

Rare variant

discovery by protein pathway analysis

Introduction

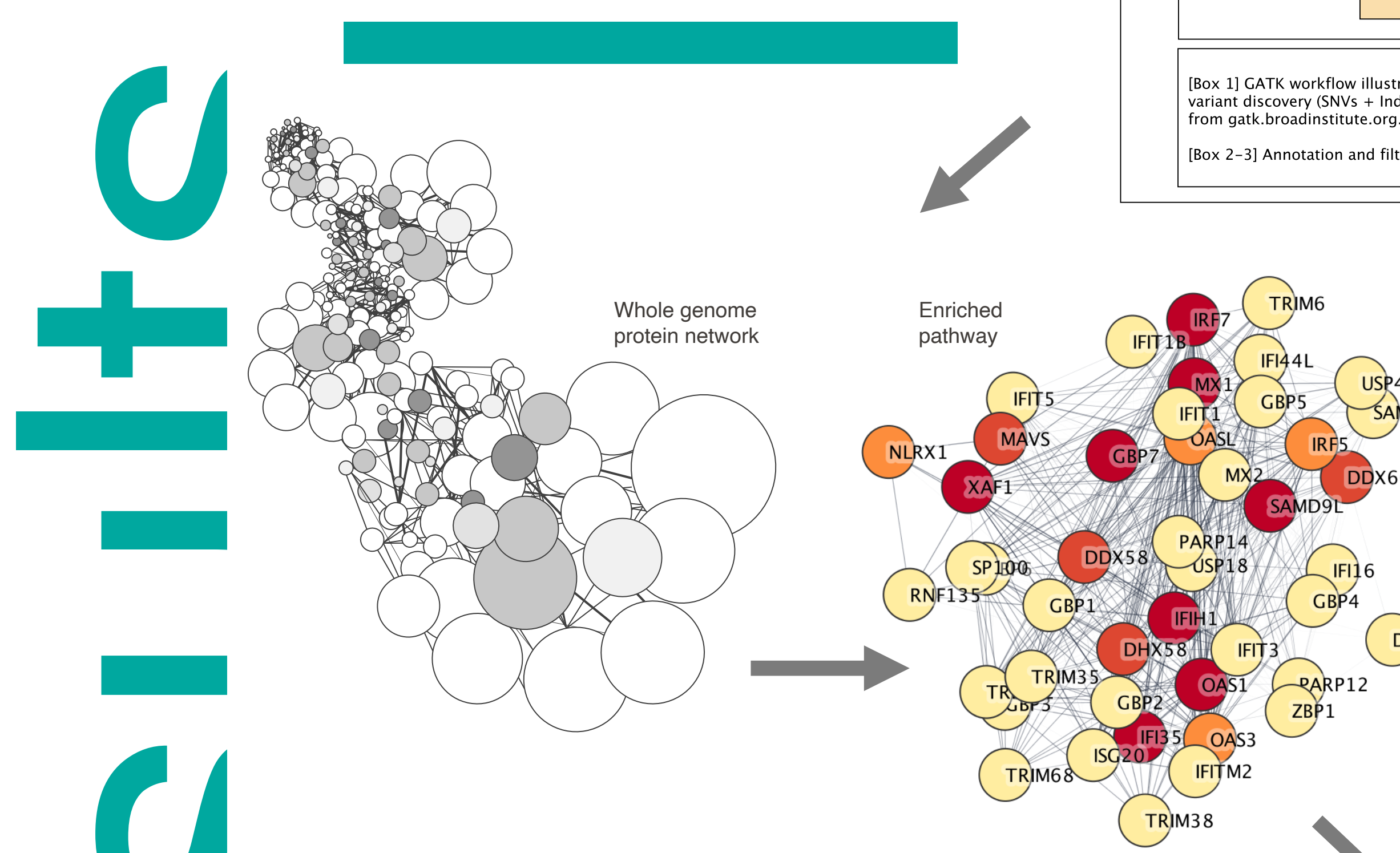
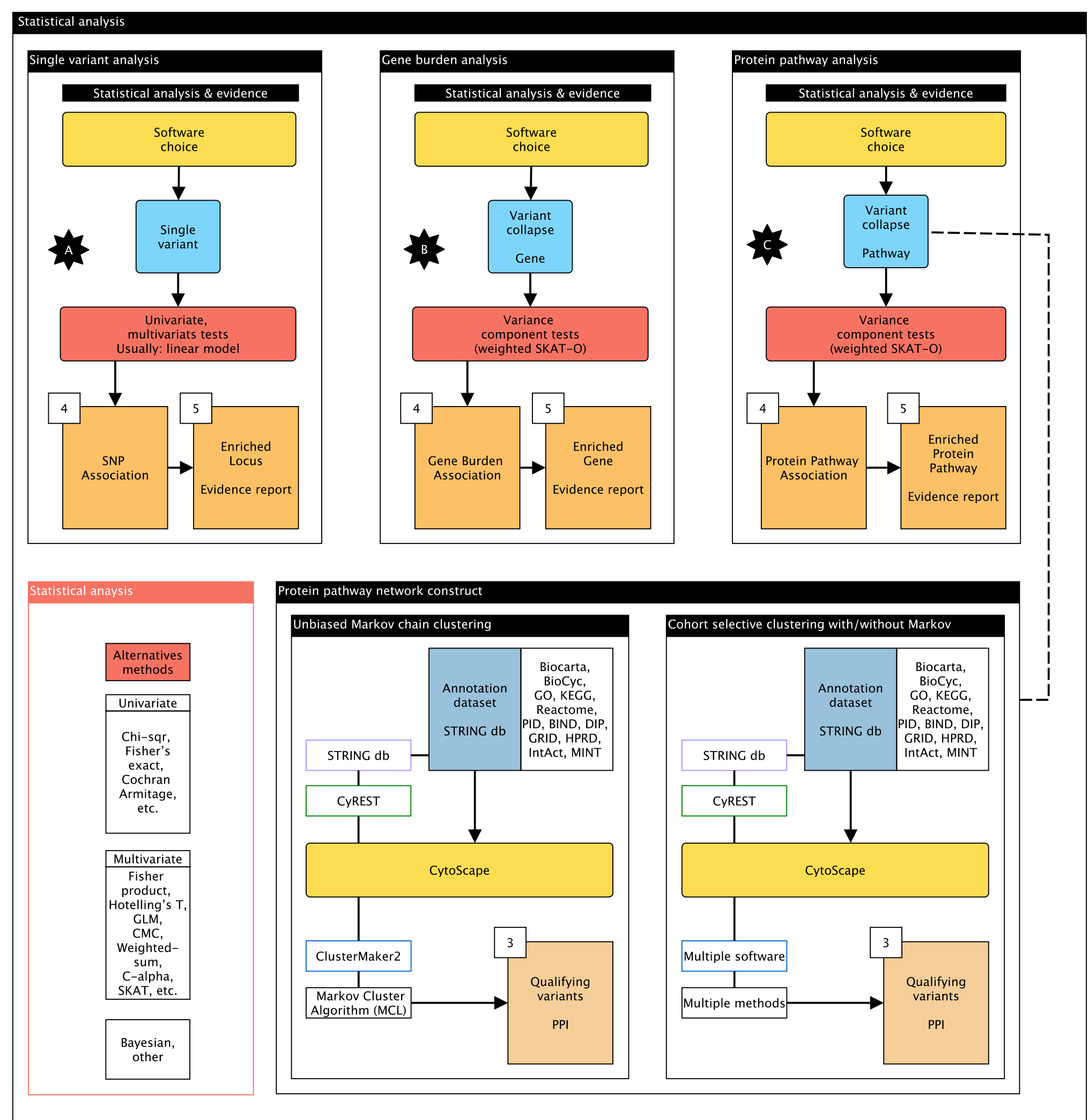
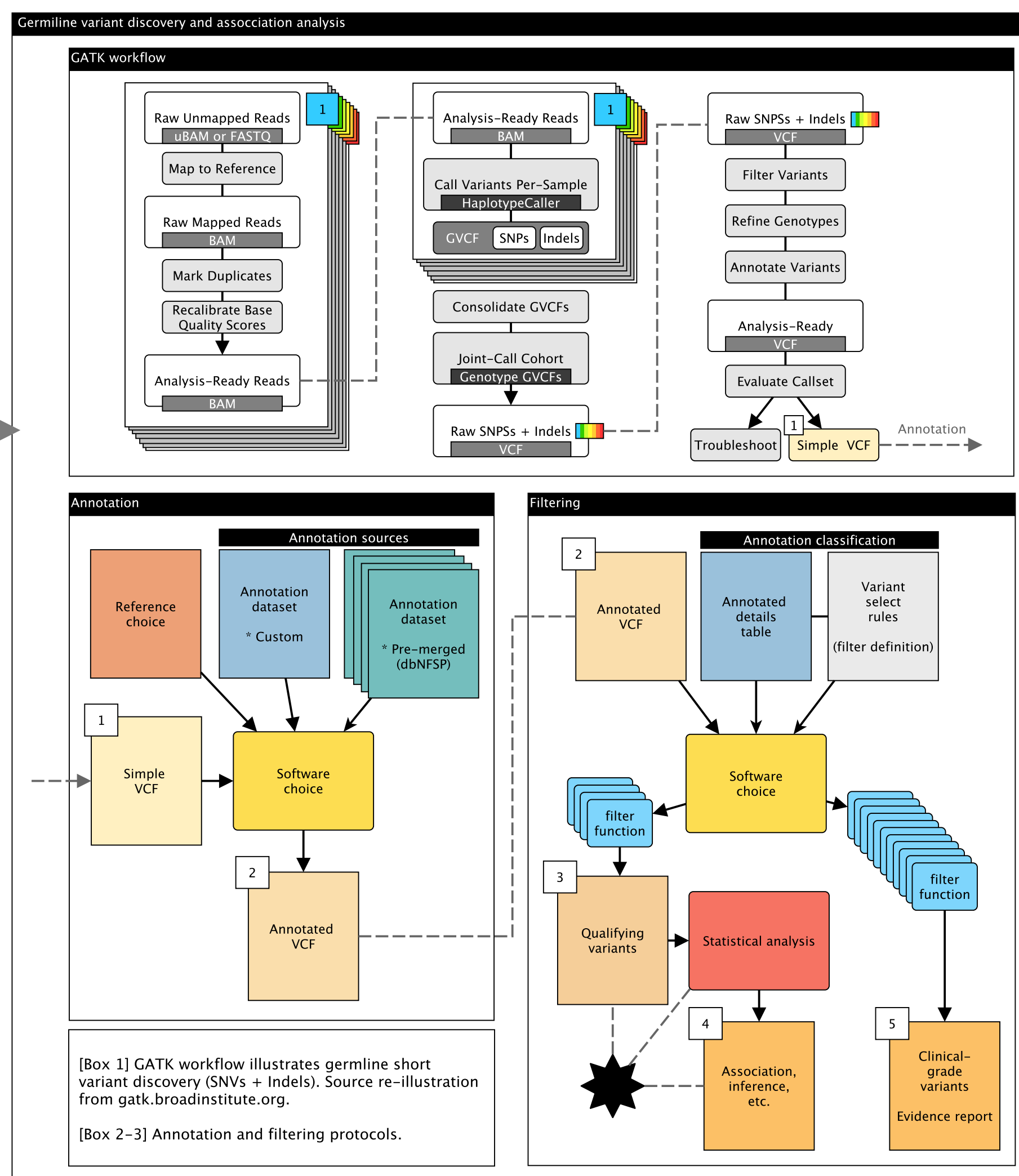
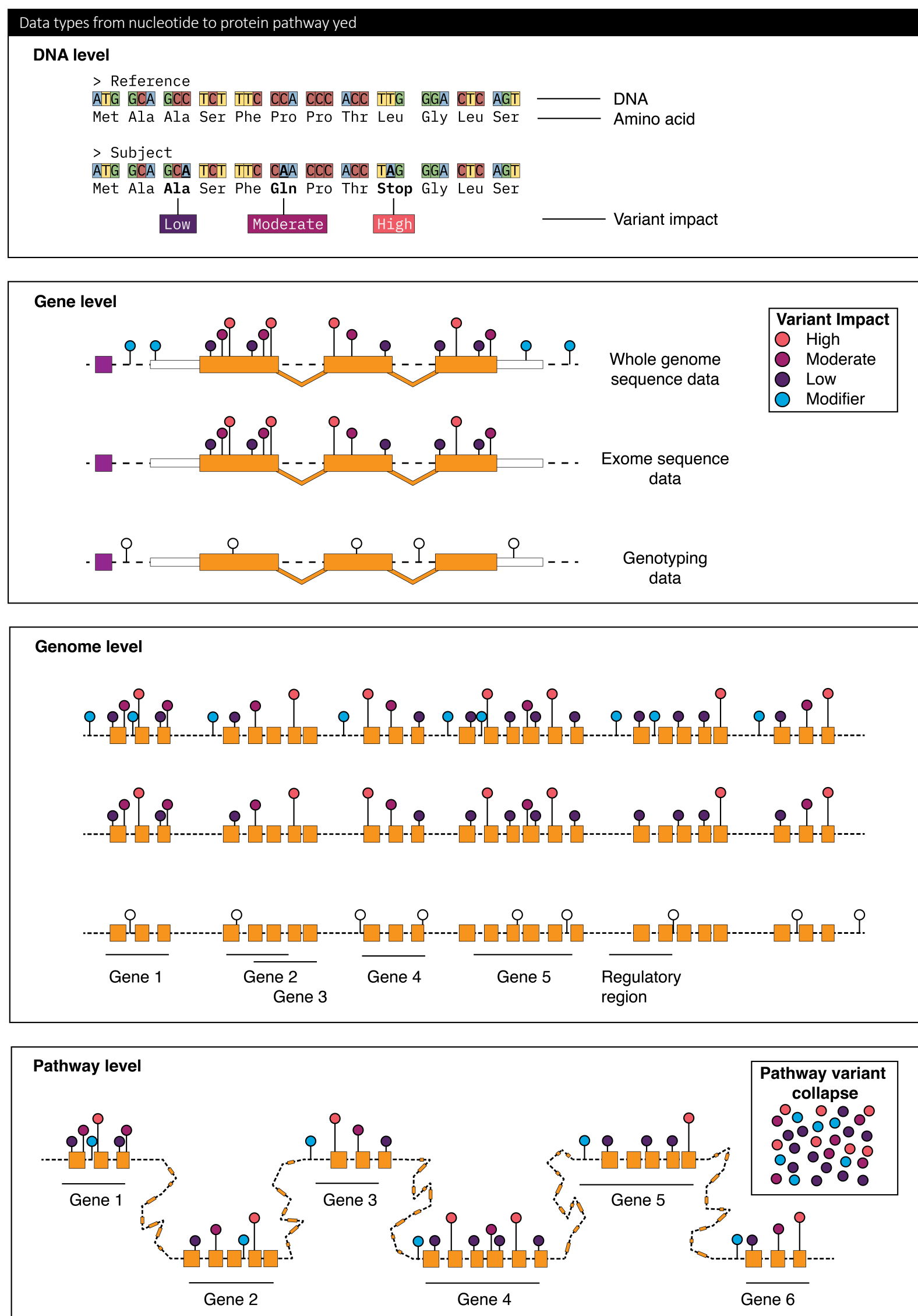
To uncover new mechanisms of immune response, we search for **rare** genetic variants in unusually severe cases of infection.

We used genome sequencing followed by protein **network** analysis to catalog rare coding variants and cluster them by known physical and functional associations.

Burden testing is also possible at the **gene level**. We present a summary of our unbiased statistical method for extending this search into the **pathway level**.

Statistical association analysis between genotype and phenotype can be performed on individual variants at **DNA level**.

By collapsing qualifying disease-causing variants into protein pathways we can perform robust association analysis.



A protein **interaction network** graph was built from qualifying variant genes. Protein pathways were identified by **clustering**.

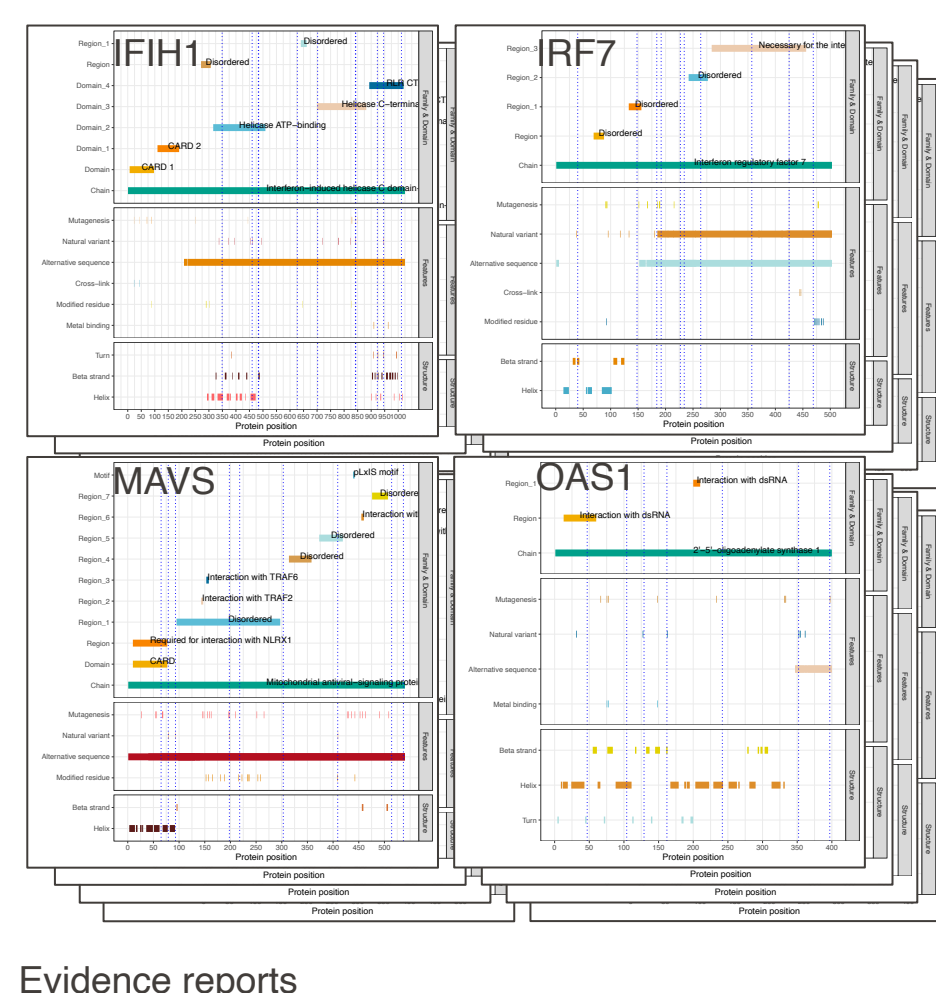
Case/control **association** testing within pathways found enrichment for inflammatory response, **viral** nucleic acid **detection**, and other IRF-related genes.

Viral **recognition** and inflammatory **response** is controlled by the key domains which were detected as damaged in these patients. Including:

- CARD effector domain
- ATP-binding helicase domain
- C-terminal repressor domain
- dsRNA binding motif

We designed a **modular** protocol. Alternative annotation datasets and statistical tests can be substituted for a **tailored** study design.

Genomic data is prepared according to a standardized format; **GATK** best practices for genome analysis¹ with joint genotyping and custom filtering².



Filtering is customizable for detecting highly reliable candidate disease-causing variants based on standardized protocols for clinically-actionable interpretation.

Genes harboring functional variants are grouped based on **protein-protein interactions**³.

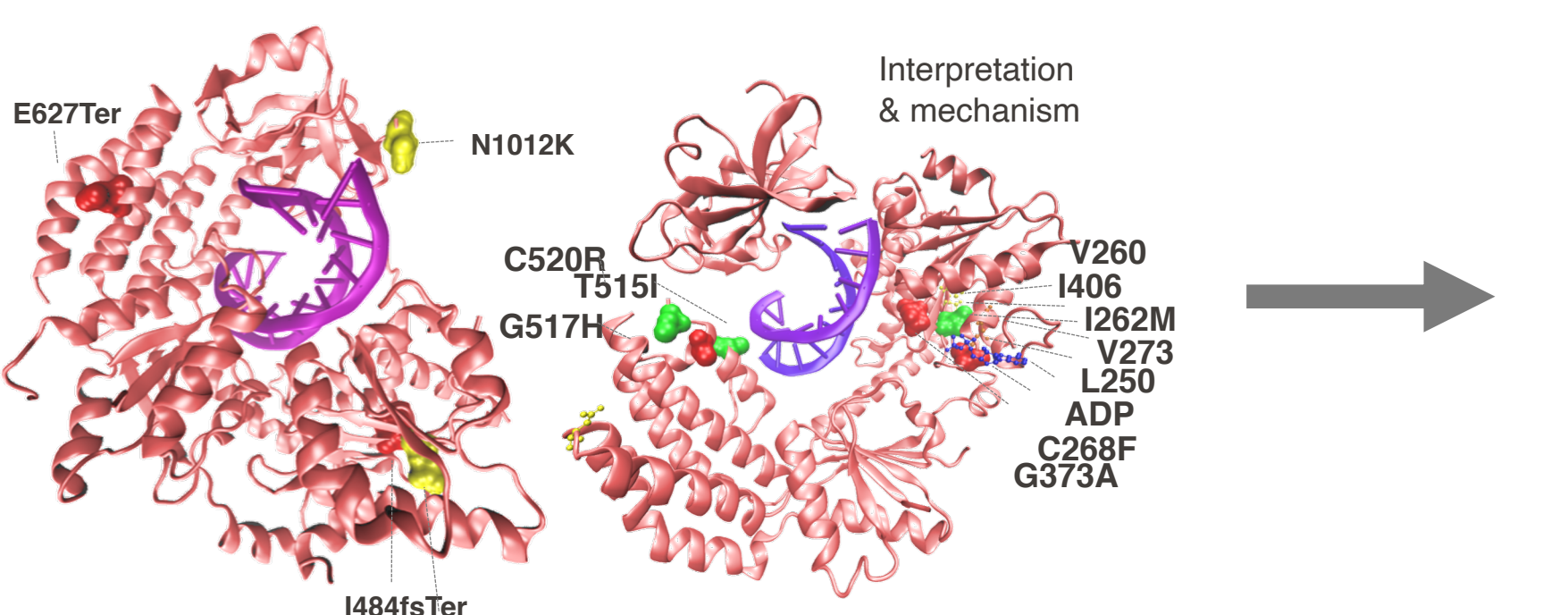
Annotation of all variants with >160 bioinformatic resources⁴: variant effect, pathogenicity, structure, etc.

Variant **collapse** is applied for qualifying variants. A Markov cluster algorithm⁵ separates protein networks.

Case-control testing is performed on each protein pathway cluster⁶.

Significantly enriched protein pathways are identified.

Evidence-based standardized variant interpretation allows for rapid explanation of disease and biological function⁷.



Evidence reports
Systematic
Open data
Reproducible
Unbiased
Statistically robust

Summary

We present a method for the unbiased detection of a protein network, driving disease, based on damaging functional variants. Easy interpretation of both raw and summary data.

Evidence reports explain clinical interpretation and biological mechanism. Detection of known damaging variants and explanation for those previously categorized as variants-unknown-significance.

Under the hood

For n subjects with that has m collapsed variants. Let X be the $n \times m$ genotype matrix for regression model: $g(E(Y_i)) = \alpha_0 + C_i\alpha + X_i\beta$, where $g(\cdot)$ can be set for continuous or dichotomous traits;

- $\alpha = (\alpha_1, \dots, \alpha_p)'$ - regression coefficients
- $C_i = (C_{i1}, \dots, C_{ip})'$ - covariates
- $X_i = (X_{i1}, \dots, X_{im})$ - vector of genotypes
- Y_i - trait value
- $\beta = (\beta_1, \dots, \beta_m)'$ - regression coefficients for the m genetic variants.

We assume that β is a random variable with $E(\beta_j) = 0$, $\text{Var}(\beta_j) = w_j^2\tau$, $\text{corr}(\beta_j, \beta_k) = \rho$ for different j and k . For testing $H_0: \beta = 0$ (no genetic effects) we use the variance-component score statistic $Q_p = (Y - \bar{\mu})' K_p (Y - \bar{\mu})$ where $K_p = XW R_p W X'$, in which $R_p = (1 - \rho)I + \rho 11'$ specifies an exchangeable correlation matrix and $W = \text{diag}(w_1, \dots, w_m)$ is a diagonal weight matrix; for a dichotomous trait, $\bar{\mu}$ is a vector of estimated probabilities of Y under the null model.

With this general class of tests we can perform analysis which best represents the biological scenario. Two commonly used tests include the burden and SKAT test. SKAT-O ideal for all tested genotype-phenotype models of rare mendelian disease.

- SKAT: $Q_p = 0 = \sum_{i=1}^n w_i^2 [\sum_{j=1}^m (Y_i - \bar{\mu}_j) X_{ij}]^2$
- Burden: $Q_p = 1 = [\sum_{i=1}^n w_i \sum_{j=1}^m (Y_i - \bar{\mu}_j) X_{ij}]^2$
- SKAT-O: $Q_p = (1 - \rho) Q_{SKAT} + \rho Q_{burden}$

- DePristo, GATK: A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Gen 2011
- Povysil, Gundula, et al. Rare-variant collapsing analyses for complex traits: guidelines and applications. Nature Reviews Genetics 2019.
- Franceschini, A et al. STRING: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res '13.
- Liu, X., et al. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome medicine 2020.
- Enright A.J., et al. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 2002.
- Lee, S., et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. The American Journal of Human Genetics 2012.
- Richards, Sue, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the ACMG and AMP. Genetics in medicine 2015