WHITEPAPER

MolecularDB: VarSome's Big Data

DATED

April 21, 2020

DOCUMENT VERSION

1.0

Contents

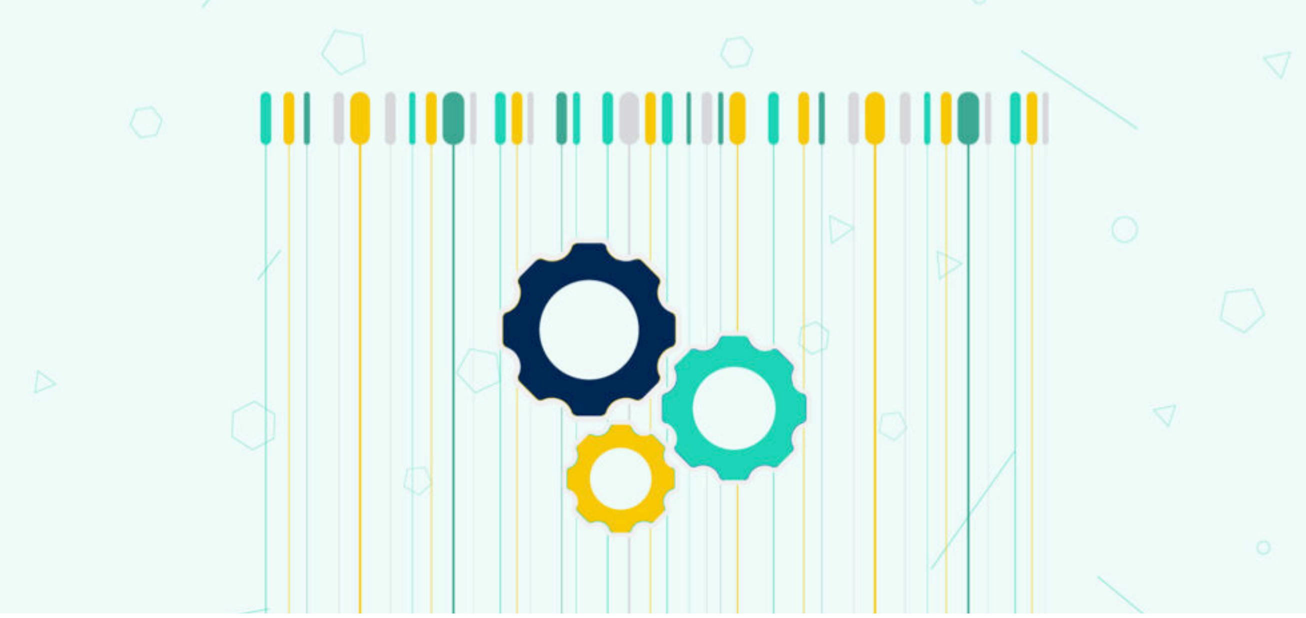
DATED
April 21, 2020

DOCUMENT VERSION
1.0

Introduction

As a result of little standardization, large amount of new scientific findings generated almost every day and an explosion of sequencing data for various purposes, the landscape of human genomics is quite fragmented, siloed, and inconsistent. We all know how frustrating the process of assessing a comprehensive information for genomic variant can be. The way forward is data integration, harmonization and cross-referencing.

However, integration of large data sets (Big Data), especially in the field of genomics, is a challenging endeavor which can be successfully tackled only by a multidisciplinary team, bringing together strong skills in both life sciences and software engineering. The data integration process can be seen as constructing a skyscraper. One can't build a skyscraper by stacking up small houses on top of each other. Building a skyscraper requires a whole new approach, and so does the integration and harmonization of genomics Big Data. In our approach, VarSome is the skyscraper, while MolecularDB is the architectural plan for it.



MolecularDB

MolecularDB is VarSome’s integration and harmonization engine for genomics Big Data. It’s a purpose-built data storage system specifically designed by our engineers since the first line of its code to meet the high demands of clinical-grade genomics applications, such as annotation of whole genomes, exomes, and gene panels.

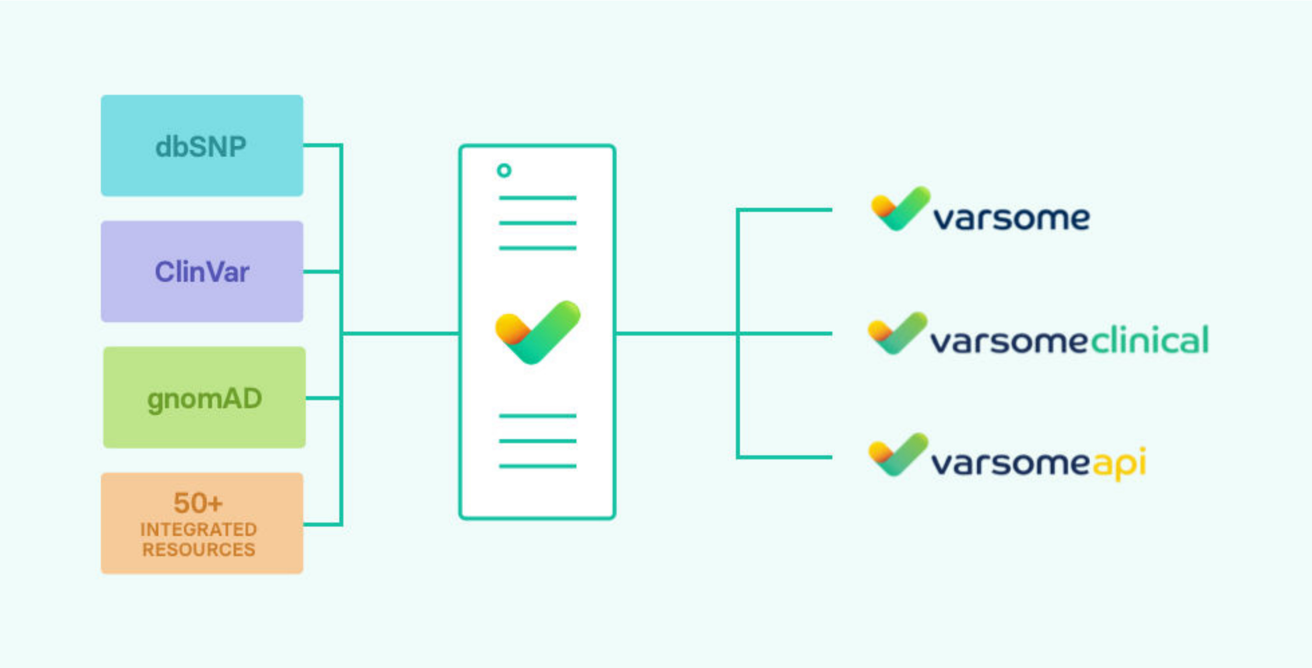
Currently, VarSome provides access through MolecularDB to over 50 public genomics-related data sets, which represents over 33 billion data points, plus contributions from a 200'000-strong global community. But there is more to it: whenever a public database is updated, MolecularDB quickly processes it and makes it available on VarSome in very short turnaround times. Apart from public data resources, MolecularDB can facilitate access to proprietary (such as your own private variant database) as well as licensed databases (such as HGMD) and cross-reference their content with data sets already available on VarSome (either in public or private manner).

Data quality is of paramount importance: MolecularDB ensures genomics data are meticulously integrated and cross-referenced, and insertions and deletions are matched consistently across all the data resource available on VarSome. MolecularDB also runs daily comprehensive data integrity checks.

50+ Integrated Resources

- ClinVar
- dbSNP
- gnomAD
- HPO
- Ensembl
- RefSeq
- GWAS
- CGD
- HGNC
- UniGene
- Orphanet
- CIViC
- genes
- GERP
- dbNSFP
- COSMIC
- IARC TP53
- ICGC
- Kaviar
- DANN scores
- CIViC mutations
- UniProt variants
- UniProt domains
- GHR
- CPIC
- DGV
- DECIPHER
- ExAC CNVs
- ExAC genes
- PanelApp
- Mondo
- PMKB
- BRAVO
- REVEL
- scSNV

We keep adding new ones!



We keep adding new data resources based on the user feedback. Let us know which database you miss on VarSome.com. The list above is valid at the date of publishing this whitepaper. For an updated list of resources available, visit varsome.com.

Did you know?

VarSome's integrated database is leveraged in VarSome Clinical, a CE-IVD-certified and HIPAA-compliant platform allowing fast and accurate variant discovery, annotation, and interpretation of NGS data for whole genomes, exomes, and gene panels. VarSome Clinical helps molecular geneticists and clinicians reach faster and more accurate diagnoses and treatment decisions for genetic conditions.

LEARN MORE ABOUT VARSOME CLINICAL →

varsomeclinical

ACMG Guidelines

One of the benefits of possessing such a massive aggregated and harmonized database is that it can be applied in further downstream processes, such as automated variant classification according to the guidelines of the American College of Medical Genetics and Genomics (ACMG). VarSome's robust implementation of ACMG guidelines contains explanations for each ACMG rule, along with why it has been triggered, or why not. If you have some additional evidence, you can manually turn on or off other ACMG rules, reach and evaluate the final verdict for your variant, and save it eventually as a manual classification for your future samples. Besides that, VarSome's ACMG receives lots of scrutiny from 200k+ users worldwide, which ensures its quality and comprehensiveness. Indeed, in our recent survey, a very large number of users claimed VarSome's ACMG is one of the main reasons for using VarSome!

WHITEPAPER

Implementation of ACMG Guidelines

LEARN MORE

Performance

Apart from aggregation and harmonization of genomics data resources, VarSome's MolecularDB ensures extremely fast data retrieval for sample annotations as performance matters a lot when it comes to annotation of large data sets, such as whole genomes and exomes, possessing easily millions of variants. Full results and functional annotations are typically generated in a few tenths of a second.

Application Programming Interface

VarSome comes with an Application Programming Interface (API), which similarly to the MolecularDB has been designed with performance in mind: in practice it can fully annotate over 1'000 variants per second. This is made possible through batch requests, where each API request can contain several thousand variants in a single call. A user-configurable allele frequency filter allows to further increase throughput up to 4x times.

Variant Search

Another consequence of the specific architecture of MolecularDB is that VarSome offers very versatile variant look up mechanisms. You can search VarSome by HGVS nomenclature (both on DNA and on protein level), rsID, gene name, transcript symbol or genomic location. VarSome can also parse single lines from VCF files to look up the variant it describes. In addition to that, the results are not limited to known variants only, you can query any possible variant, including 'abstract variants', i.e. variants defined with a range of coordinates or with specific attributes. As a consequence of this powerful search mechanism, VarSome annotates variants that no one has seen before.

VarSome's variant query examples:

- rs746753722 or CLN6 E227K or NM_017882.3(CLN6):c.679G>A or 15:68500735:C:T
- TP53:R175L or NM_000546:R175L or NM_000546(TP53):p.Arg175Leu or TP53:c.524G>T or chr17-7578406-C-A or rs28934578
- rs113488022, rs376932266
- BRAF:c.1799T>G, FTO:c.46-43098T>C,
- SYNGR1:c.607_608insACA, BAIAP2L2:c.1322_1363del
- BRAF:V600E
- 15-73027478-T-C, X 153418497 A G
- chr2-131129929-GACGGG-, chr13-38320595-AA-, 5:156479558:15: (deletions)
- chr22:39777823::CAA, 7-151945072--T (insertions)
- HAVCR1:c.487 (transcript position), 5:156479558 (genomic position)
- BRCA1, EGFR, HGNC:1097, ENTREZ:1956, UNIPROT:B7ZA85 (genes)
- NM_002482.3 (transcripts)

Full-text Search

VarSome's full-text search functions like other Internet search engines with one important difference: the search query returns entries only from the VarSome aggregated knowledge base, thus showing you the result relevant only for the genomics field. It enables you to perform targeted searches not just for variants, but over the entire contents of VarSome, such as articles, diseases, phenotypes, genes, etc. Importantly, this includes content provided by the entire VarSome global user community.

VarSome’s variant query examples:

- **shox:** finds all the genes, disease, articles etc. that mention the SHOX gene.
- **royal disease:** will return the disease, phenotypes & PubMed articles related to Haemophilia, including some ClinVar variants.
- **gene lung cancer:** will return all genes associated to lung cancer.
- **royal disease:** will return the disease, phenotypes & PubMed articles related to Haemophilia, including some ClinVar variants.
- **short stature syndrome:** will list all results referring to short stature, you can then navigate the various landing pages to find associated phenotypes, diseases, publications and genes. Alternatively, searching for “gene short stature” immediately returns the SHOX and SHOX2 genes.
- **“short stature”:** the quotes mean that this phrase, with exactly this spelling, must be found in the results.
- **clinvar "likely pathogenic" renal failure:** returns all variants classified in ClinVar as "Likely Pathogenic" and associated with renal failure.
- **uniprot "pathogenic" "Q9UMX9":** returns all variants classified by UniProt as pathogenic within protein Q9UMX9.

Results

Results are automatically ranked by relevance: this includes the words found, how often an article has been referenced in the VarSome database and the impact factor of the journal it is published in. Preference is given to genes, diseases and phenotypes. You can narrow down the search results to a given result type by clicking on the corresponding link. Clicking on a result will either take you to the standard VarSome page for a gene, or to a new dedicated page giving you all the available information for that result. This includes a list of all the items that refer to this result. For example, if you find a PubMed article, you can see genes, variants, diseases etc. that may refer to that article.

Advanced Search

As mentioned above, clicking on a type will refine the ranked results to only objects of that type, for example only genes or publications.

- Alternatively you can narrow the result type by starting your query by one of the following keywords: "gene", "disease", "phenotype", "article", "clinvar" or "uniprot".
  For example: 'gene IVF'
- Quoting words: adding quotes around a word tells the search engine to discard any entries that don't contain exactly that word. This helps narrow a search if there are too many results

Further learning resources

varsomeclinical

VarSome Clinical is a CE-IVD-certified and HIPAA-compliant platform allowing fast and accurate variant discovery, annotation, and interpretation of NGS data for whole genomes, exomes, and gene panels. VarSome Clinical helps molecular geneticists and clinicians reach faster and more accurate diagnoses and treatment decisions for genetic conditions.

LEARN MORE

PREMIUM varsome

No delays, premium features & data resources! AACT, COSMIC, Polyphen-2, CADD, OncoKB, CKB, PharmaGKB, and more.

LEARN MORE

Latest VarSome News

VarSome Documentation

Social networks

LinkedIn Twitter YouTube

Contact us

Saphetor SA – The creator of the VarSome Suite
EPFL Innovation Park - C
1015 Lausanne
Switzerland
VAT: CHE-467.115.331

support@varsome.com

DATED
April 21, 2020

DOCUMENT VERSION
1.0