Genomics 103 (2014) 169–176

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods

exomeSuite: Whole exome sequence variant filtering tool for rapid identification of putative disease causing SNVs/indels

B. Maranhao [a,b], P. Biswas [a], J.L. Duncan [c], K.E. Branham [d], G.A. Silva [a,b,e], M.A. Naeem [f], S.N. Khan [f], S. Riazuddin [f], J.F. Hejtmancik [g], J.R. Heckenlively [d], S.A. Riazuddin [f,h], P.L. Lee [a], R. Ayyagari [a,*]

[a] Department of Ophthalmology, University of California, San Diego, UC Jacobs Retina Center, 9415 Campus Point Drive, La Jolla, CA 92037-0946, USA
[b] Department of Bioengineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[c] Department of Ophthalmology, University of California San Francisco, San Francisco, CA 94143, USA
[d] Department of Ophthalmology and Visual Sciences, University of Michigan Medical School, Ann Arbor, MI 48109, USA
[e] Neurosciences Graduate Program, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[f] National Centre of Excellence in Molecular Biology, University of the Punjab, Lahore, Pakistan
[g] Ophthalmic Genetics and Visual Function Branch, National Institutes of Health, Bethesda, MD 20892, USA
[h] The Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA

ARTICLE INFO

Article history:
Received 6 March 2013
Accepted 24 February 2014
Available online 3 March 2014

Keywords:
Exome
Filtering
Software
Mendelian disease
Homozygosity mapping

ABSTRACT

Exome and whole-genome analyses powered by next-generation sequencing (NGS) have become invaluable tools in identifying causal mutations responsible for Mendelian disorders. Given that individual exomes contain several thousand single nucleotide variants and insertions/deletions, it remains a challenge to analyze large numbers of variants from multiple exomes to identify causal alleles associated with inherited conditions. To this end, we have developed user-friendly software that analyzes variant calls from multiple individuals to facilitate identification of causal mutations. The software, termed exomeSuite, filters for putative causative variants of monogenic diseases inherited in one of three forms: dominant, recessive caused by a homozygous variant, or recessive caused by two compound heterozygous variants. In addition, exomeSuite can perform homozygosity mapping and analyze the variant data of multiple unrelated individuals. Here we demonstrate that filtering of variants with exomeSuite reduces datasets to a fraction of a percent of their original size. To the best of our knowledge this is the first freely available software developed to analyze variant data from multiple individuals that rapidly assimilates and filters large data sets based on pattern of inheritance.

1. Introduction

Recent technological advances have made it economically feasible to sequence exomes or whole genomes of large numbers of individuals. As a result new analysis tools are required to handle the large sets of genetic sequencing data. To this end several sequence alignment and variant callers have been developed, however; user friendly software that systematically compiles and compares variant call files is lacking.

The software, called exomeSuite, was designed to perform multiple functions related to novel monogenic disease-associated mutation discovery. The primary function is to filter variant calls from either an individual patient or a cohort of subjects, including patients and unaffected individuals, to identify candidate disease-causing mutations inherited in either dominant or recessive fashion. Additionally, the software allows users to compile databases of variants identified from exomes they have sequenced and use these databases to annotate results or to query them by gene. Other minor functions incorporated into the software are: the ability to build and maintain databases of variants of interest and screen variant call format (VCF) files against the said databases, the ability to annotate results with publicly available databases/tools (similar to ANNOVAR [1]), the ability to perform SNP-based homozygosity mapping (similar to HomozygosityMapper2012 [2]), and the ability to perform set functions on result files.

The software was designed to accommodate a variety of organizational and structural differences in the format of input files generated by different labs; as such, extremely few formatting requirements are imposed on input data. Details of these formatting requirements are given in Section S1 of the Supplemental material.

We describe the software that we developed, and the analysis of two pedigrees that led to an improvement in the original software design.

* Corresponding author. Fax: +1 858 246 0568.
E-mail addresses: bmaranha@ucsd.edu (B. Maranhao), pbiswas@ucsd.edu (P. Biswas), duncanj@vision.ucsf.edu (J.L. Duncan), haag@med.umich.edu (K.E. Branham), gsilva@ucsd.edu (G.A. Silva), f3h@helix.nih.gov (J.F. Hejtmancik), jrheck@umich.edu (J.R. Heckenlively), plee@scripps.edu (P.L. Lee), rayyagari@ucsd.edu (R. Ayyagari).

http://dx.doi.org/10.1016/j.ygeno.2014.02.006
0888-7543/

2. Methods

2.1. Subjects

Information and blood samples were collected from members across three generations of a consanguineous Pakistani family (Pedigree 1, Fig. 1A), as well as a non-consanguineous family of European ancestry (Pedigree 2, Fig. 2A). Standard ophthalmic evaluation was performed on all available members of these pedigrees. Written informed patient consent and local institutional review board (IRB) approvals were obtained for studies involving human subjects. All clinical investigations were conducted according to the principles expressed in the Declaration of Helsinki.

The Pakistani family (Pedigree 1) is from the Punjab province of Pakistan. A detailed medical history was obtained by interviewing family members. A total of 22 individuals including 9 affected individuals were enrolled. Detailed retinal evaluation including funduscopy was carried out on 3 affected members (V:1, V:10 and V:16) at Layton Rahmatulla Benevolent Trust (LRBT) Hospital, Lahore. Blood samples were collected from affected and unaffected family members. DNA was extracted as described previously [3,4].

In the pedigree of European ancestry (Pedigree 2), funduscopy, electroretinography, and visual field measurement were performed on Patient III:3. Blood was collected from the proband as well as two additional affected brothers of the proband who were reported to have been diagnosed with retinitis pigmentosa (RP). There was no known history of hearing loss (*USH2A*) mutations.

2.2. Exome capture, sequencing and data processing

Whole exomes were captured using NimbleGen SeqCap EZ exome V3 probes (Roche NimbleGen, Madison, WI) and sequenced using the Illumina Hi Seq (Illumina, San Diego, CA) following manufacturers' protocols. Unique paired-end (2 × 100 base) DNA sequence reads that passed quality control was mapped to the human reference genome build hg19 with SAMtools [5] and variants were called using GATK [6].

Using exomeSuite, the resultant single nucleotide variant (SNV) and insertion/deletion (indel) files were first filtered for the presumed pattern of inheritance, and subsequently analyzed for variants that matched the following criteria: (1) the variant was found in the homozygous state in fewer than 0.5% of individuals in all three public databases integrated into exomeSuite (1000 Genome [7], HapMap [8], and NHLBI ESP6500 [9]), (2) if the variant is a SNV it is not predicted by MutationTaster [10], PolyPhen [11], and SIFT [12] as benign or polymorphic, and (3) if the variant is intronic, it is less than 50 bases from the intron–exon junction. Additionally, if the presumed inheritance pattern

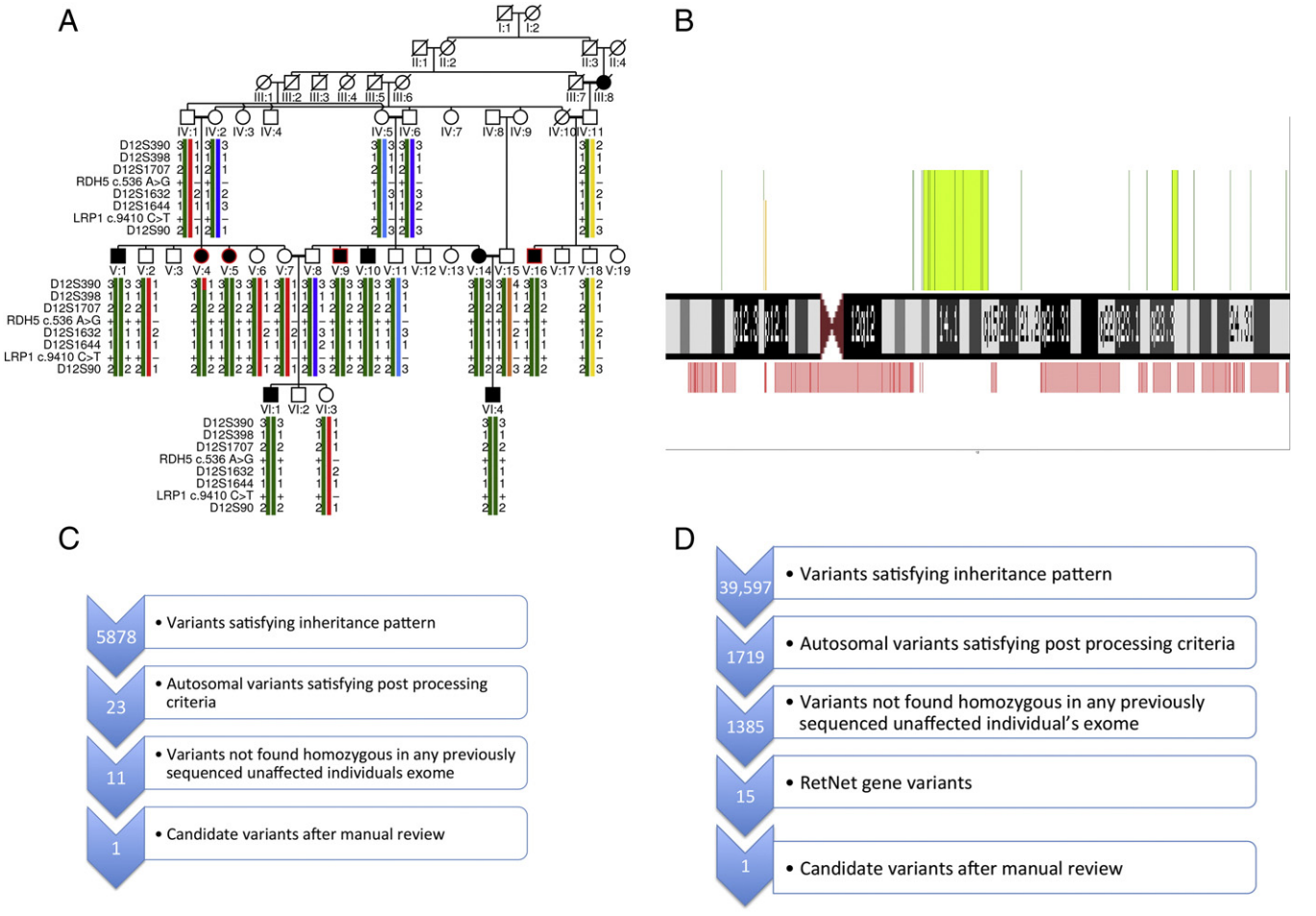

Fig. 1. Pedigree 1. (A) Results of STR analysis as well as *RDH5* c.536 and *LRP1* c.9410 genotyping are displayed below the corresponding individual adjacent to colored chromosomes depicting haplotypes. Loci are ordered top-to-bottom in a centromeric-to-telomeric fashion based on coordinates given by UCSC Genome Browser. (B) Results of in-silico homozygosity mapping performed on SNPs exome sequenced from individuals V:4, V:5, V:9, and V:16. A 14.7 MBp region at chromosome 12q14 that is homozygous in the four affected members exome sequenced in Pedigree 1 is shown. Green tally marks indicate loci that are homozygous and identical in all four individuals, red tally marks indicate loci that are not, and yellow tally marks indicate loci where there was insufficient read depth in one or more individuals. Regions between adjacent tally marks of the same color are shaded. The 12q14 homozygous region includes the *LRP1* and *RDH5* genes which are separated by approximately 1.4 MBp. Performance of exomeSuite processing when absence of a variant in a variant call file is interpreted as homozygous reference (C) and unknown genotype (D).

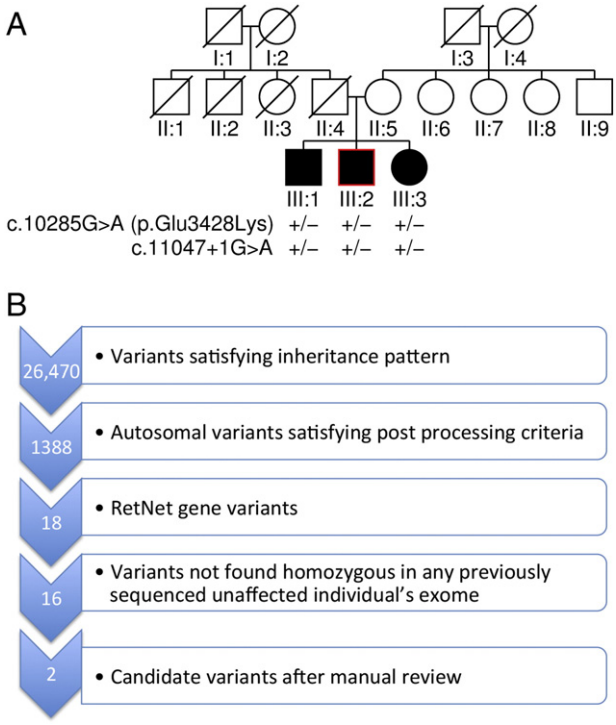

Fig. 2. Pedigree 2. (A) The individual whose exome was sequenced, patient III:2, is denoted by a red outline. *USH2A* genotypes for all individuals enrolled in this study are found below the respective subject. (B) Filtering performance of exomeSuite processing.

is recessive with underlying compound heterozygous variants then there must be at least two variants within a gene that satisfy these post-filtering criteria.

2.3. Segregation analysis

For Pedigree 1, primers (forward primer 5-CAGATGCTCCCAGGAAGAAG-3, reverse primer 5-GAGTGGGCTGCTGTTAGTCC-3 and forward primer 5-CTTCCTGCAGGTCCTACACC-3, reverse primer 5-AATGGTCACCCCAGTCTGTC-3) were designed to amplify and sequence exon 3 of retinol dehydrogenase 5 (*RDH5*) and exon 59 of low density lipoprotein receptor-related protein 1 (*LRP1*), respectively. All members of pedigree 1 for whom DNA samples were available as well as 95 unrelated Pakistani control subjects were screened for the *RDH5* c.536A>G and *LRP1* c.9410C>T variants.

For Pedigree 2, primers (forward primer 5-TGAAAGTCACAAAAGCCTACCC-3, reverse primer 5-TGGCCTCAAAGTATGATGGA-3, and forward primer 5-CAACTCTGCATGTTACTTCTGG-3, reverse primer 5-CAGATTCCACCTCAAAATGCT-3) were designed to amplify and sequence exons 52 and 56 of *USH2A* respectively.

2.4. Linkage and haplotype analyses

Microsatellite markers spanning reported loci or genes associated with RP/RD were analyzed and haplotypes were constructed as described earlier [13]. In brief, fluorescently labeled short tandem repeat markers flanking the disease locus were used for amplification. The resulting PCR products were separated in an ABI3100 DNA analyzer, and alleles were assigned by using GeneMapper software ver. 4 (Applied Biosystems, Foster City, CA). The marker order and distances between the markers were obtained from the national Center for Biotechnology Information sequence maps. Two-point linkage analysis with markers at the RDH5 locus was performed using the FASTLINK version of MLINK from the LINKAGE Program Package, whereas maximum LOD scores were calculated using ILINK. Autosomal recessive RD was analyzed as a fully penetrant trait with an affected allele frequency of 0.001 [13].

2.5. Development of exomeSuite

exomeSuite is a collection of scripts programmed in a combination of Matlab and C++ available either as source code or as pre-compiled stand-alone application with a graphical user interface.

exomeSuite requires installation of the Matlab Compiler Runtime. Mac and Linux installations additionally require X11 installation [14].

2.5.1. Screening & filtering by inheritance pattern

exomeSuite is blind to the relationship between individuals and only considers whether individuals are affected or unaffected by a condition; it operates on the premise that the reference allele does not manifest the condition studied.

The software was designed with the following processing pipeline in mind (Fig. 3). First, screen individuals for previously published mutations known to cause the disease of interest. If none are found, then screen individuals for variants segregating with the disease phenotype according to the expected mode of inheritance. After input files are filtered to identify variants segregating with the disease phenotype, variants can then be annotated based on user-specified criteria, such as allele frequency less than 0.5%, to eliminate variants which may segregate with the disease phenotype but occur at high frequency in the population. Additional annotation steps might include predicted impact of the mutation (damaging vs. benign), and expression in relevant tissue(s).

Input files should be in the form of individual tab-separated variable format text files that adhere to the variant call file (VCF) format, or some highly similar format. Acceptable deviations from the VCF format are detailed in the manual. Since exomeSuite was conceived for the purpose of germ line variant discovery, it accepts a maximum of two alleles at a given location for a given individual, and the variant(s) must be reported as either heterozygous or homozygous variants.

Screening for known variants is accomplished by comparing the genomic location and variant alleles to one or more databases of variants of interest. The most current set of dbSNP variants with clinical impact attributes is included in the software. Additional databases can be created by the user either manually or by importing a tab separated variable table that can be created in any spreadsheet viewer, e.g. Microsoft Excel. Results of screening for known mutations are output in a tab delimited text format that can be opened in any spreadsheet viewer.

When screening for known variants fails to identify the causal gene for the disease, an individual or a group of individuals can be filtered for variants that segregate by mode of inheritance: autosomal dominant, autosomal recessive with suspected homozygous inheritance, and autosomal recessive with suspected compound heterozygous inheritance. Each of these analyses are output as separate files as described below.

- dominant.txt: The mode of inheritance is presumed to be autosomal dominant. The output file includes all variants, for which each affected individual carries one variant allele and one reference allele, and each unaffected individual is homozygous reference.
- recessive-homozygous.txt: The mode of inheritance is presumed to be recessive, and the causative mutation occurs in the homozygous state. The output file includes all variants, for which each affected individual carries the variant at the same site in both alleles, and each unaffected individuals has at least one reference allele.
- recessive-compound het.txt: The mode of inheritance is presumed to be recessive, but disease in affected individuals is caused by inheritance of two potentially damaging mutations on separate alleles (compound heterozygotes). Unaffected individuals carry only a single (simple heterozygotes) or no variant alleles (homozygous reference). The compound heterozygous output file includes all variants for

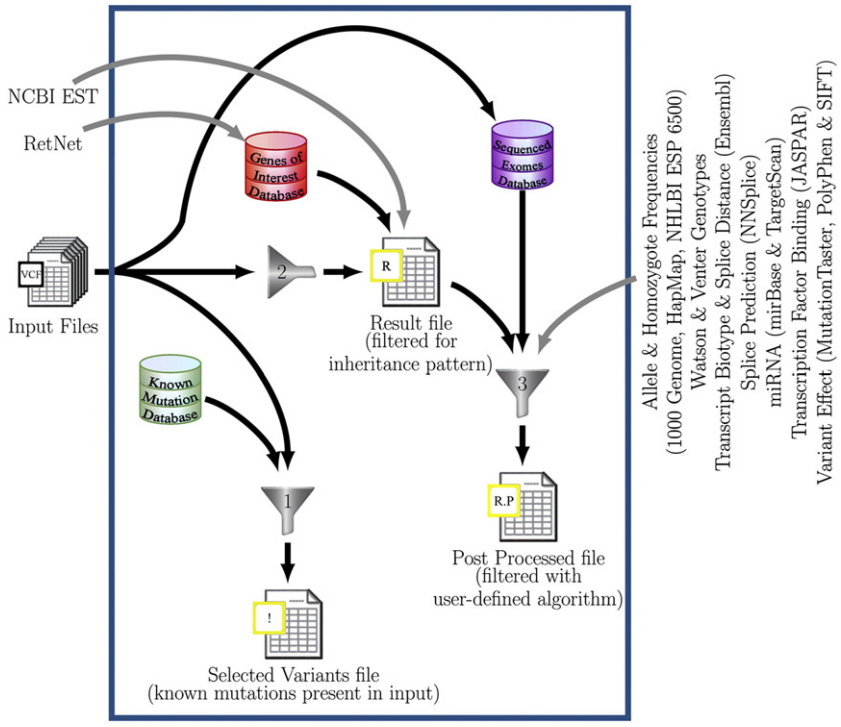

Fig. 3. Basic workflow of exomeSuite. The blue box indicates everything included in exomeSuite. Gray arrows entering the blue box indicate outside resources that exomeSuite has been designed to utilize automatically. Within the blue box databases that the user can create and manipulate are indicated by colored cylinders; the red database is the "genes of interest" database, the purple database is the database of variants identified in sequenced exomes, and the green database is a database of known mutations. Funnels represent filtering functions, and are numbered to indicate the order in which they were designed to be used.

which all affected individual are heterozygous, and for which all unaffected individuals are either homozygous reference or heterozygous, with the additional caveat that there must be two or more variants within a gene for them to be included in the output. If three or more heterozygous variants are present within a single gene all are included in the result file, and the researcher is left to decide which two variants are causative of the disease state should this gene indeed be responsible for the manifested phenotype. Complicating this analysis is the presence of mutations in cis (occurring on the same allele) and trans (occurring on different alleles) that needs to be clarified by segregation analysis.

Informal and formal descriptions of the algorithm are available in the Supplementary material.

To properly analyze sex-linked disease-causative mutations all Y chromosome SNV/indels called for female subjects are eliminated since they are due to misalignment of the sequence to the reference human genome. Likewise, all X and Y chromosome SNV/indels called for male subjects are treated as hemizygous (coded as homozygous SNV/indels) when filtering for homozygous recessive inheritance. We treat X chromosome SNV/indels in males as heterozygous when filtering for compound heterozygous recessive inheritance, since the dominant filter eliminates any variants for which an affected individual is a homozygous variant. This is not done for the Y chromosome because it would result in an output identical to the homozygous recessive file.

In addition to generating result files for the three segregation modes, exomeSuite is capable of generating the union, intersection and complement of one result file with respect to another result file. The union of two result files allows one to study mutations in unrelated individuals in the context of genes rather than variants, enabling identification of candidate genes in unrelated individuals where mutations may be homozygous recessive in some and compound heterozygous in other individuals. The intersection allows one to study specific mutations at the variant level across files of different input format. The complement of results is particularly informative for determining what variants have been filtered out by the inclusion of additional individuals in the analysis.

2.5.2. Post-filtering annotation

The software allows for the automated annotation of variants, a feature we refer to as post-filtering analysis, based on user-defined criteria. The information for this automated annotation must be provided in the input file or selected from the annotation databases/tools that exomeSuite was designed to interface with. Users may define their personal lists of genes of interest, i.e. genes for which mutations are known to be associated with the disease being studied; exomeSuite includes a list of genes curated by the RetNet website [15], as well as the ability to download newer lists when they become available on the RetNet website. exomeSuite includes the ability to annotate variants for: allele frequency by interfacing with 1000 Genome [7], HapMap [8], and NHLBI ESP 6500 [9], as well as user-defined databases created with the software; presence in Watson and Venters genomes [16]; impact as predicted by MutationTaster [10], PolyPhen [11], and SIFT [12]; distance to nearest splice site (UCSC) [17]; predicted impact on splicing (NNSplice) [18]; if the variant occurs in a microRNA gene (mirBase) [19] or region targeted by microRNA (TargetScan) [20]; and lastly, tissue expression (NCBI Unigene EST Profile [21]). exomeSuite also provides hyperlinks for each candidate gene to the HGNC site and hyperlinks for rsIDs to NCBI's SNP Cluster Report site [22].

Available disease-associated variant databases are known to contain false positives, hence results of variant screening against such databases should be interpreted carefully. For example, *ABCA4* p.His423Arg & p.Ser2255Ile are both listed in HGMD [23] but have homozygous genotype frequencies greater than 6.5% in 1000 Genome [7], HapMap [8] and NHLBI ESP 6500 [9] and so are unlikely to cause disease. Post-filtering annotation of variant screening files that remove high frequency alleles eliminates many false positives.

2.5.3. SNP-based homozygosity mapping

When analyzing consanguineous families, homozygosity mapping may be used to identify regions of the genome that are conserved across, and unique to, a group of affected individuals. Homozygosity mapping requires SAMtools [5], and access to the sequence pileup of the individuals in the form of SAM or BAM files [5]. Users are able to set a minimum mapping quality for reads and a minimum phred score for bases used to perform homozygosity mapping, as well as a minimum read depth and cutoff for making homozygosity/heterozygosity decisions. Results are displayed as colored tally marks along an image of the chromosome; an example is given for Pedigree 1 in the Supplemental material. Green tally marks designate SNPs that are identically homozygous in all affected individuals and not so in any unaffected individual, red tallies indicate regions that are not identically homozygous in all affected individuals, black tallies indicate regions that are inconclusive (all affected individuals are homozygous, but so is at least one unaffected individual), and yellow tally marks indicate SNPs for which one or more individuals lack sufficient read depth to conclude segregation. Regions between adjacent tally marks of the same color are shaded in.

Users are able to define a custom set of SNPs that they would like to use when performing homozygosity mapping, or rely upon a set of 2277 SNPs selected from those identified in the NHLBI ESP6500 database based on two criteria: (1) maximize the information the SNV imparts, and (2) maintain a maximum distance between adjacent queried SNVs of 10 megabases where possible.

3. Results

To demonstrate the capabilities of exomeSuite, we present the analysis of exomes from two different pedigrees; one representing a consanguineous family with recessive inheritance (Pedigree 1), and another representing a recessive pattern of inheritance possibly due to a pair of compound heterozygous mutations (Pedigree 2).

Exome sequencing produced an average of 13.75 gigabases of sequence per individual which when mapped yielded an average read depth of 106 within the targeted regions.

Pedigree 1 (Fig. 1A), a five-generation pedigree with 6 consanguineous marriages and 9 members affected with retinal degeneration (RD) was ascertained from the Punjab province of Pakistan. The parents of all affected members were unaffected. The pattern of inheritance of retinal degeneration was assumed to be autosomal recessive.

The initial symptom of disease was night blindness with onset in early childhood. Visual acuities of affected individuals were reduced when compared to unaffected individuals. Funduscopy of Patient V:1, V:10 and V:16 revealed findings consistent with FA, a specific form of congenital night blindness that may be associated with diffuse cone dysfunction [24,25]. At the age of 17, the fundus of patient V:1 (Fig. 4A) exhibited normal foveal reflexes with no apparent vascular abnormalities and normal optic nerves. Whitish yellow subretinal punctate deposits were present along and anterior to the arcades, and extended into the periphery as linear hypopigmented flecks organized in concentric rings. Patient V:10, at age 18, showed similar findings to patient V:1 (Fig. 4B) with whitish-yellow subretinal punctate deposits extending posteriorly within the arcades, and linear hypopigmented lesions extending into the periphery with punctate spots superimposed upon the deeper, hypopigmented, linear lesions. Foveal light reflexes were not clearly evident and there were retinal pigment epithelial abnormalities near the fovea in each eye. Patient V:16 at 30 years old exhibited similar fundus findings as V:1 and V:10 (Fig. 4C); however, the punctate deposits were less white and appeared less superficial. The foveal light reflex was not evident and there were retinal pigment epithelial abnormalities near the fovea as well as yellow subretinal lesions within the arcades and involving the fovea. The punctate deposits

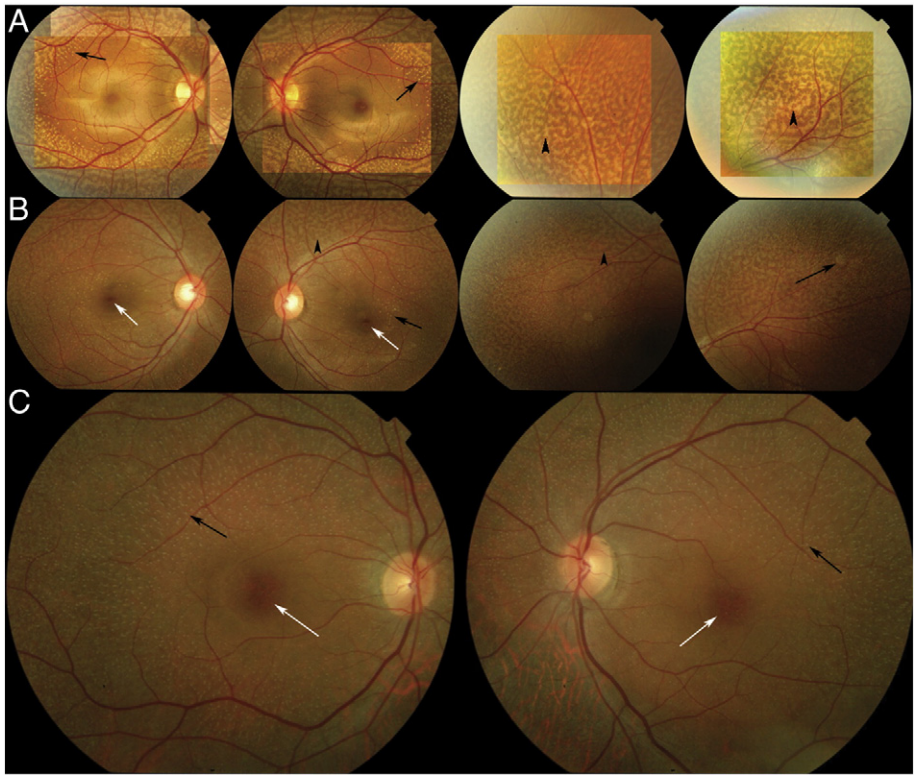

Fig. 4. Color fundus photos of affected individuals. (A) V:1 at age 17 shows whitish-yellow subretinal punctate deposits along and anterior to the arcades (black arrows), extending into the periphery as linear hypopigmented flecks organized in concentric rings (black arrowheads). (B) V:10 at age 18 shows similar punctate deposits extending posteriorly within the arcades and linear hypopigmented lesions extending into the periphery with punctate spots superimposed upon the deeper, hypopigmented, linear lesions (black arrow). Foveal light reflexes were not clearly seen and there were retinal pigment epithelial abnormalities near the fovea in each eye (white arrows). (C) V:16 at age 30 shows similar findings with small yellow subretinal lesions involving the fovea in the left eye (white arrow).

and flecks were observed together in some regions of the fundus with the puncta superficial to the deeper, hypopigmented, linear flecks. Electroretinography (ERG) of the affected individuals showed no response after dark adaptation to dim or bright stimuli, although photopic 30 Hz flicker responses were measurable but reduced from normal, consistent with diffuse rod-greater-than-cone dystrophy (data not shown). Overall the phenotype observed in affected members of this pedigree is similar to other patients described with FA [26,27,25].

The exome of individual V:5 was captured and sequenced. Analysis of these variants did not identify either homozygous or compound heterozygous, known or novel damaging variants in genes previously reported to be involved in causing RD. Subsequently, exome capture and sequencing were performed on three additional members (V:4, V:9, and V:16). We found 5878 variants in 3571 unique genes segregating with the disease phenotype (Fig. 1c). Twenty-three autosomal variants passed the post processing filters described in Section 2.2; twelve of the 23 variants were not found in the homozygous state in any of our previously sequenced unaffected individual exomes. One of the 12 variants was located in an intergenic region, eight were intronic variants predicted by NNSplice to not affect splicing, one represented an in-frame insertion of a single amino acid, and one was a silent variant found in several other exomes we've sequenced and predicted by NNSplice to not affect splicing. The remaining variant was a missense variant predicted to be damaging by PolyPhen [11] and MutationTaster [10], NM 002332.2:c.9410C>T (p.Thr3137Met), located in the LRP1 gene. This mutation is located within a YWTD domain between the second and third ligand-binding clusters of the protein. This position in the *LRP1* protein is highly conserved across species. *LRP1* is a member of the low-density lipoprotein receptor (LDLR) family, and LDLR is the most studied member of this family. In LDLR the YWTD domain is thought to be critical for pH-dependent ligand release within endosomes [28], and several missense mutations within this structure have been associated with familial hyperlipidemia [29]. LRP5, another LDLR member, has been previously associated with familial exudative vitreoretinopathy (FEVR) in a single British woman [30].

To exclude the possibility of the involvement of another causative mutation, and to confirm the involvement of *LRP1*, microsatellites linked to known RD loci associated with diminished but not absent cone responses were analyzed and tested for homozygosity. In addition homozygosity mapping was performed using exomeSuite. Analysis of microsatellite marker genotypes in the *LRP1* locus revealed a homozygous region segregating with the disease (Fig. 1A). Linkage analysis with markers at this locus confirmed linkage with significant positive LOD scores (Table S5). Similarly, analysis of exome variants using exomeSuite identified a 14 megabase homozygous region on chromosome 12 that segregated with the disease phenotype (Fig. 1B). This homozygous region on chromosome 12 encompassed the *RDH5* gene in addition to *LRP1*. Mutations in *RDH5* have been reported to be associated with FA [26,31].

Subsequent manual examination of *RDH5* variants in affected members variant call files revealed the presence of a novel SNV, NM 001199771.1:c.536A>G (p.Lys179Arg), in all affected members of the pedigree except the proband. Manual inspection of the sequence alignments suggested that the p.Lys179Arg *RDH5* mutation is present in the proband but was not called as the sequence in that region did not pass the criteria for SNP calling. Dideoxy sequencing analysis confirmed segregation of both *RDH5* c.536 A>G and *LRP1* c.9410 C>T novel variants in members of pedigree 1. These variants were not detected in 95 unrelated Pakistani controls.

To address the concern that the causative mutation might be missed in an individual by exome analysis, the exomeSuite software was augmented to allow for the absence of a variant in an individual's input file to be interpreted either as homozygous reference, or as an unknown genotype. Reanalysis of the exomes of the four affected individuals in Pedigree 1 yielded 35,957 variants found to be homozygous in one or more of the affected individuals and not heterozygous in any (Fig. 1D). Post processing with the filters described in Section 2.2 narrowed the list to 1719 autosomal variants, of which 1385 were not previously observed in any of the unaffected members in our sequenced exome database. Of these, 15 were located in RetNet genes (Table S5). Seven of the 15 were intronic variants predicted by NNSplice to not affect splicing, one was a silent coding variant likewise predicted to not affect splicing, one was a variant in a 3′ UTR that did not occur within a miRNA target region, and one was a known 5′ UTR variant found to occur at a homozygous frequency of less than 0.2% of individuals. The remaining five variants were missense variants, only one of which, a novel missense change, *RDH5* SNV NM 001199771.1:c.536A>G (p.Lys179Arg), was found to segregate with the disease phenotype when the alignment files of the four sequenced individuals were manually reviewed. Reanalysis of variants following the improvements made to exomeSuite resulted in the identification of the *RDH5* variant as a candidate.

Pedigree 2 (Fig. 2A) is of European ancestry with three siblings affected with recessive RD. The onset of symptoms in the proband (III:3) was at age 26. The patient was examined at age 46 and visual acuity was noted to be 20/30 OD, 20/50 OS. Upon fundus examination her retina showed diffuse retinal atrophy with mild bone spicules (data not shown). Her electroretinogram showed diffuse rod-cone dysfunction (performed at age 39) and visual fields showed a ring scotoma with intact central fields, consistent with a diagnosis of RP (data not shown).

Patient III:2, was selected for exome sequencing, and sequence variants were initially analyzed for compound heterozygous mutations. Patient III:2 carried 26,470 variants in 5916 genes that matched the inheritance criteria (Fig. 2B). Post filtering analysis, as described in Section 2.2, narrowed this list to 1388 autosomal variants in 334 unique genes; of these, 18 variants were identified in 7 unique genes previously associated with RD (Table S3). Removing a variant found to be homozygous in one or more unaffected individuals in our exome sequenced database resulted in just 16 variants in 6 genes that could satisfy a recessive pattern of disease inheritance assumed to be caused by two compound heterozygous mutations. Removing intronic variants and silent coding variants predicted by NNSplice to not affect splicing further narrowed the candidate list to 5 variants in 2 RetNet genes. One of the variants was located in the 3′ UTR region of PDE6B that is predicted to not reside within a miRNA targeted region. Only one gene, *USH2A*, was determined to carry two previously reported mutations: a missense mutation NM 206933.2:c.10342G>A (p.Glu3448Lys) [32], and an intronic mutation NM 206933.2:c.11047+1G>A [33]. The p.Glu3448Lys mutation was previously reported to be heterozygous in a single affected individual and absent in 80 controls [32], and categorized as a rare polymorphism. The p.Glu3448Lys mutation was predicted by SIFT [12], pMut [34] and AGVGD [35] to be benign while PolyPhen [11] predicted the variant to be damaging. On reexamination, AGVGD [35] categorized the variant as C55, very likely to be damaging. Additionally, MutationTaster [10] predicts the variant as “disease causing.” Subsequent dideoxy sequencing of the three affected members of this pedigree confirmed heterozygosity for the *USH2A* missense variant and the *USH2A* intronic SNV segregating with the disease phenotype (Fig. 2A).

4. Discussion

As the rapid pace of next generation sequencing technology continues to drive cost and sequencing times down, utilization of the technique will continue to grow. This growth however must be met with an ability to rapidly process the enormous amount of data and interpret the results. Our aim was to develop a simple-to-use, freely available application that rapidly assimilates and filters large sets of NGS data to facilitate identification of disease-associated genes. Special care was taken to ensure minimal requirements on the format of input data so that the software could process virtually all types of data sets.

We have used exomeSuite previously to analyze a consanguineous pedigree with autosomal recessive retinal degeneration in which the exomes of one unaffected and two affected siblings were sequenced. This led to the identification of a novel splice site variant in the RBP4 gene, NM 006744.3:c.111+1G>A, that segregated with the disease phenotype [36]. The two pedigrees presented in this manuscript where exomeSuite identified four mutations in three genes (*LRP1*, titRDH5 and *USH2A*), along with the previously reported pedigree [36], demonstrate the ability of exomeSuite to correctly identify causative mutations segregating with disease.

Furthermore, we have demonstrated how the application may be used to identify disease-causing mutations despite their absence in some of the input files. The built-in homozygosity mapper allowed for the identification of a novel *RDH5* variant despite a false negative SNV call in the proband of pedigree 1. Despite these limitations, and until direct sequencing (i.e. not resequencing) of the whole genome becomes available, exome sequencing continues to be widely accepted as a standard tool for screening individuals for novel disease-causing genetic mutations.

In this study exome sequencing led to the identification of the novel disease causing *RDH5* mutation in a consanguineous Pakistani pedigree. This mutation occurs in the active site of the enzyme [37] and likely eliminates or drastically reduces its activity. The fundus phenotype of the three patients examined is typical of FA. Therefore it is likely that the novel *RDH5* missense variant c.536 A>G (p.Lys179Arg) may be sufficient to cause the retinal phenotype observed in Pedigree 1. However, involvement of the novel *LRP1* change segregating with the disease in this pedigree cannot be ruled out. Additional studies are needed to establish the causative nature of the *RDH5* variant and the role of the *LRP1* variant in causing retinal pathology or modifying the phenotype in this pedigree.

Using exomeSuite we also identified two *USH2A* mutations in a pedigree of European ancestry. Both of these mutations have been previously reported. The intronic variant was recently found to occur in a homozygous state in an affected individual of a Pakistani pedigree [33], while the missense variant was identified in a single patient while screening a cohort for *USH2A* variants [32].

exomeSuite provides users with a user-friendly package of functions geared to rapidly assimilate data and narrowing several thousand variants to small lists that can easily be manually reviewed for the identification of disease-causing variants. To the best of our knowledge there is no other freely available application that provides all this functionality. In this manuscript we validated its main functions in identifying the underlying genetic cause of disease in two pedigrees, including one for which the input data was imperfect.

Exome sequencing and analysis enable discovery of variants in several genes in an unbiased manner. This has led to the identification of several novel variants in genes known to cause disease, as well as new disease-causative genes. Additionally, it can lead to the identification of rare variants in multiple genes segregating with disease, such as the *RDH5* and *LRP1* variants presented here. Further analysis of such variants may explain the heterogeneity of phenotypes observed in individuals with the same clinical diagnosis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2014.02.006.

References

[1] K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Res. 38 (2010) e164.

[2] D. Seelow, M. Schuelke, HomozygosityMapper2012 — bridging the gap between homozygosity mapping and deep sequencing, Nucleic Acids Res. (2012) 1–5.

[3] H. Kaul, S.A. Riazuddin, A. Yasmeen, S. Mohsin, M. Khan, I.A. Nasir, S.N. Khan, T. Husnain, J. Akram, J.F. Hejtmancik, S. Riazuddin, A New Locus for Autosomal Recessive Congenital Cataract Identified in a Pakistani Family, 2010. (molvis.org).

[4] H. Kaul, S.A. Riazuddin, M. Shahid, S. Kousar, N.H. Butt, A.U. Zafar, S.N. Khan, T. Husnain, J. Akram, J.F. Hejtmancik, S. Riazuddin, Autosomal Recessive Congenital Cataract Linked to EPHA2 in a Consanguineous Pakistani Family, 2010. (molvis.org).

[5] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.

[6] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. del Angel, M.A. Rivas, M. Hanna, A. McKenna, T.J. Fennell, A.M. Kernytsky, A.Y. Sivachenko, K. Cibulskis, S.B. Gabriel, D. Altshuler, M.J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data, Nat. Genet. 43 (2011) 491–498.

[7] 1000 Genomes Project Consortium, G.R. Abecasis, D. Altshuler, A. Auton, L.D. Brooks, R.M. Durbin, R.A. Gibbs, M.E. Hurles, G.A. McVean, A map of human genome variation from population-scale sequencing, Nature 467 (2010) 1061–1073.

[8] International HapMap 3 Consortium, D.M. Altshuler, R.A. Gibbs, L. Peltonen, D.M. Altshuler, R.A. Gibbs, L. Peltonen, E. Dermitzakis, S.F. Schaffner, F. Yu, L. Peltonen, E. Dermitzakis, P.E. Bonnen, D.M. Altshuler, R.A. Gibbs, P.I.W. de Bakker, P. Deloukas, S.B. Gabriel, R. Gwilliam, S. Hunt, M. Inouye, X. Jia, A. Palotie, M. Parkin, P. Whittaker, F. Yu, K. Chang, A. Hawes, L.R. Lewis, Y. Ren, D. Wheeler, R.A. Gibbs, D.M. Muzny, C. Barnes, K. Darvishi, M. Hurles, J.M. Korn, K. Kristiansson, C. Lee, S.A. McCarrol, J. Nemesh, E. Dermitzakis, A. Keinan, S.B. Montgomery, S. Pollack, A.L. Price, N. Soranzo, P.E. Bonnen, R.A. Gibbs, C. Gonzaga-Jauregui, A. Keinan, A.L. Price, F. Yu, V. Anttila, W. Brodeur, M.J. Daly, S. Leslie, G. McVean, L. Moutsianas, H. Nguyen, S.F. Schaffner, Q. Zhang, M.J.R. Ghori, R. McGinnis, W. McLaren, S. Pollack, A.L. Price, S.F. Schaffner, F. Takeuchi, S.R. Grossman, I. Shlyakhter, E.B. Hostetter, P.C. Sabeti, C.A. Adebamowo, M.W. Foster, D.R. Gordon, J. Licinio, M.C. Manca, P.A. Marshall, I. Matsuda, D. Ngare, V.O. Wang, D. Reddy, C.N. Rotimi, C.D. Royal, R.R. Sharp, C. Zeng, L.D. Brooks, J.E. McEwen, Integrating common and rare genetic variation in diverse human populations, Nature 467 (2010) 52–58.

[9] W. NHLBI GO Exome Sequencing Project (ESP), Seattle, exome variant server, http://evs.gs.washington.edu/EVS/ 2013 (Accessed: 03/18/2013).

[10] J.M. Schwarz, C. Rödelsperger, M. Schuelke, D. Seelow, MutationTaster evaluates disease-causing potential of sequence alterations, Nat. Publ. Group 7 (2010) 575–576.

[11] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov, S.R. Sunyaev, Nat Methods. 7 (4) (2010) 248–249.

[12] P. Kumar, S. Henikoff, P.C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, Nat. Protoc. 4 (2009) 1073–1081.

[13] L. Li, N. Nakaya, V.R.M. Chavali, Z. Ma, X. Jiao, P.A. Sieving, S. Riazuddin, S.I. Tomarev, R. Ayyagari, S.A. Riazuddin, J.F. Hejtmancik, A mutation in ZNF513, a putative regulator of photoreceptor development, causes autosomal-recessive retinitis pigmentosa, Am. J. Hum. Genet. 87 (2010) 400–409.

[14] M. forge, XQuartz, http://xquartz.macosforge.org/trac 2013 (Accessed: 12/13/2013).

[15] L. for the Molecular Diagnosis of Inherited Eye Diseases, RetNet: retinal information network, http://sph.uth.tmc.edu/retnet/ 2012 (Accessed: 08/03/2012).

[16] Ensembl, homo_sampiens VCF, ftp://ftp.ens0065mbl.org/pub/release-74/variation/vcf/homo_sapiens .

[17] UCSC, CSC Genome Browser: Downloads, http://hgdownload.soe.ucsc.edu/downloads.html#human 2013 (Accessed: 09/7/2013.).

[18] M.G. Reese, F.H. Eeckman, D. Kulp, D. Haussler, Improved splice site detection in Genie, J. Comput. Biol. 4 (1997) 311–323.

[19] A. Kozomara, S. Griffiths-Jones, miRBase: integrating microRNA annotation and deep-sequencing data, Nucleic Acids Res. 39 (2010) D152–D157.

[20] R.C. Friedman, K.K.-H. Farh, C.B. Burge, D.P. Bartel, Most mammalian mRNAs are conserved targets of microRNAs, Genome Res. 19 (2009) 92–105.

[21] NCBI, Expressed sequence tags database, http://www.ncbi.nlm.nih.gov/dbEST/ 2013 (Accessed: 08/27/2013).

[22] NCBI, dbSNP short genetic variations, http://www.ncbi.nlm.nih.gov/SNP/ 2013 (Accessed: 12/13/2013).

[23] I. of Medical Genetics in Cardiff, The Human Gene Mutation Database, http://www.hgmd.org/ 2013 (Accessed: 12/13/2013).

[24] A.V. Cideciyan, F. Haeseleer, R.N. Fariss, T.S. Aleman, G.F. Jang, C.L. Verlinde, M.F. Marmor, S.G. Jacobson, K. Palczewski, Rod and cone visual cycle consequences of a null mutation in the 11-cis-retinol dehydrogenase gene in man, Vis. Neurosci. 17 (2000) 667–678.

[25] Y. Niwa, M. Kondo, S. Ueno, M. Nakamura, H. Terasaki, Y. Miyake, Cone and rod dysfunction in fundus albipunctatus with RDH5 mutation: an electrophysiological study, Invest. Ophthalmol. Vis. Sci. 46 (2005) 1480–1485.

[26] H. Yamamoto, A. Simon, U. Eriksson, E. Harris, E.L. Berson, T.P. Dryja, Mutations in the gene encoding 11-cis retinol dehydrogenase cause delayed dark adaptation and fundus albipunctatus, Nat. Genet. 22 (1999) 188–191.

[27] M. Nakamura, Y. Hotta, A. Tanikawa, H. Terasaki, Y. Miyake, A high association with cone dystrophy in Fundus albipunctatus caused by mutations of the RDH5 gene, Invest. Ophthalmol. Vis. Sci. 41 (2000) 3925–3932.

[28] J. Herz, Deconstructing the LDL receptor — a rhapsody in pieces, Nat. Struct. Biol. 8 (2001) 476–478.

[29] H. Jeon, W. Meng, J. Takagi, M.J. Eck, T.A. Springer, S.C. Blacklow, Implications for familial hypercholesterolemia from the structure of the LDL receptor YWTD-EGF domain pair, Nat. Struct. Biol. 8 (2001) 499–504.

[30] C. Toomes, H.M. Bottomley, R.M. Jackson, K.V. Towns, S. Scott, D.A. Mackey, J.E. Craig, L. Jiang, Z. Yang, R. Trembath, G. Woodruff, C.Y. Gregory-Evans, K. Gregory-Evans, M.J. Parker, G.C.M. Black, L.M. Downey, K. Zhang, C.F. Inglehearn, Mutations in

LRP5 or FZD4 underlie the common familial exudative vitreoretinopathy locus on chromosome 11q, Am. J. Hum. Genet. 74 (2004) 721–730.
[31] F. Gonzalez-Fernandez, D. Kurz, Y. Bao, S. Newman, B.P. Conway, J.E. Young, D.P. Han, S.C. Khani, 11-cis retinol dehydrogenase mutations as a major cause of the congenital night-blindness disorder known as fundus albipunctatus, Mol. Vis. 5 (1999) 41.
[32] T.L. McGee, B.J. Seyedahmadi, M.O. Sweeney, T.P. Dryja, E.L. Berson, Novel mutations in the long isoform of the USH2A gene in patients with Usher syndrome type II or non-syndromic retinitis pigmentosa, J. Med. Genet. 47 (2010) 499–506.
[33] P. Le Quesne Stabej, Z. Saihan, N. Rangesh, H.B. Steele-Stallard, J. Ambrose, A. Coffey, J. Emmerson, E. Haralambous, Y. Hughes, K.P. Steel, L.M. Luxon, A.R. Webster, M. Bitner-Glindzicz, Comprehensive sequence analysis of nine Usher syndrome genes in the UK National Collaborative Usher Study, J. Med. Genet. 49 (2012) 27–36.
[34] C. Ferrer-Costa, J.L. Gelpí, L. Zamakola, I. Parraga, X. de la Cruz, M. Orozco, PMUT: a web-based tool for the annotation of pathological mutations on proteins, Bioinformatics 21 (2005) 3176–3178.
[35] S.V. Tavtigian, Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral, J. Med. Genet. 43 (2005) 295–305.
[36] C. Cukras, T. Gaasterland, P. Lee, H.V. Gudiseva, V.R.M. Chavali, R. Pullakhandam, B. Maranhao, L. Edsall, S. Soares, G.B. Reddy, P.A. Sieving, R. Ayyagari, Exome analysis identified a novel mutation in the RBP4 gene in a consanguineous pedigree with retinal dystrophy and developmental abnormalities, PLoS ONE 7 (2012) e50205.
[37] H. Joernvall, B. Persson, M. Krook, S. Atrian, R. Gonzalez-Duarte, J. Jeffery, D. Ghosh, Short-chain dehydrogenases/reductases (SDR), Biochemistry 34 (1995) 6003–6013.