data

MDPI

Data Descriptor

Database for Gene Variants and Metabolic Networks Implicated in Familial Gastroschisis

Víctor M. Salinas-Torres [1,*], Hugo L. Gallardo-Blanco [1,*], Rafael A. Salinas-Torres [2] and Laura E. Martínez de Villarreal [1,*]

[1] Department of Genetics, School of Medicine and University Hospital "Dr. José Eleuterio González", Universidad Autónoma de Nuevo León, Ave. Madero y Gonzalitos S/N Col. Mitras Centro, Monterrey CP 64460, Nuevo León, Mexico

[2] Department of Systems and Computing, Instituto Tecnológico de Tijuana, Calzada del Tecnológico S/N Fracc. Tomas Aquino, Tijuana CP 22414, Baja California, Mexico

* Correspondence: vm_salinas7@hotmail.com (V.M.S.-T.); hugo.gallardobl@uanl.edu.mx (H.L.G.-B.); laelmar@yahoo.com.mx (L.E.M.d.V.); Tel.: +52-81-8329-4217 (V.M.S.-T. & H.L.G.-B. & L.E.M.d.V.); Fax: +52-81-8348-3509 (V.M.S.-T. & H.L.G.-B. & L.E.M.d.V.)

Received: 22 May 2019; Accepted: 9 July 2019; Published: 11 July 2019

Abstract: Gastroschisis is one of the most prevalent human birth defects concerning the ventral body wall development. Recent research has given a better understanding of gastroschisis pathogenesis through the identification of multiple novel pathogenetic pathways implicated in ventral body wall closure. Deciphering the underlying genetic factors segregating among familial gastroschisis allows better detection of novel susceptibility variants than the screening of pooled unrelated cases and controls, whereas bioinformatic-aided analysis can help to address new insights into human biology and molecular mechanisms involved in gastroschisis. Technological advances in DNA sequencing (Next Generation Sequencing), computing power, and machine learning techniques provide opportunities to the scientific communities to assess significant gaps in research and clinical practice. Thus, in an effort to study the role of gene variation in gastroschisis, we employed whole exome sequencing in a Mexican family with recurrence for gastroschisis. Stringent bioinformatic analyses were implemented to identify and predict pathogenetic networks comprised of potential gastroschisis predispositions. This is the first database for gene variants and metabolic networks implicated in familial gastroschisis. The dataset provides information on gastroschisis annotated genes, gene variants, and metabolic networks and constitutes a useful source to enhance further investigations in gastroschisis.

Dataset: Data presented here is deposited in doi.org/10.5281/zenodo.3270692

Keywords: alleles; bioinformatics; gastroschisis; genes; genetics; recurrence; whole exome sequencing

1. Introduction

Gastroschisis represents one of the leading human birth defects affecting ~1:2500 live births with an alarming increase in its prevalence [1]. In addition to risk factors such as maternal smoking and young maternal age [1], there is emerging evidence for a genetic component in gastroschisis etiology [2–6]. Heritable factors in gastroschisis were estimated to be 3% adjusted for probands, 4.3% in gastroschisis cases followed by a subsequent affected pregnancy, and overall recurrence risk of 5.7% [3]. Moreover, candidate gene analysis has been performed identifying gene variants and pathways related to xenobiotic metabolism, regulation of cell adhesion, regulation of gene expression, inflammatory response, regulation of vascular development, keratinization, left-right symmetry, epigenetic, ubiquitination, and regulation of protein synthesis [4–6].

Within this framework, the aim of the present study is to collect genes, gene variants, and metabolic networks implicated in gastroschisis by employing whole exome sequencing and bioinformatic analysis in a Mexican family with recurrence for gastroschisis. Additionally, the primary aim for collecting this database is to contribute to research and diagnostic of gastroschisis in future studies as well as to promote relevant gene variants and metabolic networks that can be validated in studies based on cases and controls.

As a result of this work, gene variants and metabolic networks showing plausibility with gastroschisis from two affected half-sisters, mother, and father of the proband have been described and fully available to scientific communities to review and verify our proposed model [6,7]. Furthermore, our dataset could be useful to enhance the data transparency or reusability for further investigations and can also be utilized for studying and addressing new insights into human biology and molecular mechanisms involved in gastroschisis.

2. Methods

2.1. Ethics Statement

The present study was approved by the Institutional Ethics Committee from the School of Medicine and University Hospital "Dr. José Eleuterio González", Universidad Autónoma de Nuevo León, México (Approval: 21 September 2017, GN17-00002). Written informed consent was obtained from the parents.

2.2. Experiment Design

This work represents a family-based study involving two affected half-sisters with gastroschisis, each patient with different father and the same mother (the parents were unaffected). Centogene AG® (Rostock, Germany) performed a trio-based whole exome sequence (WES) for the affected index patient and her parents, whereas LC Sciences® (Houston, TX, USA) performed the WES for the affected half-sister (her father was not assessed).

From these reports, we retrieved the processed gene variants describing genomic data from each family member, which allowed building databases for further independent bioinformatic analyses:

- Prioritization of genes and gene variants by phenotypic relevance [8] as well as functional and impact properties [9].
- Gene functional enrichment analysis (computational model of biological systems) [10].
- Protein–protein interaction (PPI) network analysis (physical/functional network modeling) [11].
- Classification of gene variants by inheritance pattern (model scoring prediction) [8].
- Gene functional similarity analysis (candidate gene prioritization) [12].

Figure 1 illustrates the workflow for the experimental design and bioinformatic analysis in the present study.

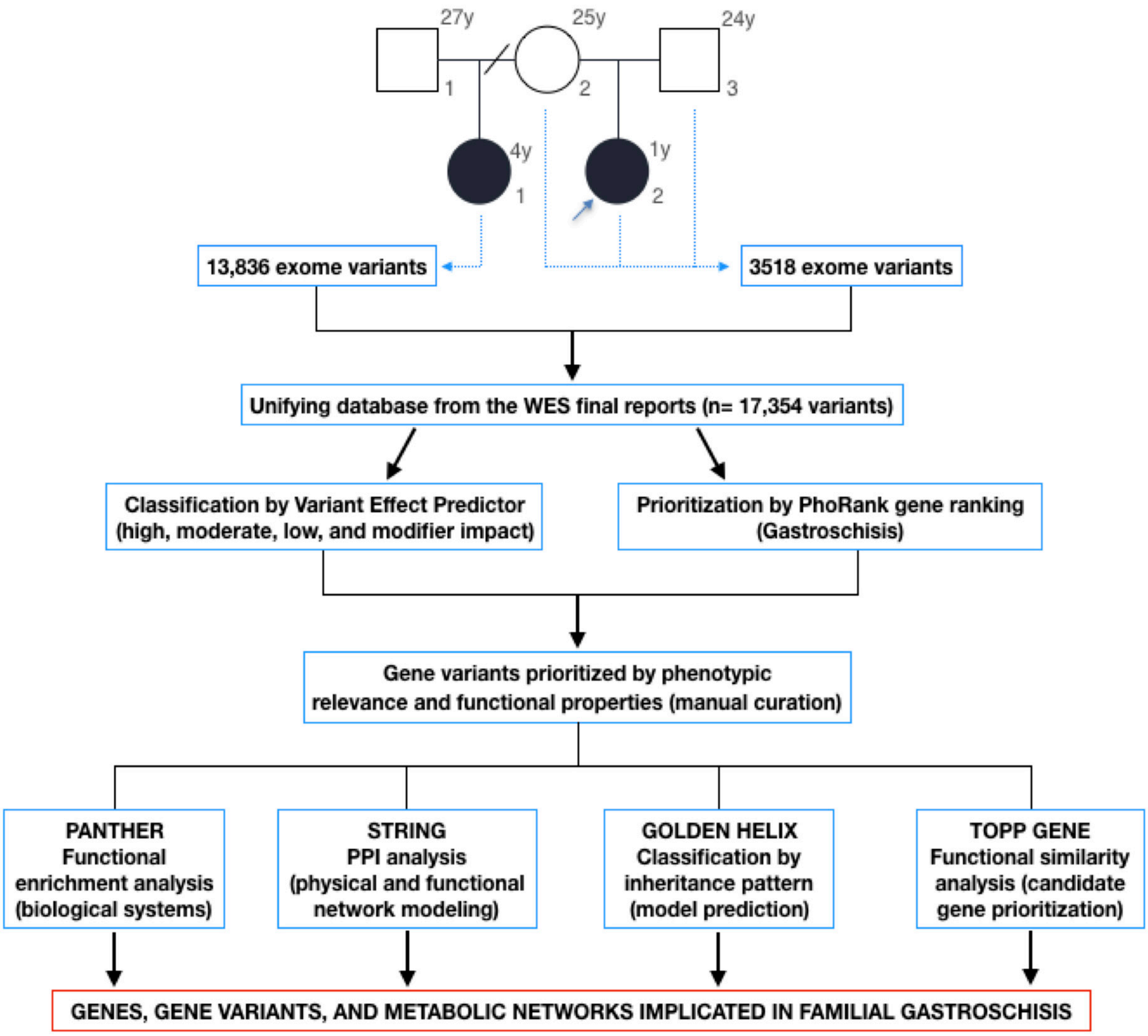
27y
25y
24y
1
2
3
4y
1
1y
2
13,836 exome variants
3518 exome variants
Unifying database from the WES final reports (n= 17,354 variants)
Classification by Variant Effect Predictor (high, moderate, low, and modifier impact)
Prioritization by PhoRank gene ranking (Gastroschisis)
Gene variants prioritized by phenotypic relevance and functional properties (manual curation)
PANTHER Functional enrichment analysis (biological systems)
STRING PPI analysis (physical and functional network modeling)
GOLDEN HELIX Classification by inheritance pattern (model prediction)
TOPP GENE Functional similarity analysis (candidate gene prioritization)
GENES, GENE VARIANTS, AND METABOLIC NETWORKS IMPLICATED IN FAMILIAL GASTROSCHISIS

Figure 1. Workflow of the experimental design and bioinformatic analysis in the present study. The retrieved whole exome sequence (WES) gene variants from the Mexican family with recurrence for gastroschisis were prioritized based on phenotypic relevance and functional properties by independent bioinformatics platforms collecting genes, gene variants, and metabolic networks implicated in familial gastroschisis. PPI: Protein–protein interaction.

2.3. *Bioinformatic Analysis*

2.3.1. Database Generation from Exome Variants

Once we retrieved all reported gene variants previously filtered (quality controls) and curated by Centogene AG® (Rostock, Germany) and LC Sciences® (Houston, TX, USA), we generated a unifying database including 17,354 gene variants.

Quality controls from Centogene AG® (Rostock, Germany), included end-to-end inhouse bioinformatics pipelines including base calling, primary filtering of low-quality reads and probable artefacts, and annotation of variants. For LC Sciences® (Houston, TX, USA) the following quality controls and bioinformatics pipeline were applied:

- Quality control: FastQC software (version 0.10.1).
- Alignment: Burrows-Wheeler Alignment software (version 0.7.10) and Sequence Alignment/Map tools software (version 0.1.19).
- Duplicates remove: Picard software (version 1.119).

- Single nucleotide variants (SNV) and insertions and deletions (INDELS) calling: Genome Analysis Toolkit software (version 3.7).
- SNV and INDELS annotation: SnpEff (version 4.1).

Annotation of genes and gene variants was based on the Human Genome annotation GRCh37/hg19.

2.3.2. Prioritization of Gene Variants by Phenotypic Relevance and Functional Properties

Ranking genes from the investigated phenotype was inquired by PhoRank Gene Ranking algorithm [8], which ranks genes based on their relevance to the specified phenotype "Gastroschisis" (only annotate mRNA transcripts, ontologies used; human phenotype ontology (HPO) 2017-12-12, gene ontology (GO) 2017-12-15, and online mendelian inheritance in man (OMIM) phenotype ontology 2017-06-15). Functional properties were inquired by variant effect predictor (VEP) (2018-08-31) [9], which assessed the potential deleterious effects (high, moderate, low, and modifier impact) of the exome variants based on a combined effect of the following algorithms: SIFT (sorting tolerant from intolerant), PolyPhen (polymorphism phenotyping), dbNSFP (database for nonsynonymous single nucleotide polymorphisms' (SNP) functional predictions), condel (Consensus deleteriousness score of missense single nucleotide variant), LoFtool (Loss-of-function mutations), MaxEntScan (splicing prediction), and BLOSUM62 (blocks of amino acid substitution matrix 62, conservation prediction). Then, manual curation and prioritization of genes and gene variants was based on the following:

- Genes and gene variants phenotypically relevant according to PhoRank Gene Ranking [8].
- Genes and gene variants classified as high, moderate, or modifier impact according to VEP [9].
- Genes and gene variants segregating among both affected half-sisters and the mother.

2.3.3. PPI Gene Network Modeling

PPI gene network modeling was based on a two-step process. First, a functional enrichment analysis was performed using the GO Consortium and Panther Classification System databases, which contain comprehensive information on the evolution and function of protein-coding genes from 104 completely sequenced genomes [10]. Second, gene pairs detected in two or more of the PPI including coexpression, protein homology, curated databases, gene neighborhood, or experimentally determined data sets were selected and included in the network modeling using String database 10.5 [11]. A gene–gene pairwise network was constructed using the PPI with a confidence score of 0.4 [6]. These databases include a "hierarchical view" as a structure of the most significant classifications and ontologies of the human genes and E-value statistics (p values of less than 0.05 false discovery rate (FDR)-adjusted) [10,11].

2.3.4. Classification by Inheritance Pattern

The annotation and filtration of gene variants, including de novo candidate variants, recessive and dominant models variant score reports, heterozygous compound genes score report, and inheritance classification report were generated based on Golden Helix SNP Variation Software (SVS) version 8.8 (With RefSeq Genes 105 Interim v1, NCBI, and ClinVar 2018-06-07, NCBI) [8]:

- De novo candidate variants. With the script "de novo candidate variants," the variants were classified by de novo candidate variants.
- Recessive model variant score report. With the script "score variants by recessive model, 2014-03-05" the variants were scored based on the expected recessive model inheritance pattern.
- Dominant model variant score report. With the script "score variants by dominant model, 2014-07-01" the variants were scored based on the expected dominant model inheritance pattern.
- Heterozygous compound genes score report. With the script "score compound heterozygous regions (2014-12-01)" the variants were scored based on the compound heterozygous regions by recessive model (Gene Track: RefSeqGenes63-UCSC_2014-02-16_GRCh_37_Homo_sapiens.tsf:1).

- Inheritance classification. With the script "classify by inheritance pattern, 2014-03-05" the variants were classified by inheritance.

Recessive model variant score report, was obtained based on how well each variant follows the expected recessive model inheritance pattern, generating one score for each nuclear family. The outcomes are addressed by a recessive model score formula:

$$\text{score} = \text{numHetParents} + \text{numAltAffected} + \text{numRefUnaffected}/\text{numGenotypes}$$

The total number of genotypes that follow the recessive model pattern is divided by the total number of considered genotypes. If missing values are treated as reference then missing values are included in the denominator. Otherwise, only called genotypes are included in denominator [8].

Dominant model variant score report was obtained based on how well each variant followed the expected dominant model pattern based on case/control status, generating, in the case of a pedigree spreadsheet, one score per family. The outcomes are addressed by dominant model score formulas:

$$S = x + y/\text{number of samples}$$

The dominant model score: x = number of heterozygous cases; y = number of homozygous reference controls.

$$S_{wt} = x + y - 0.5 * z/\text{number of samples} \quad (1)$$

If the option to treat missing genotypes as reference is not selected then a weighted score is also included in the output (wt = weighted, x = number of heterozygous cases, y = number of homozygous reference controls, z = number of missing genotypes). The dialog presented will depend on whether or not the spreadsheet has pedigree information [8].

Finally, the heterozygous compound genes score report calculates the number compound heterozygous inheritance events within each gene region. The generation of this report requires (a) children have a heterozygous genotype, (b) one parent has a copy of the alternate allele, (c) the parental source of the alternate allele is known and not ambiguous, (d) a child has two heterozygous genotypes within the same gene where the alternate allele is inherited from each parent at a minimum of two different loci, and (e) optionally, a child's heterozygous genotype can be counted when one of the parent has two copies of the alternate allele [8].

2.3.5. Candidate Gene Model Generation

Candidate gene model generation was based on a three-step process. First, manual curation of genes identified as (a) high impact, (b) segregating among both affected half-sisters and the mother, (c) highly significant functional enrichment, (d) high connectivity/direct PPI, and (e) highest score for dominant, recessive, and heterozygous compound models [6]. Second, gene functional similarity analysis as well as candidate gene prioritization was performed using ToppGene Suite database, which combines an overall score using statistical meta-analysis including *p*-value (FDR-adjusted) of each annotation of a test gene derived by random sampling from the whole genome [12]. Third, a final manual curation of GO-biological processes and pathways were selected based on their proximity and plausibility to the phenotype, including previous pathways associated to gastroschisis [4,5]. These implemented analyses allowed us to identify and predict pathogenetic networks comprised by potential gastroschisis predispositions [6].

3. Data Description

3.1. Dataset Characteristics and Format

The released dataset comprises a set of fifteen tables available in CSV files, which provides annotation of genes and gene variants based on the Human Genome annotation GRCh37/hg19 [7]. Table 1 summarizes the available data including its description and data/file types.

Table 1. List of data available in the created dataset.

Table	Data	Description	Data Type	File Format
1	Processed exome variants database for the index case, mother, and father	Genotypes from 3518 exome variants [1]	Tabular	CSV
2	Processed exome variants database for the affected half-sister	Genotypes from 13,836 exome variants [2]	Tabular	CSV
3	Exome variants database in the family with gastroschisis	Unifying database of 17,354 exome variants	Tabular	CSV
4	Exome variants database cosegregating in the family with gastroschisis	Unifying database of 214 exome variants	Tabular	CSV
5	Exome variants database classified by Variant Effect Predictor	Annotation impact of exome variants [3]	Tabular	CSV
6	Exome variants database classified by PhoRank gene ranking	Annotation of 189 exome variants [4]	Tabular	CSV
7	Exome variants database prioritized by phenotypic relevance and functional properties	Database of 428 gene variants [3,4]	Tabular	CSV
8	Exome gene list input for functional enrichment and protein–protein interaction analysis	Database of 432 genes [5]	Tabular	CSV
9	Exome variants database classified by dominant model score	Database of 212 exome variants [4]	Tabular	CSV
10	Exome variants database classified by recessive model score	Database of 212 exome variants [4]	Tabular	CSV
11	Exome variants database classified by heterozygous compound model score	Database of 276 exome variants [4]	Tabular	CSV
12	Exome variants database classified by inheritance	Database of 212 exome variants [4]	Tabular	CSV
13	Exome gene database by candidate gene prioritization	Database of 2670 GO-biological processes and pathways [6]	Tabular	CSV
14	Whole exome sequencing final report for the index case, mother, and father	List of 3518 exome variants [1]	Tabular	CSV
15	Whole exome sequencing final report for the affected half-sister	List of 13,836 exome variants [2]	Tabular	CSV

[1] Exome variants reported by Centogene AG® (Rostock, Germany). [2] Exome variants reported by LC Sciences® (Houston, TX, USA). [3] Based on Ensembl—VEP [9]. [4] Based on SVS—PhoRank gene ranking by Golden Helix® [8]. [5] Gene list input for analysis based on Panther and String [10,11]. [6] Based on ToppGene Suite [12]. GO (gene ontology).

3.2. Dataset Description

The file "Table 1.csv" provides processed information about 3518 exome variants including 3054 SNV and 464 INDELS for the trio-based WES. Column "A" provides information that correspond to chromosome, start position, reference allele, and alternative allele. Columns "B," "C," and "D" describe the genotypes involving the index case, father, and mother, respectively.

The file "Table 2.csv" provides processed information about 13,836 exome SNV for the affected half-sister. Column "A" provides information that correspond to chromosome, start position, reference allele, and alternative allele. Column "B" describes the genotypes for the affected half-sister.

The file "Table 3.csv" provides a unifying database of the 17,354 exome variants including 16,890 SNV and 464 INDELS reported in the family with gastroschisis. Columns "A," "B," and "C" provide info that correspond to chromosome, start position, reference allele, and alternative allele, respectively. Column "D," "E," "F," "G," and "H" provides information that correspond to dbSNP (single nucleotide polymorphism database) variant ID, gene, reference and variant allele, and DNA and protein change, respectively. Data described as "." or "NA" was not available.

The file "Table 4.csv" provides a unifying database for 214 exome SNV cosegregating in the family with gastroschisis. Row 1 (intersecting with columns "A" to "G") provides info that correspond to family members, family ID, patient ID, father ID, mother ID, sex, and affection status, respectively. Cosegregating gene variants were described in row 1 intersecting with columns "H" to "HM". These columns provide information that correspond to genotypes for each family member (half-sister in row 2, index case in row 3, mother in row 4, and father in row 5). Data for family ID "1" involves one family and "?" implies unknown. Data for sex "1" and "0" stands for female and male respectively, whereas data for affected status "1" and "0" stands for affected and unaffected respectively.

The file "Table 5.csv" provides a database for 1810 annotation impact variants classified by VEP [9]. This database includes the following impact annotations: High 22, moderate 293, modifier 1178, and low 317. Row 1 (intersecting with columns "A" to "BU") provides information that correspond to uploaded variation, location, allele, consequence, impact, symbol, gene, feature type, feature, biotype, exon, intron, HGVS (human genome variation society) DNA and protein change, DNA position, coding sequence, protein position, amino acids, codons, existing variations, distance, strand, flags, symbol source, HGNC (HUGO gene nomenclature committee) ID, TSL (transcript support level), APPRIS (annotating principal splice isoforms), consensus coding sequence, ENSP (ensemble protein), Swissprot, Translated EMBL, UniProt archive, SIFT, PolyPhen, domains, HGVS offset, AF (allele frequency), African AF, American AF, East Asian AF, European AF, South Asian AF, African American AF, European American AF, gnomAD (genome aggregation database) AF, gnomAD African AF, gnomAD American AF, gnomAD Ashkenazi Jewish AF, gnomAD East Asian AF, gnomAD Finnish AF, gnomAD non-Finnish AF, gnomAD other AF, gnomAD South Asian AF, ClinVar significance, somatic, phenotype, PubMed, motif name, motif position, high information position, motif score change, MaxEntScan alt, diff, and ref, ENSP DNA and protein change, LoFtool, ClinVar ID, amino acid, miRNA, BLOSUM62, ada score, rf score, and Condel, respectively.

The file "Table 6.csv" provides a database for 189 genes classified by SVS—PhoRank gene ranking by Golden Helix® [8]. Columns "A" to "J" provide information that correspond to gene, chromosome, start position, stop position, number of markers, ranks, scores, rank scores, and pathways, respectively. Column "H" depicts the rank score obtained from ranks and scores for each gene (column "F" × "G"). Column "I" depicts the rank score obtained from the number of markers, ranks, and scores for each gene (column "E" × "F" × "G"). Data described as "?" implies unknown.

The file "Table 7.csv" provides a database for 428 gene variants prioritized by phenotypic relevance and functional properties [8,9]. This database includes the following SNV impact annotations: High 9, moderate 100, and modifier 319. Columns "A," "B," and "C" provide information that correspond to variant ID, gene/location, and impact, respectively.

The file "Table 8.csv" provides a database for 432 genes for functional enrichment and protein–protein analysis based on Panther and String platforms [10,11]. This database includes 41 genes with significance by Golden Helix SVS [8], 9 genes with high impact annotation, 97 genes with moderate impact annotation, and 285 genes with modifier impact annotation [9]. Columns "A" and "B" provide information that correspond to gene and impact, respectively.

The file "Table 9.csv" provides a database for 212 gene variants classified by dominant model score based on SVS—PhoRank gene ranking by Golden Helix® [8]. Columns "A" to "Q" provide information that correspond to variant ID, chromosome, start position, dbSNP variant ID, marker, gene, reference allele, alternative allele, reference alleles, total number of dominant families, sum score, number of carriers in affected, affected samples, dominant model score, respectively. Data described as "?" implies unknown.

The file "Table 10.csv" provides a database for 212 gene variants classified by recessive model score based on SVS—PhoRank gene ranking by Golden Helix® [8]. Columns "A" to "N" provide information that correspond to variant ID, chromosome, start position, dbSNP variant ID, marker, gene, reference allele, alternative allele, reference alleles, drop, total number of recessive families, sum score, and recessive model score, respectively. Data described as "?" implies unknown.

The file "Table 11.csv" provides a database for 276 genes classified by heterozygous compound model score based on SVS—PhoRank gene ranking by Golden Helix® [8]. Columns "A" to "K" provide information that correspond to genes, chromosome, start position, stop position, gene name, transcript name, strand, known gene ID, known gene description, reference sequence summary, and total compound heterozygous, respectively. Columns "L" to "W" describe compound heterozygous, heterozygous from father, heterozygous from mother, total inherited heterozygous, ambiguous heterozygous, and Mendelian error, for the half-sister and the index case, respectively. Data described as "?" implies unknown.

The file "Table 12.csv" provides a database for 212 gene variants classified by inheritance models based on SVS—PhoRank gene ranking by Golden Helix® [8]. This database include one heterozygous de novo variant, one maternal de novo variant, 7 homozygous both variants, 183 heterozygous maternal variants, 6 heterozygous paternal variants, and 9 heterozygous either variants. Columns "A" to "J" provide information that correspond to variant ID, chromosome, start position, dbSNP variant ID, marker, gene, reference allele, alternative allele, reference alleles, and model, respectively. Data described as "?" implies unknown.

The file "Table 13.csv" provides a database for 2670 terms including 2095 GO-biological processes and 575 pathways based on ToppGene Suite platform [12]. The described terms include gene functional similarity analysis and candidate gene prioritization of genes identified as: (a) high impact, (b) segregating among both affected half-sisters and the mother, (c) highly significant functional enrichment, (d) high connectivity/direct PPI, and (e) highest score for dominant, recessive, and heterozygous compound models. Columns "A" to "K" provide information that correspond to category, GO and pathway ID, name, source, p-value, q-value Bonferroni, q-value FDR (false discovery rate) B&H, q-value FDR B&Y, hit count in query list, hit count in genome, and hit in query list, respectively.

The file "Table 14.csv" includes the whole exome sequencing final report by Centogene AG® (Rostock, Germany) for the index case, mother, and father. The 3518 exome variants available in the report also include information from the variant call format (VCF) file as well as in silico and bioinformatics pipeline for which columns "A" to "FP" provide info that correspond to chromosome-start position-reference-alternative, CompHetFlag, chromosome, genomic position, reference, variant, members with zygosity, index frequency, index read number, index quality, approved symbol, location, transcript, cDNA change, protein change, distance from exon, exonic function reference gene, population frequency max, exac03, mutation's phenotype, CentoMD clinical significance, CentoMD id/curator/date, PGMD (PharmacoGenomics mutation database) accession, PGMD gene, PGMD drug, PGMD description, ACMG (American college of medical genetics) recommendation (of gene), in silico summary, caddgt10 (combined annotation dependent depletion), phyloP46way_placental (phylogenetic p-values), HPO (human phenotype ontology) match terms, OMIM (online mendelian inheritance in man) phenotype (of gene), all phenotypes, WES (whole exome sequence) statistics, WES family count, genome family count, members affected, family members, zygosity, index VC info, index VC count, variant caller count, frequency, read number, quality, homopoly length, gene, variant ID, number within family, WES patcount, WES total patients, WES total families, WES 5 patients, WES 5 families, CES (clinical exome sequencing) statistics, CES patcount, CES total patients, CES family count, CES total families, CES 5 patients, CES 5 families, genome statistics, genome patcount, genome total patients, genome total families, genome 5 patients, genome 5 families, proton statistics, proton patcount, proton total patients, proton family count, proton total families, proton 5 patients, proton 5 families, Roche statistics, number of Roche patients, CentoMD statistics, SNP ID, snp138, avsnp142, clinvar_20150330, gwascatalog, cosmic70 (catalogue of somatic mutations in cancer), HGMD (human gene mutation database) accession, mutation type, HGMD list, HPO match ids, HPO ids, HPO terms, disease_description, HGMD phenotype (of gene), trait_association(gwas), function_description, targetscans, wgRna, MIM_id, MIM_phenotype_id, MIM_disease, GO_slim_biological_process, GO_slim_cellular_component, GO_slim_molecular_function, essential_gene, func_refgene, gene_refgene, genedetail_refgene, AAchange_refgene (Aminoacid), func_knowngene, gene_knowngene, genedetail_knowngene, exonicfunc_knowngene, AAchange_knowngene, func_ensgene, gene_ensgene, genedetail_ensgene, exonicfunc_ensgene, AAchange_ensgene, genomicsuperdups, dgvmerged, 1000 genomes_all, 1000genomes_African, 1000genomes_American, 1000genomes_East Asian, 1000genomes_European, 1000genomes_South Asian, esp6500siv2_all, esp6500siv2_ African American, esp6500siv2_European American, exac_African, exac_American, exac_East Asian, exac_Finnish, exac_non Finnish, exac_other, exac_South Asian, CG46, cg69, nci60, SIFT_score, SIFT_pred, Polyphen2_hdiv_score, Polyphen2_hdiv_pred, Polyphen2_hvar_score, Polyphen2_hvar_pred, LRT_score (likelihood ratio test), LRT_pred, MutationTaster_score, MutationTaster_pred, MutationAssessor_score, MutationAssessor_pred, FATHMM_score (functional analysis through hidden markov models), FATHMM_pred, RadialSVM_score (radial support vector machine), RadialSVM _pred, LR_score (logistic regression), LR_pred, vest3_score (variant effect scoring tool), cadd_raw, cadd_phred, gerp_rs (genomic evolutionary rate profiling), phylop100way_vertebrate, siphy_29way_logodds (site-specific phylogenetic analysis), gene_full_name, pathway (uniprot), pathway (consensuspathdb), expression (egenetics), expression (gnf/atlas), p (hi), p (rec), known_rec_info, vcf_pos_normalized, vcf_ref_normalized, and vcf_alt_normalized, respectively. Data described as "." or "NA" was not available.

The file "Table 15.csv" includes the whole exome sequencing final report by LC Sciences® (Houston, TX, USA) for the affected half-sister. The 13,836 exome SNV available in the report provides info from the VCF file for which columns "A" to "S" correspond to chromosome, start position, reference allele, alternative allele, quality, filter, information, annotation, annotation impact, gene name, feature type, transcript biotype, rank, HGVS DNA and protein change, description, format, and statistics, respectively. Data described as "." or "NA" was not available.

3.3. *Clinical Data for Assessed Family Members*

Figure 1 illustrate the pedigree of the assessed family members. The index case, a one-year-old female, was born at the 37th week of gestation by cesarean section after prenatal diagnosis of gastroschisis. Apgar score was 9, birth weight 2050 g (<3rd centile), and length 48 cm (15–50th centile). Her half-sister, a four-year-old child, was born at the 36th week of gestation by cesarean section after prenatal diagnosis of gastroschisis. Apgar score was 9, birth weight 1670 g, and length 42 cm (both were <3rd centile). Postnatally, in both cases, gastroschisis was confirmed as an isolated anomaly through the open right side of the umbilical ring and a primary closure of the abdominal wall was performed without further clinical complications.

Regarding the parents, consanguinity was ruled out; however, change in paternity was detected. The mother was 20- and 24-years-old at conception from her first and second pregnancy affected with fetal gastroschisis, respectively (prepregnancy body mass index was 18.4 and 19.2, respectively). The mother referred genitourinary infections during the second trimester of pregnancy in both cases as well as a limited preconceptional care (less than four clinical follow-ups in each pregnancy), and there was no consumption of folic acid-containing supplements. Tobacco smoking (1–2 cigarettes per day) and alcohol consumption (1–2 beer cans of 355 mL per week) at preconception and during the first trimester of pregnancy were detected in the mother and her 24-years-old husband (father of the index case).

4. Conclusions

- This work represents the first family-based study for recurrence in gastroschisis-deciphering novel susceptibility gene variants and metabolic networks.
- Genes and gene variants from WES data were prioritized by a multistep bioinformatics process including (a) phenotypic relevance by PhoRank Gene Ranking algorithm [8], (b) functional and impact properties by VEP [9], (c) GO functional enrichment analysis by Panther [10], (d) PPI network modeling analysis by String [11], (e) classification of gene variants by inheritance pattern by SVS—Golden Helix® [8], and (f) gene functional similarity analysis and candidate gene prioritization by ToppGene [12].
- Stringent bioinformatic analyses identified and predicted pathogenetic networks comprised of potential gastroschisis predispositions, addressing new insights into human biology and molecular mechanisms involved in gastroschisis [6].
- The dataset provides information on gastroschisis annotated genes, gene variants, and metabolic networks and constitutes a useful source to enhance further investigations in gastroschisis.

Author Contributions: Conceptualization, V.M.S.-T., and H.L.G.-B.; Recruitment of patients, V.M.S.-T., and R.A.S.-T.; Formal analysis and interpretation of data, V.M.S.-T., R.A.S.-T., and H.L.G.-B.; Writing original draft, V.M.S.-T.; Review and editing, H.L.G.-B., R.A.S.-T., and L.E.M.d V.; All authors read and approved the final manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors are grateful to the participating family for their valuable cooperation.

Conflicts of Interest: The authors declared no potential conflicts of interest and received no financial support concerning the research, authorship, or publication of this article.

Abbreviations

AA	Amino acid
ACMG	American college of medical genetics
AF	Allele frequency
APPRIS	Annotating principal splice isoforms
BLOSUM62	Blocks of amino acid substitution matrix 62
Caddgt10	Combined annotation dependent depletion
CES	Clinical exome sequencing
Condel	Consensus deleteriousness score of missense single nucleotide variant
Cosmic70	Catalogue of somatic mutations in cancer
dbNSFP	Database for nonsynonymous SNPs' functional predictions
dbSNP	Single nucleotide polymorphism database
ENSP	Ensembl protein
FATHMM	Functional analysis through hidden markov models
FDR	False discovery rate
GERP	Genomic evolutionary rate profiling
gnomAD	Genome aggregation database
GO	Gene ontology
HGMD	Human gene mutation database
HGNC	HUGO gene nomenclature committee
HGVS	Human genome variation society
HPO	Human phenotype ontology
Indels	Insertions and deletions
LoFtool	Loss-of-function mutations
LR	Logistic regression
LRT	Likelihood ratio test
OMIM	Online mendelian inheritance in man
PGMD	PharmacoGenomics mutation database
PhyloP46way_placental	Phylogenetic p-values

PolyPhen	Polymorphism phenotyping
PPI	Protein-protein interaction
RadialSVM	Radial support vector machine
SIFT	Sorting tolerant from intolerant
Siphy	Site-specific phylogenetic analysis
SNV	Single nucleotide variant
SVS	SNP Variation Software
TSL	Transcript support level
VCF	Variant call format
VEST	Variant effect scoring tool
VEP	Variant effect predictor
WES	Whole exome sequence

References

1. Salinas-Torres, V.M.; Salinas-Torres, R.A.; Cerda-Flores, R.M.; Martínez-de-Villarreal, L.E. Prevalence, mortality, and spatial distribution of gastroschisis in Mexico. *J. Pediatr. Adolesc. Gynecol.* **2018**, *31*, 232–237. [CrossRef] [PubMed]
2. Salinas-Torres, V.M.; Salinas-Torres, R.A.; Cerda-Flores, R.M.; Martínez-de-Villarreal, L.E. Evaluation of familial factors in a Mexican population-based setting with gastroschisis: Further evidence for an underlying genetic susceptibility. *J. Pediatr. Surg.* **2018**, *53*, 521–524. [CrossRef] [PubMed]
3. Salinas-Torres, V.M.; Salinas-Torres, R.A.; Cerda-Flores, R.M.; Martínez-de-Villarreal, L.E. Familial occurrence of gastroschisis: A population-based overview on recurrence risk, sex-dependent influence, and geographical distribution. *Pediatr. Surg. Int.* **2018**, *34*, 277–282. [CrossRef] [PubMed]
4. Salinas-Torres, V.M.; Salinas-Torres, R.A.; Cerda-Flores, R.M.; Martínez-de-Villarreal, L.E. Genetic variants conferring susceptibility to gastroschisis: A phenomenon restricted to the interaction with the environment? *Pediatr. Surg. Int.* **2018**, *34*, 505–514. [CrossRef] [PubMed]
5. Salinas-Torres, V.M.; Salinas-Torres, R.A.; Cerda-Flores, R.M.; Gallardo-Blanco, H.L.; Martínez-de-Villarreal, L.E. A clinical-pathogenetic approach on associated anomalies and chromosomal defects supports novel candidate critical regions and genes for gastroschisis. *Pediatr. Surg. Int.* **2018**, *34*, 931–943. [CrossRef] [PubMed]
6. Salinas-Torres, V.M.; Gallardo-Blanco, H.L.; Salinas-Torres, R.A.; Cerda-Flores, R.M.; Lugo-Trampe, J.J.; Villarreal-Martínez, D.Z.; Martínez de Villarreal, L.E. Bioinformatic analysis of gene variants from gastroschisis recurrence identifies multiple novel pathogenetic pathways: Implication for the closure of the ventral body wall. *Int. J. Mol. Sci.* **2019**, *20*, 2295. [CrossRef] [PubMed]
7. Salinas-Torres, V.M.; Gallardo-Blanco, H.L.; Salinas-Torres, R.A.; Martínez de Villarreal, L.E. Dataset for Genes and Gene Variants from Familial Gastroschisis (Version 3). Available online: http://doi.org/10.5281/zenodo.3270692 (accessed on 6 July 2019).
8. SNP & Variation Suite™ (Version 8.x). Golden Helix, Inc: Bozeman, MT, USA. Available online: http://www.goldenhelix.com (accessed on 31 August 2018).
9. Zerbino, D.R.; Achuthan, P.; Akanni, W.; Amode, M.R.; Barrell, D.; Bhai, J.; Billis, K.; Cummins, C.; Gall, A.; Girón, C.G.; et al. Ensembl 2018. *Nucleic Acids Res.* **2018**, *46*, D754–D761. [CrossRef] [PubMed]
10. Mi, H.; Huang, X.; Muruganujan, A.; Tang, H.; Mills, C.; Kang, D.; Thomas, P.D. PANTHER version 11: Expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* **2017**, *45*, D183–D189. [CrossRef] [PubMed]
11. Szklarczyk, D.; Morris, J.H.; Cook, H.; Kuhn, M.; Wyder, S.; Simonovic, M.; Santos, A.; Doncheva, N.T.; Roth, A.; Bork, P.; et al. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **2017**, *45*, D362–D368. [CrossRef] [PubMed]
12. Chen, J.; Bardes, E.E.; Aronow, B.J.; Jegga, A.G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **2009**, *37*, W305–W311. [CrossRef] [PubMed]