# 3billion's state-of-the-art molecular diagnostic test for rare Mendelian diseases

3billion

# Table of Contents

# 3billion's Genomic Test Services

3billion's genomic test menu includes 3B-GENOME for genome sequencing test, 3B-EXOME for exome sequencing test and 3B-VARIANT for searching variants reported from 3B-GENOME or 3B-EXOME in related family members. 3B-GENOME and 3B-EXOME are based on next-generation sequencing (NGS) technology while 3B-VARIANT uses a traditional Sanger sequencing method. Both 3B-GENOME and 3B-EXOME are comprised of four main parts:

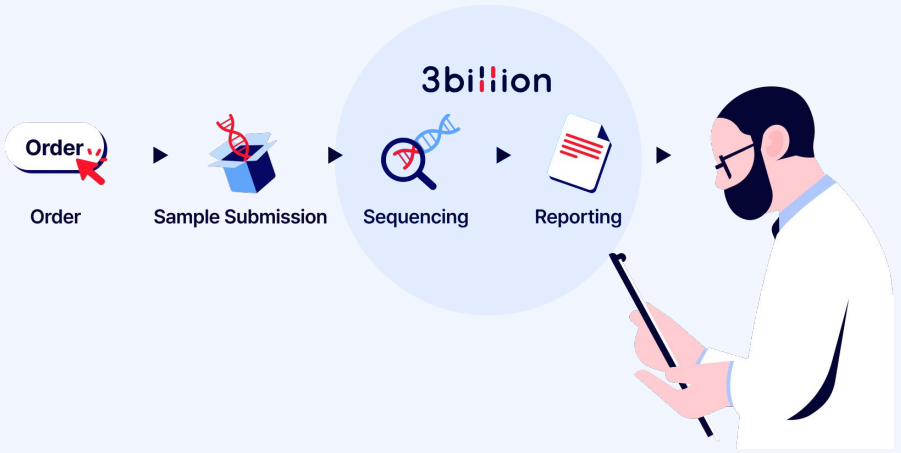



Figure 1. Schematic diagram of 3billion's genomic test service workflow

### 1. High-quality sequencing:

Sequencing library preparation and sequencing are performed using clinically validated Standard Operating Procedures (SOPs). 3billion's laboratory (3billion Co LTD Molecular Diagnostics Laboratory) is accredited by CAP (College of American Pathologists) and CLIA (Clinical Laboratory Improvement Amendments).

### 2. Sequencing data analysis:

Once the sequencing data is generated, 3billion's bioinformatics workflow is run on each sample, following the clinically validated SOPs.

### 3. Variant annotation and prioritization by EVIDENCE following the ACMG/AMP guidelines:

EVIDENCE is 3billion's state-of-the-art, highly automated and cost-effective analytical system developed in-house. Through its annotation, classification, and phenotype matching process only a handful of variants are left for the expert to interpret.

### 4. Variant interpretation in the context of the patient's symptoms and reporting of disease-causing variants:

Once EVIDENCE prioritizes the top candidate variants/genes, 3billion's trained medical geneticists manually curate each variant to identify the disease-causing variant for reporting.

# Introduction

Exome sequencing and genome sequencing are now being routinely used to diagnose suspected rare genetic (Mendelian) disorders by rapidly identifying the disease-causing-variants in an unbiased way. Identifying the molecular diagnosis for patients with rare genetic disorders is extremely important as it not only provides the patients with personalized clinical care and management plan but also opens genetic counseling opportunities for their family members.

Nevertheless, a substantial number of patients with suspected rare genetic diseases remain undiagnosed. A few of the reasons are: 1) limited access to genomic tests because of a relatively high cost and challenges with insurance coverage, 2) limited knowledge of gene-disease association, and 3) technical limitations with sequencing data analysis and variant interpretation. However, with increasing amounts of sequencing data being generated every day from a number of laboratories, and significant efforts to further advance analytical and interpretation skills, some of these challenges are getting resolved.

3billion has joined this global effort since October 2016, with the vision of providing an affordable test to patients with suspected rare genetic disorders and maximizing the variant interpretation skills and speed to ensure every patient who walks into 3billion's system can promptly get a clear molecular diagnosis.

# 3B-GENOME

Genome sequencing libraries are generated using TruSeq DNA PCR-Free Low Throughput Library Prep Kit (Illumina, San Diego, CA, USA) and sequencing is performed on NovaSeq X (Illumina, San Diego, CA, USA). Currently, the minimum depth-of-coverage (DOC) of autosomes per genome is 30x with a minimum 95% of the autosomes covered at 20x DOC.

Once sequencing is complete, the binary base call (BCL) sequence files generated by NovaSeq X are converted and demultiplexed to FASTQ files using bcl2fastq v2.20.0.422 [1]. Sequence reads in the FASTQ files are aligned to the human reference genome (GRCh38.p14 from NCBI, February 2022) and revised Cambridge Reference Sequence for mitochondrial genome (GenBank accession number: NC_012920) using BWA-mem 2.2.1 [2] to generate BAM files. BAM files are processed following the Rovaca v1.0.1 [3] and the GATK best practices (GATK v4.4.0, Genome Res. 2010;20:1297-303) [4] for single nucleotide variants (SNV) and small insertions/deletions (INDEL) variant calling to generate VCF files [5, 6]. Mutect2 is used for calling lower level heteroplasmic SNV/INDEL in the mitochondrial genome [7]. Structural variants(SV), including copy number variants (CNVs), inversions, translocations, repeat expansions and mobile element insertions, are also called from the BAM files using 3bCNV (in-house), MANTA (v1.6.0) [8], ExpansionHunter (v5.0.0) [9] and MELT (v2.2.2) [10]. AutoMap (v1.2) [11] is used for region of homozygosity (ROH) detection from the VCF files .

Various quality control metrics such as Q30, mapping rate, PCR duplication rate, total number of variants, heterozygous/homozygous (het/hom), and transition/transversion (ts/tv) ratios are used to ensure the sequencing data is within an acceptable range for a clinical test.

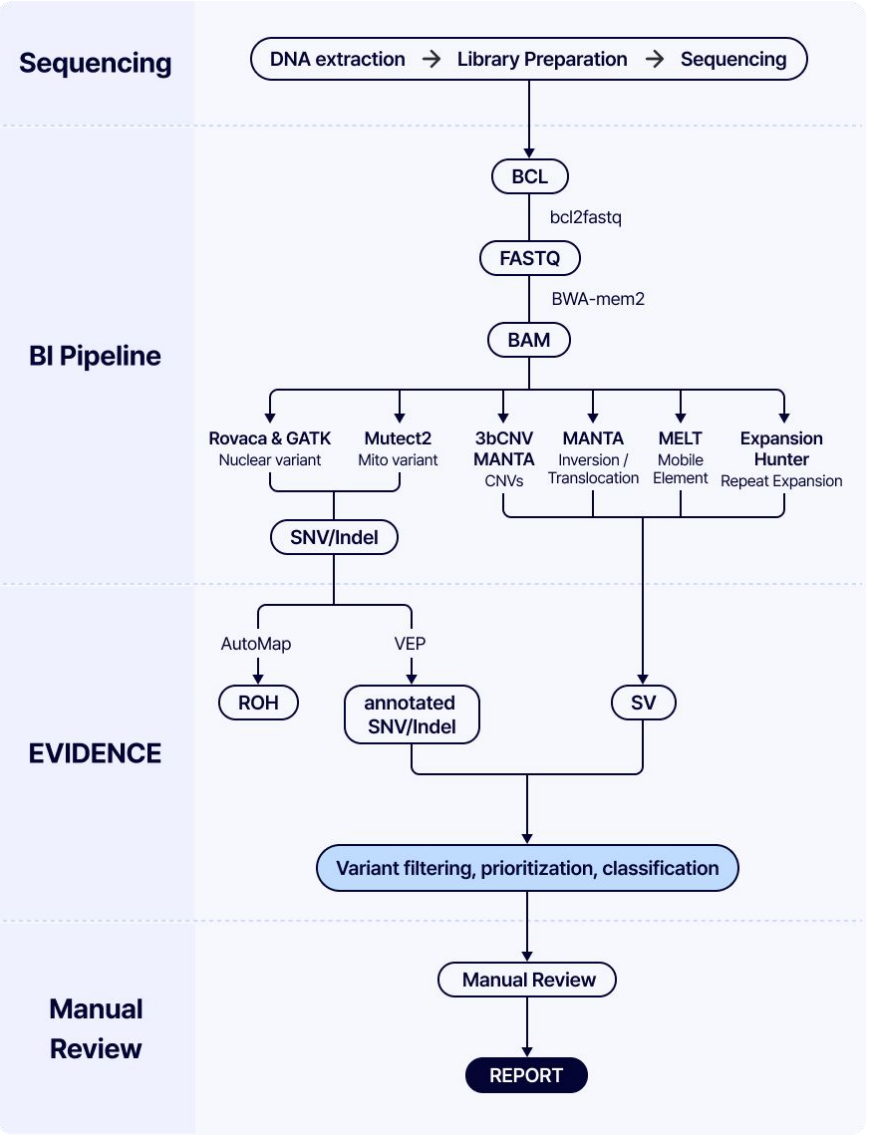



Figure 2. Schematics of 3B-GENOME analysis workflow

# 3B-EXOME

3billion performs exome capture with IDT xGen Exome Research Panel v2, supplemented with xGen human mtDNA panel and xGen Custom Hyb Panel v1 (Integrated DNA Technologies, Coralville, Iowa, USA) and sequencing on NovaSeq X (Illumina, San Diego, CA, USA). The IDT panel was selected after a thorough evaluation of the coverage statistics in comparison with other commercially available capture kits. Currently, the minimum DOC per exome is 100x with a minimum 98% of the targeted region covered at 20x DOC.

Once the sequencing is complete, the binary base call (BCL) sequence files generated by NovaSeq X are converted and demultiplexed to FASTQ files using bcl2fastq v2.20.0.422 [1]. Sequence reads in the FASTQ files are aligned to the human reference genome (GRCh38.p14 from NCBI, February 2022) and revised Cambridge Reference Sequence for mitochondrial genome (GenBank accession number: NC_012920) using BWA-mem 0.7.17 [2] to generate BAM files. BAM files are processed following the Rovaca v1.0.1 [3] and the GATK best practices (GATK v4.4.0, Genome Res. 2010;20:1297-303) [4] for SNV and small INDEL variant calling to generate VCF files [5, 6]. Mutect2 is used for calling lower level heteroplasmic SNV/INDEL in the mitochondrial genome [7]. 3bCNV is used for CNV calling based on DOC data and MANTA (v1.6.0) is used for CNV calling based on paired-end information [8]. Due to the lack of sequencing data between exons, the resolution of CNV calls is minimum 3 consecutive exons and for most of the CNVs, exact breakpoints are not identifiable. ExpansionHunter (v5.0.0) is used for repeat expansion variants[9]. MELT (v2.2.2) is used for calling mobile element insertion variants [10]. AutoMap (v1.2) is used for ROH detection from the VCF file [11].

Various quality control metrics such as Q30, mapping rate, PCR duplication rate, capture efficiency, total number of variants, het/hom, and ts/tv ratios are used to ensure the sequencing data is within an acceptable range for a clinical test.

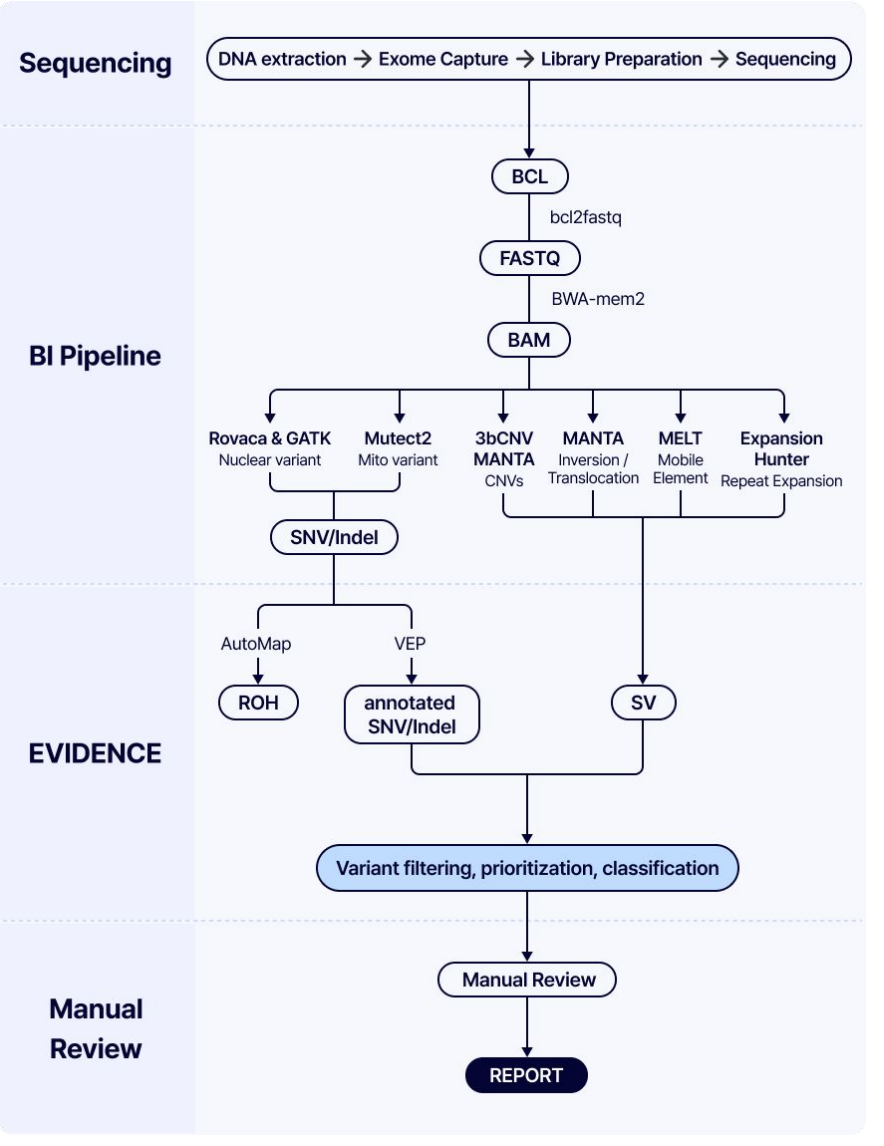



Figure 3. Schematics of 3B-EXOME analysis workflow

# 3B-INTERPRETER

3B-INTERPRETER is a service designed to provide you a comprehensive report within 2 weeks upon receiving your FASTQ or VCF data. You can use 3B-INTERPRETER if you want to perform the sequencing yourself but need the analysis and clinical report afterwards, or if you want to reanalyze undiagnosed genomic data produced after WES tests by a different laboratory. The FASTQ/VCF files were required 1) Good quality (Q30≥85%) Illumina 150bp paired-end sequencing data. 2) Sufficient and uniform coverage across the exome (Mean depth of coverage ~100x, Targeted regions covered at 20X ≥98%) 3) No indication of sample contamination.

Sequence reads in the FASTQ files are aligned to the human reference genome (GRCh38.p14 from NCBI, February 2022) using BWA-mem 0.7.17 [2] to generate BAM files. BAM files are processed following the Rovaca v1.0.1 [3] and the GATK best practices (GATK v4.4.0, Genome Res. 2010;20:1297-303) [4] for SNV and small indels variant calling to generate VCF files [5, 6]. Mutect2 is used for calling lower level heteroplasmic SNV/INDEL in the mitochondrial genome. [7] 3bCNV is used for CNV calling based on DOC data however, its only available when sufficient number of samples from same sequencing methods are submitted. Due to the lack of sequencing data between exons, the resolution of CNV calls is minimum 3 consecutive exons and for most of the CNVs, exact breakpoints are not identifiable. MANTA is used for CNV calling based on paired-end information [8]. ExpansionHunter (v5.0.0) is used for repeat expansion variants[9]. MELT (v2.2.2) is used for calling mobile element insertion variants [10]. AutoMap (v1.2) is used for ROH detection from the VCF file [11].

Various quality control metrics such as Q30, mapping rate, PCR duplication rate, capture efficiency, total number of variants, het/hom, and ts/tv ratios are used to ensure the sequencing data is within an acceptable range

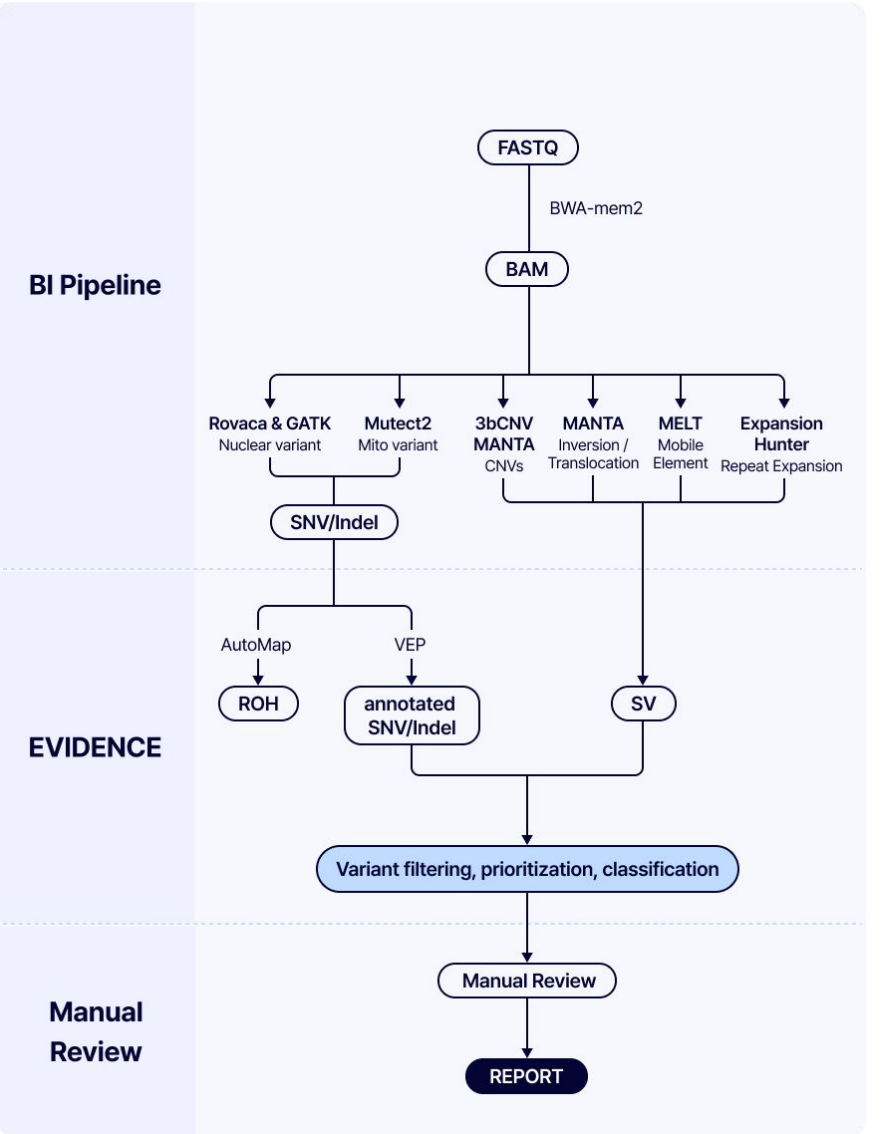



Figure 4. Schematics of 3B-INTERPRETER analysis workflow

# 3B-VARIANT

3B-VARIANT, also called variant specific test (VST), uses Sanger sequencing for genotyping a specific variant position in family members. Once a proband is reported with a variant by 3B-GENOME or 3B-EXOME, the presence of the same variant in proband's parents or other family members can be tested with 3B-VARIANT. The test provides a cost-effective method for determining whether the proband's variant is inherited or not, which is often crucial for evaluating its pathogenicity. Extending the test to other family members can also enable genetic counseling, expanding to other family members by either confirming the diagnosis in other affected members or informing potential disease risk.

Genomic DNA is extracted from whole blood, buccal swab or dried blood spot (DBS) samples, using QIAamp blood (QIAGEN, GmbH, Germany), AccuBuccal DNA Prep kit (AccuGene, Incheon, Korea), and AccuFAST DBS Prep Kit (AccuGene, Incheon, Korea), respectively. PCR primers are designed using Primer3 (v.0.4.0), [12, 13] and NCBI GenBank reference sequence. PCR amplification and Sanger sequencing are performed following the standard protocol using PCR Master Mix Kit (ThermoFisher Scientific, Waltham, MA, USA) and SeqStudio Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). The sequencing results are manually analyzed using Sequence Scanner Version 1.0 (Applied Biosystems, Foster City, CA, USA).

Each case is then comprehensively reviewed by our clinical team of physicians, geneticists and informaticists.

# Quality Metrics for Sanger Validation of Identified Variants

Even though NGS has settled down to be a robust technology for molecular diagnostic tests, because Sanger sequencing is oftentimes still considered as the gold standard in the field, variants identified by NGS have been subject to Sanger confirmation prior to being reported. This confirmation process results in delayed turnaround time and increased cost. Multiple groups, including 3billion, have investigated the needs of Sanger confirmation for NGS-based tests to uniformly report that Sanger confirmation is not necessary for variants with 'good' quality scores as long as sufficient validation and quality control measures are implemented [15, 16, 17]. 3billion has performed a thorough validation study to determine a conservative threshold using the variant quality score generated by Rovaca, GATK and variant allele frequency (VAF) to define 'good' variants that do not require Sanger confirmation. This reduced the number of variants requiring Sanger confirmation by more than 90%.

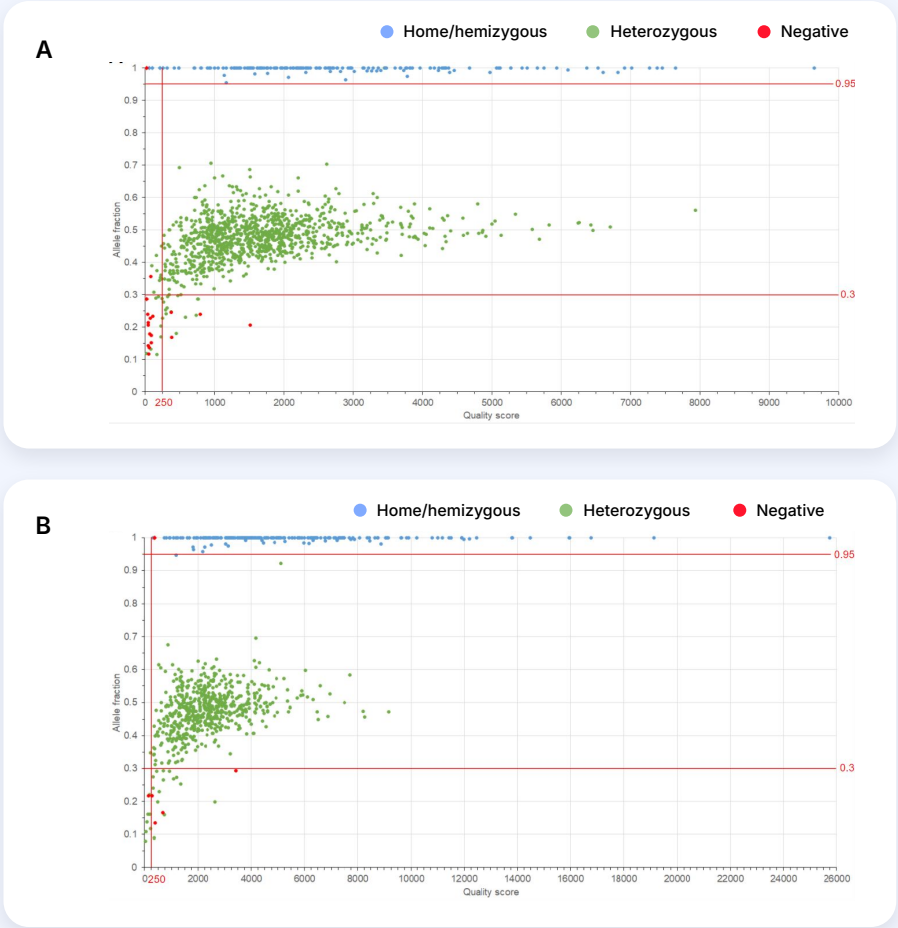



Figure 5. Variants are plotted by their quality score and VAF.

A. SNV, B. INDEL. Blue dots are variants called homozygous or hemizygous by WES and Sanger sequencing, green dots are variants called heterozygous by WES and Sanger sequencing and red dots are variants called as homozygous/hemizygous or heterozygous by WES but not confirmed by Sanger sequencing. Variants with (quality score> 250) and (VAF> 0.3 (heterozygous) or >0.95 (homozygous)) and (read depth>=10) were determined to be defined as 'good' variants without the need of Sanger confirmation.

# 3billion's Reports

## 3B-GENOME / EXOME / INTERPRETER report

3billion's NGS test report consists of 16 sections: 'Patient Information' which includes demographic information of the patient; 'Clinical information' which includes symptoms of the patient provided by the clinician; 'Result Summary' which includes information of clinically significant variants if exist (if not, it will includes comments for Negative result); 'Result Interpretation' which includes various evidences of the reported variant information; 'Additional findings' which includes variants that could not be reported as primary findings due to limited evidence of pathogenicity even though they may explain the patient's symptoms; 'Secondary findings' (if opted in); 'Resources' which includes description of databases used for analysis; 'References' which include a list of publications which was referenced for the variant interpretation purpose; 'Notes' which are for brief interpretation of 'Results summary' section and 'Variant Classification' section; 'Recommendations' which include both recommendation for the provider and limitation of the test; 'Methods' which include pipeline of the analysis and detailed QC value of the NGS test of the patient; 'Additional Note' which includes additional comment about the patient provided by the clinician; 'Disclaimer'; 'Accreditations and Certifications' which includes CAP License # and CLIA ID #; Director's Signature; 'Appendix. Requested gene(s) findings' which show cov20X value of suspected genes.

Variant information is mainly described in the Results, Interpretation and Secondary findings section as described below.

## 3B-VARIANT report

3B-VARIANT report consists of 4 sections: order information, result, methods, and references. Test result is described in the result section, consisting of 2 types of results; Detected and Not detected.

# 3B-GENOME / EXOME / INTERPRETER report

## 1. Results

Results can be positive, inconclusive, or negative. For positive and inconclusive reports, a variant table(s) is shown with the variant, gene and disease information as shown below.

**POSITIVE**

A heterozygous likely pathogenic variant was identified in *NIPBL*. *NIPBL* is associated with autosomal dominant 'Cornelia de Lange syndrome 1 (OMIM: 122470)'. As this variant has never been reported in other patients, clinical correlation is recommended. Parental testing is also recommended to check if the variant is de novo or inherited.

**Cornelia de Lange syndrome 1** (OMIM: 122470)

| Gene | Variant | | Classification |
|---|---|---|---|
| *NIPBL* | Genomic Position | 5-36985791-AGG-A (GRCh38) | Likely pathogenic |
| | cDNA | NM_133433.4:c.2612_2613del | |
| | Protein | NP_597677.2:p.Arg871ThrfsTer2 | |
| | Zygosity | Heterozygous | |
| | Inheritance | Unknown | |

Figure 6. An example of a positive test result

Positive reports are issued when the report contains only pathogenic or likely pathogenic variant(s) that fully explain(s) inheritance pattern of the disease. For example, a report with a likely pathogenic variant in an autosomal recessive disorder is reported as positive.

**INCONCLUSIVE**

A heterozygous variant of uncertain significance was identified in *PTPN11*. *PTPN11* is associated with autosomal dominant 'Noonan syndrome 1 (OMIM: 163950)'. Currently available evidence is insufficient to classify the variant as pathogenic or likely pathogenic. Clinical correlation may provide further evidence to reclassify the variant. Parental testing is also recommended to check if the variant is de novo or inherited.

**Noonan syndrome 1** (OMIM: 163950)

| Gene | Variant | | Classification |
|---|---|---|---|
| *PTPN11* | Genomic Position | 12-112453317-G-A (GRCh38) | VUS |
| | cDNA | NM_002834.5:c.455G>A | |
| | Protein | NP_002825.3:p.Arg152His | |
| | Zygosity | Heterozygous | |
| | Inheritance | Unknown | |

Figure 7. An example of an inconclusive test result

Inconclusive reports are issued when VUS variant is included in the report or reported variant(s) cannot explain the inheritance pattern of the reported disease. For example, a report with a pathogenic variant in an autosomal recessive disorder is reported as inconclusive.

**NEGATIVE**

No clinically significant variant relevant to the patient's phenotype as provided in the Human Phenotype Ontology (HPO) terms, additional memo and attached documents was identified.

Figure 8. An example of a negative test result

Negative reports are issued when clinically significant variant was not identified from disease that would fit the patient's phenotype.

## 2. Interpretation

As shown below in an example, the interpretation section provides detailed information of the variants being reported in the context of the ACMG guidelines: population data, predicted consequence and location of the variant, segregation data if family members were tested, computation and functional data from in silico prediction programs and literature, previous reports on the variant if available, disease association, Sanger validation results, and variant classification.

**A**

**RESULTS INTERPRETATION**

| ***NIPBL*** **NM_133433.4:c.2612_2613del (NP_597677.2:p.Arg871ThrfsTer2)** | |
|---|---|
| **Population Data** | The variant is not observed in the gnomAD v4.0.0 dataset. |
| **Predicted Consequence / Location** | Frameshift: predicted to result in a loss or disruption of normal protein function through nonsense-mediated decay (NMD) or protein truncation. Multiple pathogenic variants are reported downstream of the variant. |
| **Segregation Data** | None |
| **Computation and Functional Data** | None |
| **Previously Reported Variant Data** | None |
| **Disease Association** | Cornelia de Lange syndrome 1 (OMIM: 122470) |
| **Validation** | Not performed as the variant was considered high-quality |
| **Variant Classification** | Likely pathogenic |

**B**

**RESULTS INTERPRETATION**

| **NC_000016.10:g.(?_47641035)_(47641692_?)del (GRCh38)** |
|---|
| The homozygous deletion NC_000016.10:g.(?_47641035)_(47641692_?)del spans exon 15-16 of the PHKB (NM_000293.3 transcript). The variant is not observed in the gnomAD SVs v2.1.1 dataset. PHKB is subject to loss of function. Other pathogenic variants of the same consequence have been reported in this exon(s). Therefore, this variant was classified as likely pathogenic. |

Figure 9. An example of interpretation for A. a SNV and B. a SV (in this case CNV)

## 3. Additional finding

The additional finding section describes a list of variants that could not be reported as primary findings due to paucity of evidence for pathogenicity, even though there is possibility of explaining the patient's symptoms.

**ADDITIONAL FINDINGS**

No additional variants were identified, including variants of uncertain significance (VUSs) that could not be reported as primary findings due to limited evidence of pathogenicity, even though they may explain the patient's symptoms; pathogenic, likely pathogenic variants or VUSs that may partially explain the patient's symptoms, regardless of whether they fit the mode of inheritance; or variants associated with the family history provided by the healthcare provider, regardless of the patient's current symptoms.

Figure 10. An example of additional findings

## 4. Secondary finding (*if opted in*)

The secondary findings section describes the variant identified in one or more of the 84 genes that were selected by ACMG [14] as medically actionable and recommended to be reported if a pathogenic or likely pathogenic variant is found (details vary by gene). This section will be included only when the patient opts in to receive the information.

**SECONDARY FINDINGS**

Breast-ovarian cancer, familial 2 (OMIM: 612555) is an autosomal dominant, multifactorial disorder. Individuals with pathogenic variants in *BRCA2* (OMIM: 600185) have an increased risk for developing breast cancer and ovarian cancer (includes fallopian tube and primary peritoneal cancers) and other cancers such as prostate cancer, pancreatic cancer, and melanoma to a lesser extent. Genetic counseling and clinical management are warranted.

| **Breast-ovarian cancer, familial, 2** (OMIM: 612555) | | | |
|---|---|---|---|
| **Gene** | **Variant** | | **Classification** |
| *BRCA2* | **Genomic Position** | 13-32362596-A-T (GRCh38) | Pathogenic |
| | **cDNA** | cDNA: NM_000059.4:c.7879A>T | |
| | **Protein** | NP_000050.3:p.Ile2627Phe | |
| | **Zygosity** | Heterozygous | |
| | **Inheritance** | Unknown | |

Figure 11. An example of secondary findings

# 3B-VARIANT report

**Detected:**

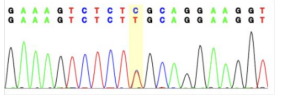
Detected result is designated when the variant previously identified by 3B-EXOME or 3B-GENOME is also found in the sample ordered for 3B-VARIANT.

**Not Detected:**

Not Detected result is designated when the sample does not carry the variant of interest.

**PRIMARY FINDING**

| Variant 1: NC_000023.11:g.25013650NGC[23] (GRCh38) | Detected |
|---|---|
| **Results** | Variant was observed |
| Gene | *ARX* |
| cDNA | NM_139058.3:c.298GCN[23] (p.Ala100[23]) |
| Zygosity | Not applicable/Heteroygous/Hemizygous/Homozygous |
| Disease [Inheritance Mode] | Developmental and epileptic encephalopathy 1 (OMIM: 308350) |
| Classification | Pathogenic |



NC_XXXX:g.XXX-XXXX

Chromatogram showing +/-10bp flanking region of the variant breakpoint (vertical line).

Figure 12. An example of 3B-VARIANT report

For more information, check sample reports at our Resources page.

# 3billion's State-of-the-Art Technology

3billion’s genomic data generation, interpretation and variant classification is a multistep process involving automated variant annotation, AI-machine learning prediction model, phenotype assessment and manual case level interpretation. Following variant interpretation guidelines provided by ACMG/AMP, our team have refined and modified individual criteria in order to provided comprehensive and consistent variant interpretation while maintaining the most up-to-date information and minimizing inter-laboratory discordancy.

# Exome Boosting

Whole exome sequencing technology is most widely used in the clinical setting as it is more easily accessible and cost-effective for physicians and patients. However, there are technical limitations on whole exome sequencing.

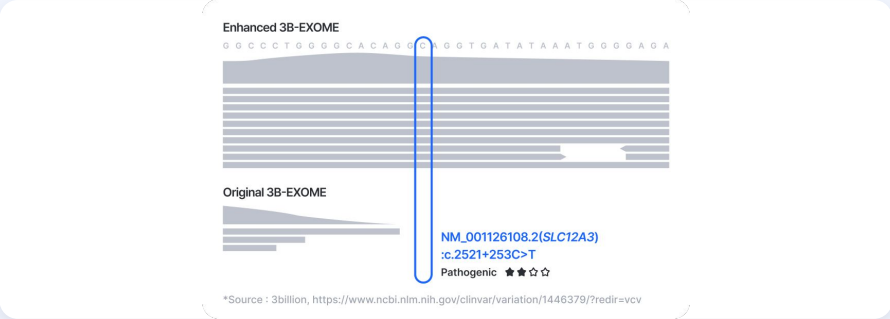
- The coverage of WES is not as uniform as WGS especially within low complexity region.
- Limited coverage of mitochondrial genomes.
- Introns with known pathogenic variants.
- Reduced sensitivity of smaller exonic copy-number-variants.

Although WGS solves most of the limitations, increase in cost prevents many patients from achieving diagnosis. To solve this problem, 3billion regulatory updates capture kits to boost previously uncaptured regions of interests.

## Selection of Boosted Region

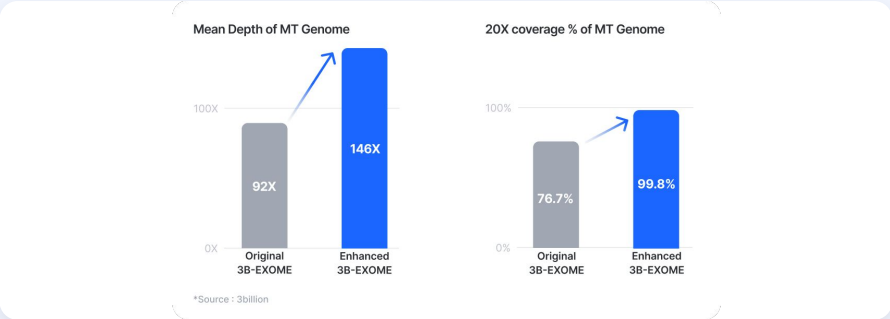
### 1.Non-coding disease-causing variant positions

The ~570 non-coding variant positions are captured and sequenced with sufficient coverage and we will be continuously updating for newly identified non-coding disease causing variant regions.



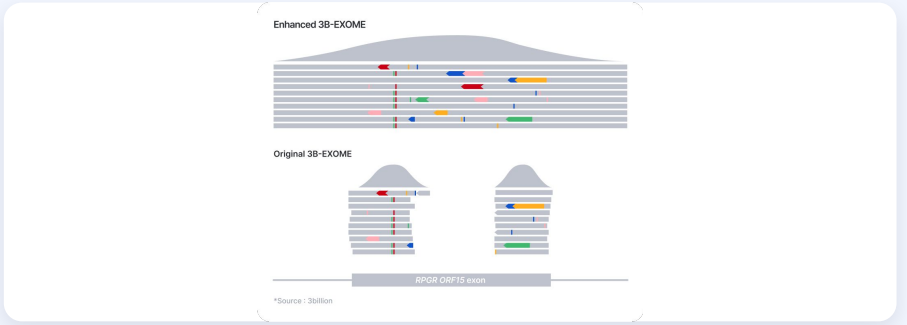

### 2. Mitochondrial DNA

The mean depth increased by ~70% and almost all mitochon -drial genome consistently have >20x coverage.



### 3. Difficult to capture exonic regions

e.g *RPGR* ORF15 exon, is a well-known exome coverage drop-out region despite many disease-causing variants being reported within. ORF15 exon is completely sequenced with sufficient coverage.




### 4. Intronic regions of *GLA*, and *RPE65*

*GLA* and *RPE65*, associated with Fabry disease and Leber congenital amaurosis, for which treatment options are available, are captured in all the intronic regions. We now can capture breakpoints of small CNVs within these regions.

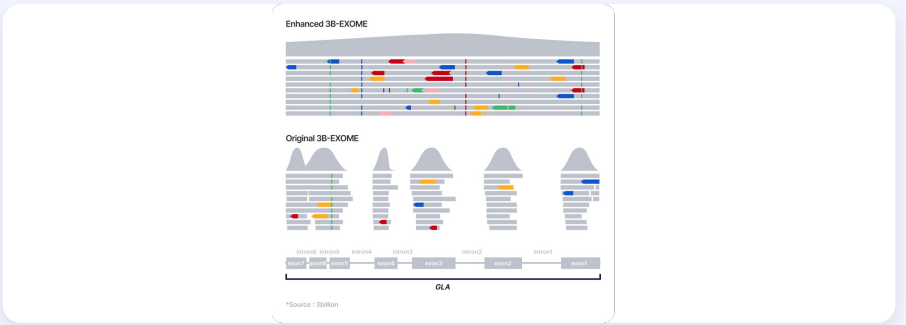



Figure 13. Example of Exome boosted region to include detection of a known intronic pathogenic variants.

# EVIDENCE:

## Automatic Variant Prioritization System

EVIDENCE is an automated variant prioritization system that has been developed to facilitate genomic sequencing analysis.

EVIDENCE is composed of 3 key modules:
1. variant annotation module with daily updated database
2. customized variant classification module
3. phenotype similarity scoring module

## 1. Variant annotation module with daily updated database

Annotating each variant with public and private (in-house) data is the first step of variant analysis as this collective annotation data is used as supporting evidence for the variant classification. As new information on genes, variants, and disorders become available everyday, it is important to update and integrate various databases such as ClinVar, HGMD (Human Gene Mutation Database) professional, OMIM (Online Mendelian Inheritance in Man), ENSEMBL Genes, NCBI Genes, HGNC (HUGO Gene Nomenclature Committee) PubMed, in-house database, etc as often as possible. The more information on each variant we can access, the more accurate molecular diagnosis we can make. Various databases are available at the variant level, gene level, and disease level. Insufficient or outdated information for variant interpretation can lead to an incorrect molecular diagnosis with incorrect variant classification. To minimize this risk, 3billion checks for any updates on each database every single day. The newer version of the updated database is downloaded and internally validated before it is applied to the variant analysis. See below Table 1 for the database list currently used at 3billion.

| Category | Database | Source | Version |
|---|---|---|---|
| **Sequence** | GRCh37/19<br>GRCh38/hg38 | https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/<br>https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40 | GRCh37.p13<br>GRCh38.p14 |
| **Population frequency** | gnomAD (variant and SV) | https://gnomad.broadinstitute.org/downloads (GRCh37)<br>https://gnomad.broadinstitute.org/downloads (GRCh38) | v2.1.1<br>v4.1.0 |
| **Gene** | HGNC | https://ftp.ebi.ac.uk/pub/databases/genenames/new/tsv/hgnc_complete_set.txt | Daily up-to-date |
| | NCBI gene | https://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz | Daily up-to-date |
| **Transcript** | RefSeq | https://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/105.20220307/GCF_000001405.25_GRCh37.p13/GCF_000001405.25_GRCh37.p13_genomic.gff.gz<br>https://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/110/GCF_000001405.40_GRCh38.p14/GCF_000001405.40_GRCh38.p14_genomic.gff.gz | GRCh37.p13<br>GRCh38.p14 |
| | Ensembl | https://ftp.ensembl.org/pub/grch37/release-87/gtf/homo_sapiens/Homo_sapiens.GRCh37.87.gtf.gz<br>https://ftp.ensembl.org/pub/release-109/gtf/homo_sapiens/Homo_sapiens.GRCh38.109.gtf.gz | GRCh37.87<br>GRCh38.109 |
| | GTEx | https://www.gtexportal.org/home/datasets | V8 |
| **Disease** | OMIM | https://www.omim.org/downloads | Daily up-to-date |
| | Orphanet | https://www.orpha.net/consor/cgi-bin/index.php | 2022.12 |
| | CGD | https://research.nhgri.nih.gov/CGD/download/txt/CGD.txt.gz | 2022.10 |
| | HPO | https://raw.githubusercontent.com/obophenotype/human-phenotype-ontology/master/hp.obo,<br>http://purl.obolibrary.org/obo/hp/hpoa/phenotype.hpoa | 2023.01 |
| | In-house database | | Daily up-to-date |
| **Variant** | ClinVar | https://ftp.ncbi.nlm.nih.gov/pub/clinvar/xml/weekly_release/ClinVarFullRelease_00-latest_weekly.xml.gz | Weekly up-to-date |
| | UniProt | https://www.uniprot.org/downloads | 2022.12 |
| | DGV | http://dgv.tcag.ca/dgv/docs/DGV.GS.March2016.50percent.GainLossSep.Final.hg19.gff3<br>http://dgv.tcag.ca/dgv/docs/DGV.GS.hg38.gff3 | 2016.05.13 |
| | HGMD | https://www.hgmd.cf.ac.uk/ | version 2022.4 |
| | In-house database | | Daily up-to-date |
| **Domain** | UniProt | https://www.uniprot.org/downloads | 2022.12 |
| **Prediction tool** | dbNSFP (REVEL, GERP++RS) | http://database.liulab.science/dbNSFP | v4.3a |
| | dbscSNV (ADA_score, RF_score) | http://www.liulab.science/dbscsnv.html | v1.1 |
| | Splice AI | https://github.com/Illumina/SpliceAI | v1.3.1 |
| | 3Cnet | In-house database | |
| | RepeatMasker | https://www.repeatmasker.org/RepeatMasker/ | 4.1.4 |
| | REVEL | https://sites.google.com/site/revelgenomics/ | May 3, 2021 |
| | GERP | https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1001025 | |
| **Scientific literature** | PubMed and Google Scholar | | |

Table 1. Database list

## 2. Customized variant classification module

The American College of Medical Genetics and Genomics (ACMG) and the American Molecular Pathology (AMP) have put together standards and guidelines for variant interpretation in 2015 initially [18]. These guidelines and any updates followed are commonly adopted by many diagnostic laboratories. However, it is also known that even when the same guidelines are used, a variant can be given different classifications by different laboratories due to condensed/vague descriptions of various rules in the guidelines [19, 20]. 3billion tried to scrutinize and customize each rule in the guidelines to make them more precise based on existing knowledge gathered from the public databases and the internal database. This effort was developed into the variant classification module of EVIDENCE.

Variants are classified as pathogenic (P), likely pathogenic (LP), variants of uncertain significance (VUS), likely benign (LB), or benign (B) based on the guidelines suggested by the American College of Medical Genetics and Genomics and Association for Molecular Pathology (ACMG/AMP). The ACMG/AMP guidelines have provided a framework for assessing the pathogenicity of genetic variants by considering a wide range of evidence. Various information such as variant type, predicted consequence, variant frequency, segregation, *in silico* prediction, and in vitro functional effect are integrated to determine the pathogenicity of each variant.

Nevertheless, the interpretation of genetic variants may result in discrepancies, leading to divergences between distinct testing facilities and even within a given laboratory, resulting in inconsistent classifications of the variants. 3billion has customized the guideline embodying each criteria with more specific rules and strengths so that at least within 3billion, variants are classified more consistently across different interpreters or timepoints.

This is described in more detail in Seo et al., 2020 [21].

### a) SNVs and INDELs

ACMG/AMP guidelines proposed 28 criteria that can be assessed when determining variant pathogenicity.

## 1) Pathogenic criteria

### PVS1

Null variant (nonsense, frameshift, canonical +/-1 or 2 splice sites, initiation codon, single or multi-exon deletion) in a gene where the loss of function (LOF) is a known mechanism of disease.

- PVS1 criteria have been modified with reference to two articles [22]
- Exception
  PVS1 could be claimed when the absence of gene expression or protein production is experimentally proven through methods such as RNA sequencing, RT-PCR for mRNA expression, etc.
- Start loss variant: an alternative start codon should not be present in a near downstream region as in-frame or in another transcript (alternate transcript). Our system monitors the presence of previously reported pathogenic variants upstream of the new potential start codon. Classification is upgraded or downgraded accordingly.

### PS1

Same Amino acid change as the previously established pathogenic variant, regardless of the nucleotide change.

- Variant type: missense variants.
- Definition of the established pathogenic variant: variants with P/LP determined by the ACMG guidelines' criteria, referenced from the reputable variant database (Table 1). Furthermore, medical geneticists perform a manual review of all previously documented pathogenic variants in order to verify their consistent pathogenicity.

### PS2

De novo (maternity and paternity confirmed) variant, with matching highly specific symptoms from the disease and with no previous family history of the disease.

- Variant type: all types
- PS2 can be claimed for a previously reported de novo variant, with matching, highly specific symptoms. The variants reported as de novo in literature or in the in-house database have been manually curated by medical geneticists. Strength can be increased for recurrent de novo variants.

**PVS1 evaluation**

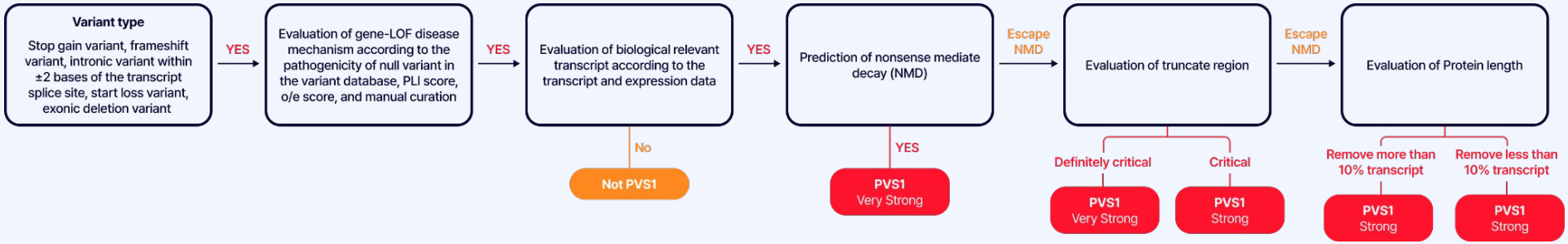



Figure 14. Schematic of PVS1 evaluation

### PS3
Well-established in vitro or in vivo functional studies supporting a damaging effect on the gene or gene product.

- Variant type: all types
- PS3 can be applied if there is solid functional study data on the variant. Our Medical geneticists manually review the functional study data from external resources to determine if it was performed robustly.

### PS4
Variant prevalence in the affected individuals is significantly higher than in the controls.

- Variant type: all types
- For exceedingly rare variants, a moderate level of evidence may be used: 1) insufficient case-control studies may be available to obtain statistical significance; 2) the variants for the identical phenotype are found in multiple unrelated patients, but not in the general population. The strength would be upgraded depending on the number of reports of variants in unrelated families [23].

### PM1
Variant located in a mutational hot spot and/or a critical and well-established functional domain (e.g., the active site of an enzyme) without benign variation.

- Variant type: missense variants and in-frame variants
- Domain and variant databases are utilized to evaluate “well-established functional domains without benign variants”.
  A mutational hot spot is determined by the distribution of pathogenic variants extracted from reputable databases.

### PM2
Variant is absent from controls (or at extremely low frequency if recessive; see Table 6) in the Exome Sequencing Project, 1000 Genomes, or ExAC.

- Variant type: all variants
- The population frequency database evaluates the PM2, BA1, BS1, and BS2.
- The disease-specific allele frequency threshold (dMAF) is used to estimate the rarity of variants based on prevalence and penetrance [24]. If the prevalence of the disease is unknown, the prevalence is assumed to be 1/1,000,000.

**Dominant disease**

$$\text{dMAF} = \frac{Prevalence(d)}{2 * Penetrance(d)}$$

**Recessive disease**

$$\text{dMAF} = \sqrt{\frac{Prevalence(d)}{Prevalence(d)}}$$

### PM3
Variant detected in trans with another Pathogenic variant for recessive disorders. Parental testing is required to determine a phase.

- Variant type: all types
- PM3 can be claimed for a previously reported variant in the trans phase with highly specific, matching symptoms. Phases of the variants from the literature and the in-house database are reviewed and updated manually by medical geneticists. The strength would be adjusted for recurrent occurrences.
- Markedly, variants found within 200 base pairs are assessed for phase status by each read, indicating that the interpretation of variants includes potential phase results.

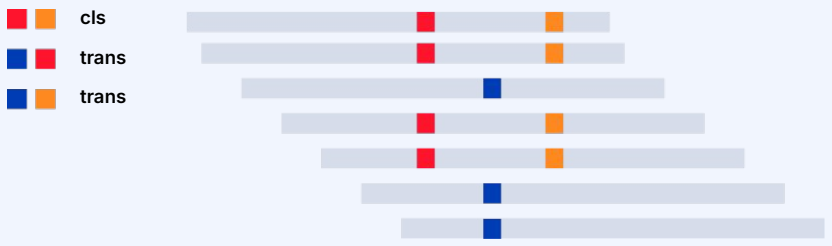



Figure 15. Phasing analysis based on reads

**PM4**

Changes in protein length due to in-frame deletions/insertions in a non-repeat region or stop-loss variants.

- Variant type: in-frame deletion/insertions, stop loss variants
- The repeat region is determined by RepeatMasker.
- To avoid double-counting the same evidence, PM4 will not be claimed for variants already issued with PVS1.

**PM5**

Novel missense changes in amino acid residues where an alternative missense change has been previously reported to be pathogenic.

- Variant type: missense variants
- Definition of the established pathogenic variant: variants with P/LP determined by the ACMG guidelines' criteria, referenced from the reputable variant database (Table 1). In addition, medical geneticists review every previously reported pathogenic variant to confirm the established pathogenicity.

**PM6**

Assumed de novo, but without any confirmation of paternity and maternity.

- Variant type: all variants
- PM6 can be claimed for variants previously reported as assumed de novo variants if highly specific symptoms are matched. The assumed de novo variants in literature or in the in-house database will also be updated by medical geneticists.

**PP1**

Co-segregation of a causative gene and disease in multiple affected family members.

- Variant type: all types
- PP1 can be claimed for co-segregated variants with a previously reported disease in multiple affected family members. The updated variants would be manually curated by medical geneticists. The strength can be increased by the number of meiosis and affected relatives.

**PP2**

Missense variants in a gene where missense variants are observed as a common disease mechanism.

- Variant type: Missense variants

**PP3**

Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.).

- Variant type: missense variants, splice region variants outside +/-2-bp of a splicing junction, synonymous variants, and intron variants
- The functional effect of missense variants is predicted using programs such as REVEL [25], and 3Cnet [26].
- Splice region variants outside +/-2-bp of a splicing junction, synonymous variants, and intron variants are analyzed to predict the functional effect using ADA, and RF scores [28].

### PP4

Patient's phenotype or family history is highly specific for a disease with a single genetic etiology.

- Variant type: all types
- PP4 requires a similarity score >5 between the patient's phenotype and disease symptoms. Attention must be paid to applying this rule, as the symptoms provided may not be sufficient.

### PP5

Variants reported as pathogenic in reputable sources, but the evidence might not be available for laboratories to perform an independent evaluation.

- Variant type: all types
- In 2018, ACMG/AMP made a recommendation to discontinue the use of PP5, due to the risk of possible double-counting [29]. However, external databases such as ClinVar are still actively used as important evidence for variant classification. To avoid the risk of missing such important evidence, 3billion applies the PP5/BP6 rules based on the level of evidence, after extensive review and evaluation of the variant by medical geneticists.

## 2) Benign criteria

### BA1

Allele frequency is above 5% in the Exome Sequencing Project, 1000 Genomes, or ExAC.

- Variant type: all types
- Allele frequency is >0.05 in any general continental population dataset of at least 2,000 observed alleles. Non-continental populations (Jewish and Finnish groups) were excluded.
- A BA1 exception list has also been integrated [30].

### BS1

Allele frequency is greater than expected for a disorder

- Variant type: all types
- Applied to variants with an allele frequency 10-fold or more in PM2 threshold.

### BS2

Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder, with full penetrance expected at an early age.

- Variant type: all types
- BS2 is applied depending on the inheritance pattern. Diseases with adult-onset and/or incomplete penetrance were excluded.

### BS3

Well-established in vitro or in vivo functional studies showing no damaging effects on protein function or splicing.

- Variant type: all types
- Functional studies would be validated and proven by solid reproducibility in well-established clinical laboratory settings. Medical geneticists review the functional study data related to the variants.

**BS4**
Lack of segregation in affected family members.

- Variant type: all types
- BS4 can be claimed when disease variants are not segregated in the previously reported multiple affected family members. The updated variants are manually reviewed by medical geneticists.

**BP1**
Missense variant in a gene where premature termination variant is an expected mechanism of pathogenicity.

- Variant type: missense variants

**BP2**
Observed in trans with a pathogenic variant for a fully penetrant dominant gene/disorder or observed in cis with a pathogenic variant in any inheritance pattern.

- Variant type: all types
- Variants located within 200 base pairs are evaluated for phase status read by read. BP2 can be accepted as a label when separate variants are confirmed to be located in the cis phase.

**BP3**
In-frame deletions/insertions in a repetitive region without known function.

- Variant type: in-frame deletion/insertion variants
- The repeat region is selected using the RepeatMasker.

**BP4**
No expected impact on gene or gene product (conservation, evolutionary, splicing impact, etc.) measured by computational tools.

- Variant type: missense variant, splice region variant outside +/-2-bp of a splicing junction, synonymous variant, and intron variant
- The functional effect of missense variants is predicted by programs such as REVEL [25] and 3Cnet [26].
- Splice region variants outside +/-2-bp of a splicing junction, synonymous variants, and intron variants are analyzed to predict the functional effect using ADA, and RF scores [28].

**BP5**
Variants found with a disease that has an alternate molecular basis.

- Not applicable

**BP6**
Variants reported as benign in reputable sources, but the evidence might not be available for laboratories to perform an independent evaluation.

- Variant type: all types
- refer to comments on PP5

**BP7**
A synonymous (silent) variant predicted to have no impact on the splice consensus sequence or the creation of a new splicing site by splicing prediction algorithms, AND the nucleotide is not highly conserved.

- Variant type: synonymous variants
- ADA, RF score, and GERP++RS are used to predict the functional effects of synonymous variants.

(The criteria strength could be upgraded or downgraded via a manual review of our expert panel)

## 3) Rules for Combining Criteria to classify variants

| | |
|---|---|
| **Pathogenic** | **1. Very Strong (PVS1) *AND***<br>a. ≥ 1 Strong (PS1 - PS4) *OR*<br>b. ≥ 2 Moderate (PM1-PM6) *OR*<br>c. 1 Moderate (PM1-PM6) and 1 Supporting (PP1-PP5) *OR*<br>d. ≥ 2 Supporting (PP1-PP5)<br>**≥ 2 Strong (PS1-PS4) *OR***<br>**1. Strong (PS1-PS4) *AND***<br>a. ≥ 3 Moderate (PM1-PM6) *OR*<br>b. 2 Moderate (PM1-PM6) *AND* ≥ 2 Supporting (PP1-PP5) *OR*<br>c. 1 Moderate (PM1-PM6) *AND* ≥ 4 Supporting (PP1-PP5) |
| **Likely Pathogenic** | **1. Very Strong (PVS1) AND 1 Moderate (PM1-PM6) *OR***<br>**1. Strong (PS1-PS4) *AND* 1-2 Moderate (PM1-PM6) *OR***<br>**1. Strong (PS1-PS4) *AND* ≥ 2 Supporting (PP1-PP5) *OR***<br>**≥ 3 Moderate (PM1-PM6) *OR***<br>**2 Moderate (PM1-PM6) *AND* ≥ 2 Supporting (PP1-PP5) *OR***<br>**1 Moderate (PM1-PM6) *AND* ≥ 4 Supporting (PP1-PP5)** |
| **Benign** | **1. Stand-Alone (BA1) *OR***<br>**≥ 2 Strong (BS1-BS4)** |
| **Likely Benign** | **1. Strong (BS1-BS4) and 1 Supporting (BP1-BP7) *OR***<br>**≥ 2 Supporting (BP1-BP7)** |

** Variants should be classified as Uncertain Significance if other unmet or benign and pathogenic criteria are contradictory.

The 2015 ACMG/AMP guidelines marked all variants with conflicting evidence as VUSs. It would be reasonable if the level of evidence for pathogenicity and strength is comparable. However, the level of VUS can differ depending on the number and strength of criteria claimed to support pathogenicity. Notably, for SNV and small indel, a Bayesian framework is used to quantify the variant pathogenicity and make a final decision to determine accuracy by overcoming the limitations of the 2015 ACMG/AMP guidelines [31]. 3billion exploits the original guidelines along with the Bayesian scores and professional judgment for accuracy and validity in analyzing variants.

### b) CNVs

ACMG/AMP guidelines proposed a semi-quantitative point-based scoring metric for CNV classification when determining variant pathogenicity. Separate scoring criteria have been developed for copy-number-loss and copy-number-gain and are interpreted using 5 different sections [32].

#### Section 1. Genomic content evaluation

Section 1 evaluates the genomic content in the affected CNV area. Based on reputable databases (Table 1), each CNV is checked if it contains any protein-coding regions, protomers, enhancers, or other regulatory regions. CNVs only containing non-coding/non-regulatory regions (UTR, intron, pseudogene) are more likely to be benign than pathogenic.

#### Section 2. Gene dosage evaluation

Section 2 evaluates individual genes that are inside the affected CNV region and determines whether the genes are known to be haploinsufficient or triplosensitive from reputable databases. Tools that predict haploinsufficiency or triplosensitivity are also used to support their pathogenicity. If the breakpoints are located inside the genes of interest and expected to result in loss of function is also vetted.

### Section 3. Gene number evaluation

CNV is evaluated based on the number of genes within. CNVs that encompass a larger number of genes are expected to be more pathogenic than smaller ones.

### Section 4. Evaluation of literature and public databases.

Section 4 compares a CNV to previously reported CNVs in the literature and reputable databases that overlap. Evidence such as the number of previously reported cases, reported segregation data, phenotype similarities alongside how unique they are, and, if possible, the prevalence of reported CNVs are all used to determine the pathogenicity.

### Section 5. Evaluation of Patient Being Studied

In the final section, proband specific case-level information is evaluated. Segregation information and specificity of patient phenotypes are used to determine the pathogenicity of a given CNV.

## C) SVs

Although current ACMG guidelines lacks incorporation of copy number neutral structural variants such as inversion and translocation, 3billion have incorporated our own asserted interpretation guidelines for interpretation of copy-number neutral structural variants.

Structural variants are analyzed following similar guidelines that are used to classify single nucleotide variants and small insertions/deletions. Once structural variants are identified, interpretations are separated based on the identity: Inversion , breakend and insertion .

Interpretation of structural variants are only available for whole genome.

### Interpretation of Inversions

Inversions are evaluated based on assumption that genes located in the breakpoints will results in loss-of-function of the the gene. Based on the exact breakpoint of the gene, predictions are made following the PVS1 guidelines provided by ACMG. Population frequency, in-house frequency, location of breakpoints in gene are taken into consideration in interpretation of inversion.. If available, segregation information is also taken into account.

### Interpretation of Breakends

Breakends are evaluated based on the assumption that they are part of balanced chromosomal translocation. Population frequency, in-house frequency, location of breakpoints in gene are taken into consideration in interpretation of breakend. Loss-of-function predictions are made following the PVS1 guidelines provided by ACMG. If available, segregation information is also taken into account during interpretation of breakends.

### Interpretation of Insertions

Insertions are evaluated based on assumption that insertion will results in loss-of-function of gene they are inserted. This includes mobile element insertion such as LINE or SINE. Location of insertion are taken into account during interpretation as exonic insertion will results in full loss-of-function of genes while insertions in introns are less likely to results in loss-of-function. Population frequency, in-house frequency, are also taken into consideration in interpretation of breakend. If available, segregation information is evaluated in interpretation of the insertion.

### 3. Symptoms similarity scoring module

Ultimately, the variant interpretation is carried on in the context of the patient's phenotype. EVIDENCE uses a 'symptom similarity scoring' module that scores how well the symptoms between the patient's phenotype and disease phenotype match. The symptom of each patient is converted to the corresponding standardized Human Phenotype Ontology (HPO) term, which in turn is used to compare to the HPO terms for each of the ~7,000 rare genetic disorders. The similarity between the patient's symptoms and the reported phenotypes of a certain disease is evaluated and presented as a similarity score ranging from 0 to 10. Empirical data suggests that a gene with 3billion's symptom similarity score >=6 has a significantly higher chance of being the diagnosis.

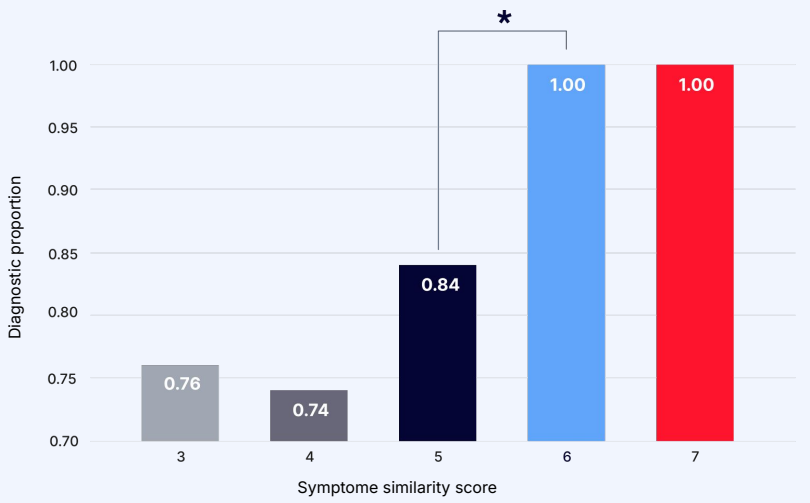



Figure 16. Probability of a gene being the diagnosis according to its symptom similarity score. There is a significant jump between the score of 5 and 6 (*$P < 0.05$).

# AI-based Variant Interpretation Algorithms

# 3Cnet:

## Pathogenicity Prediction Tool for Variants

Missense variants are common, corresponding to 83% of nonsynonymous variants in the population, and many genetic disorders are caused by missense variants. According to dbNSFP, the possible number of missense variants within the human genome is 82,755,468. However, less than 2,000,000 missense variants are known to be pathogenic or benign with strong confidence, leaving the pathogenicity of most of the variants unknown. The number is nearly infinite for other types of variants, such as insertions and deletions. Therefore, various attempts have been made to develop artificial intelligence (AI)-based diagnostics using the rapidly increasing volume of genomic data.

3billion developed 3Cnet, which employs deep neural networks to predict pathogenicity based on the protein sequence, evolutionary constraints and physicochemical features of the variant [26]. This AI model can identify disease-causing variants of patients 3 times more sensitively. For the interpretation of variants, 3Cnet is only used to evaluate missense variants following the ACMG guideline. With its recent update to version 2, its capability of predicting the pathogenicity covers 99.99% of variants including start-loss, stop-gain, stop-loss, in-frame deletion, frameshift, in-frame insertion, delins, duplication, 5' extension, and 3' extension.

3Cnet makes use of 3 different genomic databases to train pathogenicity of variants effectively, and to avoid overfitting of the model network. 1) Clinical data which consists of pathogenic and benign variants from ClinVar database, 2) Common variants observed in the general human population from gnomAD database, 3) Conservation data, which refers to the simulated variants that we generated based on evolutionary conservation using UniRef database. The network architecture of 3Cnet is composed of two modules, feature extractor and pathogenicity classifier.

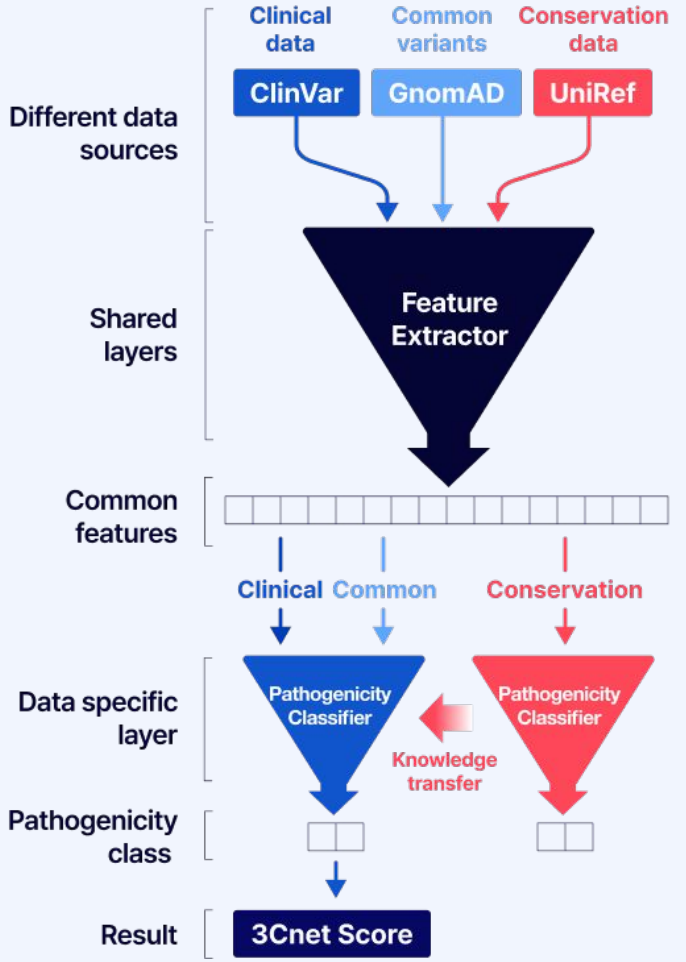



Figure 17. The deep learning network of 3Cnet. Multi-task learning using various genomic databases to avoid overfitting of the network to a small number of clinical data. conservation data reflects the evolutionary constraints given on genes.

3Cnet can classify pathogenic and benign variants the most accurately compared to other methods including REVEL, VEST4, SIFT, Polyphen2, PrimateAI, CADD, FATHMM, and DANN. Also, it can discover disease-causing variants in patient genomes with 3 times greater sensitivity than currently available tools, thereby improving diagnosis rates.

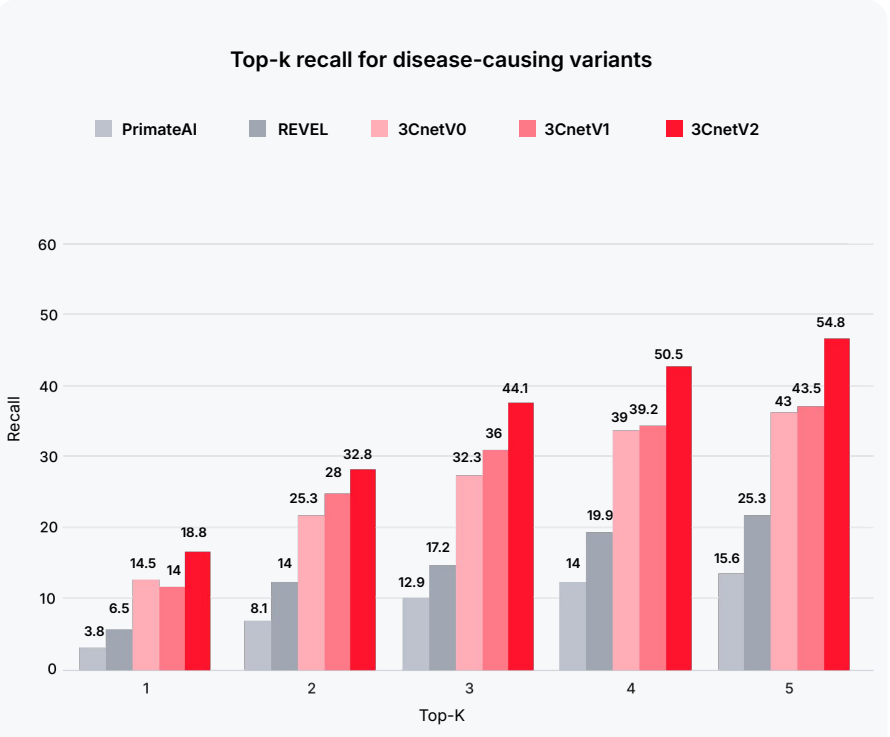


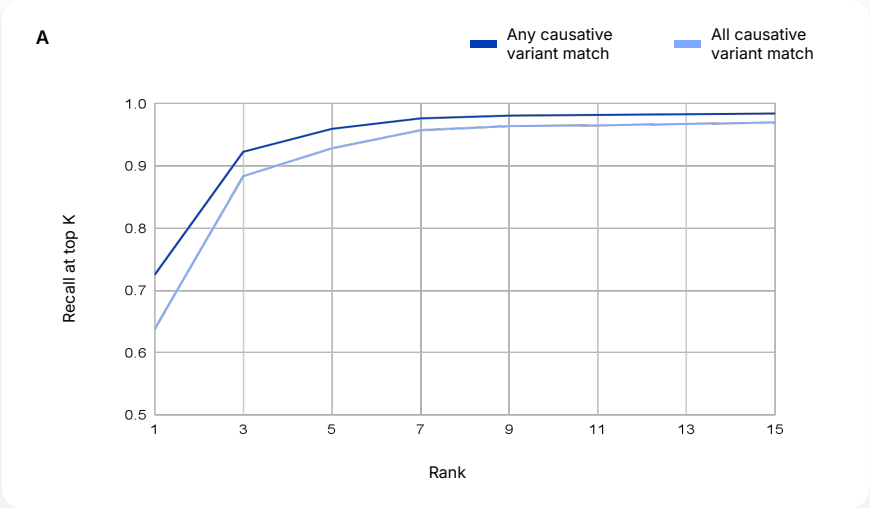

**Figure 18. Prediction sensitivity of 3Cnet to discover disease-causing variants.** The top-k recall rate implies the probability of determining the true disease-causing variant(s) among the top ranked variants using prediction scores.

# 3ASC:
## Variant Recommendation System

While NGS genomic tests have become routine, analyzing and interpreting the vast amounts of data they produce remains a significant challenge, consuming considerable time and resources. Existing variant prioritization tools aim to expedite this process but often fall short due to limited capabilities and incomplete integration of crucial data. Recognizing the need for a more robust solution, we developed 3ASC [33]—a cutting-edge, data-driven machine learning model that revolutionizes variant interpretation.

3ASC leverages up to 41 features to predict the likelihood of each variant being disease-causing. It integrates patient symptoms, disease inheritance patterns, number of variants, population allele frequencies, annotated 28 ACMG criteria, and more to provide a holistic approach to variant prioritization. Trained on genomic data from over 20,000 patients using advanced deep learning techniques like attention-gated multiple instance learning, 3ASC excels in prioritizing SNVs, INDELs, and CNVs.

Demonstrating superior performance compared to other models such as LIRICAL and Exomiser, 3ASC successfully identifies disease-causing variants within the top five candidates 97% of the time. With its high efficiency and accuracy, 3ASC empowers our medical geneticists to interpret exome and genome results more effectively, ultimately accelerating diagnosis and improving patient outcomes.



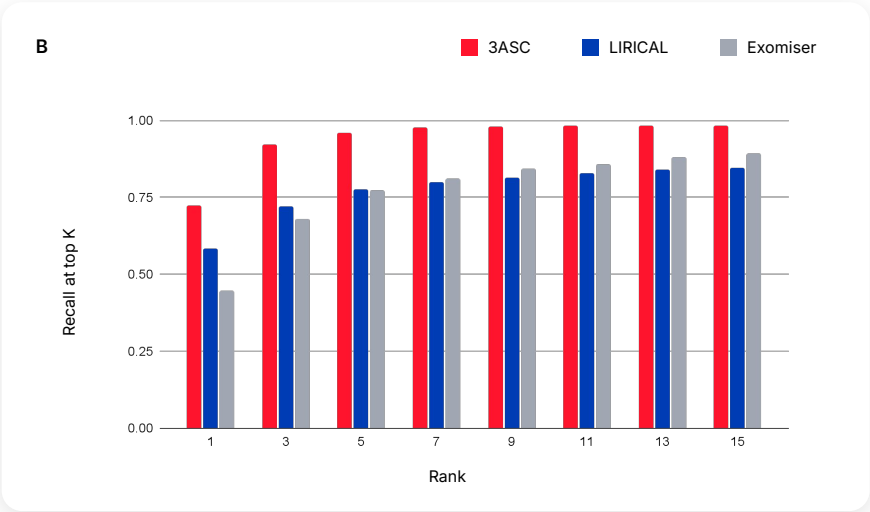
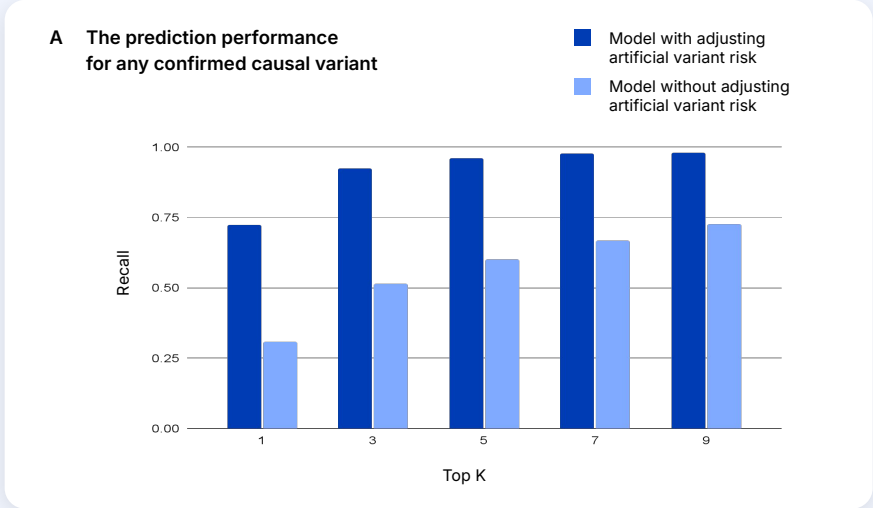




Figure 19. A. Model performance based on single match (any causative variant) and full match (all causative variants match) B. Comparison of recall of Exomiser, LIRICAL, and proposed model by gene-level match



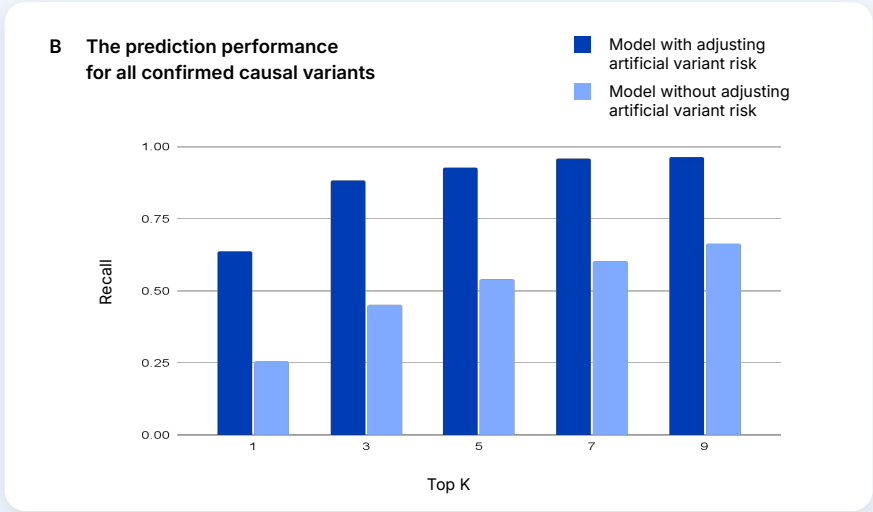




Figure 20. For the prediction of any confirmed causal variant, Figure 20-A showed that the model with adjusting the artificial variant outperformed than the model without leveraging this risk Also, Figure 20-B consistently showed the model with adjusting the artificial variant outperformed for the prediction of all confirmed causal variants.

# Automated reanalysis system

It is reported that approximately 30% of exome-negative patients receive diagnosis through reanalysis service (interval: 2–3 years), with a considerable increase of 10–15% in the overall diagnostic rate [34, 35, 36, 37]. It also indicates an over five- and three-fold increase in the diagnostic rate compared with the chromosomal microarray technique and all genetic tests in clinical practice. Diagnosis through reanalysis reduces costs, as patients can avoid unnecessary redundant diagnostic testing. Moreover, patients and family members have a better chance of being involved in making the right treatment decisions.

3billion performs reanalysis of the NGS sequencing data on all patients who did not receive a clear molecular diagnosis for their chief complaints. Patients have the option to opt-out from receiving the reanalysis. An updated report is generated at no cost if a clinically significant variant is identified or a previously reported variant is reclassified through the reanalysis.

3billion's reanalysis is performed through EVIDENCE using the latest supporting evidence downloaded by the automated database updating system. To estimate the molecular diagnostic rate from reanalysis, we tracked 1,064 patients with a neurodevelopmental delay between April, 2018-Feb, 2022 who were referred as part of a research project.

31 patients received a new diagnosis through reanalysis. The time interval between the initial analysis and the reanalysis that yielded a new diagnosis was 1.2 ± 0.9 years (from a minimum of 1 month to a maximum of 3.3 years[38]. Most of the diagnosis from reanalysis were due to novel genes discovered in between the initial analysis and reanalysis.

*** Available only for cases with reanalysis consent provided during 3billion portal ordering.**

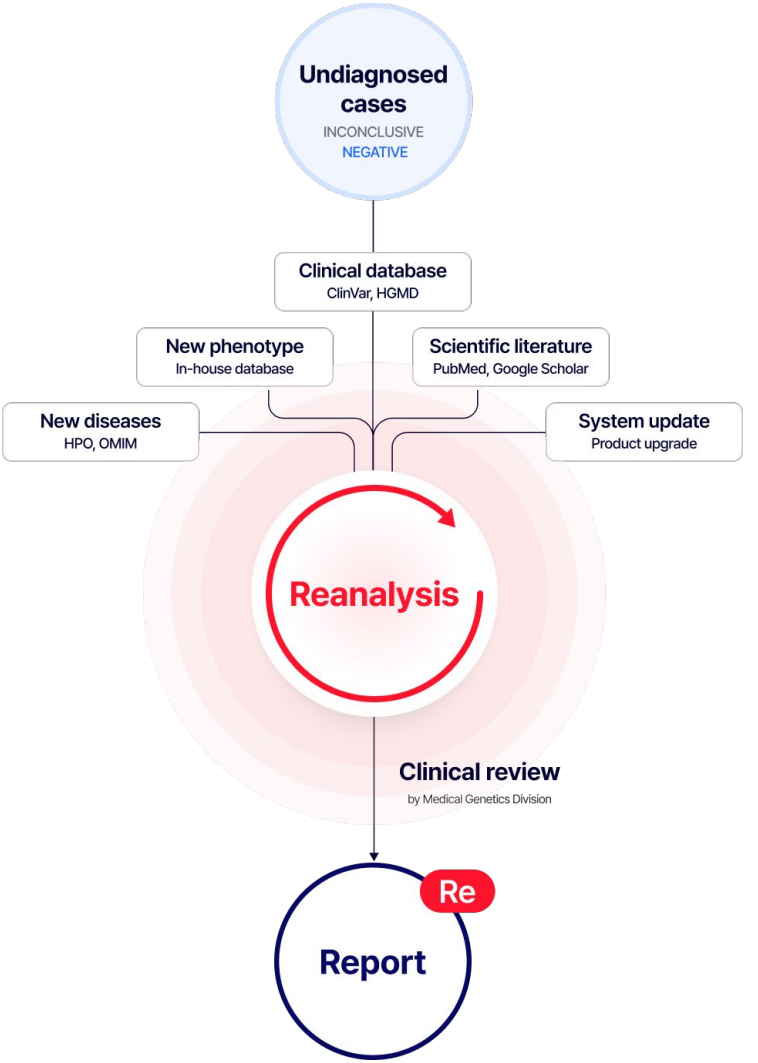



Figure 21. Reanalysis system.
For patients with no clinically significant variants, EVIDENCE is run with the most recent annotation information. All variants reclassified as pathogenic or likely pathogenic in genes that could fit the patient's phenotype are reviewed by 3billion's medical geneticists.

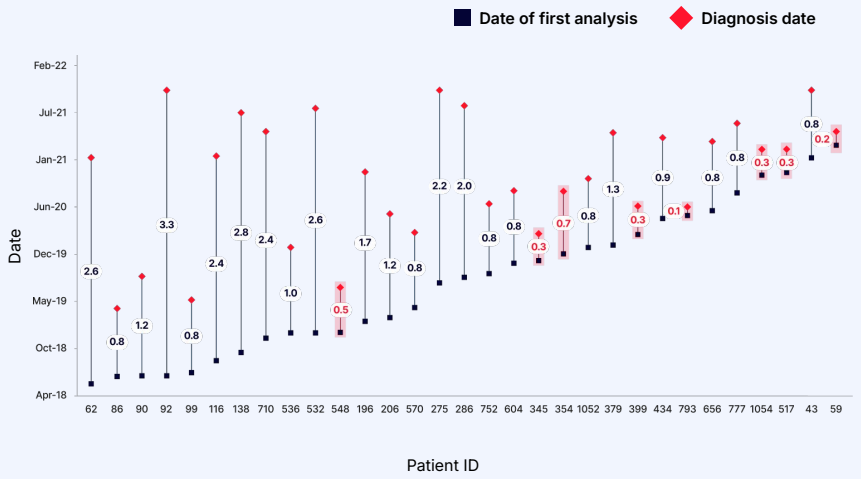



**Figure 22. The time interval between the first analysis (blue dot) and the reanalysis yielded a molecular diagnosis (red dot) for 31 patients who received a diagnosis through reanalysis.** X-axis: individual patients. Y-axis: Analysis date. The number shown for each patient: time interval in years

3billion's reanalysis process involves re-annotation of all variants that were identified from initial analysis and selects the variant based on patient's symptom similarity to disease, variant's previous bayesian score to current bayesian score, disease inheritance, OMIM disease updates, in-house data, and any new bioinformatics annotation including *in silico* predictions.

The selected variants are then presented to medical geneticists for an review. Medical geneticists will then review the data and decide if variant needs to be reported or not. Once decision is made a reanalysis report will be generated and sent to the ordering physician.

## Reanalysis Case

**NEGATIVE report**

**Feb. 2021 :**
No clinically significant
SNVs/INDELs were identified.

**INCONCLUSIVE report**
NM_015001.3:c.5806C>T
(p.Arg1936Ter)

**Mar. 2021 :**
**Am J Hum Genet. 2021:108(3):502-516**
SPEN haploinsufficiency causes a Neurodevelopmental disorder overlapping proximal 1p36 deletion
syndrome with an episiqnature of X chromosomes in females

**POSITIVE report**

**May. 2021 :**
**New Disease update**
Radio-Tartaglia syndrome
(OMIM: 619312) – SPEN gene

Figure 23. Case example of a patient's timeline from test order to diagnosis through reanalysis.

# Conclusions

Over 790+ medical institutions across 70+ countries have used our service to diagnose 90,000+ suspected rare genetic disease patients.

The overall diagnostic rate of all tested patients is approximately 31%. The diagnostic rate varies among different disease categories.

The accumulated genomic and clinical data are invaluable sources to make the more accurate diagnosis achievable, for which we do research collaborations with physicians and investigators worldwide. 3billion is also committed to contribute in discovering drug targets using AI and genomic data, which paves the path to a new drug for various rare diseases yet immedicable.

3billion is always here to help patients suffering from an undiagnosed rare genetic disorder until their diagnostic odyssey ends. We vision that no undiagnosed patient is left behind without access to genetic testing. Join us to work together to explore the world of rare genetic disorders.

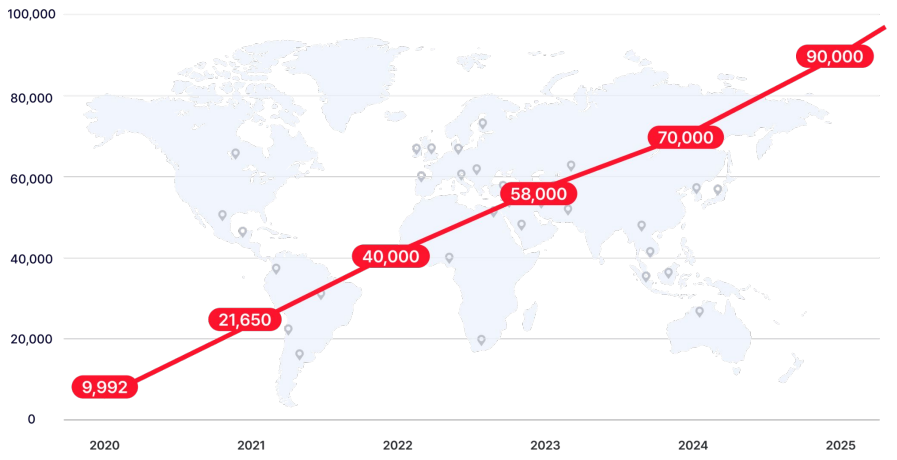



Figure 24. Accumulated number of patient between 2020–2024

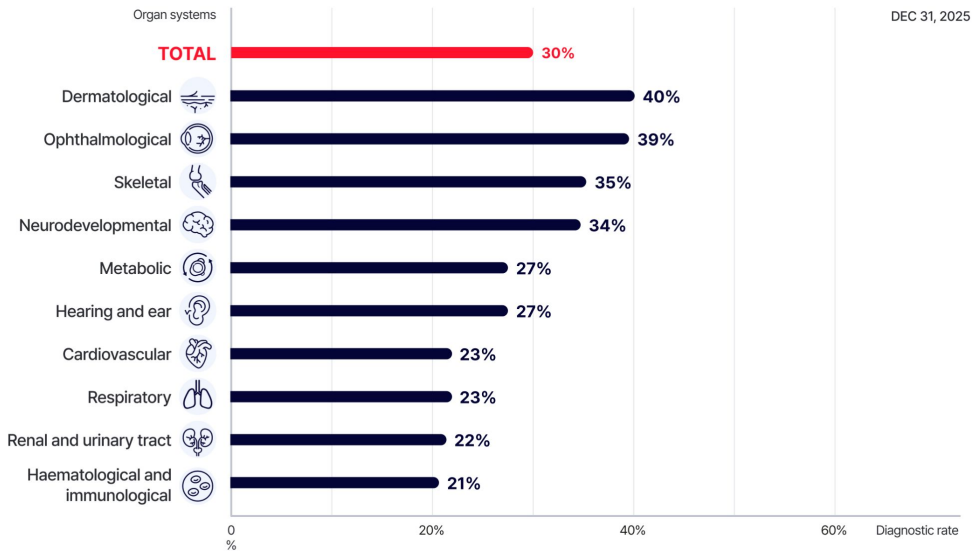



Figure 25. Diagnostic rate for different disease categories

# References


1. BCL2FASTQ: bcl2fastq Conversion Software

2. BWA-mem arXiv: 1303.3997 [q-bio.GN]

3. Rovaca v1.0.1 (doi.org/10.1101/2025.10.19.677660)

4. GATK Best Practices Workflows – GATK (broadinstitute.org)

5. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-1303.

6. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-498.

7. Benjamin, David, et al. "Calling somatic SNVs and indels with Mutect2." BioRxiv (2019): 861054.

8. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016 Apr 15;32(8):1220-2. doi: 10.1093/bioinformatics/btv710. Epub 2015 Dec 8. PMID: 26647377.

9. Dolzhenko E, Deshpande V, Schlesinger F, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics. 2019;35(22):4754-4756.

10. Gardner EJ, Lam VK, Harris DN, et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. 2017;27(11):1916-1929.

11. Quinodoz M, Peter VG, Bedoni N, et al. AutoMap is a high performance homozygosity mapping tool using next-generation sequencing data. Nat Commun. 2021;12(1):518.

12. Untergasser A, Cutcutache I, Koressaar T, et al. Primer3--new capabilities and interfaces. Nucleic Acids Res. 2012;40(15):e115.

13. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics. 2007;23(10):1289-1291.

14. Lee, K., Abul-Husn, N.S., Amendola, L.M. et al. ACMG SF v3.3 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med. 2025 un 23;27(8):101454 PMID 40568962.

15. Strom SP, Lee H, Das K, et al. Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. Genet Med. 2014;16(7):510-515.

16.De Cario R, Kura A, Suraci S, et al. Sanger Validation of High-Throughput Sequencing in Genetic Diagnosis: Still the Best Practice?. Front Genet. 2020;11:592588. Published 2020 Dec 2.

17. Arteche-López A, Ávila-Fernández A, Romero R, et al. Sanger sequencing is no longer always necessary based on a single-center validation of 1109 NGS variants in 825 clinical exomes. Sci Rep. 2021;11(1):5697.

18. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-424. doi:10.1038/gim.2015.30

19. Harrison SM, Dolinsky JS, Knight Johnson AE, et al. Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. Genet Med. 2017;19(10):1096-1104.

20. Tavtigian SV, Greenblatt MS, Harrison SM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet Med. 2018;20(9):1054-1060.

21. Seo GH, Kim T, Choi IH, et al. Diagnostic yield and clinical utility of whole exome sequencing using an automated variant prioritization system, EVIDENCE. Clin Genet. 2020;98(6):562-570.

22. Abou Tayoun AN, Pesaran T, DiStefano MT, et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Hum Mutat. 2018;39(11):1517-1524.

23. Harrison SM, Biesecker LG, Rehm HL. Overview of Specifications to the ACMG/AMP Variant Interpretation Guidelines. Curr Protoc Hum Genet. 2019;103(1):e93.

24. Shah N, Hou YC, Yu HC, et al. Identification of Misclassified ClinVar Variants via Disease Population Prevalence. Am J Hum Genet. 2018;102(4):609-619. doi:10.1016/j.ajhg.2018.02.019

25. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99(4):877-885.

26. Won DG, Kim DW, Woo J, Lee K. 3Cnet: pathogenicity prediction of human variants using multitask learning with evolutionary constraints [published online ahead of print, 2021 Jul 16]. Bioinformatics. 2021;37(24):4626-4634.

27. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176(3):535-548.e24.

28. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014;42(22):13534-13544.

29. Biesecker LG, Harrison SM, ClinGen Sequence Variant Interpretation Working Group. The ACMG/AMP reputable source criteria for the interpretation of sequence variants. Genet Med. 2018;20(12):1687-1688.

30. Ghosh R, Harrison SM, Rehm HL, Plon SE, Biesecker LG, ClinGen Sequence Variant Interpretation Working Group. Updated recommendation for the benign stand-alone ACMG/AMP criterion. Hum Mutat. 2018;39(11):1525-1530.

31. Tavtigian SV, Greenblatt MS, Harrison SM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet Med. 2018;20(9):1054-1060.

32. Riggs ER, Andersen EF, Cherry AM, et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Genet Med. 2020;22(2):245-257.

33. Kim, H. H., Woo, J., Kim, D. W., Lee, J., Seo, G. H., Lee, H., & Lee, K. (2022). Disease-causing variant recommendation system for clinical genome interpretation with adjusted scores for artefactual variants. bioRxiv, 2022-10.

34. Fung JLF, Yu MHC, Huang S, et al. A three-year follow-up study evaluating clinical utility of exome sequencing and diagnostic potential of reanalysis. NPJ Genom Med. 2020;5(1):37.

35. Machini K, Ceyhan-Birsoy O, Azzariti DR, et al. Analyzing and Reanalyzing the Genome: Findings from the MedSeq Project. Am J Hum Genet. 2019;105(1):177-188.

36. Liu P, Meng L, Normand EA, et al. Reanalysis of Clinical Exome Sequencing Data. N Engl J Med. 2019;380(25):2478-2480.

37. Costain G, Jobling R, Walker S, et al. Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing. Eur J Hum Genet. 2018;26(5):740-744.

38. Seo GH, Lee H, Lee J, et al. Diagnostic performance of automated, streamlined, daily updated exome analysis in patients with neurodevelopmental delay. Mol Med. 2022;28(1):38.

**Web.** 3billion.io

**Order.** portal.3billion.io

**Email.** support@3billion.io

3billion