Screening Biases in Inherited Diseases

Alper Karagöl

Department of Medicine

MUTATION ≠ DISEASE

- Yes, even if inherited.

%30 of hemophilia B caused by a de novo mutation. = 1/60000 of males have a de novo mutation that causes hemophilia.

Meanwhile general mutation frequency is 0.05, thus statistically 3000/600000 males already had mutation independent of the clinical phenotype. So, even if the mutation causes the disease, not ALL mutations cause the disease. In fact, the ratio is just 1/3000.

According to Mendelian inheritance of the X chromosome, if family history exists (a brother has the disease) the screened person will have a 35% chance of being affected (mother is carrier). Meanwhile the general F9 mutation frequency is 0.30. This makes genetic screening for hemophilia B extremely inefficient.

Our analysis revealed an unacceptable false-positive rate (36%) in genes without genotyping data provided.

- Providing 6500 case reports of the patients, this screening can be successfully confirmed.

- Even without confirmed cases, our server provides SNPs with high frequencies. Excluding them, the probability of a healthy person having at least one variant is reduced to ≈ 0.05. **Thus, drastically increasing screening accuracy to 92%**
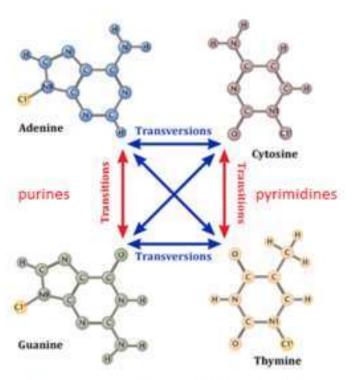
MISSENSE VARIANT ≠ STRUCTURAL ALTERATION

- The improved tools started(2023) to utilize structural models rather than solely depending on sequence data. We proposed implementing them on hemophilia research. The results were in line with old methods but much more diverged and show statistically significant differences (V = 240466, p-value < 0.01).

- We identified a clustering problem in older method that causes false positives and it may have serious implications in hemophilia B research.

NON-DAMAGING VARIANT ≠ PUBLICATION

- Publication bias

- False positives produce publications but proving them false don't.

Transitions vs. Transversions

Adenine

purines

Transversions

Transitions

Transitions

Cytosine

pyrimidines

Transversions

Guanine

Thymine

There are 8 ways to produce a transversion and only 4 ways to produce a transition, yet in nature transition mutations are much more common than transversions.

Why?

DNA substitution mutations are of two types. Transitions are interchanges of two-ring purines (A/G) or of one-ring pyrimidines (C/T): they therefore involve bases of similar shape. Transversions are interchanges of purine for pyrimidine bases, which therefore involve exchange of one-ring and two-ring structures.

- Mutations are mostly resulted from codon properties
- Transition>Transversion

- Observed mutations passively reflect mutational properties of the genetic code, as distributions are roughly in line with expectations. In other words, instead of the distributions from natural selection acting and producing observational biases, the codonic characteristics(non-Darwinian evolution) may more affected the studied variant distributions. It is also possible that structural similarities of amino acids resulted in the presentation we see, as fatal mutations would produce sample biases.

- Non-GENETIC causes have drastic effects
  - Even if environmental effects only accounted for %10

- Meanwhile general mutation frequency is 0.05, thus statistically 3000/600000 males already had mutation independent of the clinical phenotype. So, even if the mutation causes the disease, not ALL mutations cause the disease. In fact, the ratio is just 1/3000.
- if environmental effects only accounted %1 This makes only 0.9/30 of hemophilia B is resulted from a mutation.
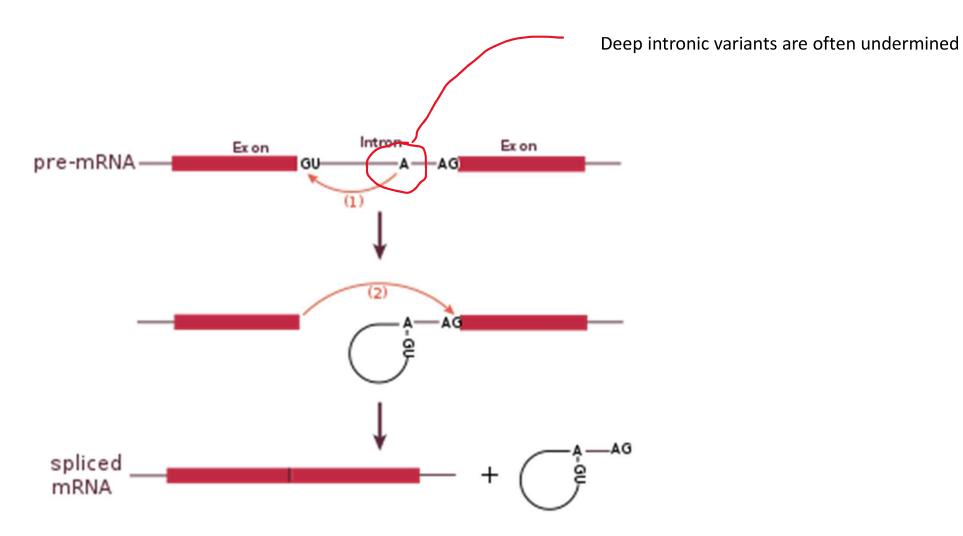
Environmental effects only accounted %0.03 (we spesificly selected Hemophilia B):

By using Bayesian statistics, we calculated that if a single case variant in the database is causal to the disease by 49.3%; for a similar variant that has three cases, the probability goes to %87. This statistic has potential implications, the genetic databases are historically treated as diagnostic and monitoring tools, but without sufficient case reports they are far from accurate.

Genes are not deterministic

- RNA processing, gene to gene interactions, epistasis, nonMendelian inheritance

- Non-coding mutations as likely as coding mutations to cause the disease
- Synonymous mutations can alter mRNA processing

Deep intronic variants are often undermined

Apart from the majority of the variants that were located in canonical splice sites. We calculated the deep intronic variant of c.392-864T>G was to disrupt splicing ($\Delta = 0.3$).

Our physical analysis showed that this variant facilitates donor gain, and its clinical phenotype has been confirmed by a case report in 7 patients.

Epistasis

- Although rare, some variants in the database become gradually less severe over time and some are even not reported anymore, indicating compensatory mechanisms for F9 may have evolved in the population. A good example is Arg43Trp variant which has a significant negative relationship between Severity and Year, ($r(50) = -0.291$, $p = .036$). Alongside, p.Ile443Thr is becoming less and less submitted ($r(32) = -0.582$, $p < .001$). The evolution of the disease indicates that frequent updates in the database are not only useful but in fact necessary.

Genetic drifts

- Variants based on populations can not be directly generalizable to artificial variants as natural selection and genetic drifts in the non-Hardy-Weinberg equilibrium population introduce a sample selection bias.

- It should be noted that, the human reference genome may not fully represent the diversity of humanity, due to the great diversity found in some parts of the human genome, compounded by historical bias in the selection of participants in genomic studies.

- We implemented patient data from Turkey and Central Asia.

References

- van Iterson, M., van Zwet, E.W. et al. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol* 18, 19 (2017). https://doi.org/10.1186/s13059-016-1131-9

- Fitipaldi H, Franks PW. Ethnic, gender and other sociodemographic biases in genome-wide association studies for the most burdensome non-communicable diseases. *Hum Mol Genet*. 2023. doi: 10.1093/hmg/ddac245

- **Karagöl T, Karagöl A.** Zhang S. Structural design of altered dopamine transporters and uncovering the evolutionary dynamics of QTY-code pairs. *Scientific Reports.* (Under review, Sept 2023)

- **Karagöl et al.** Structural bioinformatics studies of glutamate transporters and uncovering the natural mutations of L->Q, I->T, F->Y and Q->L, T->I and Y->F. *PLOS ONE* (2023)

- **Karagöl et. al.** A genomic upgrade to 2301 genetic variants in the interactive coagulation Factor IX database provides enhanced insights into haemophilia B. *Journal of Thrombosis and Haemostasis.* (Under review)

- Perkisn S.J, **Karagöl T., Karagöl A.,** Spencer H. Factor IX Variant Database: Genetic variants in the human coagulation factors F9 as a service for clinicians and biochemists. *© Copyright 2003-2023, Structural Immunology Group , University College London (2003- 2023)*

- **Karagöl et al.** Inhibitory effects of truncated isoforms on Glutamate Transporter trimerization. *In prep*

.