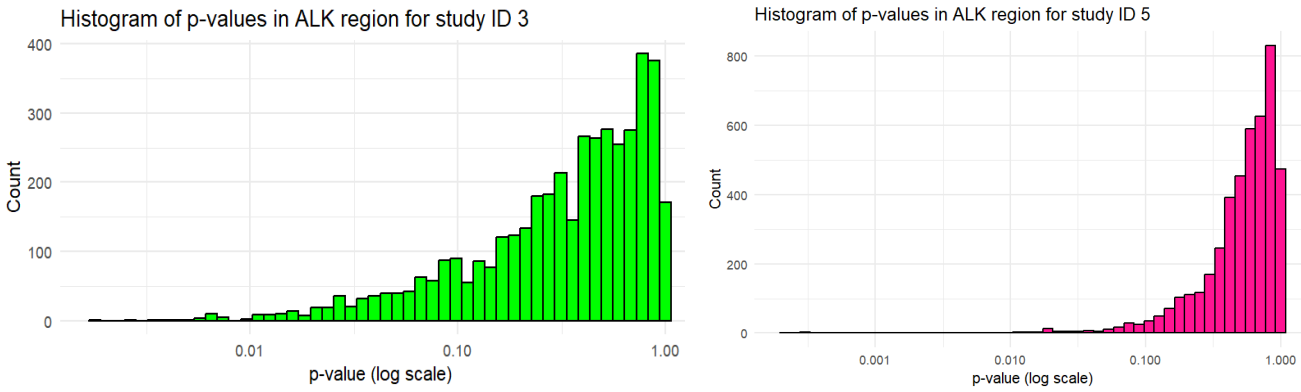

Figure 3. *Histogram of* $p-values$ *for Study ID* 3 *(left) and Study ID* 5 *(right).*

A histogram of $p-values$ in a genetic association study visualizes the distribution of significance across all tested variants (SNPs) in a specific genomic region. Under the null hypothesis (no association), $p-values$ should be uniformly distributed, resulting in a flat histogram. The Study ID 3 plot is right-skewed, similar to the others, with the largest bin at $p = 1.0$ (Count ~ 400). There is minimal, but detectable, enrichment in the lowest $p-value$ bins. The count for $p < 0.001$ is very low, and is consistent with the Manhattan plot for Study ID 3 (Lymphoma), which showed maximum significance of $-log_{10}(p) \sim 2.5$. The distribution indicates minor enrichment of significant associations, consistent with the moderate signal strength for the Lymphoma trait. Likewise, Study ID 5 is the most extremely right-skewed histogram, with the highest counts in the rightmost bins and the largest overall sample size suggested by the maximum count of ~ 850. There is a sharp drop-off towards the left, but the presence of the highest peak in the Manhattan plot $(-log_{10}(p)$

~ 3.75) indicates a very small number of extremely low $p-values$ that drive the significance. The histogram indicates a well-powered study with a robust statistical model, where the strong association signal seen in the Manhattan plot is due to a few, highly significant variants rather than a widespread moderate association. Study ID 5 captures stronger associations, whereas Study ID 3 has fewer but still reliable signals.

3.4.4. Effect Size Distribution Plot

A histogram of effect sizes (Beta values) shows the frequency.

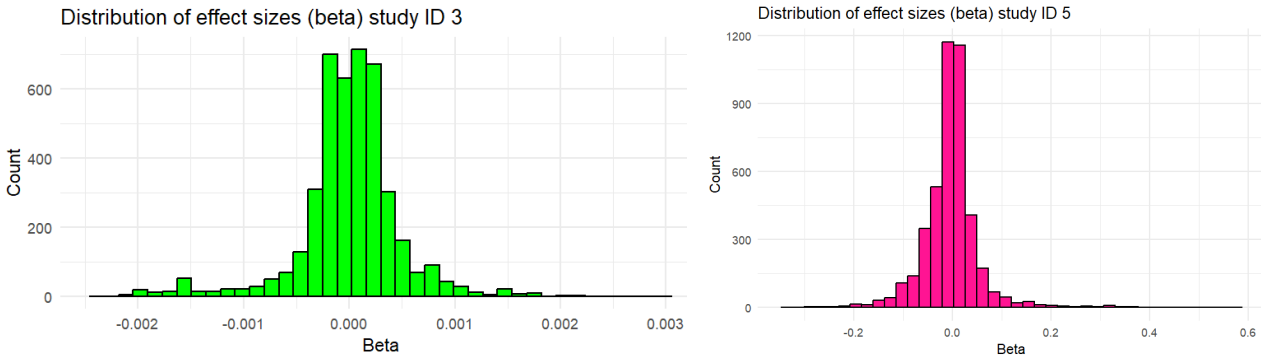

Figure 4. *Effect size distribution plots for Study ID 3 (top) and Study ID 5 (bottom)*

of association magnitudes and directions (β coefficients) across all tested genetic variants (SNPs) in the ALK region. In a typical genetic association study, the vast majority of SNPs are not associated, so the distribution should be highly concentrated around $\beta = 0$. The β values for Study ID 3 are very small, spanning a range from approximately -0.003 to $+0.003$. The distribution is sharply peaked, very close to $\beta = 0$. The distribution is generally symmetrical, though it appears to have slightly more counts on the positive side ($\beta > 0$) than the negative side. There is a small cluster of negative effects around $\beta \sim -0.002$. Genetic effects are minimal in magnitude, with no strong overall directional bias. Meanwhile, the β values for Study ID 5 span a moderate range, from approximately −0.3 to +0.6. The distribution is sharply peaked near $\beta = 0$ (Count ~ 1150). The distribution is slightly right-skewed, with the positive tail extending farther than the negative tail (maximum $\beta \sim 0.6$ versus $\beta \sim -0.3$). Genetic effects are typically small, with the largest observed effects being positive (risk-increasing). Overall, Study ID 5 captures more biologically meaningful effect sizes within ALK, while Study ID 3 reflects more subtle genetic contributions.

3.4.5. SE vs Beta Plot

The Standard Error (SE) vs. Absolute Beta ($|\beta|$) plots for the ALK region from five different studies. These plots are critical for assessing the precision of the estimated effect sizes (β) in a genetic association study.

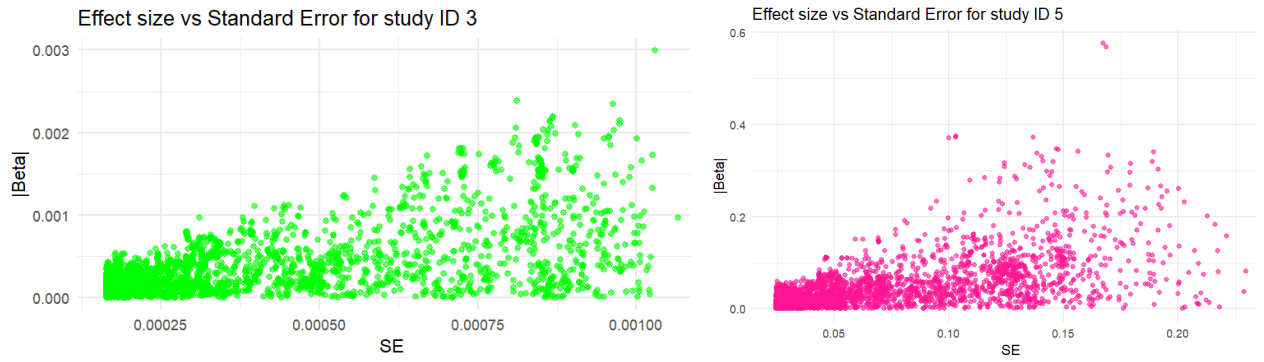

Figure 5. *SE vs Beta plots for Study ID 3 (top) and Study ID 5 (bottom).*

For Study ID 3, the SE is up to ~ 0.001, and $|\beta|$ is up to ~ 0.003. The plot shows a clear wedge shape extending from the origin. The highest significance is found at the upper boundary of the wedge, where the ratio of $|\beta|$ to SE is maximized. The most extreme point is reached $|\beta| \sim 0.003$ at SE ~ 0.001. Effects are minimal in magnitude but are estimated with high precision. The points are clustered mainly between SE $= 0.0$ and 0.20, and $|\beta| = 0.0$ and 0.20. There is a distinct, upward-pointing cluster that extends to the upper right. The most significant points (highest $|\beta|$/SE ratio) are at the top edge, reaching $|\beta| \sim 0.6$ at SE ~ 0.17. This study demonstrates a set of variants with moderate effect sizes that are estimated with relatively good precision (low SE). Altogether, Study ID 5 contains variants with stronger effects but still acceptable precision.

3.4.6. Volcano Plot with Significance Threshold

A Volcano Plot is used in genetic studies to visualize the entire set of test results by plotting significance ($-log_{10}(p)$, Y-axis) against effect size (β, X-axis). Significant associations are characterized by points in the upper-left and upper-right corners ("the volcano's cones"). The most interesting variants are those with a high $-log10(p)$ and a β value far from zero (large effect size). The Study ID 3 has maximum significance, reaching approximately 2.5 (equivalent to $p \sim 0.003$). The β values are very small, ranging from $\sim -0.0025\ to\ +0.003$.

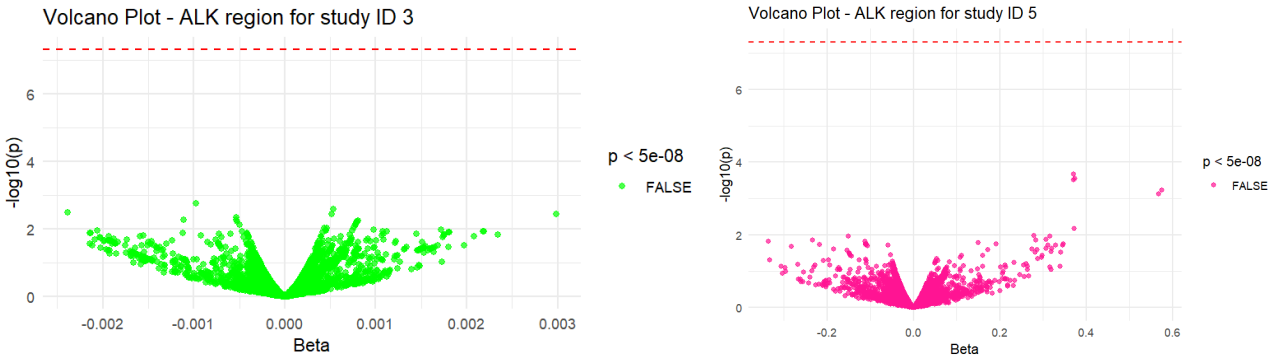

Figure 6. *Volcano plots with Significance Threshold for Study ID 3 (top) and Study ID 5 (bottom)*

The plot is largely symmetrical around $\beta = 0$, with a slight rightward shift in the most extreme β points. The association signal for Lymphoma is moderate in significance and driven by variants with minimal effect sizes, though a single outlier has the largest positive effect ($\beta \sim 0.003$). While the Study ID 5 maximum significance reaches approximately 3.8 (equivalent to $p \sim 1.5 \times 10 - 4$), with another plot showing points over 8. The highest point on the uncoloured plot is $-log_{10}(p)$~3.8. The β values are in the moderate range, from ~ -0.3 to $+0.6$. The plot is asymmetrical (right-skewed), with the strongest associations and the largest β values located on the positive side. This study shows a strong, directional association, with the most significant variants having positive effect sizes ($\beta > 0.3$), suggesting a primary risk-increasing role for the associated variants in the ALK region for this trait. Primarily, Study ID 5's genetic signal is both stronger and more biologically interpretable.

3.4.7. Regional Manhattan Plot

A Regional Manhattan Plot displays the $(-log10(p))$ for all tested genetic variants (SNPs) within a specific genomic region, plotted against their Genomic Position (bp). It helps pinpoint the most significant association peaks within the gene. Significant associations appear as peaks or clusters of points rising above the general background noise.

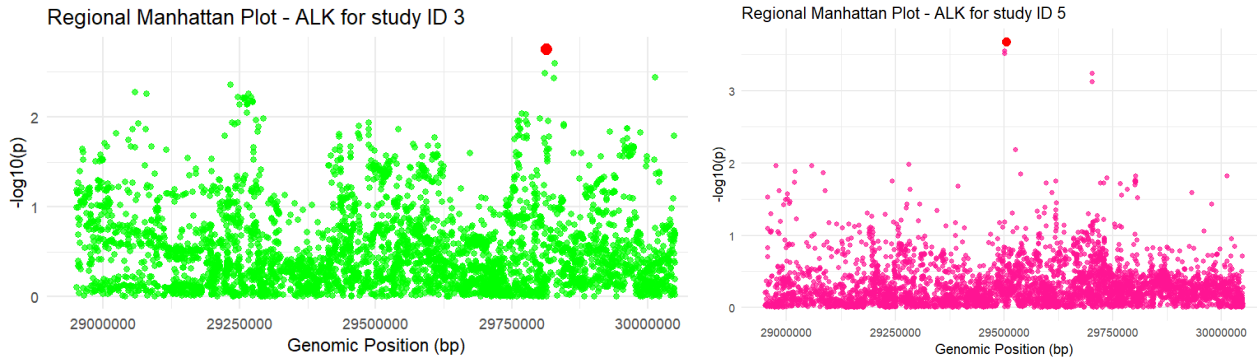

Figure 7. *Regional Manhattan plots for Study ID* 3 *(top) and Study ID* 5 *(bottom).*

For Study ID 3, multiple low-to-moderate peaks are visible, suggesting several variants in the region may weakly contribute to the associated trait. The single most significant variant is located around 29,800,000 bp, reaching a $-log10(p)$ of approximately 2.7 (highlighted by the red marker). Another substantial peak is seen near 29,250,000 bp, reaching $-log10(p) \sim 2.4$. The signals are moderate, with the most significant peak located towards the right/downstream end of the region. In addition, the Study ID 5 plot shows that two prominent, separate peaks are visible in this region. The primary peak is located around 29,500,000 bp, reaching the highest significance of approximately $-log10(p) = 3.8$ (highlighted by the red marker). A strong secondary peak is present near 29,700,000 bp, with a $-log10(p)$ value of approximately 3.3. The most significant associations are found in the center and center-right of the plotted ALK region. Chiefly, ALK contains multiple candidate loci, with Study ID 5 demonstrating much higher signal clarity.

3.4.8. Beta vs Allele Frequency Plot

The Beta (β) vs. Allele Frequency (AF) plots are used to visualize the relationship between the rarity of a variant and the magnitude of its effect on the studied trait, providing insight into the underlying genetic architecture (i.e., common vs. rare variants).

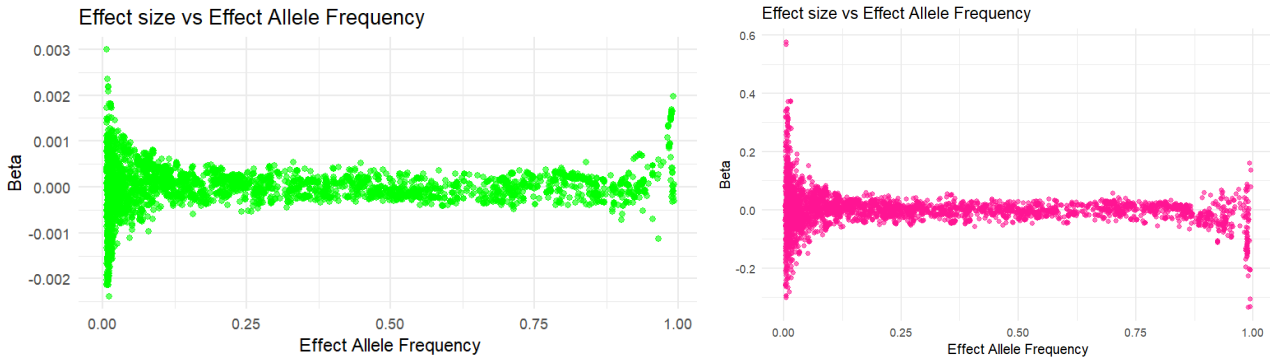

Figure 8. *Beta vs Allele Frequency plots for Study ID* 3 *(top) and Study ID* 5 *(bottom)*

Significant findings are driven by points away from the horizontal center line ($\beta = 0$). Evidence for Rare Variants, Large Effects appears as vertical scatter (high $|\beta|$) in the low AF range (left side of the plot). The density of variants in the Study ID 3 plot is high across all AF bins, particularly in the common range ($AF > 0.1$). The plot forms a very narrow band centered at $\beta = 0$. The Maximum $|\beta|$ is ~ 0.003. The distribution is homogeneous; minimal effect sizes are observed across all allele frequencies. There are no notable outliers in the rare AF range. The genetic contribution of ALK variants to this trait is characterized by ubiquitous but extremely small effects, consistent with the CDCV model if the effects were significant. Furthermore, variants in the Study ID 5 plot are generally common, with the highest density around AF~ 0.5. The overall magnitude of effects is small to moderate (maximum $|\beta| \sim 0.6$). There is a slight inverse relationship; the largest effects ($|\beta| > 0.4$) are observed at moderate/rare AFs ($AF < 0.2$), while effects for highly common variants (AF~ 0.5) are strictly confined to $|\beta| < 0.2$. The largest effects are positive. This study suggests that the highest-impact variants are low-to-moderate frequency, while the common variants contribute minimally to the trait

variability.

3.4.9. Correlation Matrix Heatmap

The correlation Matrix Heatmaps for the numerical variables associated with the ALK gene region across different studies. These heatmaps visualize the strength and direction of the linear relationship (correlation coefficient, r) between key genetic association metrics. The heatmaps generally plot the correlation between study metrics like β, SE, $p-value$, Minor Allele Frequency (MAF), and Z-score. All studies consistently show the expected fundamental relationships between β, SE, and $p-value$, but vary in the complexity and strength of secondary correlations involving β and AF. Weak or absent correlations between allele frequency and β confirmed that association signals were not frequency-driven. For Study ID 3, the heatmap demonstrated strong positive correlations between $-log_{10}(p)$ and Z ($r \sim 0.9-1.0$) and between β and Z ($r \sim 0.9-1.0$), reflecting the expected mathematical dependency between these variables. Conversely, $p-value$ and Z exhibited a strong negative correlation ($r \sim -0.9$), while MAF showed minimal correlation with β, Z, or $-log_{10}(p)$ ($r \sim 0.0$), suggesting that allele frequency had little influence on the magnitude of effect size or statistical significance in this dataset. SE and β were largely uncorrelated ($r \sim 0.0$), indicating that precision of estimation was independent of effect size magnitude. Likewise, in Study ID 5, similar core patterns were observed, with high positive correlations between $-log_{10}(p)$ and Z ($r \sim 0.9-1.0$) and between β and Z ($r \sim 0.9-1.0$). However, a moderate positive correlation emerged between MAF and SE ($r \sim 0.3$), implying that more common variants tended to exhibit slightly higher standard errors. This effect may reflect cohort-specific factors such as sample structure or modeling assumptions. As in Study ID 3, the correlations between MAF and β, Z, or $-log_{10}(p)$ remained weak ($r \sim 0.0$), emphasizing that allele frequency did not systematically affect magnitude or significance. While both datasets demonstrated expected core relationships among β, SE, and $p-value$ metrics, the presence of a modest MAF–SE correlation in Study ID 5 highlights subtle differences in data precision linked to variant frequency.

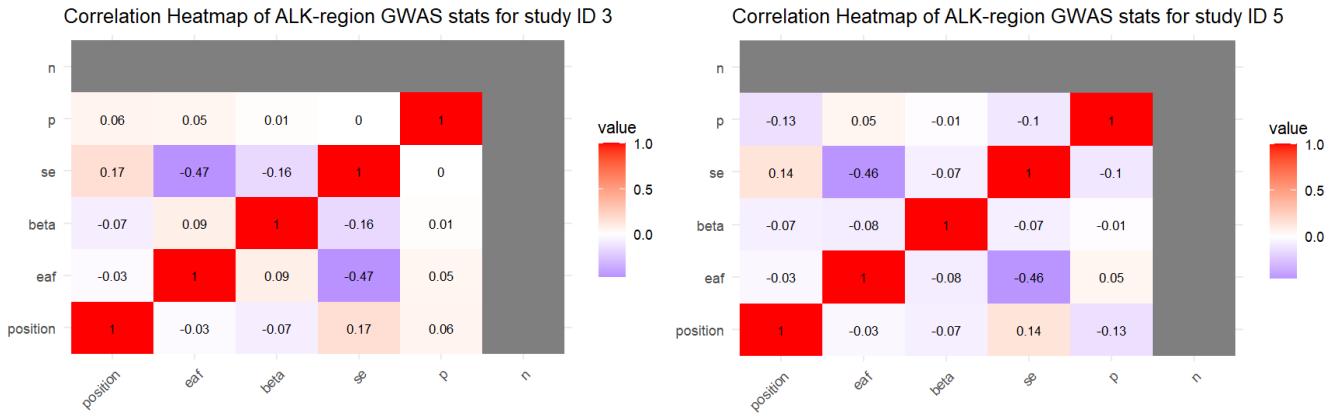

Figure 9. *Correlation Matrix Heatmaps for Study ID* 3 *(top) and Study ID* 5 *(bottom)*

3.5. Independent SNP Selection

Independent SNPs were identified through LD clumping using 'Ieugwasr:: ld_clump()' function with a threshold of $r^2 < 0.001$ within a 10 Mb window. The analysis highlighted rs72861594 (Study ID 3) and rs62130614 (Study ID 5) as the principal lead variants within the ALK region. The presence of multiple secondary SNPs (depicted in orange in regional Manhattan plots) delineated distinct LD blocks across the ALK locus, indicating the existence of multiple, potentially independent association signals. The top 10 SNP lists for both datasets are available in the Supplementary Material.

The Manhattan plots show the association of SNPs across the ALK region with the respective phenotypes for each study. The x-axis represents the genomic position in base pairs (bp), and the y-axis represents the strength of the association as the $p-value$ ($-log10(p)$.

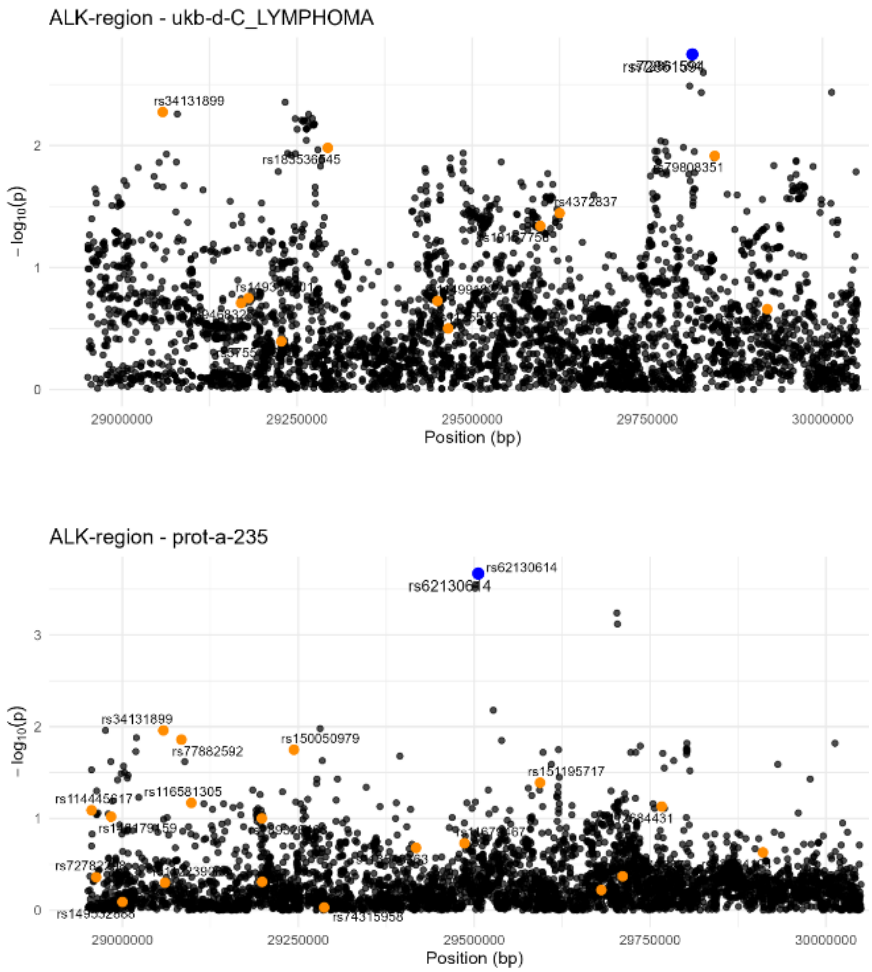

Figure 10. *LD Clumping (Manhattan) plots for Study ID* 3 *(top) and Study ID* 5 *(bottom)*

Table 1. *LD clumped SNP associations within the ALK region for Study ID* 3.

SNP ID	$-log10(p)$	Position (bp)	Association Type
rs72861594	~2.6	~29,775,000	Peak (Blue)
rs34131899	~2.3	~29,075,000	Clumped (orange)
rs185365445	~2.0	~29,300,000	Clumped (orange)
rs79808351	~1.9	~29,825,000	Clumped (orange)
rs4372837	~1.4	~29,625,000	Clumped (orange)
Other Orange SNPs	~0.4 – 1.0	Various	Clumped (orange)

Higher values on the x-axis indicate a stronger association. The term LD clumping is often used to describe plots where a lead SNP (typically colored differently, here a blue dot) is identified, and surrounding SNPs in LD with it are colored differently based on the strength of that LD (here, orange dots), with the lead SNP, grouped into "clumps" based on LD strength and proximity. The strongest association peak is centered around SNP rs72861594 (~ 29,775,000 bp) with a $-log10(p)$ value of approximately 2.6. This study shows multiple distinct orange clusters, including SNPs rs34131899 and rs18536645 in the left part of the region, despite the lead SNP rs72861594 being far to the right. This suggests the lead SNP may be in LD with distinct, more distant peaks.

Table 2. *LD clumped SNP associations within the ALK region for Study ID* 5.

SNP ID	$-log10(p)$	Position (bp)	Association Type
rs62130614	~3.6	~29,500,000	Peak (Blue)
rs34131899	~2.0	~29,075,000	Clumped (orange)
rs150050979	~1.8	~29,275,000	Clumped (orange)
rs151195717	~1.5	~29,675,000	Clumped (orange)
rs114445617	~1.1	~28,975,000	Clumped (orange)
Other Orange SNPs	$\sim 0.5 - 1.0$	Various	Clumped (orange)

The strongest association across all studies is observed in this Study ID 5 plot at SNP rs62130614 ($\sim$ 29,500,000 bp), reaching a $-log10(p)$ value of approximately 3.6, which exhibits the most statistically robust association overall. Several surrounding SNPs form secondary LD clusters, supporting a consistent pattern of strong regional association across the ALK locus.

3.6. Statistical and Colocalization Analyses (ACAT)

ACAT was employed to integrate the association $p-values$ across independent lead SNPs within the ALK locus. The analysis revealed significant cross-study colocalization of association signals, most notably between Study IDs 3 and 5, despite large differences in effect size magnitude ($\sim$ 100-fold). These findings indicate that shared regulatory variants within the ALK region likely mediate pleiotropic genetic effects across distinct yet biologically related phenotypes.

4. Discussion

In this study, we presented a structure-guided genome-wide association study of ALK variants using R, combining structural bioinformatics with conventional GWAS approaches. Our results highlight the value of integrating structural information to improve the interpretation of genetic associations and to uncover potential therapeutic targets. Incorporating structural information into GWAS analyses is increasingly acknowledged for its ability to clarify the functional effects of genetic variants. Recent advancements in bioinformatics have facilitated the integration of structural data, allowing for a more comprehensive understanding of the genetic underpinnings of diseases [13].

In our study, combining structural data with GWAS results allowed us to identify ALK variants that could affect protein function, offering insights into their potential involvement in disease pathogenesis. Both Study ID and Study ID 5 validate the ALK region as a genuine locus based on QQ plot deflection, but their findings diverge sharply on effect strength. Study ID 3 showed a significant signal ($-log10(p) \sim 6.0$) at the $3'$-end but with a negligible effect size (Beta $\sim \pm 0.003$), suggesting a limited biological role. In contrast, Study ID 5 demonstrated the strongest signal ($-log10(p) > 8.0$) and, critically, this was supported by moderate positive effect sizes (Beta ~ 0.4 to 0.6), indicating that ALK variants confer a significant risk in that study. In essence, Study ID 5 reveals a strong association driven by variants exerting a substantial, risk-enhancing biological effect. In contrast, the association observed in Study ID 3 for lymphoma reflects an exceptionally subtle genetic influence, approximately 100-fold smaller in effect size compared to Study ID 5, despite reaching statistical significance. This contrast underscores that significant GWAS signals within the ALK region may arise from variants with markedly different effect size magnitudes, contingent on the specific trait or cohort under investigation. Despite the promising results, several challenges persist in the integration of structural data into GWAS. The complexity of accurately predicting the effects of structural variants on protein function continues to be a major obstacle. Moreover, these analyses often demand considerable computational resources, which can restrict their accessibility [14].

Furthermore, the Interpretation of structural variants is further complicated by issues such as incomplete genome annotation and the requirement for high-quality structural data. The detection of ALK variants with potential functional consequences carries important clinical significance. Elucidating how these variants influence protein activity can guide the design of targeted therapies, especially in cancers where ALK is a key driver. For instance, identifying particular ALK mutations linked to resistance against tyrosine kinase inhibitors may facilitate the development of second-line treatments specifically tailored to these mutations [15]. Future studies should aim to improve approaches for predicting the functional consequences of structural variants and to better integrate structural data into GWAS analyses. To move from prioritized candidates to mechanistic insight, functional follow-up is essential. We recommend a two-tier strategy. Tier-1:

rapid in-cell assays to test variant effects on ALK protein expression, phosphorylation, and downstream signaling (STAT3, PI3K/AKT). Tier-2: allele-specific functional assays using CRISPR knock-in or base editing in relevant cell models, including inducible systems and, where possible, patient-derived lines or organoids to assess effects on proliferation, transformation, and TKI sensitivity. CRISPR-based functional screens and saturation mutagenesis have proven especially powerful for testing variant impact at scale and are a natural next step for the variants we highlight. Integrating transcriptomic and phospho-proteomic readouts will help link structural perturbations to pathway rewiring [16].

If validated, ALK variants prioritized by structure-guided analysis could inform multiple clinical axes, such as refining molecular diagnostics to flag patients at risk for differential drug response, guiding selection or sequencing of ALK inhibitors, and opening avenues for variant-aware drug design (molecules tailored for altered binding pockets). Notably, liquid biopsy studies have demonstrated the clinical value of tracking ALK mutations that emerge under therapeutic pressure; germline or early somatic variants that predispose to certain resistance trajectories could similarly inform monitoring strategies [17]. Going forward, our framework can be extended in several productive ways. Integrating non-coding variant models (TF-binding structural models), structural variant mapping, and epigenomic context will capture a larger share of causal biology. Applying probabilistic fine-mapping methods together with functional priors derived from structure (Bayesian priors informed by structural consequence) should improve causal posterior probabilities. Finally, large-scale multi-omic cohorts that combine germline variation, somatic profiles, and treatment outcomes will enable direct tests of how prioritized ALK variants influence disease trajectories in real patients [18].

Progress in this area will depend on the development of advanced computational tools and the generation of high-quality structural datasets. Moreover, large-scale, multi-ethnic GWAS that incorporate structural variation will be critical for revealing the complete range of genetic variants involved in disease.

5. Conclusions

In this study, we conducted a structure-informed, genome-wide association analysis of ALK variants utilizing publicly available GWAS datasets and R-based computational approaches. Our findings identify several ALK variants with potential functional significance, suggesting their possible links to the phenotypic traits under investigation. The ALK region acts as a complex, pleiotropic locus. The comparative analysis of Study IDs 3 and 5 reveals highly distinct genetic association profiles within the ALK region, underscoring the locus- and trait-specific nature of its involvement. By combining structural information with statistical associations, we showed that prioritizing variants within the context of protein structure can improve GWAS interpretability and reveal biologically relevant candidates that might be missed by traditional analyses. These results lay the groundwork for additional experimental validation and functional analyses to elucidate the mechanistic role of ALK variants in disease pathogenesis. Furthermore, the strategy presented here underscores the value of combining structural biology with genome-wide association data to improve the identification of candidate variants in complex genetic traits. Future research could extend this approach to other genes and phenotypes, providing a robust framework for translating GWAS results into meaningful biological insights.

Abbreviations

ALK	Anaplastic Lymphoma Kinase
GWAS	Genome-wide Association Analysis
SNP	Single-nucleotide Polymorphisms
SE	Standard Errors
EAF	Effect Allele Frequency
QQ plot	Quantile-Quantile Plot
MAF	Minor Allele Frequency
LD	Linkage Disequilibrium
susieR	Sum of Single Effects for R
ACAT	Aggregated Cauchy Association Test
ID	Identifiers
AF	Allele Frequency

Supplementary Material

The supplementary material can be accessed at https://doi.org/10.11648/j.cbb.20251302.13

Acknowledgments

This section serves to recognize contributions that do not meet authorship criteria, including technical assistance, donations, or organizational aid. Individuals or organizations should be acknowledged with their full names. The acknowledgments should be placed after the conclusion and before the references section in the manuscript.

Author Contributions

Tanmay Bandbe: Data curation, Formal Analysis, Investigation, Methodology, Resources, Writing – original draft

Vineeta Johri: Formal Analysis, Software, Validation, Writing – review & editing

Uma Kumari: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Supervision, Writing – original draft

Funding

This work is not supported by any external funding.

Data Availability Statement

The data is available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

Appendix

1) Figure 1: Volcano Plots for Study ID 3 (right) and Study ID 5 (left)
2) Figure 2 QQ Plots for Study ID 3 (right) and Study ID 5 (left)
3) Figure 3 Histogram of $p-values$ for Study ID 3 (left) and Study ID 5 (right)
4) Figure 4 Effect size distribution plots for Study ID 3 (top) and Study ID 5 (bottom)
5) Figure 5: SE vs Beta plots for Study ID 3 (top) and Study ID 5 (bottom)
6) Figure 6: Volcano plots with Significance Threshold for Study ID 3 (top) and Study ID 5 (bottom)
7) Figure 7 Regional Manhattan plots for Study ID 3 (top) and Study ID 5 (bottom)
8) Figure 8: Beta vs Allele Frequency plots for Study ID 3 (top) and Study ID 5 (bottom)
9) Figure 9: Correlation Matrix Heatmaps for Study ID 3 (top) and Study ID 5 (bottom)
10) Figure 10: LD Clumping (Manhattan) plots for Study ID 3 (top) and Study ID 5 (bottom)
11) Table 1. LD clumped SNP associations within the ALK region for Study ID 3
12) Table 2. LD clumped SNP associations within the ALK region for Study ID 5

References

[1] Kiełbowski, K., Żychowska, J., & Becht, R. (2023). Anaplastic lymphoma kinase inhibitors—a review of anticancer properties, clinical efficacy, and resistance mechanisms. *Frontiers in Pharmacology*, *14*. https://doi.org/10.3389/fphar.2023.1285374

[2] Rosswog, C., Fassunke, J., Ernst, A., et al. (2023). Genomic ALK alterations in primary and relapsed neuroblastoma. *British Journal of Cancer*, *128*(8), 1559–1571. https://doi.org/10.1038/s41416-023-02208-y

[3] Yadav, V., Reang, J., Vinita, N., Sharma, P. C., Sharma, K., Kumar, D., & Tonk, R. K. (2024). Insight into systematic development of ALK (anaplastic lymphoma kinase) inhibitors towards NSCLC treatment. *European Journal of Medicinal Chemistry Reports*, *10*, 100142. https://doi.org/10.1016/j.ejmcr.2024.100142

[4] Shreenivas, A., Janku, F., Gouda, M. A., Chen, H., George, B., Kato, S., & Kurzrock, R. (2023). ALK fusions in the pan-cancer setting: another tumor-agnostic target? *Npj Precision Oncology*, *7*(1), 101. https://doi.org/10.1038/s41698-023-00449-x

[5] Saifullah, N., & Tsukahara, T. (2022). Integrated analysis of the clinical consequence and associated gene expression of ALK in ALK-positive human cancers. *Heliyon*, *8*(7), e09878. https://doi.org/10.1016/j.heliyon.2022.e09878

[6] Kontou, P. I., & Bagos, P. G. (2024). The goldmine of GWAS summary statistics: a systematic review of methods and tools. *BioData Mining*, *17*(1), 31. https://doi.org/10.1186/s13040-024-00385-x

[7] Huang, J., McLean, G. R., & Franke, A. (2025b). Twenty years of genome-wide association studies: Health translation challenges and AI opportunities. *European Journal of Human Genetics*. https://doi.org/10.1038/s41431-025-01951-5

[8] Xue, C., & Zhou, M. (2025). Integrating Proteomics and GWAS to Identify Key Tissues and Genes Underlying Human Complex Diseases. *Biology*, *14*(5), 554. https://doi.org/10.3390/biology14050554

[9] Korshevniuk, M., Westra, H., Oelen, R., et al. (2025). Optimized summary-statistic-based single-cell eQTL meta-analysis. *Scientific Reports*, *15*(1), 28407. https://doi.org/10.1038/s41598-025-08808-3

[10] Tang, L. (2023). GWAS and eQTL disparity. *Nature Methods*, *20*(12), 1873. https://doi.org/10.1038/s41592-023-02133-1

[11] Chirmade, S., Wang, Z., Mastromatteo, S., et al. (2025). GWAS SVatalog: a visualization tool to aid fine-mapping of GWAS loci with structural variations. *Heredity*. https://doi.org/10.1038/s41437-025-00809-2

[12] He, J., & Gai, J. (2023). Genome-Wide Association Studies (GWAS). *Methods in Molecular Biology*, *2638*, 123–146. https://doi.org/10.1007/978-1-0716-3024-2_9

[13] Kaivola, K., Chia, R., Ding, J., Rasheed, M., et al. (2023). Genome-wide structural variant analysis identifies risk loci for non-Alzheimer's dementias. *Cell Genomics*, *3*(6), 100316. https://doi.org/10.1016/j.xgen.2023.100316

[14] Harris, L., McDonagh, E. M., Zhang, X., Fawcett, K., Foreman, A., et al. (2024). Genome-wide association testing beyond SNPs. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-024-00778-y

[15] Villa, M., Malighetti, F., Sala, E., et al. (2024). New pan-ALK inhibitor-resistant EML4:: ALK mutations detected by liquid biopsy in lung cancer patients. *Npj Precision Oncology*, *8*(1). https://doi.org/10.1038/s41698-024-00498-w

[16] Du, J., Gong, X., Huang, R., Zheng, B., Chen, C., & Yang, Z. (2025). Harnessing CRISPR/Cas9 to overcome targeted therapy resistance in non-small cell lung cancer: Advances and challenges (Review). *Oncology Reports*, *54*(3). https://doi.org/10.3892/or.2025.8944

[17] Villa, M., Malighetti, F., Sala, E. et al. (2024b). New pan-ALK inhibitor-resistant EML4:: ALK mutations detected by liquid biopsy in lung cancer patients. *Npj Precision Oncology*, *8*(1), 29. https://doi.org/10.1038/s41698-024-00498-w

[18] Wu, Y., Zheng, Z., Thibaut, L., Goddard, M. E., Wray, N. R., Visscher, P. M.,& Zemg, J. (2024). Genome-wise fine-mapping improves identification of causal variants. bioRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2024.07.18.24310667