BMC Medical Genomics

SOFTWARE Open Access

Genome analysis and knowledge-driven variant interpretation with TGex

Dvir Dahary[1*], Yaron Golan[1], Yaron Mazor[1], Ofer Zelig[1], Ruth Barshir[2], Michal Twik[2], Tsippi Iny Stein[2], Guy Rosner[3,4], Revital Kariv[3,4], Fei Chen[5], Qiang Zhang[5], Yiping Shen[5,6,7], Marilyn Safran[2], Doron Lancet[2*] and Simon Fishilevich[2*]

Abstract

Background: The clinical genetics revolution ushers in great opportunities, accompanied by significant challenges. The fundamental mission in clinical genetics is to analyze genomes, and to identify the most relevant genetic variations underlying a patient's phenotypes and symptoms. The adoption of Whole Genome Sequencing requires novel capacities for interpretation of non-coding variants.

Results: We present TGex, the Translational Genomics expert, a novel genome variation analysis and interpretation platform, with remarkable exome analysis capacities and a pioneering approach of non-coding variants interpretation. TGex's main strength is combining state-of-the-art variant filtering with knowledge-driven analysis made possible by VarElect, our highly effective gene-phenotype interpretation tool. VarElect leverages the widely used GeneCards knowledgebase, which integrates information from > 150 automatically-mined data sources. Access to such a comprehensive data compendium also facilitates TGex's broad variant annotation, supporting evidence exploration, and decision making. TGex has an interactive, user-friendly, and easy adaptive interface, ACMG compliance, and an automated reporting system. Beyond comprehensive whole exome sequence capabilities, TGex encompasses innovative non-coding variants interpretation, towards the goal of maximal exploitation of whole genome sequence analyses in the clinical genetics practice. This is enabled by GeneCards' recently developed GeneHancer, a novel integrative and fully annotated database of human enhancers and promoters. Examining use-cases from a variety of TGex users world-wide, we demonstrate its high diagnostic yields (42% for single exome and 50% for trios in 1500 rare genetic disease cases) and critical actionable genetic findings. The platform's support for integration with EHR and LIMS through dedicated APIs facilitates automated retrieval of patient data for TGex's customizable reporting engine, establishing a rapid and cost-effective workflow for an entire range of clinical genetic testing, including rare disorders, cancer predisposition, tumor biopsies and health screening.

Conclusions: TGex is an innovative tool for the annotation, analysis and prioritization of coding and non-coding genomic variants. It provides access to an extensive knowledgebase of genomic annotations, with intuitive and flexible configuration options, allows quick adaptation, and addresses various workflow requirements. It thus simplifies and accelerates variant interpretation in clinical genetics workflows, with remarkable diagnostic yield, as exemplified in the described use cases.
TGex is available at http://tgex.genecards.org/

Keywords: Next generation sequencing analysis, Clinical variant interpretation and classification, Exome sequencing, Whole genome sequencing, Non-coding variants, Biomedical knowledgebase, Rare genetic diseases, Hamartomatous polyposis

* Correspondence: dvird@lifemapsc.com; doron.lancet@weizmann.ac.il; simon.fishilevich@weizmann.ac.il
[1]Clinical Genetics, LifeMap Sciences Inc., Marshfield, MA 02050, USA
[2]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel
Full list of author information is available at the end of the article

Background

Clinical genetics has progressed remarkably in the last decade, moving rapidly from genotyping selected mutations to whole exome sequencing (WES) and whole genome sequencing (WGS) [1–3]. Improvements in technology and analysis capabilities, accompanied by reduced costs, have revolutionized genomics, enabling one to pinpoint relevant genetic variations within millions of variants in sequenced patients. These advances have extraordinary impact on medical care, clinical diagnostics of rare diseases, discovery of novel pathogenic variants and gene-disease relationships, prenatal testing, genetic counseling, prediction of cancer predisposition, pharmacogenomics and personalized medicine [4–7].

The fundamental mission of a clinical genetics platform is to analyze thousands to millions of genetic variants, and to identify the relevant, typically one or two, genetic variations most likely to underlie the patient's phenotypes and symptoms. The first applications of scaled clinical exome sequencing applied to undiagnosed patients with suspected genetic conditions yielded a molecular diagnosis rate of ~ 25% [8, 9]. More recently, higher yields were described, with only a handful exceeding the 50% barrier [10, 11]; typically the reports are within the modest range of 25–40% [12–14]. Each technological and informatics enhancement offers an opportunity to improve the diagnostic yield, necessitating optimal variant interpretation as a key avenue to pursue.

Clinical adoption of WGS faces many challenges, including cost, speed of delivery and expert time [15], ambiguities and errors in variant calling and annotation [15–17], undiscovered variant- and gene-disease associations, incomplete views of disease associations within databases [15, 18], genetic and phenotypic heterogeneity [15, 18], and the difficulties posed by incidental findings [19]. For optimal diagnosis rates, clinical genetics analysis requires a knowledge-driven analysis platform, based on a comprehensive and regularly updated knowledgebase, and complying with guidelines for reporting recommendations [5, 20, 21].

Systematic re-analysis of un-solved exomes using up-to-date databases was shown to improve the diagnostic yield [15, 22]. The inclusion of newly discovered variant- and gene-disease associations is a key factor in maximizing the diagnosis rate. However, the wealth of relevant biological information, extremely valuable for that purpose, is typically scattered in numerous databases and tools encompassing genomics, bioinformatics, systems biology and systems medicine. Moreover, browsing and extracting the most relevant pieces of data and reaching comprehensive genetic diagnosis poses an overwhelming challenge.

Our widely used GeneCards Suite [23] provides a comprehensive solution. It constitutes a searchable, integrated biomedical knowledgebase, containing comprehensive information on all human genes and diseases. It includes GeneCards, the human gene database, with consolidated gene-centric data from over 150 sources, including genomic, transcriptomic, proteomic, genetic, clinical, and functional information. It also encompasses MalaCards [24], the human disease companion database which integrates more than 60 sources. This knowledgebase represents an extensive network of annotations and mutual relationships, together with the infrastructure needed for rapid biological interpretation of clinical genetics data. The recent augmentation of the knowledgebase to include an extensive collection of functional non-coding regions (non-coding RNA (ncRNA) genes, enhancers and promoters) provides solid grounds for the analysis of typically un-explored out-of-exome variants in WGS [25, 26].

GeneCards is fortified by wide ranging search capabilities, allowing users to enter any Boolean expression with disease-relevant keywords in order to identify the most relevant genes. To cater to variant disease interpretation, the Suite provides VarElect [27], a leading phenotype-based gene prioritization tool [28]. Gene-based prioritization uses broad information to identify and rank likely damaged genes associated with one or more phenotypes, as opposed to simply identifying potentially damaging variants, facilitating the interpretation of novel variants of known disease-genes [29]. The strength of VarElect lies in its capacity to perform automated GeneCards searches on a long list of Next Generation Sequencing (NGS) candidate variant-containing genes and output a scored, prioritized gene list according to disease phenotype and symptom relationships, using the comprehensive GeneCards information. VarElect not only scores and ranks the genes, but also provides detailed evidence of the associations across sections of molecular and genetic data, which is critical for reviewing results and selecting relevant genes and candidate variants. These features of VarElect are indispensable for prioritization in analyses of the millions of variants detected by WGS.

Variant interpretation tools are evolving from simple command-line-based programs and expert excel-sheet-based reviews to interactive, web-based decision support frameworks. In such platforms, variant and gene prioritization are only one component of a dynamic, multifactorial approach to discovery and diagnosis [29]. In this paper, we describe our recently established TGex (Translational Genomics expert), the GeneCards Suite knowledge-driven clinical genetics analysis platform. TGex combines VarElect's strength with comprehensive variant annotation and filtering capabilities, within a consolidated user interface that supports browsing, viewing, filtering and interpretation interactively, facilitating review and examination by the genetic analyst. The reporting system of TGex leverages the capabilities of VarElect and the vast amount of structured data available in GeneCards to automatically generate full and comprehensive clinical reports. TGex effectively enables biomedical professionals and scientists, without any prerequisite of bioinformatics skills, to perform genome analysis, all the way from raw patient genetic data in VCF (Variant Call Format) files to detailed reports. TGex's key innovation and strength is the combination of a comprehensive biomedical knowledgebase with broad variant annotation and gene-phenotype prioritization, and a powerful, interactive, user friendly, and adaptable interface, allowing evidence exploration, decision making, and automatic reporting.

Implementation

The general workflow of genetic labs handling sequencing-based genetic tests typically starts from processing and annotating variant files (usually VCF, including all of the variant calls of a certain sample), followed by clinical genetic analysis, and ends with generating a report summarizing the relevant findings. TGex is a clinical genetics analysis platform, providing an end-to-end solution for genetic labs as illustrated in Fig. 1. TGex supports virtually all of the VCF file formats generated by the variety of sequencing machines and primary analysis pipelines found in genetic labs and clinical genetics centers. In addition, TGex accepts patient metadata, sample information (details in Additional file 1: Fig. S1), and clinical details for incorporation in its reporting system. After analysis and interpretation, TGex outputs a report file (PDF or Word) together with a detailed variant annotation file (Excel).

TGex annotation process

Once a VCF file is uploaded as part of a new analysis in TGex (Fig. 1), it undergoes initial processing, including: (i) ensuring the inclusion of the required fields (validating the VCF format specifications; adding missing VCF headers, e.g. "DP", "DP4"); (ii) cleaning irrelevant information (e.g. filtering out non-canonical chromosomes; extracting only raw INFO fields if the VCF file is already annotated); (iii) filtering according to relevant genomic regions (using a selected BED file, Exome by default). Following the processing of the VCF file, TGex launches its annotation engine to enrich the VCF with specific information regarding each variant. The first step in the annotation is the prediction of the functional effect with SnpEff [30]. This is followed by applying SnpSift [31] and BCFtools [32] to extract detailed information from dozens of data sources as summarized in Table 1, as well as using proprietary tools developed by LifeMap Sciences (LMS) to split multi-allelic variants and merge identical variants on different transcripts according to the ACMG interpretation guidelines. The next stage is the clinical interpretation of each variant according to the ACMG/AMP 2015 guidelines [51] using all of the available data, including the allele frequency in control populations, clinical significance as provided by specified databases and prediction of the effect of the variant on protein function and transcript integrity using a variety of well-established prediction and scoring tools (detailed in Table 1). Next, TGex uses the data accumulated in the user account (or in a group of accounts sharing information, if applicable) to add in-house information per variant, including in-house allele frequency and user interpretation on variant and gene levels. Finally, additional custom annotations may also be added, including local population-specific frequency data and previous variant interpretations.

Gene-phenotype interpretation

For gene-phenotype prioritization, TGex leverages VarElect, the GeneCards Suite gene phenotyper [27]. VarElect is fully integrated within the TGex analysis screen and reporting system, via the VarElect application programming interface (API). The API input includes user-defined free-text keywords submitted to the GeneCards Knowledgebase search engine (this might include but is not limited to disease names or symptoms in any nomenclature (e.g. HPO [52], UMLS [53] terms)), along with a list of variant-containing genes from the TGex analysis screen. The API output is embedded within the analysis screen, including the gene-keyword score, and the "MiniCards" evidence showing the context of the hits. The "MiniCards" are automatically incorporated in TGex reports, and include extensive gene-phenotype evidence, with dedicated hyperlinks to source databases.

Versions and data updates

Each report generated in TGex includes documentation of the specific version of the knowledgebase and its annotation databases, which enables tracking, traceability and reproducibility. TGex and the GeneCards knowledgebase are frequently updated with the newest version of dozens of relevant data sources. Knowledgebase updates ensure that the analysis is performed using up-to-date biological knowledge, and often include new annotation sources and new system features. Having a frequently and regularly updated knowledgebase is the basis for our planned reanalysis feature. The genetic and clinical information of each case is stored throughout the lifetime of the TGex account, enabling data querying and case re-analysis using updated knowledgebase versions. This will enable automatic case re-analysis which will trigger sending alerts for outstanding novel findings.

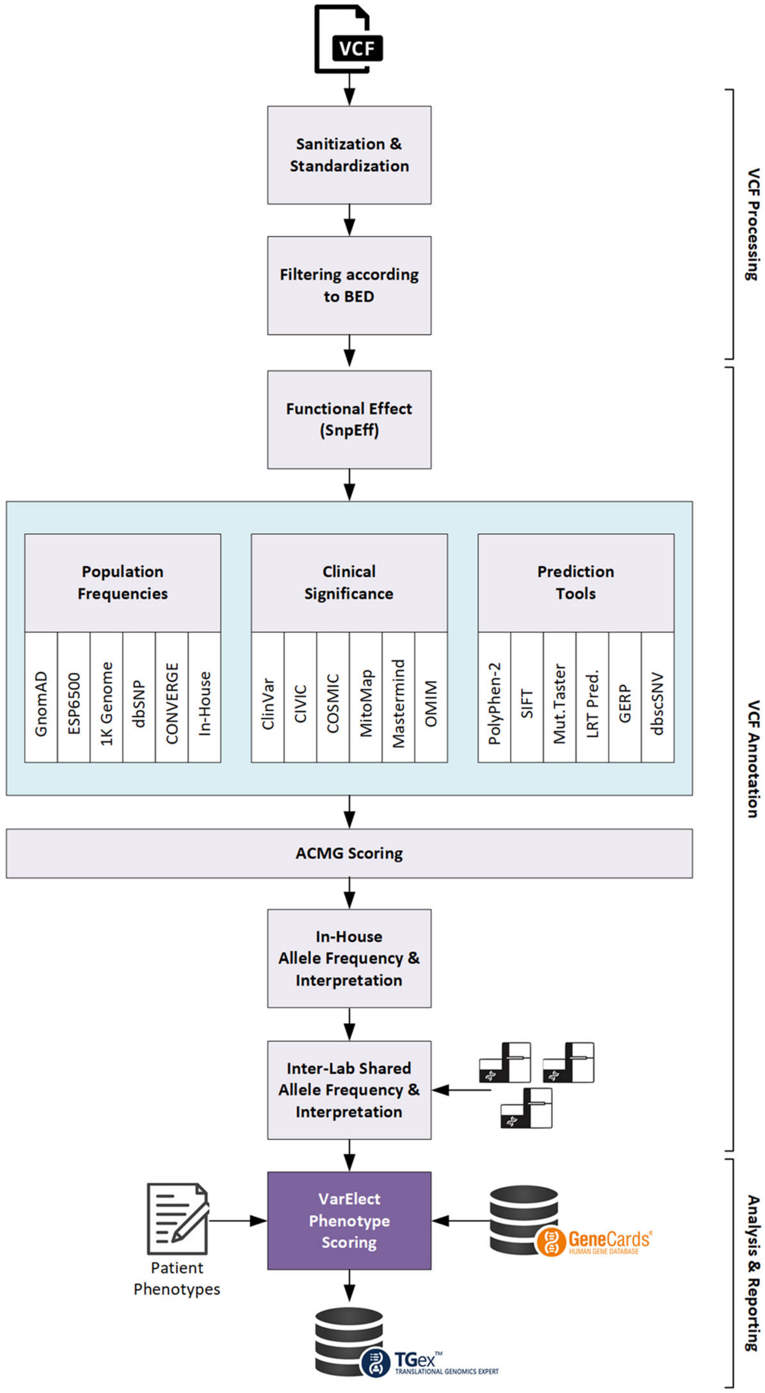


Fig. 1 Architecture of variant interpretation in TGex

Automation and APIs

TGex supports customizable, template-based reports with multiple export formats, including Excel, PDF, Word, HTML and JSON. TGex also supports JSON based exports of report data to external reporting engines, which are implemented on-site. The downloading of reports is also available via APIs, enabling laboratory information management systems (LIMS) and electronic health record (EHR) systems to automatically access reports from TGex. The TGex API allows integrators to optimally control the interaction with TGex within broader use-case contexts, including: (i) Integration with primary and secondary analysis pipelines, allowing automated upload and annotation of VCF files; and (ii)

Table 1 Summary of annotation databases and tools used in TGex

Data Source	Category	Reference
SnpEff	Functional Effect	[30]
ExAC (including GnomAD)	Frequency	[33]
ESP6500	Frequency	[34]
1000 Genomes Project	Frequency	[35]
dbSNP	Frequency	[36]
CONVERGE	Frequency	[37]
ClinVar	Evidence and clinical significance	[38]
CiVIC	Evidence and clinical significance	[39]
COSMIC	Evidence and clinical significance	[40]
MitoMap	Evidence and clinical significance	[41]
Mastermind	Evidence and clinical significance	[42]
OMIM	Evidence and clinical significance	[43]
PolyPhen-2	Effect and Prediction	[44]
SIFT	Effect and Prediction	[45]
MutationTaster	Effect and Prediction	[46]
LRT Prediction	Effect and Prediction	[47]
GERP	Effect and Prediction	[48]
dbscSNV	Effect and Prediction	[49]
RepeatMasker	Genomic repeats	[50]

Integration with LIMS or EHR systems, by enabling the creation of automated analyses of patient clinical information, and the streamlining of reports from TGex to the LIMS/EHR. TGex also supports fully automated analyses, enabling sophisticated screening protocols to be implemented easily by private and hospital labs. These may include pharmacogenomics, cancer and carrier screening, and newborn screening, among others.

Data protection compliances

TGex is HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation) compliant.

Software implementation

TGex is and can be deployed on the cloud (Azure, AliCloud, Amazon and others) or on-premises behind an organization's firewall. Today, two public cloud-based solutions are offered:

(1) In the Microsoft Azure East US server farm, serving all territories but China, and

(2) In the 21ViaNet Azure Server Farm in Shanghai, serving Chinese users.

The platform consists of four main components:

TGex Web server – Based on ASP.NET and utilizing an MS SQL Server and an Elastic Search server. The TGex server component serves both the TGex web client and other API-based automation clients.

TGex Web Client – A web-based client written in Angular, a popular JavaScript framework. The web client provides the user interface for management of samples, analyses and report lifecycles.

TGex Annotation server – Based on .NET, this server manages the annotation of variant data (in VCF or TSV formats). It includes sanitization, validation and annotation (see below). The TGex annotation server is optimized to rapidly annotate VCF files in under 5 min per whole Exome using proprietary acceleration technologies, and is scalable for large installations.

TGex Reporting server – This service manages report templates for customers and generates patient reports on demand or automatically. Importantly, this server allows each lab to build its own customized reports, which may include any information from the VCF, its annotations, data from the GeneCards knowledgebase, and/or customer proprietary data.

Minimum requirements for using the TGex Web Client are:

- A modern browser (Chrome, FireFox, Safari or Edge)
- An Intel i5, i7 or i9 7th generation or newer processor
- At least 4GB of RAM
- An internet or intranet connection of at least 10Mbit

Results

TGex is a novel patient-driven web platform for management of clinical genetic tests. It includes annotation, filtering, analysis and interpretation of clinical genetics data. TGex serves as a holistic solution for clinical genetics workflow integration, including management, analysis and reporting of genetic tests, starting from uploading VCF files and going all the way to report generation.

Clinical genetics workflow with TGex

In this section we describe the clinical genetics workflow within TGex, focusing on identification of rare germline genetic variants, one of the variety of types of analyses that can be performed using TGex (Table 2). The platform is comprised of three main components. The first is the management module, represented by the TGex dashboard, where the user can easily review and access current analyses, or create a new case by uploading a new VCF together with all relevant patient details

(Additional file 1: Fig. S1). The second and main module is the analysis component. Following the creation of a new case in TGex, the uploaded VCF file goes through the annotation process described in the Implementation section above. The result is a fully annotated table of variants to be analyzed via the main analysis screen (Fig. 2). The third module is the reporting engine, which collects all relevant patient information, the clinical details, the samples, and most importantly the selected variants and accompanying genetic, biomedical and molecular information from the GeneCards knowledgebase, and consolidates them into a fully automated, comprehensive and customizable report (Additional File 2).

The annotation pipeline

The general flow of clinical genetic tests starts from getting the full list of genetic variants, whether by genotyping specific positions in the genome or, more commonly nowadays, by sequencing large regions in the human genome (e.g. gene panels or Exomes). Recently, even full genomes using NGS and additional technologies are provided. Handling thousands to millions of variants with the aim of identifying the single or perhaps a few mutations that are causal for specific symptoms, clinical conditions, or genetic disorders requires four main steps – annotation, filtering, analysis and interpretation.

Once uploaded, the VCF goes through the annotation pipeline of TGex (Fig. 1), which enriches each variant with dozens of pieces of information using various available data sources and software tools (see also Table 1 and Implementation). These annotations are grouped into 7 categories:

1. **Genomic and Genetic Data**: genomic location, affected gene, reference allele, alternative allele, amino acid change and genotype (zygosity).
2. **Variant Calling Quality and Reliability (Q&R)**: combined quality score, absolute read counts, and the percentage of reads showing the alternative allele.
3. **Evidence**: the VarElect score for the association between the gene and the phenotype terms, the number of matched phenotypes, matching COSMIC [40], CiVIC [39] and ClinVar [38] entries, and publications associated with the variant.
4. **Effect and Prediction**: the effect on the gene, the severity of the effect (combining several prediction algorithms), and the calculated ACMG pathogenicity assignment.
5. **Frequency**: the allele frequency observed in the following control datasets: 1000 Genomes [35], ESP6500 [34], ExAC (including GnomAD) [33] and CONVERGE [37].
6. **In-House**: allele frequency within all of the cases in the account, pointers to previously selected matching variants and genes in all analyzed cases and their interpretations.
7. **Inter-Lab sharing**: allele frequency within all cases in accounts sharing data with this account, pointers to previously selected matching variants and genes in all analyzed cases in the sharing group and their annotations.

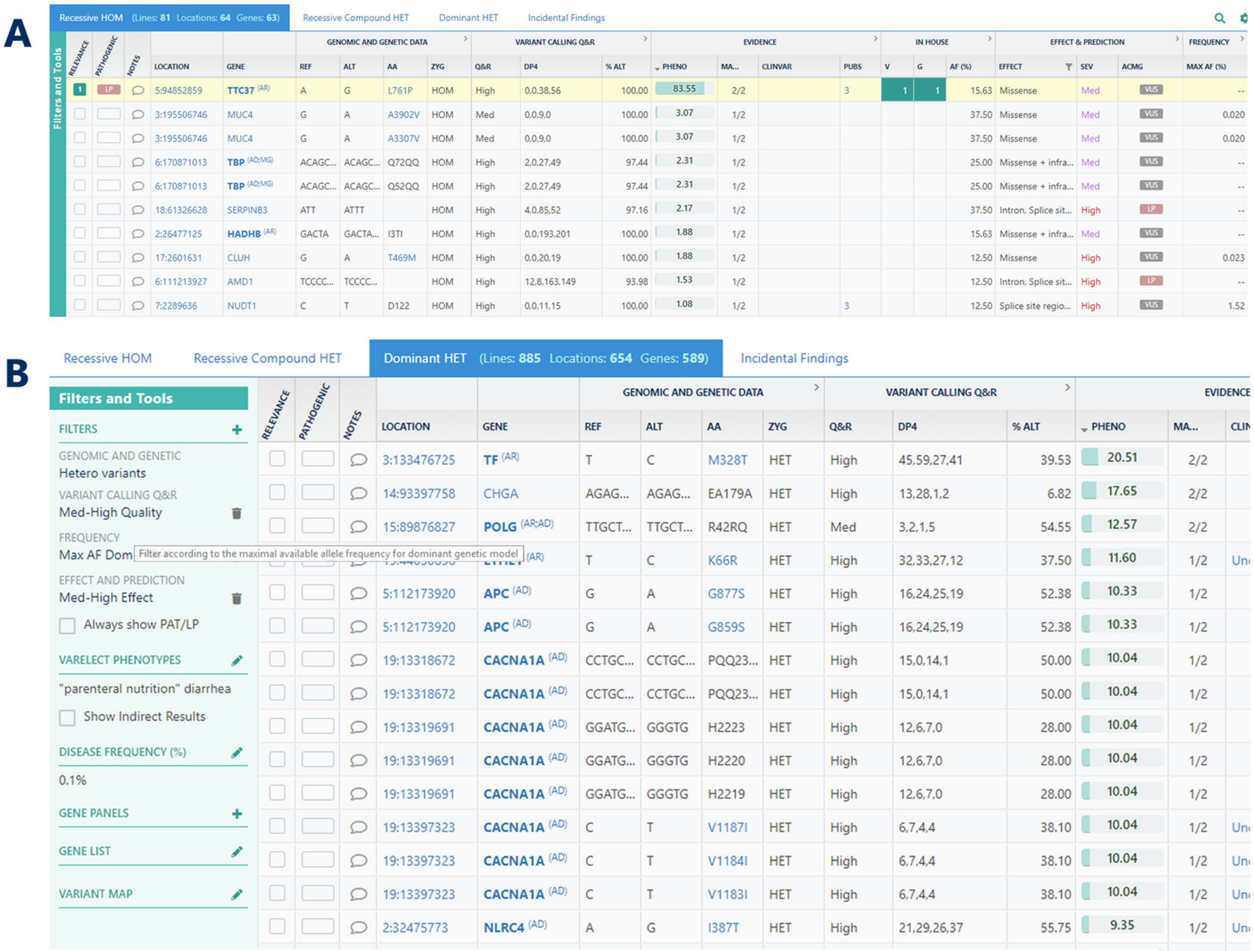
The user interface

The main analysis screen of TGex is essentially a detailed interactive table, where each row represents a single genomic position with a variation, and each column is populated with the relevant information gathered during the annotation process (Fig. 2a). The user can search or apply filters on any column, and the resulting list or variants can be sorted according to any column, in order to examine the remaining list of variants by their relevance or by their probability of being the causal variants.

One of the important strengths of TGex is the ability to create protocols (Table 2). A protocol in TGex can, if applicable, define which set of genetic models should be analyzed, include combinations of simple or more complex filters on any set of annotation entities, be restricted to certain lists of genes or genomic regions, define a template for a report and the data that should be within it, and much more. The resulting interface is represented by a tab for each genetic model, and a set of filters that are explicitly shown in a collapsible pane on the left side (Fig. 2b). Moreover, each column filter that is applied by the user is also documented in the filter pane. One may

Table 2 Examples of protocols in TGex

Analysis Type (Protocol)	Description	Main Sample	Associated Samples
Single Sample Exome	Rare genetic disorders	Proband	N/A
Trio Exome	Rare genetic disorders	Proband	Mother; Father
Tumor Biopsy	Cancer genetics	Tumor biopsy	Matched germline
Carrier Screening	Mendelian disorders	Virtual offspring	Mother; Father
Health Screening	Cancer risk assessment	Patient	N/A
PGx	PharmacoGenomics	Patient	N/A



A

Recessive HOM (Lines: 81 Locations: 64 Genes: 63) | Recessive Compound HET | Dominant HET | Incidental Findings

Filters and Tools

					GENOMIC AND GENETIC DATA				VARIANT CALLING Q&R			EVIDENCE				IN HOUSE			EFFECT & PREDICTION			FREQUENCY
RELEVANCE	PATHOGENIC	NOTES	LOCATION	GENE	REF	ALT	AA	ZYG	Q&R	DP4	% ALT	PHENO	MA...	CLINVAR	PUBS	V	G	AF (%)	EFFECT	SEV	ACMG	MAX AF (%)
1	LP	💬	5:94852859	TTC37 (AR)	A	G	L761P	HOM	High	0,0,38,56	100.00	83.55	2/2		3	1	1	15.63	Missense	Med	VUS	--
□	□	💬	3:195506746	MUC4	G	A	A3902V	HOM	Med	0,0,9,0	100.00	3.07	1/2					37.50	Missense	Med	VUS	0.020
□	□	💬	3:195506746	MUC4	G	A	A3307V	HOM	Med	0,0,9,0	100.00	3.07	1/2					37.50	Missense	Med	VUS	0.020
□	□	💬	6:170871013	TBP (AD;MG)	ACAGC...	ACAGC...	Q72QQ	HOM	High	2,0,27,49	97.44	2.31	1/2					25.00	Missense + infra...	Med	VUS	--
□	□	💬	6:170871013	TBP (AD;MG)	ACAGC...	ACAGC...	Q52QQ	HOM	High	2,0,27,49	97.44	2.31	1/2					25.00	Missense + infra...	Med	VUS	--
□	□	💬	18:61326628	SERPINB3	ATT	ATTT		HOM	High	4,0,85,52	97.16	2.17	1/2					37.50	Intron, Splice sit...	High	LP	--
□	□	💬	2:26477125	HADHB (AR)	GACTA	GACTA...	I3TI	HOM	High	0,0,193,201	100.00	1.88	1/2					15.63	Missense + infra...	Med	VUS	--
□	□	💬	17:2601631	CLUH	G	A	T469M	HOM	High	0,0,20,19	100.00	1.88	1/2					12.50	Missense	High	VUS	0.023
□	□	💬	6:111213927	AMD1	TCCCC...	TCCCC...		HOM	High	12,8,163,149	93.98	1.53	1/2					12.50	Intron, Splice sit...	High	LP	--
□	□	💬	7:2289636	NUDT1	C	T	D122	HOM	High	0,0,11,15	100.00	1.08	1/2		3			12.50	Splice site regio...	High	VUS	1.52

B

Recessive HOM | Recessive Compound HET | Dominant HET (Lines: 885 Locations: 654 Genes: 589) | Incidental Findings

Filters and Tools

FILTERS

GENOMIC AND GENETIC
Hetero variants

VARIANT CALLING Q&R
Med-High Quality

FREQUENCY
Max AF Dom Filter according to the maximal available allele frequency for dominant genetic model

EFFECT AND PREDICTION
Med-High Effect

□ Always show PAT/LP

VARELECT PHENOTYPES

"parenteral nutrition" diarrhea

□ Show Indirect Results

DISEASE FREQUENCY (%)

0.1%

GENE PANELS

GENE LIST

VARIANT MAP

					GENOMIC AND GENETIC DATA				VARIANT CALLING Q&R			EVIDENCE		
RELEVANCE	PATHOGENIC	NOTES	LOCATION	GENE	REF	ALT	AA	ZYG	Q&R	DP4	% ALT	PHENO	MA...	CLIN
□	□	💬	3:133476725	TF (AR)	T	C	M328T	HET	High	45,59,27,41	39.53	20.51	2/2	
□	□	💬	14:93397758	CHGA	AGAG...	AGAG...	EA179A	HET	High	13,28,1,2	6.82	17.65	2/2	
□	□	💬	15:89876827	POLG (AR;AD)	TTGCT...	TTGCT...	R42RQ	HET	Med	3,2,1,5	54.55	12.57	2/2	
□	□	💬	[illegible]	[illegible] (AR)	T	C	K66R	HET	High	32,33,27,12	37.50	11.60	1/2	Un
□	□	💬	5:112173920	APC (AD)	G	A	G877S	HET	High	16,24,25,19	52.38	10.33	1/2	
□	□	💬	5:112173920	APC (AD)	G	A	G859S	HET	High	16,24,25,19	52.38	10.33	1/2	
□	□	💬	19:13318672	CACNA1A (AD)	CCTGC...	CCTGC...	PQQ23...	HET	High	15,0,14,1	50.00	10.04	1/2	
□	□	💬	19:13318672	CACNA1A (AD)	CCTGC...	CCTGC...	PQQ23...	HET	High	15,0,14,1	50.00	10.04	1/2	
□	□	💬	19:13319691	CACNA1A (AD)	GGATG...	GGGTG	H2223	HET	High	12,6,7,0	28.00	10.04	1/2	
□	□	💬	19:13319691	CACNA1A (AD)	GGATG...	GGGTG	H2220	HET	High	12,6,7,0	28.00	10.04	1/2	
□	□	💬	19:13319691	CACNA1A (AD)	GGATG...	GGGTG	H2219	HET	High	12,6,7,0	28.00	10.04	1/2	
□	□	💬	19:13397323	CACNA1A (AD)	C	T	V1187I	HET	High	6,7,4,4	38.10	10.04	1/2	Un
□	□	💬	19:13397323	CACNA1A (AD)	C	T	V1184I	HET	High	6,7,4,4	38.10	10.04	1/2	Un
□	□	💬	19:13397323	CACNA1A (AD)	C	T	V1183I	HET	High	6,7,4,4	38.10	10.04	1/2	Un
□	□	💬	2:32475773	NLRC4 (AD)	A	G	I387T	HET	High	21,29,26,37	55.75	9.35	1/2	Un

Fig. 2 The TGex analysis screen (SNVs). The example shown here (and in Fig. 3) is a case of non-syndromic congenital diarrhea [54]. Following WES, the variant with the strongest phenotype implication for "diarrhea" was within *TTC37* (L761P), a known gene for trichohepatoenteric syndrome. The discovery of this novel homozygous damaging missense variant was significant for providing an effective diagnosis for a misdiagnosed case. **a** The main analysis screen is designed to optimally provide the analyst with information and user-interface options. The main analysis area is divided into dedicated tabs for each genetic model used for the analysis, and an additional tab for incidental findings. Each tab is an interactive table where each row represents a variant, and each column depicts a particular variant attribute. The attributes are divided into 7 categories, each category is collapsed by default, showing a subset of critical attributes, with an option to expand. Each column has two interactive functionalities – sorting (by a click on the header) and filtering (clicking on the filter icon to the right). **b** The Filters and Tools pane summarizes all applied filters for a specific tab in a given analysis. Via this pane, or alternatively via each of the attribute columns in the main analysis screen, filters can be easily added, edited or removed while reviewing the variants. All applied filters are also documented in the Methods section of the final report. In addition to the column filters described above, the pane includes advanced filter options, including predefined gene panels, manually entered gene list filters, VarElect terms used for phenotype prioritization, and Disease frequency used for the allele frequency filter

apply and/or remove filters during the analysis, and examine the resulting instantly updated list of variants.

To create a consolidated view of the most relevant annotations, the default view hides the additional annotations of each category; users can then choose to expand and review any of the dozens of available annotations. For example, by default, the 'Effect and Prediction' category contains 3 columns (Fig. 2a), which includes the effect as provided by SnpEff [30], while the expanded view for this category presents the actual scores from many individual sources and predictions tools, e.g. GERP [48], SIFT [45], PolyPhen-2 [44] and MutationTaster [46]. Importantly, the composition of columns presented in the consolidated view is fully customizable as part of the protocol definitions.

Workflow examples

WES analyses

We start with a common example of a rare congenital genetic disorder case, where the DNA sample of the patient underwent WES. Following a standard primary analysis pipeline, typically BWA-GATK [55], the resulting VCF file lists between 20,000 and 50,000 short nucleotide variants (SNVs) representing substitutions and short insertions/deletions (indels) [35, 56]. The featured protocol in TGex in this case typically includes 4 genetic models – recessive homozygote (HOM), recessive compound heterozygote (HET), dominant HET, and incidental findings (based on ACMG guidelines [20]). The dominant HET genetic model, for instance, automatically applies 4 default filters (Fig. 2b):

1. Genomic and Genetic: Includes HET variants only
2. Variant Calling Q&R: Excludes low reliability variants
3. Frequency: Excludes common variants (using the cutoff set by the user; 0.1% in this case)
4. Effect and Prediction: Excludes variants with low or no predicted effect on the protein function

Even after applying stringent filters, several hundred variants typically remain. The analysis step deals with the challenge of browsing and examining numerous variants, in order to select the best candidates for in-depth review and interpretation. Notably in TGex, the resulting list of variants is sorted according to the VarElect score of the affected gene, which reflects the strength of the association between the gene and the list of user-defined keywords [27], in this case, the list of phenotypes exhibited by the patient. The list of keywords can be defined as part of a customized user protocol, entered manually while initiating a new analysis, and/or modified during the analysis.

The next stage of the recommended analysis workflow is the interpretation and examination of potential candidate variants. The key entity to explore is the phenotype association, hence the default sorting of the candidate variants to review is their VarElect score. Clicking on the VarElect score, one can review a popup containing all of the evidence gathered from various data sources in GeneCards and MalaCards, represented by texts from the knowledgebase, and clearly highlighting matched keywords within their original context (Fig. 3a). The evidence popup also includes links to the relevant sections in the Suite's websites, together with links to external sites such as PubMed [57], OMIM [43], and others. Subsequently, the user may perform a thorough examination of the candidate variant's characteristics, such as its reliability (coverage, reads distribution etc.), its predicted effect on the protein (selected prediction tools), its allele frequency in control populations, and clinical information of the gene, which are also consolidated into a single ACMG-guidelines based score (Fig. 3).

Following this close examination of candidate variants, the user can select the most plausible ones and annotate them according to their relevance and pathogenicity, add free text comments, and select the evidence sections provided by VarElect to be integrated into the report (Fig. 3b). The user may then proceed to other genetic models, and end by exploring the variants found in the 59 recommended ACMG incidental findings genes [20].

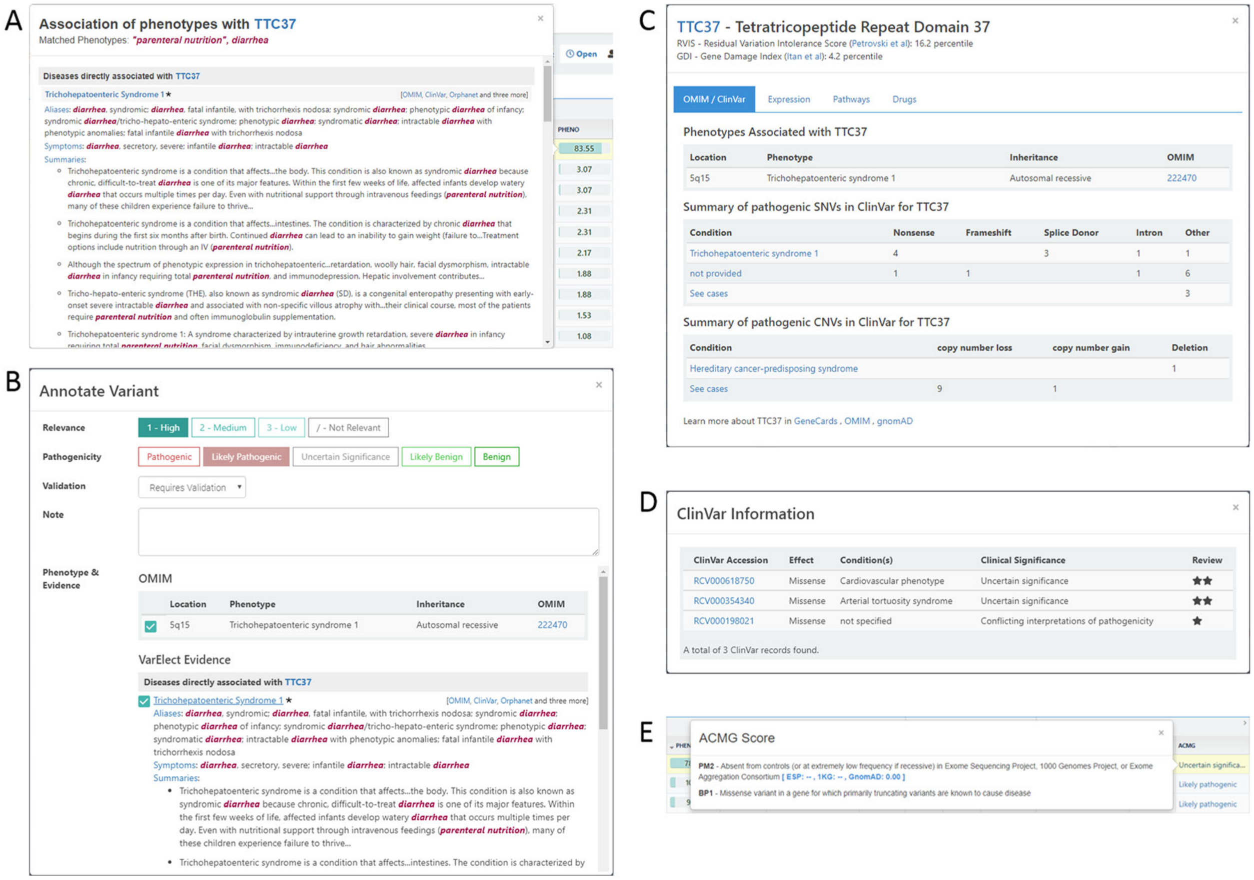
At this stage, clicking the 'Report Preview' button extracts all of the selected variants with their annotation and evidence sections, allowing one to review the current status of the analysis. Once satisfied with the preview, clicking the 'Generate Report' button launches the reporting system to generate the final report in the selected format (Word or PDF), and a supplementary Excel table listing all of the variants in each genetic model for future documentation.

Whole genome sequence analyses

There are a growing number of large-scale sequencing projects performing WGS [58, 59], and a growing number of hospitals and genetic laboratories that are now transitioning to WGS for interpretation of genetic diseases. WGS can characterize various types of genetic variation in all parts of the genome [19], making the data much more complex for interpretation. A critical example is structural variants (SVs), known to be a major source of pathogenicity [60–62]. The disease-related mechanism of SVs might not involve any overlap with a disease-associated coding gene. Rather, it might act by influencing genes over large distances by altering non-coding functional units such as regulatory elements (promoters and enhancers) and ncRNA genes. Evaluation of the impact of non-coding variants for disease interpretation is a great challenge, and requires novel approaches and increasingly sophisticated software solutions [29].

For this aim, TGex leverages GeneHancer [26], the GeneCards Suite database of regulatory elements and their gene targets. GeneHancer provides a unique non-redundant and comprehensive genome-wide map of scored ~ 400,000 enhancers and promoters ("GeneHancers"), and their gene associations. The combination of GeneHancer and VarElect enables translating the finding of an SV or SNV variant in a non-coding region into a variant-to-gene-to-phenotype annotation, enabling prioritization of phenotype associations of variant-containing elements via the elements' gene targets (Fig. 4).

To address the WGS interpretation challenges, we developed a new interface of TGex for WGS data analysis with the focus on complex CNV and SV data. In the GeneHancer-augmented TGex SV module, SVs are mapped to both genes and regulatory elements, followed by VarElect interpretation of the mixed list of genes and enhancers/promoters. VarElect processes GeneHancer element identifiers, performing gene-phenotype prioritization for GeneHancer element gene targets. The phenotype prioritization score in this workflow is calculated by combining the VarElect phenotype score of the element gene target with the



Association of phenotypes with TTC37
Matched Phenotypes: *"parenteral nutrition"*, *diarrhea*

Diseases directly associated with TTC37

Trichohepatoenteric Syndrome 1 ★ [OMIM, ClinVar, Orphanet and three more]

Aliases: ***diarrhea***, syndromic; ***diarrhea***, fatal infantile, with trichorrhexis nodosa; syndromic ***diarrhea***; phenotypic ***diarrhea*** of infancy; syndromic ***diarrhea***/tricho-hepato-enteric syndrome; phenotypic ***diarrhea***; syndromatic ***diarrhea***; intractable ***diarrhea*** with phenotypic anomalies; fatal infantile ***diarrhea*** with trichorrhexis nodosa

Symptoms: ***diarrhea***, secretory, severe; infantile ***diarrhea***; intractable ***diarrhea***

Summaries:

- Trichohepatoenteric syndrome is a condition that affects...the body. This condition is also known as syndromic ***diarrhea*** because chronic, difficult-to-treat ***diarrhea*** is one of its major features. Within the first few weeks of life, affected infants develop watery ***diarrhea*** that occurs multiple times per day. Even with nutritional support through intravenous feedings (***parenteral nutrition***), many of these children experience failure to thrive...
- Trichohepatoenteric syndrome is a condition that affects...intestines. The condition is characterized by chronic ***diarrhea*** that begins during the first six months after birth. Continued ***diarrhea*** can lead to an inability to gain weight (failure to...Treatment options include nutrition through an IV (***parenteral nutrition***).
- Although the spectrum of phenotypic expression in trichohepatoenteric...retardation, woolly hair, facial dysmorphism, intractable ***diarrhea*** in infancy requiring total ***parenteral nutrition***, and immunodepression. Hepatic involvement contributes...
- Tricho-hepato-enteric syndrome (THE), also known as syndromic ***diarrhea*** (SD), is a congenital enteropathy presenting with early-onset severe intractable ***diarrhea*** and associated with non-specific villous atrophy with...their clinical course, most of the patients require ***parenteral nutrition*** and often immunoglobulin supplementation.
- Trichohepatoenteric syndrome 1: A syndrome characterized by intrauterine growth retardation, severe ***diarrhea*** in infancy requiring total ***parenteral nutrition***, facial dysmorphism, immunodeficiency, and hair abnormalities

Annotate Variant

Relevance 1 - High | 2 - Medium | 3 - Low | / - Not Relevant

Pathogenicity Pathogenic | Likely Pathogenic | Uncertain Significance | Likely Benign | Benign

Validation Requires Validation ▾

Note

Phenotype & Evidence

OMIM

	Location	Phenotype	Inheritance	OMIM
☑	5q15	Trichohepatoenteric syndrome 1	Autosomal recessive	222470

VarElect Evidence

Diseases directly associated with TTC37

☑ Trichohepatoenteric Syndrome 1 ★ [OMIM, ClinVar, Orphanet and three more]

Aliases: ***diarrhea***, syndromic; ***diarrhea***, fatal infantile, with trichorrhexis nodosa; syndromic ***diarrhea***; phenotypic ***diarrhea*** of infancy; syndromic ***diarrhea***/tricho-hepato-enteric syndrome; phenotypic ***diarrhea***; syndromatic ***diarrhea***; intractable ***diarrhea*** with phenotypic anomalies; fatal infantile ***diarrhea*** with trichorrhexis nodosa

Symptoms: ***diarrhea***, secretory, severe; infantile ***diarrhea***; intractable ***diarrhea***

Summaries:

- Trichohepatoenteric syndrome is a condition that affects...the body. This condition is also known as syndromic ***diarrhea*** because chronic, difficult-to-treat ***diarrhea*** is one of its major features. Within the first few weeks of life, affected infants develop watery ***diarrhea*** that occurs multiple times per day. Even with nutritional support through intravenous feedings (***parenteral nutrition***), many of these children experience failure to thrive...
- Trichohepatoenteric syndrome is a condition that affects...intestines. The condition is characterized by

TTC37 - Tetratricopeptide Repeat Domain 37
RVIS - Residual Variation Intolerance Score (Petrovski et al): 16.2 percentile
GDI - Gene Damage Index (Itan et al): 4.2 percentile

OMIM / ClinVar | Expression | Pathways | Drugs

Phenotypes Associated with TTC37

Location	Phenotype	Inheritance	OMIM
5q15	Trichohepatoenteric syndrome 1	Autosomal recessive	222470

Summary of pathogenic SNVs in ClinVar for TTC37

Condition	Nonsense	Frameshift	Splice Donor	Intron	Other
Trichohepatoenteric syndrome 1	4		3	1	1
not provided	1	1		1	6
See cases					3

Summary of pathogenic CNVs in ClinVar for TTC37

Condition	copy number loss	copy number gain	Deletion
Hereditary cancer-predisposing syndrome			1
See cases	9	1	

Learn more about TTC37 in GeneCards , OMIM , gnomAD

ClinVar Information

ClinVar Accession	Effect	Condition(s)	Clinical Significance	Review
RCV000618750	Missense	Cardiovascular phenotype	Uncertain significance	★★
RCV000354340	Missense	Arterial tortuosity syndrome	Uncertain significance	★★
RCV000198021	Missense	not specified	Conflicting interpretations of pathogenicity	★

A total of 3 ClinVar records found.

ACMG Score

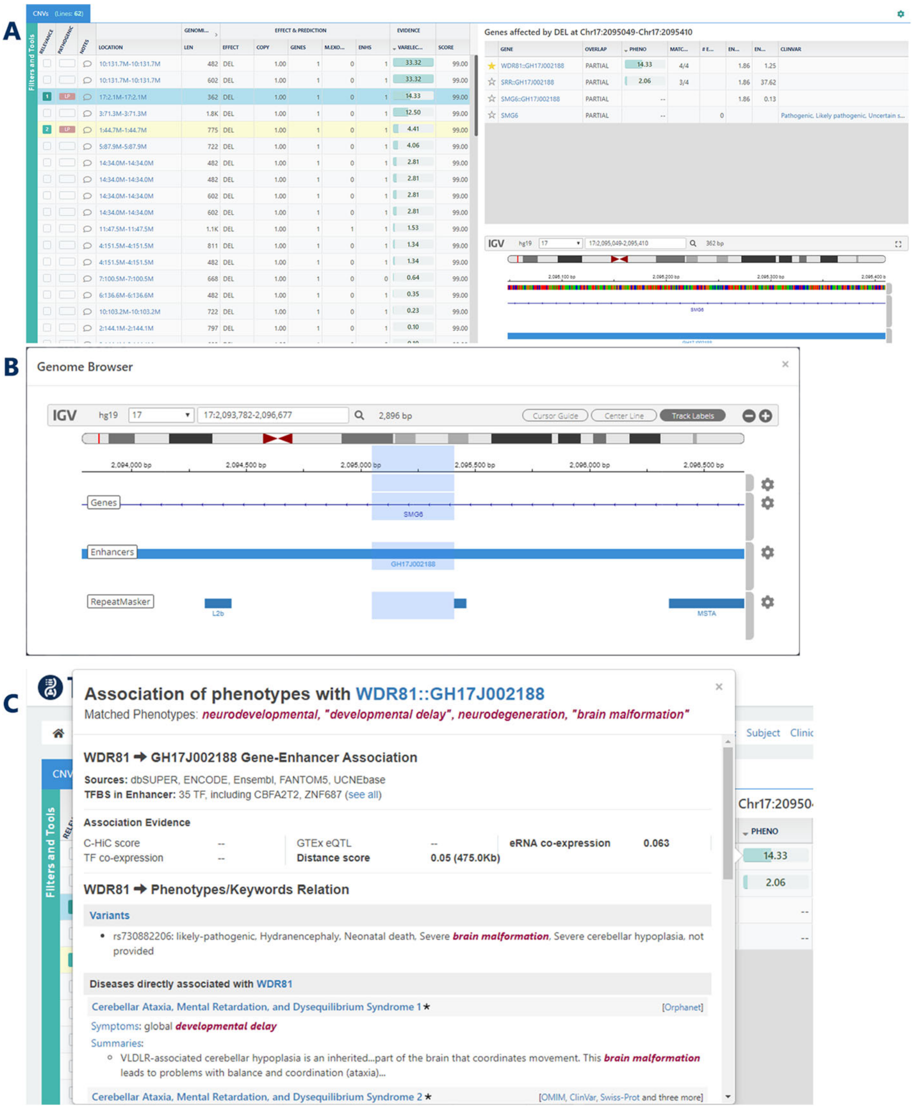
PM2 - Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium [ESP: -- , 1KG: -- , GnomAD: 0.00]

BP1 - Missense variant in a gene for which primarily truncating variants are known to cause disease

Fig. 3 Variant analysis and interpretation. This figure shows several views in TGex providing detailed information and useful links to source data, with the goal to improve and hasten expert variant interpretation. **a** VarElect MiniCards. The extensive gene-phenotype hit-context evidence from the GeneCards knowledgebase is portrayed in the MiniCards. This figure shows selected parts of the MiniCard for the gene *TTC37* and the phenotypes used in the congenital diarrhea case. A list of matched phenotypes is shown in red in the top part, followed by extensive gene-centric evidence for queried phenotype association from various GeneCards sections. This is combined with MalaCards-based evidence, similarly showing queried phenotype associations in diseases associated with the gene *TTC37*, from various MalaCards sections. Search terms are highlighted throughout the text, and links to specific GeneCards/MalaCards webcard positions enable further scrutiny via more detailed evidence exploration within the knowledgebase. **b** Variant and evidence selection. Several types of marks can be defined per candidate variant by the analyst, upon clicking the 'Annotate variant' button located to the left of each variant row. This includes relevance (High, Med or Low), the pathogenicity of the variant, and a free text note. Below, information pieces regarding the variant/gene pathogenicity can be selected, based on VarElect MiniCards and OMIM disease records. The selected variants and their annotations are propagated to the report. **c** Gene view. A gene-centric summary for the gene *TTC37*, including associated diseases, mode of inheritance, and pathogenic variants summary, based on OMIM and ClinVar records. **d** ClinVar information – ClinVar records matching a given variant, including the condition and clinical significance. **e** ACMG score – Clinical significance based on the ACMG score. Clicking upon the variant clinical significance value shows a detailed view of the data used for the classification.

GeneHancer element and element-gene association confidence scores.

The CNV/SV analysis screen of TGex is divided into three sections (Fig. 4A). On the left, the user can browse the reported genetic events and use their features, such as the genomic length, the number of affected genes, the copy number, and others, to filter and sort the list of events, quite similarly to the SNV analysis screen. The top right section of the screen lists the genes and the GeneHancers that are affected by the highlighted event and their VarElect score for the input phenotype keywords. The events in the main screen are sorted by default by the highest VarElect score for each event, and therefore increase the probability that the more relevant events appear higher in the main list, markedly reducing the time spent on performing such non-trivial analyses. Clicking on the VarElect phenotype score shows the MiniCards, with evidence for the element-gene-phenotype association (Fig. 4c). The bottom right panel provides an expandable genomic view (Fig. 4b), where the analyst can visually examine the genomic context and assess the relevance of the genetic event in question.



CNVs (Lines: 62)

RELEVANCE	PATHOGENIC	NOTES	LOCATION	GENOMI… LEN	EFFECT & PREDICTION EFFECT	COPY	GENES	M.EXO…	ENHS	EVIDENCE VARELEC…	SCORE
□			10:131.7M-10:131.7M	482	DEL	1.00	1	0	1	33.32	99.00
□			10:131.7M-10:131.7M	602	DEL	1.00	1	0	1	33.32	99.00
1	LP		17:2.1M-17:2.1M	362	DEL	1.00	1	0	1	14.33	99.00
□			3:71.3M-3:71.3M	1.8K	DEL	1.00	1	0	1	12.50	99.00
2	LP		1:44.7M-1:44.7M	775	DEL	1.00	1	0	1	4.41	99.00
□			5:87.9M-5:87.9M	722	DEL	1.00	1	0	1	4.06	99.00
□			14:34.0M-14:34.0M	482	DEL	1.00	1	0	1	2.81	99.00
□			14:34.0M-14:34.0M	482	DEL	1.00	1	0	1	2.81	99.00
□			14:34.0M-14:34.0M	602	DEL	1.00	1	0	1	2.81	99.00
□			14:34.0M-14:34.0M	602	DEL	1.00	1	0	1	2.81	99.00
□			11:47.5M-11:47.5M	1.1K	DEL	1.00	1	1	1	1.53	99.00
□			4:151.5M-4:151.5M	811	DEL	1.00	1	0	1	1.34	99.00
□			4:151.5M-4:151.5M	482	DEL	1.00	1	0	1	1.34	99.00
□			7:100.5M-7:100.5M	668	DEL	1.00	1	0	0	0.64	99.00
□			6:136.6M-6:136.6M	482	DEL	1.00	1	0	1	0.35	99.00
□			10:103.2M-10:103.2M	722	DEL	1.00	1	0	1	0.23	99.00
□			2:144.1M-2:144.1M	797	DEL	1.00	1	0	1	0.10	99.00

Genes affected by DEL at Chr17:2095049-Chr17:2095410

	GENE	OVERLAP	PHENO	MATC…	# E…	EN…	EN…	CLINVAR
★	WDR81::GH17J002188	PARTIAL	14.33	4/4		1.86	1.25	
☆	SRR::GH17J002188	PARTIAL	2.06	3/4		1.86	37.62	
☆	SMG6::GH17J002188	PARTIAL	--			1.86	0.13	
☆	SMG6	PARTIAL	--		0			Pathogenic, Likely pathogenic, Uncertain s…

Association of phenotypes with WDR81::GH17J002188

Matched Phenotypes: *neurodevelopmental, "developmental delay", neurodegeneration, "brain malformation"*

WDR81 ➔ GH17J002188 Gene-Enhancer Association

Sources: dbSUPER, ENCODE, Ensembl, FANTOM5, UCNEbase
TFBS in Enhancer: 35 TF, including CBFA2T2, ZNF687 (see all)

Association Evidence

C-HiC score	--	GTEx eQTL	--	**eRNA co-expression**	**0.063**
TF co-expression	--	**Distance score**	**0.05 (475.0Kb)**		

WDR81 ➔ Phenotypes/Keywords Relation

Variants

- rs730882206: likely-pathogenic, Hydranencephaly, Neonatal death, Severe ***brain malformation***, Severe cerebellar hypoplasia, not provided

Diseases directly associated with WDR81

Cerebellar Ataxia, Mental Retardation, and Dysequilibrium Syndrome 1★ [Orphanet]

Symptoms: global ***developmental delay***

Summaries:

- VLDLR-associated cerebellar hypoplasia is an inherited...part of the brain that coordinates movement. This ***brain malformation*** leads to problems with balance and coordination (ataxia)...

Cerebellar Ataxia, Mental Retardation, and Dysequilibrium Syndrome 2★ [OMIM, ClinVar, Swiss-Prot and three more]

Fig. 4 The TGex analysis screen (SVs). SV analysis is exemplified by a list of recurrently mutated regulatory elements discovered in a cohort of patients with neurodevelopmental disorders [63]. The highlighted element overlaps the GH17J002188 GeneHancer, an intronic enhancer of the gene *SMG6*. Remarkably, this enhancer also targets the *WDR81* gene (over ~ 476 kb), with a higher VarElect score for the relevant phenotype (neurodevelopmental, "developmental delay", etc') than *SMG6*. **a** The main analysis area for SVs is divided into 3 sections, including the main section listing the SV events (left), the detailed event section (top, right) presenting a detailed view of the list of genes and GeneHancer regulatory elements that are affected by the event, and the genomic view section (bottom, right) allowing visual examination of the genomic context of each event. **b** Expanded view of the event genomic context. **c** Clicking on the Phenotype score for a given GeneHancer opens the VarElect MiniCard for the element-gene-phenotype association. At the top part of the MiniCard, evidence describing the GeneHancer and its association with the gene target is detailed. This includes a list of sources for the identification of the element; a list of transcription factors found to have binding sites within the element; a detailed view of the evidence for element-gene associations. Below the GeneHancer details appear the classic gene-phenotype MiniCards as described in Fig. 3. Importantly, the score used for prioritization in the SV module is calculated by combining the GeneHancer confidence score of the element and of the element-gene association, with the classic VarElect gene-phenotype score of the element target gene

TGex advantages and comparison with other tools

Platforms for clinical genetics analyses are highly diverse, with numerous academic and commercial tools available, as recently reviewed [29, 64, 65]. Some facets of such workflows are essentially identical across different platforms (e.g. basic variant annotation steps, allele frequency databases, and variant damage prediction). Less standard features, e.g. phenotype-interpretation, must rely on the availability and development of complex databases, and thus often differ significantly among tools. Another aspect which naturally varies among platforms is the user interface capabilities and interactive features. The individual or organizational user would be inclined to select tools that provide a robust workflow with rapid turnover and high diagnostic yield. In this section we summarize the key properties of TGex that make it a leading analysis platform in this respect, supported by literature comparisons and reviews, TGex use-cases, and in-house data.

Phenotype-driven interpretation with VarElect

VarElect is our comprehensive phenotype-dependent gene prioritizer [27] used world-wide, with thousands of users and tens of thousands of analyses per year. VarElect takes advantage of the wealth of information in the GeneCards Suite knowledgebase, jointly exploiting the gene- and disease-centric GeneCards [23] and MalaCards [24], as well as the Suite's free-text Boolean search and scoring capabilities. VarElect thus proficiently matches variant-containing genes to user-submitted disease/symptom keywords. The VarElect algorithm infers direct as well as indirect ("guilt by association") links between genes and phenotypes. The indirect mode benefits from GeneCards' diverse gene-to-gene data links, including the broad pathway information in PathCards, the Suite's unified pathway database that integrates 12 pathway sources [66]. In addition to scoring genes against user-defined keywords, VarElect provides extensive gene-phenotype evidence (via the "MiniCards"), with such evidence hyperlinked to source databases.

VarElect was shown to outperform four other tools (Phenolyzer [67], Exomiser [68], IVA [69] and Phevor [70]) in the original VarElect paper [27]. More recently, Tomar et al. [28] compared the performance of three gene prioritization tools – VarElect, Phenolyzer [67] and OMIMExplorer [71] on a collection of 50 cases, for which the disease causing gene had already been resolved, and on a simulated real life scenario of having only partial phenotype lists. The authors reported that VarElect outperformed both OMIMExplorer and Phenolyzer. Moreover, when omitting critical keywords used for gene ranking, VarElect remained practically unaffected, while both competing tools showed a marked reduction in performance [28].

Phenotype prioritization of non-coding variants

To address one of the critical challenges in the analysis of WGS, we developed novel databases and approaches paving the way to interpretation of non-coding variants (see 'Whole genome sequence analyses' section above). With GeneHancer, the regulatory element database, and the inclusive collection of ncRNA genes in GeneCards, TGex greatly expands the genomic scope of phenotype-driven analysis, from the commonly used 2% (exome) to ~ 20%. This is a key prospect in the exploitation of WGS to increase the diagnostic yield.

Addressing a wide spectrum of genetic analyses

TGex is designed to judiciously leverage the vast GeneCards Suite knowledgebase to address various clinical genetics requirements and workflows. TGex was shown to quickly identify causal mutations of rare disease cases, both when the causal variant was previously documented [72], and in the discovery of novel mutations through prioritization of potential variants of uncertain significance (VOUSes) [73]. Even for patients diagnosed with a known genetic disorder and an identified mutation, TGex has been used to examine phenotypic variability and identify modifier mutations and genes on top of established findings [74]. TGex was also useful in the analysis of novel disorders and syndromes, establishing the evidence for the clinical validity of the association between genes and emerging disorders [75, 76], and in discovery of predisposition variants to complex diseases (e.g. Parkinson) [77], as well as hereditary cancer (see Clinical use cases section below) and molecular profiling of tumor biopsies [78].

Robustness and standardization of analysis

Platforms like TGex support performing analyses under robust, consistent, reproducible and standardized conditions, compliant with community best practices recommendations such as HIPAA standards and to ACMG guidelines.

Accounts using TGex benefit from having all analyses stored in a structured database, allowing the organization to benefit from analyzing the case statistics, workflows, bottlenecks, disease and variant trends, etc., as shown in the 'Large scale account' section below.

In-house allele frequency

TGex handles thousands of samples in specific accounts or group of collaborating accounts and automatically calculates the 'in-house' allele frequency which can be crucial for variant selection especially in highly specific ethnic groups. The user interface also includes and highlights the former annotations and interpretations as entered by the analysts, assisting in applying the accumulated in-house knowledge to new cases.

True end-to-end all-in-one platform

TGex provides a complete workflow, starting from a VCF file, performing the analysis and interpretation accompanied with evidence scrutinizing, and concluding the findings in the report, all via a user-friendly interface. The immediate consequences are the high diagnostic yield and the fast and cost-effective analysis with the intuitive interface encompassing a broad knowledgebase, optimizing the time it takes an analyst to interpret the data. Together, these features have the potential to markedly increase the volume of cases to be analyzed in large organizations.

Suwinski et al. [79] reviewed how application of biological databases and bioinformatics tools can address the bottleneck in clinical genetics data processing and analysis. Focusing on four currently available web-based interface platforms that include clinical prioritization of variants in VCF files, they conclude that in terms of innovation, depth of knowledge and the ease of generating clinical reports, TGex is the top scorer and is by far the most clinician-friendly WES analysis pipeline and reporting platform [79].

Clinical use cases

In the past few years, TGex and VarElect have been widely adopted for clinical genetics analysis in various academic institutions, genetic medical centers and hospitals world-wide, with usage volumes ranging from research groups focusing on a handful of patients to genetic centers routinely analyzing hundreds of cases per month [54, 72–78, 80–95]. Our interactions with a diversity of clinical genetics users impel us to deliver frequent community-driven improvements. In this section, we describe three representative studies that illustrate recent use of TGex in different clinical genetics contexts.

Cerebral creatine deficiency syndrome-1

Cerebral creatine deficiency syndrome-1 (CCDS1, MIM: 300352 [43], MalaCards ID:CRB151 [24]) is an X-linked disorder of creatine transport characterized by mental retardation, severe speech delay, behavioral abnormalities and seizures. Defects in the creatine transporter gene *SLC6A8* have been reported to cause CCDS1 [96]. A 5 year old male patient from a Chinese family was referred for genetic evaluation of development and speech delay and intellectual disabilities at the genetic counselling clinic in the Shenzen Maternal and Child Healthcare Hospital (China) [73]. Following targeted exome sequencing and data analysis with TGex, a novel candidate missense variant, c.1181C > A (p.Thr394Lys) in the *SLC6A8* gene (NM_005629.3) was identified, with high probability as a candidate mutation. Sanger sequencing validation confirmed that the father was not a carrier; the mutation was inherited from the heterozygous carrier mother, and also to the hemizygous similarly affected brother. The diagnosis was further confirmed by biochemical measurements, as well as by brain magnetic resonance spectroscopy. The proband's mother became pregnant with a 3rd sibling, for whom Sanger sequencing showed a negative result for this variant.

As concluded by the authors [73], this case shows that "The combination of targeted exome sequencing with systematic clinical evaluation of patients used in suspected genetic disorders may improve diagnostic yield, assist in the medical care of patients and offer genetic counseling and prenatal diagnosis for family members."

Large scale account

One remarkable example of extensive use of TGex in a large organization is provided at the Maternal and Child Health Hospital of Guangxi Zhuang Autonomous Region, China. The clinical genetics team at Guangxi has analyzed with TGex more than 3500 samples since 2017. Generating summary statistics and usage trend analyses, we focus on WES analyses of ~ 1300 singletons, and ~ 200 trios (proband and parents). We consider the 'High' or 'Medium' variant relevance marks, as submitted by the analysts, as an acceptable proxy for a resolved case. We note that the percentage of cases with marked candidates in TGex is comparable to the over-all diagnostic yield as documented by the Guangxi team: about 42% for proband only, and up to 50% in the trio cases (48 and 55% according to TGex relevance marks, respectively).

Notably, since TGex uses keywords for the analysis of rare genetic disorders, we were able to examine the nature of the keyword search as entered by the analyst in each case, and compare between classes of clinical symptoms that were used for the exome analyses. The four main classes of phenotypes (with minor overlap between them) are "Growth Retardation", "Developmental Delay", "Epilepsy" and "Genitalia symptoms". Table 3 summarizes the diagnostic yield in each phenotype class (using the aforementioned approximation marks), showing a much higher yield for the first three classes (~ 60%) while only 23% of the "Genitalia symptoms" cases were resolved. The highest yield in the "Epilepsy" cases could be attributed to the broader clinical genetics knowledge that was gathered during the last decade with hundreds of epilepsy-associated genes and validated mutations. This is in line with previous studies also showing that the diagnostic yield significantly varies among diseases, a phenomenon that might be related to a combination of several factors, including the degree of phenotype complexity, the depth of biomedical knowledge regarding the known causative genes of the specific disease, and others [12, 97, 98].

In order to examine the multi-year resolution rate, we compared the work done in 2017 and 2018 (regardless of the search keywords). We observe a significant increase in the number of cases with selected candidate variants, from ~ 42% in 2017 (599 cases) to ~ 65% in 2018 (552 cases), which might be explained by several possible reasons:

- The improved accumulation of clinical genetics data in the source databases integrated within the constantly updated GeneCards Suite knowledgebase.

- Improved adaptation of TGex within the organization and deeper experience of the analysts.
- Accumulation of data in the highly specific in-house database, which assists in variant filtering according to in-house allele frequency.

To summarize, a strong advantage of using a platform in clinical genetics practice, is the standardization of methods and protocols, allowing for simple generation of statistics related to all analyzed samples and cases. Moreover, it will enable automatic re-annotation and re-analysis of unresolved cases, and highlighting the ones worth re-examining due to novel information specifically associated with each case.

WGS for Hamartomatous polyposis syndromes

In the last decade, genetic tests of hereditary cancer has rapidly progressed from genotyping germline mutations by single gene Sanger sequencing or mutation panels to large scale sequencing of germline multi-gene panels and WES for diagnostic and prognostic applications [95, 99]. These complex genetic tests can detect more pathogenic genetic alterations, thus enabling better treatment decisions and personally tailored long-term surveillance for mutation carriers in the family. Gastrointestinal (GI) polyps and cancer have a very strong genetic component, with known genes that could be screened in high-risk families. Up to 10% of colorectal cancer (CRC) cases occur due to hereditary genetic syndromes, with even higher numbers for early-onset cases [100, 101]. Rare pathogenic mutations and common genetic variants contribute to personal and familial CRC risk.

In a cohort (with 74 patients in 52 families) presenting with hamartomatous polyposis phenotype with corresponding family history, at the Tel-Aviv Medical Center, the GI team conducted a comprehensive mutational search. Screening, with either cancer multi-gene panels or Sanger sequencing of suspected mutated genes, identified causal mutations in only ~ 50% of the families. The team selected 5 probands in which the genetic evaluation produced no significant findings, and performed a much wider search using WGS in TGex, envisioning that some mutations would not necessarily be SNVs but rather CNVs and SVs which could be identified by whole genome analysis. For each sample in this set, variant calling for SNVs and SVs was conducted, and both variant files were uploaded to TGex to a combined SNV/SV protocol. The analysis of these cases using the relevant keywords was simple and efficient, quickly pointing out the most relevant candidate variants, whether SNVs or SVs.

Remarkably, the genetic culprit was detected in all 5 cases following rapid analysis with TGex, presenting the subsequently validated causal events at the highest ranks, out of thousands of called variants. This included a loss-of-function SNV in *BMPR1A* (for this proband the suspected gene was *SMAD4* for which Sanger sequencing produced no clinically significant finding) and three SV events – two distinct cases of inversions, one affecting *BMPR1A* and the other affecting *STK11*, and a deletion in *BMPR1A* identified in two unrelated probands of a common ethnic origin, a possible founder mutation. *BMPR1A* (Bone Morphogenetic Protein Receptor Type 1A) is a cancer predisposing gene, related to polyposis, e.g. Juvenile polyposis syndrome (MIM:174900 [43], MalaCards ID:JVN014 [24]), an autosomal dominant GI cancer. All *BMPR1A* events were validated among the probands and affected family members by PCR and MLPA.

These results, although based on a small number of cases with a unique clinical phenotype, imply the potential of WGS, specifically with using the accurate and simplified TGex CNV/SV analysis, to markedly increase the diagnostic yield of genetic tests, leading the way to accurate genetic diagnosis in a timely and cost effective manner.

Future perspectives

The near future holds great promise for clinical genetics. Recent advances have made significant impact, however, analysis and interpretation of genome variation still remain challenging. Clinical genetics platforms like TGex are expected to be continually augmented with ever-growing variant- and gene-disease phenotype association knowledge, stronger variant frequency catalogs, and improved algorithms. This will be complemented with essential efforts aiming at comprehensive variant detection

Table 3 Comparison between phenotype classes in Guangxi Maternal Hospital

Phenotype class	Common associated keywords	Total cases	Resolved cases	% Resolved
Growth Retardation	Short stature	107	66	61.7
Developmental Delay	Mental retardation, Delayed speech, Motor delay	174	101	58.0
Epilepsy	Seizures, Convulsion, Spasm	191	121	63.4
Genitalia symptoms	Scrotum, Micropenis, Hypogonadism, Hypospadia	131	30	22.9

To compare between phenotype classes, highly abundant keywords in all of the cases of the account were selected. Those keywords were grouped into four main classes of phenotypes, and the statistics for all of the cases of each phenotype class were calculated (with minor case overlap between classes) of the whole gamut of variant classes (e.g. mobile elements, tandem repeats), and at improved ability to interpret non-coding variants within functional genomic regions (e.g. regulatory elements and ncRNA genes).

The GeneCards Suite is a leading biomedical knowledgebase, serving as a solid foundation for the clinical genetics variant interpretation capacities of TGex. Our effort to characterize the genomic "dark matter" arena of non-coding regions is focused on enhancing the non-coding variants interpretation capacities within TGex. With the aim of continuing our innovative development towards improved variant interpretation, the GeneCards Suite future effort will constitute significantly enhanced annotation of genome-wide functional non-coding elements, so as to allow TGex to find direct and indirect phenotype associations of those regions.

Conclusions

TGex is a powerful tool for the annotation, analysis and prioritization of coding and non-coding genomic variants. It provides access to an extensive knowledgebase of genomic annotations, with intuitive and flexible configuration options, allowing quick adaptation, and addressing various workflow requirements, simplifying and accelerating variant interpretation. TGex can be used in the various scenarios typically found in clinical organizations, e.g. by an analyst who creates an intermediate report for the clinical geneticist, or by a genetics team generating the final clinical report, based on a gene panel, Exome, or whole genome analysis. For all of these situations, TGex has a great potential to markedly reduce turn-around time by enabling methodical and faster analysis for primary analysts, followed by efficient review by geneticists. We have shown how the unique combination of TGex's strengths are increasingly useful for clinicians and researchers, and expect TGex to open new vistas for WGS in clinical genetics.

Availability and requirements

Project name: TGex

Project home page: https://tgex-app.genecards.org/ or https://tgex-app.genecards.cn (China)

Operating system(s): Platform independent (any operating system)

Programming language: .NET (back-end) and HTML/JavaScript (front-end)

Other requirements: A modern browser and processor; an internet or intranet connection.

License: Free academic research use

Any restrictions to use by non-academics: License required

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12920-019-0647-8.

Additional file 1: Figure S1. The TGex dashboard - account management module.

Additional file 2. TGex report for the trichohepatoenteric syndrome Demo example

Abbreviations

API: Application Programming Interface; CCDS1: Cerebral Creatine Deficiency Syndrome-1; CRC: Colorectal Cancer; EHR: Electronic Health Record; GDPR: General Data Protection Regulation; GI: GastroIntestinal; HET: Heterozygote; HIPAA: Health Insurance Portability and Accountability Act; HOM: Homozygote; indels: Insertions/deletions; LIMS: Laboratory Information Management Systems; LMS: LifeMap Sciences; ncRNA: Non-coding RNA; NGS: Next Generation Sequencing; Q&R: Quality and Reliability; SNVs: Short Nucleotide Variants; SVs: Structural Variants; TGex: Translational Genomics expert; VCF: Variant Call Format; VOUS: Variant Of Uncertain Significance; WES: Whole Exome Sequencing; WGS: Whole Genome Sequencing; WIS: Weizmann Institute of Science

Acknowledgements

Not applicable.

Authors' contributions

TGex was conceived and implemented by LMS and incorporates the WIS GeneCards Suite knowledgebase and gene prioritization/interpretation algorithms. The manuscript was jointly written by the LMS and WIS teams. YG and DD designed and supervised the development of TGex; YG and YM developed the software; GR, RK and DD performed the analyses and interpreted the data of the CRC use-case; YG, YM, DD, MT, RB, OZ, MS, DL and SF designed the integration of the GeneCards Suite knowledgebase and interpretation algorithms within TGex; TIS provided scientific guidance and quality assurance for the software; FC, QZ and YS supervise the clinical genetics workflow at Guangxi; DD and SF wrote the manuscript; YG, RB, GR, RK, MS and DL revised the manuscript; All authors read and approved the final manuscript.

Funding

Support for the GeneCards Suite was provided by a Research and Licensing grant from LMS. Additional funding was provided from the WIS Crown Human Genome Center, and an EU grant from PIONEER, a European Network of Excellence for Big Data in Prostate Cancer. Guangxi's project is supported by the National Key Research and Development Program (2018YFC1002501) and Major Research Plan of the Provincial Science and Technology Foundation of Guangxi (AB16380214).
The funding body LMS participated in the design of the study and in the collection, analysis, and interpretation of data, as well as in writing the manuscript, while all other funding bodies had no such roles. There are four authors from LMS, YG, YM, DD and OZ, all of whom had an active role in the project. YG and DD conceived and designed the TGex software; YG and YM developed the TGex code; YG, DD, YM and OZ collaborated with Weizmann institute authors in the integration of the Weizmann's GeneCards Suite knowledgebase and interpretation algorithms into TGex; DD collaborated with other authors in the analysis and interpretation of the use-cases data; DD and YG collaborated with SF from Weizmann in writing and revising the manuscript.

Availability of data and materials

The SNV example (Fig. 2, Fig. 3) is available for each new TGex account as a demo case. The latest TGex documentation is available at http://tgex.genecards.org/user-guide/

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

YG and YM are CEO and Vice President respectively (and co-founders) of LifeMap Sciences Inc., California, USA (LMS), OZ is the senior software engineer and DD is the clinical genetics product manager at LMS. LMS holds an exclusive license from the Yeda Research and Development Company Ltd., the commercial arm of the Weizmann Institute of Science (WIS), to market GeneCards Suite products, which are derived from the research of DL's group at WIS, which is supported, among others, by a grant from LMS. SF, MS, RB, MT, and TIS are members of DL's group. All authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Author details

[1]Clinical Genetics, LifeMap Sciences Inc., Marshfield, MA 02050, USA. [2]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel. [3]Department of Gastroenterology, Tel-Aviv Sourasky Medical Center, Tel-Aviv, Israel. [4]Faculty of Medicine, Tel Aviv University, Tel-Aviv, Israel. [5]Genetic and Metabolic Central Laboratory, Birth Defect Prevention Research Institute, Maternal and Child Health Hospital, Children's Hospital of Guangxi Zhuang Autonomous Region, Nanning 530002, China. [6]Department of Medical Genetics and Molecular Diagnostic Laboratory, Shanghai Children's Medical Center, Shanghai Jiao Tong University School of Medicine, Shanghai 200127, China. [7]Department of Neurology, Harvard Medical School, Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA 02115, USA.

Received: 16 July 2019 Accepted: 15 December 2019
Published online: 30 December 2019

References

1. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013;155(1):27–38.
2. Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun. 2015;6.
3. Lappalainen T, Scott AJ, Brandt M, Hall IM. Genomic analysis in the age of human genome sequencing. Cell. 2019;177(1):70–84.
4. Ormondroyd E, Mackley MP, Blair E, Craft J, Knight JC, Taylor JC, et al. "not pathogenic until proven otherwise": perspectives of UK clinical genomics professionals toward secondary findings in context of a genomic medicine multidisciplinary team and the 100,000 Genomes Project. Genet Med. 2018;20(3):320–8.
5. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 2013;15(7):565–74.
6. Jelin AC, Vora N. Whole exome sequencing: applications in prenatal genetics. Obstet Gynecol Clin N Am. 2018;45(1):69–81.
7. Deciphering Developmental Disorders S. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2015;519(7542):223–8.
8. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med. 2013;369(16):1502–11.
9. Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, Quintero-Rivera F, et al. Clinical exome sequencing for genetic identification of rare Mendelian Disorders. Jama-J Am Med Assoc. 2014;312(18):1880–7.
10. Tan TY, Dillon OJ, Stark Z, Schofield D, Alam K, Shrestha R, et al. Diagnostic impact and cost-effectiveness of whole-exome sequencing for ambulant children with suspected monogenic conditions. JAMA Pediatr. 2017;171(9):855–62.
11. Stark Z, Tan TY, Chong B, Brett GR, Yap P, Walsh M, et al. A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. Genet Med. 2016;18(11):1090–6.
12. Trujillano D, Bertoli-Avella AM, Kumar Kandaswamy K, Weiss ME, Koster J, Marais A, et al. Clinical exome sequencing: results from 2819 samples reflecting 1000 families. Eur J Hum Genet. 2017;25(2):176–82.
13. Dragojlovic N, Elliott AM, Adam S, van Karnebeek C, Lehman A, Mwenifumbo JC, et al. The cost and diagnostic yield of exome sequencing for children with suspected genetic disorders: a benchmarking study. Genet Med. 2018;20(9):1013–21.
14. Shashi V, McConkie-Rosell A, Rosell B, Schoch K, Vellore K, McDonald M, et al. The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. Genet Med. 2014;16(2):176–82.
15. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genet Med. 2017;19(2):209–14.
16. Cornish A, Guda C. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. Biomed Res Int. 2015.
17. McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier JB, et al. Choice of transcripts and software has a large effect on variant annotation. Genome Med. 2014;6.
18. Schmitz-Abe K, Li Q, Rosen SM, Nori N, Madden JA, Genetti CA, et al. Unique bioinformatic approach and comprehensive reanalysis improve diagnostic yield of clinical exomes. Eur J Hum Genet. 2019.
19. Taylor JC, Martin HC, Lise S, Broxholme J, Cazier JB, Rimmer A, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet. 2015;47(7):717–26.
20. Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet Med. 2017;19(2):249–55.
21. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing (vol 15, pg 565, 2013). Genet Med. 2017;19(5):606.
22. Auffray C, Griffin JL, Khoury MJ, Lupski JR, Schwab M. Ten years of genome medicine. Genome Med. 2019;11(1):7.
23. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinformatics. 2016;54:1301–13.
24. Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, et al. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res. 2017;45(D1):D877–D87.
25. Belinky F, Bahir I, Stelzer G, Zimmerman S, Rosen N, Nativ N, et al. Non-redundant compendium of human ncRNA genes in GeneCards. Bioinformatics. 2013;29(2):255–61.
26. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford). 2017;2017. bax028, https://doi.org/10.1093/database/bax028.
27. Stelzer G, Plaschkes I, Oz-Levi D, Alkelai A, Olender T, Zimmerman S, et al. VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. BMC Genomics. 2016;17(Suppl 2):444.
28. Tomar S, Sethi R. Lai PS. European Journal of Human Genetics: Specific phenotype semantics facilitate gene prioritization in clinical exome sequencing; 2019.
29. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nat Rev Genet. 2017;18(10):599–612.
30. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly Austin 2012;6(2):80–92.
31. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a model for Genotoxic chemical mutational studies with a new Program. SnpSift Front Genet. 2012;3:35.
32. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93.
33. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285.
34. Fu WQ, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493(7431):216–20.
35. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.

36. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29(1):308–11.
37. Cai N, Bigdeli TB, Kretzschmar WW, Li Y, Liang J, Hu J, et al. 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. Sci Data. 2017;4:170011.
38. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062–D7.
39. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49(2):170–4.
40. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45(D1): D777–D83.
41. Lott MT, Leipzig JN, Derbeneva O, Xie HM, Chalkia D, Sarmady M, et al. mtDNA Variation and Analysis Using Mitomap and Mitomaster. Curr Protoc Bioinformatics. 2013;44:1.23.1–6.
42. Mastermind Genomic Search Engine. https://www.genomenon.com/mastermind.
43. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM (R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(D1):D789–D98.
44. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
45. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc 2016;11(1):1–9.
46. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7(8):575–6.
47. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553–61.
48. Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15(7):901–13.
49. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum Mutat. 2016;37(3):235–41.
50. Smit A, Hubley, R & Green, P. RepeatMasker Open-4.0. updated 2013–2015. http://www.repeatmasker.org.
51. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, 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):405–24.
52. Kohler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine JP, et al. Expansion of the human phenotype ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2019;47(D1):D1018–D27.
53. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32:D267–D70.
54. Oz-Levi D, Weiss B, Lahad A, Greenberger S, Pode-Shakked B, Somech R, et al. Exome sequencing as a differential diagnosis tool: resolving mild trichohepatoenteric syndrome. Clin Genet. 2015;87(6):602–3.
55. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11 0 1–33.
56. Chen R, Mias GI, Li-Pook-Than J, Jiang LH, Lam HYK, Chen R, et al. Personal Omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012;148(6):1293–307.
57. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, et al. Database resources of the National Center for biotechnology information. Nucleic Acids Res. 2019;47(D1):D23–D8.
58. Turnbull C. Introducing whole-genome sequencing into routine cancer care: the genomics England 100 000 Genomes Project. Ann Oncol. 2018;29(4):784–7.
59. Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ. 2018;361:k1687.
60. Hurles ME, Dermitzakis ET, Tyler-Smith C. The functional impact of structural variation in humans. Trends Genet. 2008;24(5):238–45.
61. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013;14(2):125–38.
62. Mizuguchi T, Suzuki T, Abe C, Umemura A, Tokunaga K, Kawai Y, et al. A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. J Hum Genet. 2019;64(5):359–68.
63. Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, et al. De novo mutations in regulatory elements in neurodevelopmental disorders. Nature. 2018;555(7698):611–6.
64. Caspar SM, Dubacher N, Kopps AM, Meienberg J, Henggeler C, Matyas G. Clinical sequencing: from raw data to diagnosis with lifetime value. Clin Genet. 2018;93(3):508–19.
65. Hintzsche JD, Robinson WA, Tan AC. A survey of computational tools to analyze and interpret whole exome sequencing data. Int J Genomics. 2016.
66. Belinky F, Nativ N, Stelzer G, Zimmerman S, Iny Stein T, Safran M, et al. PathCards: multi-source consolidation of human biological pathways. Database (Oxford). 2015;2015.
67. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods. 2015;12(9):841–3.
68. Robinson PN, Kohler S, Oellrich A, Wang K, Mungall CJ, Lewis SE, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. 2014;24(2):340–8.
69. IVA. http://www.ingenuity.com/products/variant-analysis.
70. Singleton MV, Guthery SL, Voelkerding KV, Chen K, Kennedy B, Margraf RL, et al. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Am J Hum Genet. 2014;94(4):599–610.
71. James RA, Campbell IM, Chen ES, Boone PM, Rao MA, Bainbridge MN, et al. A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics. Genome Med. 2016;8.
72. Castori M, Fiorillo C, Agolini E, Sacco M, Minetti C, Novelli A, et al. Primary muscle involvement in a 15-year-old girl with the recurrent homozygous c. 362dupC variant in FKBP14. Am J Med Genet A. 2019;179(2):317–21.
73. Wang Q, Yang JX, Liu Y, Li XP, Luo FW, Xie JS. A novel SLC6A8 mutation associated with intellectual disabilities in a Chinese family exhibiting creatine transporter deficiency: case report. Bmc Med Genet. 2018;19.
74. Morlino S, Alesi V, Cali F, Lepri FR, Secinaro A, Grammatico P, et al. LTBP2-related "Marfan-like" phenotype in two Roma/gypsy subjects with the LTBP2 homozygous p.R299X variant. Am J Med Genet A. 2019;179(1):104–12.
75. Chen F, Yuan H, Wu W, Chen S, Yang Q, Wang J, et al. Three additional de novo CTCF mutations in Chinese patients help to define an emerging neurodevelopmental disorder. Am J Med Genet C Semin Med Genet. 2019.
76. Pascolini G, Agolini E, Majore S, Novell A, Grammatico P, Digilio MC. Helsmoortel-Van der Aa syndrome as emerging clinical diagnosis in intellectually disabled children with autistic traits and ocular involvement. Eur J Paediatr Neuro. 2018;22(3):552–7.
77. Gambardella S, Ferese R, Scala S, Carboni S, Biagioni F, Emiliano G, et al. Mitochondrial Serine Protease HTRA2 p.G399S in a Female with Di George Syndrome and Parkinson's Disease. Parkinsons Dis. 2018;2018:5651435.
78. Hazkani I, Motiei M, Betzer O, Sadan T, Bragilovski D, Lubimov L, et al. Can molecular profiling enhance radiotherapy? Impact of personalized targeted gold nanoparticles on radiosensitivity and imaging of adenoid cystic carcinoma. Theranostics. 2017;7(16):3962–71.
79. Suwinski P, Ong C, Ling MHT, Poh YM, Khan AM, Ong HS. Advancing personalized medicine through the application of whole exome sequencing and big data analytics. Front Genet. 2019;10.
80. Alkelai A, Olender T, Haffner-Krausz R, Tsoory MM, Boyko V, Tatarskyy P, et al. A role for TENM1 mutations in congenital general anosmia. Clin Genet. 2016;90(3):211–9.
81. Alkelai A, Olender T, Dode C, Shushan S, Tatarskyy P, Furman-Haran E, et al. Next-generation sequencing of patients with congenital anosmia. Eur J Hum Genet. 2017;25(12):1377–87.
82. Heimer G, Eyal E, Zhu X, Ruzzo EK, Marek-Yagel D, Sagiv D, et al. Mutations in AIFM1 cause an X-linked childhood cerebellar ataxia partially responsive to riboflavin. Eur J Paediatr Neurol. 2018;22(1):93–101.
83. Heimer G, Oz-Levi D, Eyal E, Edvardson S, Nissenkorn A, Ruzzo EK, et al. TECPR2 mutations cause a new subtype of familial dysautonomia like hereditary sensory autonomic neuropathy with intellectual disability. Eur J Paediatr Neurol. 2016;20(1):69–79.
84. Yang C, Xu Y, Yu M, Lee D, Alharti S, Hellen N, et al. Induced pluripotent stem cell modelling of HLHS underlines the contribution of dysfunctional

NOTCH signalling to impaired cardiogenesis. Hum Mol Genet. 2017;26(16): 3031–45.
85. Einhorn Y, Weissglas-Volkov D, Carmi S, Ostrer H, Friedman E, Shomron N. Differential analysis of mutations in the Jewish population and their implications for diseases. Genet Res. 2017;99.
86. Ekhilevitch N, Kurolap A, Oz-Levi D, Mory A, Hershkovitz T, Ast G, et al. Expanding the MYBPC1 phenotypic spectrum: a novel homozygous mutation causes arthrogryposis multiplex congenita. Clin Genet. 2016;90(1):84–9.
87. Jia Z, Mao FB, Wang L, Li MZ, Shi YY, Zhang BR, et al. Whole-exome sequencing identifies a de novo mutation in TRPM4 involved in pleiotropic ventricular septal defect. Int J Clin Exp Patho. 2017;10(5):5092–104.
88. Bafunno V, Firinu D, D'Apolito M, Cordisco G, Loffredo S, Leccese A, et al. Mutation of the angiopoietin-1 gene (ANGPT1) associates with a new type of hereditary angioedema. J Allergy Clin Immunol. 2018;141(3):1009–17.
89. Zhang L, Jia Z, Mao F, Shi Y, Bu RF, Zhang B. Whole-exome sequencing identifies a somatic missense mutation of NBN in clear cell sarcoma of the salivary gland. Oncol Rep. 2016;35(6):3349–56.
90. Azim MK. Mehnaz A. Mujtaba G. Exome sequencing identifies a novel frameshift variant causing hypomagnesemia with secondary hypocalcemia. CEN Case Rep: Ahmed JZ; 2018.
91. Carneiro TN, Krepischi AC, Costa SS. Tojal da Silva I, Vianna-Morgante AM, Valieris R, et al. Utility of trio-based exome sequencing in the elucidation of the genetic basis of isolated syndromic intellectual disability: illustrative cases Appl Clin Genet. 2018;11:93–8.
92. Syama A, Sen S, Kota LN, Viswanath B, Purushottam M, Varghese M, et al. Mutation burden profile in familial Alzheimer's disease cases from India. Neurobiol Aging. 2018;64:158 e7–e13.
93. Homma TK, Krepischi ACV, Furuya TK, Honjo RS, Malaquias AC, Bertola DR, et al. Recurrent copy number variants associated with Syndromic Short stature of unknown cause. Horm Res Paediatr. 2018;89(1):13–21.
94. Fidalgo F, Rodrigues TC, Silva AG, Facure L, de Sa BC, Duprat JP, et al. Role of rare germline copy number variation in melanoma-prone patients. Future Oncol. 2016;12(11):1345–57.
95. Feliubadalo L, Tonda R, Gausachs M, Trotta JR, Castellanos E, Lopez-Doriga A, et al. Benchmarking of whole exome sequencing and ad hoc designed panels for genetic testing of hereditary Cancer. Sci Rep. 2017;7:37984.
96. Valayannopoulos V, Boddaert N, Chabli A, Barbier V, Desguerre I, Philippe A, et al. Treatment by oral creatine, L-arginine and L-glycine in six severely affected patients with creatine transporter defect. J Inherit Metab Dis. 2012; 35(1):151–7.
97. Neveling K, Feenstra I, Gilissen C, Hoefsloot LH, Kamsteeg EJ, Mensenkamp AR, et al. A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. Hum Mutat. 2013;34(12):1721–6.
98. Sawyer SL, Hartley T, Dyment DA, Beaulieu CL, Schwartzentruber J, Smith A, et al. Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. Clin Genet. 2016;89(3): 275–84.
99. Weitzel JN, Blazer KR, MacDonald DJ, Culver JO, Offit K. Genetics, genomics, and cancer risk assessment: state of the art and future directions in the era of personalized medicine. CA Cancer J Clin. 2011;61(5):327–59.
100. Jasperson KW, Tuohy TM, Neklason DW, Burt RW. Hereditary and familial Colon Cancer. Gastroenterology. 2010;138(6):2044–58.
101. Valle L, Vilar E, Tavtigian SV, Stoffel EM. Genetic predisposition to colorectal cancer: syndromes, genes, classification of genetic variants and implications for precision medicine. J Pathol. 2019;247(5):574–88.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.