BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 22 2014, pages 3215–3222
doi:10.1093/bioinformatics/btu508

Genetics and population analysis

Advance Access publication July 30, 2014

Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases

Damian Smedley[1,†], Sebastian Köhler[2,†], Johanna Christina Czeschik[3], Joanna Amberger[4], Carol Bocchini[4], Ada Hamosh[4], Julian Veldboer[2,5], Tomasz Zemojtel[2,6] and Peter N. Robinson[2,5,7,8,*]

[1]Mouse Informatics Group, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK, [2]Institute for Medical Genetics and Human Genetics, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, [3]Genome Informatics Department, Institute of Human Genetics, University Hospital Essen, University of Duisburg-Essen, Hufelandstr. 55, 45122 Essen, Germany, [4]McKusick-Nathans Institute of Genetic Medicine, John Hopkins University School of Medicine, Baltimore, MD 21205, USA, [5]Department of Mathematics and Computer Science, Institute for Bioinformatics, Freie Universität Berlin, Takustrasse 9, 14195 Berlin, Germany, [6]Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-701 Poznan, Poland, [7]Berlin-Brandenburg Center for Regenerative Therapies, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin and [8]Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Whole-exome sequencing (WES) has opened up previously unheard of possibilities for identifying novel disease genes in Mendelian disorders, only about half of which have been elucidated to date. However, interpretation of WES data remains challenging.

Results: Here, we analyze protein–protein association (PPA) networks to identify candidate genes in the vicinity of genes previously implicated in a disease. The analysis, using a random-walk with restart (RWR) method, is adapted to the setting of WES by developing a composite variant-gene relevance score based on the rarity, location and predicted pathogenicity of variants and the RWR evaluation of genes harboring the variants. Benchmarking using known disease variants from 88 disease-gene families reveals that the correct gene is ranked among the top 10 candidates in $\geq$50% of cases, a figure which we confirmed using a prospective study of disease genes identified in 2012 and PPA data produced before that date. We implement our method in a freely available Web server, ExomeWalker, that displays a ranked list of candidates together with information on PPAs, frequency and predicted pathogenicity of the variants to allow quick and effective searches for candidates that are likely to reward closer investigation.

Availability and implementation: http://compbio.charite.de/ExomeWalker

Contact: peter.robinson@charite.de

Received on March 18, 2014; revised on June 30, 2014; accepted on July 18, 2014

1 INTRODUCTION

The identification of causative disease genes in Mendelian disorders has contributed greatly to our understanding of gene functions and biological pathways in rare and common disease (Antonarakis and Beckmann, 2006). With the development of whole-exome sequencing (WES), the pace of identification of novel disease genes has accelerated (Gilissen *et al.*, 2011) to the extent that groups such as the International Rare Disease Research Consortium has set out the goal of comprehensive discovery of the molecular etiologies of all rare diseases to enable molecular diagnosis for all affected individuals by the year 2020 (Baxter and Terry, 2011).

Before WES, most gene discovery projects made use of linkage analysis or association studies, which typically identified genomic intervals of 0.5–10 cm containing up to 300 genes (Botstein and Risch, 2003; Glazier *et al.*, 2002). Numerous computational procedures have been developed to prioritize candidate genes in the intervals and guide DNA sequencing efforts (reviewed in Moreau and Tranchevent, 2012). Although WES provides sequence information for the great majority of targeted exon sequences, the need for prioritization remains. An individual exome typically contains >30 000 variants as compared with the genomic reference sequence, thousands of which are predicted to lead to nonsynonymous amino acid substitutions, alterations of conserved splice site residues or small insertions or deletions. Typical analysis strategies have relied on the characteristics of the variants, focusing on rare variants that are predicted to be pathogenic (Robinson *et al.*, 2011), but even after such filtering, around ~100–1000 candidate disease-causing variants are found in a single WES dataset, and additional methods are needed to predict which of them may have serious functional consequences and prioritize them for validation (Li *et al.*, 2013; Pelak *et al.*, 2010). Because each genome harbors ~100 genuine loss-of-function (LOF) variants with ~20 genes completely inactivated (MacArthur *et al.*, 2012), a purely variant-based prioritization of candidate genes in WES studies will be limited in its ability to correctly identify the true disease gene.

Previous gene prioritization strategies for prioritizing genes in linkage studies evaluated one or more characteristics of the

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
*To whom correspondence should be addressed.

genes, including functional annotation, gene-expression data or sequence-based features (Tranchevent *et al*., 2011). Strategies to prioritize candidate genes in exome sequencing studies can also exploit the variant data itself in an attempt to improve prioritization of Mendelian disease genes, somatic mutations in cancer and others. A number of tools and pipelines have been developed that exploit sophisticated variant filtering strategies. The tools combine filtering steps that exclude common variants and retain only variants that are predicted likely pathogenic using tools such as MutationTaster (Schwarz *et al*., 2010), and then exploit sequences from multiple unrelated individuals with the sought-after disease to search for genes mutated in most or all of the individuals, as well as linkage or pedigree analysis (Coutant *et al*., 2012; Li *et al*., 2012; Santoni *et al*., 2014; Sifrim *et al*., 2012; Yandell *et al*., 2011; Zhang *et al*., 2013). Recently, approaches have been introduced that combine variant impact prediction with gene prioritization. The eXtasy algorithm uses genomic data fusion to integrate variant impact prediction, haploinsufficiency prediction and phenotype-specific gene prioritization (Sifrim *et al*., 2013). The Exomiser implements PHIVE, PHenotypic Interpretation of Variants in Exomes, an algorithm that integrates the calculation of phenotype similarity between human diseases and genetically modified mouse models, with evaluation of the variants according to allele frequency, pathogenicity and mode of inheritance (Robinson *et al*., 2014). FunSeq intersects regions of the genome that are likely to be sensitive to mutations with an analysis for variants that disrupt transcription-factor binding sites (Khurana *et al*., 2013). Each of these algorithms essentially seeks genes or genomic regions that are both relevant to the disease under investigation and also harbor variants likely to be pathogenic. We therefore reasoned that a key factor in exome prioritization algorithms is to intersect the results of variant analysis with a method that can prioritize genes according to their potential relevance.

The analysis of protein interaction networks has been widely used for computational analysis of human disease (Barabási, 2007; Gonzalez and Kann, 2012). Typically, proteins do not act in isolation, but rather perform their functions cooperatively within a network of functionally related proteins. That is, groups of functionally related proteins may physically interact with one another and thereby form a 'molecular nanomachine' that mediates a particular biological function at cellular or systems level. A protein–protein interaction (PPI) may be defined as a specific physical contact with molecular docking between proteins that occurs in cells or in a living organism *in vivo* (De Las Rivas and Fontanillo, 2010). Currently, data on >100 000 PPIs in humans are available (Schaefer *et al*., 2013), derived from experimental methods including the yeast two-hybrid system and tandem affinity purification. In this work, we make use of data from the search tool for the retrieval of interacting genes/proteins (STRING) (Franceschini *et al*., 2013), which contains not only PPIs but also indirect (functional) associations derived from genomic context, high-throughput experiments, conserved coexpression and text-mining. We will refer to this network as the protein–protein association (PPA) network. The complete set of all such interactions and associations has been referred to as the interactome, and with the increased quantity and quality of such data, analysis of the protein interactome offers an important resource for systems-level understanding of cellular processes.

The interactome has also become an important resource for the computational prioritization of disease genes (Moreau and Tranchevent, 2012). The main assumption of these methods is that genes linked to diseases with similar or even identical phenotypic manifestations will in many cases code for genes that interact in specific subnetworks within the larger interactome. Therefore, lists of candidate genes can be prioritized according to the vicinity of the candidates genes within the interactome to other known members of a given disease-gene family. Initial efforts to rank disease genes exploited the presence of direct interactions (Oti *et al*., 2006) or the length of the shortest path of interactions leading from a candidate gene to a known disease gene (George *et al*., 2006). We have shown that a global network measure of distance in the protein–protein interaction network obtained by random walk analysis, substantially improves candidate–gene prioritization, including the search for direct neighbors of other disease genes (Köhler *et al*., 2008). In fact, it was shown that random-walk approaches outperform other gene-prioritization methods (Navlakha and Kingsford, 2010). In this work, we test the hypothesis that random-walk analysis of the protein interactome can improve prioritization of candidate disease genes in exome sequencing studies.

2 METHODS

2.1 Protein–protein and functional interaction data

The PPA network is represented by an undirected graph with nodes representing the genes and edges representing the mapped interactions of the proteins encoded by the genes. Data were taken from STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) version 9.05, which contains experimental, predicted and transferred protein–protein interactions, together with interactions obtained through text mining (Franceschini *et al*., 2013). Only high-confidence interactions (score at least 0.7) were used.

2.2 Disease-gene families

A disease-gene family is defined here as a group of genes in which a mutation in any one of the genes leads to a clinically similar disorder. Thus, a disease-gene family comprises the genes associated with some genetically heterogeneous disease. In this work, we have used data on the phenotypic series from Online Mendelian Inheritance in Man (OMIM) (Amberger *et al*., 2011) of March 2013. Each phenotypic series provides a view of genetic heterogeneity of similar phenotypes across the genome.

2.3 Simulation of whole-exome and disease data

To validate our methodology, we developed a simulation strategy based on adding known disease-causing mutations from the Human Gene Mutation Database (HGMD) into either one of 1092 unaffected whole-exome files in variant call format (VCF) from the 1000 Genomes Project (1000 Genomes Project Consortium *et al*., 2012) or 144 in-house exomes. The 1000 Genomes Project individual whole-exome files were extracted from the integrated call sets (October 12, 2012 release) using tabix (Li, 2011) version 0.2.6 and vcftools (Danecek *et al*., 2011) version 0.1.9. From the initial 233 phenotypic series involving 1356 genes, we eventually tested 88 series that contained at least four genes and with known HGMD mutation(s) for the disease described in the phenotypic series, corresponding to 285 genes. For autosomal dominant (AD) diseases, one heterozygous mutation was added, and for autosomal recessive (AR) diseases, one homozygous mutation was added to the whole-exome files.

2.4 Whole-exome analysis and filtering

For each of the simulated exomes, we used an exome analysis pipeline to filter variants according to rarity, predicted pathogenicity and conformance with the expected mode of inheritance. To filter variants according to rarity, information concerning population minor allele frequency (MAF) of variants was derived from the database of single nucleotide polymorphisms (dbSNP) (NCBI Resource Coordinators, 2013) and from the Exome Variant Server (2013). For this work, the maximum population frequency of a variant was taken to be its maximal reported frequency in any data source. For the dbSNP data, only the reported frequencies from Phase I 1000 Genome Project variants were included. In addition, ExomeWalker scores variants according to the MAF as previously described (Robinson *et al.*, 2014) to give a frequency score between 1 and 0 for variants with a MAF between 0 and 2%, with more common variants receiving a score of 0. In all simulations reported in this work, variants with a MAF >1% were excluded.

In a typical whole-exome analysis, many of the variants have no available frequency data in public databases for assessment. Hence, for the simulations involving 1000 Genomes Project-based exomes, we did not make use of the 1000 Genomes Project frequency data, as this would lead to an unfair advantage because each of the non–disease-associated variants would have frequency data available for filtering and prioritization.

Variants in the VCF files (which are defined using chromosomal coordinates) were then annotated at transcript level using Jannovar (Jäger *et al.*, 2014). To filter variants according to predicted pathogenicity, a variant score was calculated for each variant. First of all, off-target variants (those not located in protein coding sequences or in splice sites) were given a score of zero and removed. Secondly, non-synonymous variants leading to the substitution of an amino acid residue were scored according to the most deleterious prediction of SIFT (Ng and Henikoff, 2002), Polyphen2 (Adzhubei *et al.*, 2010) or MutationTaster (Schwarz *et al.*, 2010). These predictions were extracted from dbNSFP (Liu *et al.*, 2011). Links between genes and Mendelian diseases were extracted from data of the Online Mendelian Inheritance in Man resource (Amberger *et al.*, 2011). In some cases, no predictions were available from any of these three sources, and a pathogenicity prediction of 0.6 was assigned. This value was arrived at during optimization for another exome prioritization tool (Robinson *et al.*, 2014) and represents a compromise between assuming novel variants are non-pathogenic or fully pathogenic. For other classes of variants, pathogenicity scores were assigned as previously described (Robinson *et al.*, 2014). Future versions of ExomeWalker will look to incorporate a single measure of pathogenicity for all types of variants such as CADD scores (Kircher *et al.*, 2014).

For variants that pass the filtering steps, a variant score is assigned for prioritization, which is simply the product of this pathogenicity score and the frequency score described above.

2.5 Random walk analysis

The random walk on graphs (Can *et al.*, 2005) is defined as an iterative walker's transition from its current node to a randomly selected neighbor starting at a given source node, *s*. Here, we used a variant of the random walk in which we additionally allow the restart of the walk in every time step at node s with probability *r*. Formally, the random walk with restart (RWR) is defined as

$$\mathbf{p}_{t+1} = (1-r)\mathbf{W}\mathbf{p}_t + r\mathbf{p}_0 \quad (1)$$

The transition matrix $\mathbf{W}$ is the column-normalized adjacency matrix of the graph, and $\mathbf{p}_t$ is a vector in which the i^{th} element holds the probability of being at node i at time step t.

In our application, the initial probability vector $\mathbf{p}_0$ was constructed such that equal probabilities were assigned to the nodes representing members of the disease, with the sum of the probabilities equal to 1. This is equivalent to letting the random walker begin from each of the known disease genes with equal probability. Candidate genes were ranked according to the values in the steady-state probability vector $\mathbf{p}_\infty$. While it is possible to obtain $\mathbf{p}_\infty$ by explicitly calculating Equation (1) until convergence, we instead solve the equation $\mathbf{p}_\infty = (1-r)\mathbf{W}\mathbf{p}_\infty + r\mathbf{p}_0$ to obtain

$$\mathbf{p}_\infty = r(\mathbf{I} - (1-r)\mathbf{W})^{-1}\mathbf{p}_0 \quad (2)$$

By precalculating the matrix $r(\mathbf{I} - (1-r)\mathbf{W})^{-1}$, we can perform random walk analysis as a simple matrix multiplication of the vector $\mathbf{p}_0$ in $\mathcal{O}(n^2)$ time, where n is the number of genes in the network. Therefore, denoting $r(\mathbf{I} - (1-r)\mathbf{W})^{-1}$ by $\mathbf{R}$, we can calculate the result of the random walk analysis by a simple matrix multiplication $\mathbf{p}_\infty = \mathbf{R}\mathbf{p}_0$. We can further simplify the calculations by noting that most of the elements of the vector $\mathbf{p}_0$ are zero, with only the elements representing the m seed genes having the non-zero value $\frac{1}{m}$. Denoting the set of the indices of these elements as $\{j\}$, then it is easy to see that only the corresponding columns of $\mathbf{R}$ contribute to the final values of $\mathbf{p}_\infty$, whose i^{th} element can be given as

$$\mathbf{p}_\infty[i] = \frac{1}{m}\sum_{j\in\{j\}} \mathbf{R}[j, i] \quad (3)$$

That is, to get element i in $\mathbf{p}_\infty$, we need only to take the sum of the products of the non-zero elements of $\mathbf{p}_0$ with the corresponding elements of column i of $\mathbf{R}$. The computational complexity of the random walk analysis in Equation (1) is dominated by the matrix-vector multiplications in each step, which is $\mathcal{O}(n^2)$ for an $n \times n$ matrix. In contrast, our method requires precomputation of one matrix inversion, but the actual calculation of $\mathbf{p}_\infty$ is $\mathcal{O}(mn)$ with $m \ll n$, as there are $\mathcal{O}(m)$ operations to calculate Equation (3), which has to be done for each of the n elements of $\mathbf{p}_\infty$.

$\mathbf{p}_\infty$ is a probability vector, and all its entries sum to unity. For the purposes of exome analysis, only those genes that have rare predicted pathogenic variants are considered. For the analysis described in this article, we chose a value of 0.7 for the restart probability r.

2.6 ExomeWalker score

Finally, a gene is assigned a combined ExomeWalker score, which is a combination of the random walk score and the best scoring variant in that gene. In the case of AR inheritance under a compound heterozygous model, the variant score is taken to be the average of the two highest scoring variants. Logistic regression on a training set of 20 000 disease variants and 20 000 benign variants was run through the Waikato Environment for Knowledge Analysis (WEKA) (Hall *et al.*, 2009) data-mining package to generate the optimal way of combining the variant and random walker scores into a final ExomeWalker score. A 10-fold cross validation was used to train and test the model, and the average of the 10 models was used for the final algorithm. Receiver operating characteristic analysis on the test datasets gave an average area under the curve of 0.96 for the ExomeWalker score compared with 0.78 for the variant score and 0.9 for the random walk score. This final ExomeWalker score gives a measure from 0 to 1 of how close the gene is to known disease-associated genes in the interactome and how rare and pathogenic are the variants in the gene.

2.7 Benchmarking of ExomeWalker

For the simulated exomes involving known disease variants in 285 genes from 88 phenotypic series, we performed 5000 analyses per experiment. ExomeWalker was run using the other genes in the phenotypic series as seed genes for the random walk. Genes were ranked by either the variant score or the ExomeWalker score. We then compared performance by assessing how often the known disease gene was recovered as the top hit or in the top 10 or 50 candidates. An ordinal ranking method was used where equal scoring genes are resolved arbitrarily but consistently by assigning a unique rank to each of the ties. In our case, we simply

sort the equally scored genes alphabetically and assign the ranks. This corresponds to the real life use case where a researcher would have to take each of the equally scored top candidates and investigate each one by one for causality by further experimentation or for further candidacy by reviewing the literature/databases using their expert knowledge.

3 RESULTS

In this work, we have implemented an algorithm for prioritizing candidate genes in WES studies by searching for rare variants with predicted pathogenicity in genes located in the vicinity of phenotypically related genes in a functional interaction network. We constructed a PPA network based on 210 945 associations among 12 511 human genes using high-confidence interactions in the STRING database (Franceschini *et al*., 2013). We implemented a global distance measure based on RWR to define similarity between genes within this interaction network (Köhler *et al*., 2008). The RWR algorithm ranks genes on the basis of their similarity to known or suspected disease genes. In our benchmarking, we use phenotypic series from the OMIM resource (Amberger *et al*., 2011) to define disease-gene families. Users of ExomeWalker can either use these same disease-gene families or enter their own list of genes that are known or suspected to be associated with the disease being studied.

In parallel, variants from an exome are annotated and the frequency and predicted pathogenicity are evaluated. Candidate genes with rare, potentially pathogenic variants are then prioritized using both the results of variant evaluation and the vicinity in the PPA network of the genes harboring the variants to the seed genes (see Section 2 for details).

Figure 1 shows the results of our analysis using STRING v9.05 as the source of interactome data. Analysis was performed by adding known disease variants to either in-house exomes or 1000 Genome Project exomes and with and without the appropriate inheritance model for the disease being tested.

The results show a substantial performance increase when including the random walk measure of protein–protein interactions with other genes associated with the disease as compared with either the variant score or the RWR score alone. For example, 39.4% of the tested exomes contained the known disease gene as the top hit for the 1000 Genomes Project–based simulations compared with 1.4% when just using variant pathogenicity and frequency to assess candidacy. This is out of an average of 907 postfiltered genes and 97.1% of the disease genes are kept during this filtering step. This increases to 43.5 and 67.3% for the AD and recessive models out of an average of 632 and 374 postfiltered genes, respectively. Similar performance and gains are seen when adding the disease variants to our in-house exomes that contain many more postfiltered genes (1141 on average for no inheritance model). The correct gene was present within the top 10 ranked candidates in nearly 75% of the simulations using in-house exomes. As shown in Figures 1 and 2, a large proportion of the performance comes from the random walk prioritization of the filtered exome candidates with the addition of variant pathogenicity and frequency data adding a further 5–10% increase.

STRING includes text-mined associations between genes and it is possible some of these associations may come from

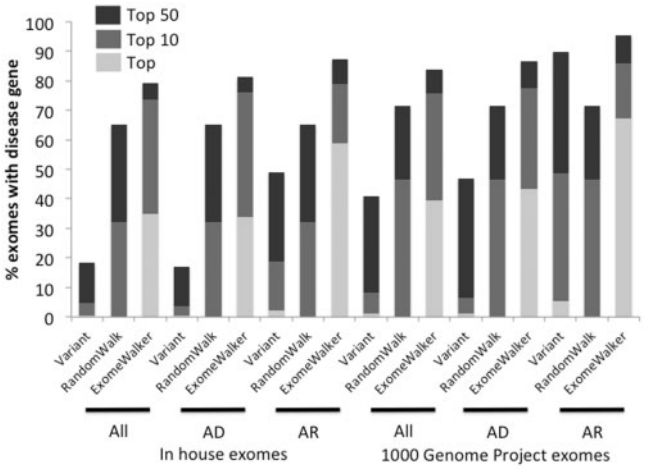

Fig. 1. Performance of ExomeWalker using STRING v9.05 as the source of interactome data. The bars show the percentage of exomes where the true disease gene is identified as the top hit or in the top 10 or 50 results. Either in-house or 1000 Genomes Project exomes were used. All exomes are filtered to remove synonymous, intergenic and intronic variants except for those in splice sites. In addition, variants with a MAF > 1% are excluded. Results are shown without (All) or with an AD or AR inheritance model applied. Ranking is either by Variant scoring that combines MAF and predicted pathogenicity, RWR analysis alone or ExomeWalker scoring that additionally includes evidence of protein–protein associations with other genes linked to the disease

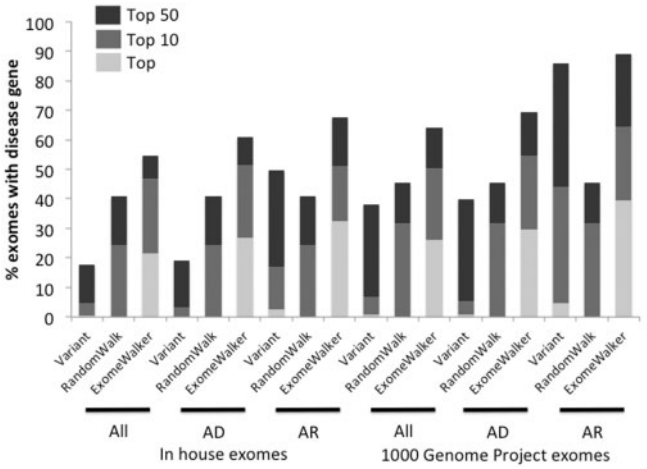

Fig. 2. Performance of ExomeWalker using STRING v9.05 without text-mined associations as the source of interactome data. Abbreviations are as in Figure 1

publications describing two genes being associated with the same disease, rather than the biological associations we are trying to detect with our simulation studies. To allow for this, we repeated the analysis with a version of STRING where all text-mined associations had been removed (Fig. 2).

As expected, there is a drop in performance compared with including the text-mined association but ExomeWalker still shows a substantial improvement over purely variant-based measures of candidacy. For example, for the 1000 Genomes

Project exomes with no inheritance model, the ExomeWalker performance drops from 39.4 to 26.1%, having the correct gene as the top hit but this is still 24.7% higher than with the variant-based scoring alone.

This strategy of removing all text-mined associations will have removed many genuine interactions that would be useful for prioritization of a novel disease–gene association. We expect the real performance of ExomeWalker in such cases to lie somewhere in-between that seen in Figures 1 and 2. To gain a realistic estimate of the performance of our method on new data, we identified 19 disease–gene associations that had been identified in 2012 that belong to one of the phenotypic series and had a known variant in HGMD. We tested the performance of our method using a PPA dataset with data before the discovery of any of these genes (STRING v9.0). The results are summarized in Table 1. The true disease-causing gene was present within the top 10 prioritized genes in 10 of 19 cases (~53%), similar to our results using large-scale simulations.

4 DISCUSSION

Computational candidate gene prioritization has matured into a field that has developed and benchmarked scores of algorithms that exploit and integrate complex and heterogeneous datasets including gene expression, sequence annotations, data mining, genetic sequences, functional annotations and protein–protein interaction networks (Aerts *et al*., 2006; Lage *et al*., 2007; Perez-Iratxeta *et al*., 2002; Tranchevent *et al*., 2011; Turner *et al*., 2003). The fundamental algorithms have been improved and extended in many ways, such as including tissue-specificity in the analysis of the protein interactome (Magger *et al*., 2012). Initial computational analysis of exome sequence data concentrated on filtering variants according to their population frequency, predicted pathogenicity and the presence of rare predicted-pathogenic mutations in multiple unrelated individuals with a certain rare disease ('intersection' strategy; Boycott *et al*., 2013; Robinson *et al*., 2011). However, it has become apparent that it remains difficult to identify novel disease genes or even known disease genes with WES because of the sheer number of candidate mutations; each genome is thought to harbor ~100 genuine LOF variants with ~20 genes completely inactivated (MacArthur *et al*., 2012). Therefore, filtering on variant characteristics alone is not effective in situations where a single affected individual or only a small number of individuals are being investigated. Therefore, just as positional cloning approaches were limited by the availability of large well-characterized families, disease-identification studies by WES are often limited by the number of individual exome sequences available for variant intersection. For this reason, candidate gene prioritization methods have recently begun to be applied to exome analysis. With positional cloning, prioritization would be applied to all genes located within the linkage interval; with exome studies, prioritization is applied to all genes that harbor rare, potentially pathogenic mutations. In both settings, the number of genes may be in the hundreds. Recently, exome prioritization methods have been introduced that exploit data fusion, phenotypic data and model organism phenotype data (Robinson *et al*., 2014; Sifrim *et al*., 2013). Random-walk analysis of protein–protein interaction data has been shown to be a powerful approach to gene prioritization in the setting of positional cloning projects (Köhler *et al*., 2008; Navlakha and Kingsford, 2010). In this work, therefore, we have adapted our previous approach and tested its utility for exome analysis.

Figure 3 illustrates the gene prioritization procedure in the case of *DDOST* and *DPM2*, components of the oligosaccharyltransferase complex that transfers a glycan chain to nascent proteins. Congenital disorders of glycosylation (CDG) are inherited AR diseases that impair *N*-glycosylation, and previously identified CDG disease genes were used to prioritize candidate genes including *DDOST* and *DPM2* in the simulations summarized in Table 1. It can be seen that *DDOST* has only two direct interactions with CDG seed genes and is at some distance from the others, leading to it only being ranked 23rd. However, *DPM2* has multiple direct and second-degree interactions with CDG genes leading to it being ranked as the top-ranked candidate in simulations.

In contrast, other genes did not achieve a high rank. For instance, *TMEM5*, mutations which were shown to be a cause of type A muscular dystrophy-dystroglycanopathy (Vuillaumier-Barrot *et al*., 2012), was placed at rank 19 by our method. This gene has only one high-confidence association in the STRING database, with *PLAU* (plasminogen activator, urokinase), which itself has 10 high-confidence associations to other genes, none of which is related to type A muscular dystrophy-dystroglycanopathy. Therefore, although mutations in *TMEM5* cause the same disease as mutations in the other genes of this family (*POMGNT1*, *POMGNT2*, *ISPD*, *FKTN*, *POMT1*, *POMT2*, *FKRP*, *LARGE*), there is little functional similarity reflecting this in STRING. Thus, although PPA analysis offers an effective way of prioritizing disease genes in many cases, there are disease genes that do not show a high random walk score.

In cases where the causative gene does not interact with previous members of the disease-gene family, or for diseases where there are no previously known genes, other approaches will have to be considered. We recently described an approach, Exomiser, that uses phenotype comparisons with model organism data to inform on candidacy (Robinson *et al*., 2014). eXtasy is another recently published solution that uses phenotype comparisons along with consideration of many other data types (Sifrim *et al*., 2013). To contrast and compare these different approaches we applied them to the same set of recently solved cases and report the performance in Table 1. Note that eXtasy does not perform any variant filtering, and so, to allow a fair comparison we used VCF files that had already been filtered in the same way as for the ExomeWalker benchmarking. Three of the diseases currently have no phenotype annotations available and are therefore not runnable through eXtasy or particularly amenable to Exomiser prioritization. eXtasy can only inform on non-synonymous variants and four of the cases involve a small deletion, which again was not assessable. Finally, for two of the cases, eXtasy removed the causative variant during analysis, so no final ranking was possible. For 2 of the 10 remaining cases, Exomiser and eXtasy performed better than ExomeWalker, with ExomeWalker outperforming them in the other cases. *KLHL3* is a good example where there is minimal evidence for interactions with previously implicated genes but where use of phenotype data allowed identification of the causative variant as the top or second best hit using eXtasy or Exomiser, respectively.

Table 1. List of 19 genes discovered during the year 2012 and for which a disease-causing mutation was listed in HGMD

Gene	ID	Disease gene family	Publication date	Variant	RandomWalk	ExomeWalker	Exomiser	eXtasy
CHMP1A	5119	Pontocerebellar hypoplasia	November 2012 (Mochida *et al.*, 2012)	3	66	3	1	[a]
NMNAT1	64802	Leber congenital amaurosis	September 2012 (Falk *et al.*, 2012)	61	14	61	52	99
CEP135	9662	Microcephaly, primary AR	May 2012 (Hussain *et al.*, 2012)	18	3	2	16	[b]
KLHL3	26249	Pseudohypoaldosteronism, type II	January 2012 (Boyden *et al.*, 2012)	4	131	5	2	1
THRA	7067	Hypothyroidism, congenital, non-goitrous	January 2012 (Bochukova *et al.*, 2012)	51	2	1	148	5
TMEM5	10329	Muscular dystrophy-dystroglycanopathy, type A	December 2012 (Vuillaumier-Barrot *et al.*, 2012)	19	298	19	19	[b]
DDOST	1650	Congenital disorders of glycosylation, type I	February 2012 (Jones *et al.*, 2012)	22	9	23	17	[b]
PNPT1	87178	Combined oxidative phosphorylation deficiency	November 2012 (Vedrenne *et al.*, 2012)	9	4	2	119	3
DPM2	8818	Congenital disorders of glycosylation, type I	October 2012 (Barone *et al.*, 2012)	3	1	1	2	4
PACS1	55690	Mental retardation, AD	December 2012 (Schuurs-Hoeijmakers *et al.*, 2012)	29	283	39	15[c]	[c]
ADAR	103	Aicardi-Goutieres syndrome	November 2012 (Rice *et al.*, 2012)	17	26	17	119[c]	[c]
CABP2	51475	Deafness, AR	October 2012 (Schrauwen *et al.*, 2012)	59	109	61	40	[a]
DST	667	Hereditary sensory and autonomic neuropathy	April 2012 Edvardson *et al.*, 2012	18	108	18	132	[b]
VPS37A	137492	Spastic paraplegia	July 2012 (Zivony-Elboum *et al.*, 2012)	11	318	11	9	10
HOXC13	3229	Ectodermal dysplasia	November 2012 (Lin *et al.*, 2012)	19	5	17	131	48
KANSL1	284058	Mental retardation, AD	April 2012 (Koolen *et al.*, 2012; Zollino *et al.*, 2012)	88	63	45	222	178
GUCY2C	2984	Diarrhea, congenital	April 2012 (Fiskerstrand *et al.*, 2012; Wu *et al.*, 2012)	18	4	1	124	13
PFN1	5216	Amyotrophic lateral sclerosis	August 2012 (Wu *et al.*, 2012)	46	68	50	166	91
CHD8	57680	Autism	December 2012 (O'Roak *et al.*, 2012a and b)	80	95	80	192	[b]

Notes. The first column shows the gene symbol and the second column shows the NCBI Entrez Gene ID. The third column provides the OMIM phenotypic series to which the gene was assigned after its identification as being causative for the disease. The next three columns show the rank obtained in ExomeWalker analysis using STRING v9.0, 1000 Genomes Project exomes and mode of inheritance filtering sorted by the variant, random walk or combined ExomeWalker score. Finally, we show the ranks obtained from the Exomiser and eXtasy tools that take an alternative approach of prioritizing by phenotype. The columns Variant, RandomWalk, ExomeWalker, Exomiser and eXtasy show the ranks of the gene by each method. The eXtasy rank was obtained after preprocessing.

[a]Variant lost during eXtasy prioritization.

[b]Deletion so not suitable for eXtasy prioritization.

[c]No phenotype annotations.

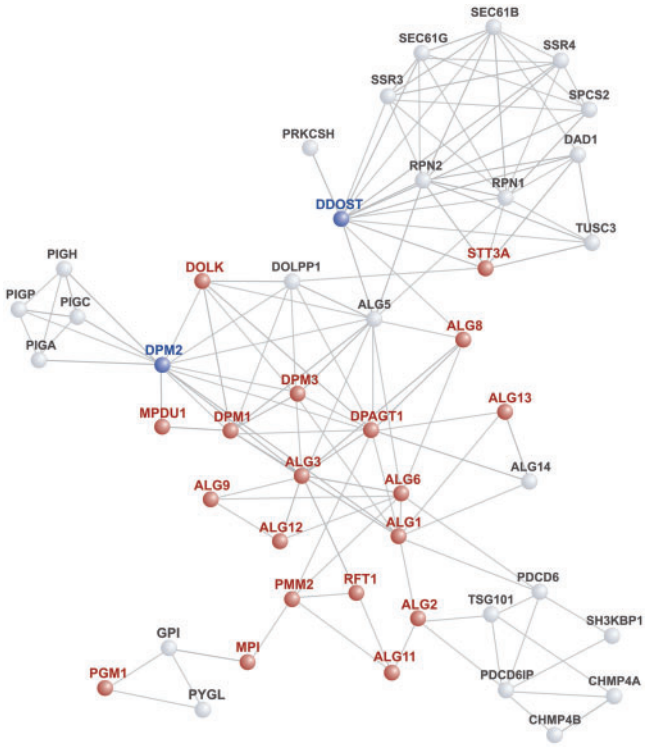

Fig. 3. PPA network derived from congenital disorders of glycosylation, type I (CDG-I) seed genes. The candidate genes *DDOST* and *DMP2* (shown in blue) interact with multiple other CDG-I genes (shown as red nodes in the network) via paths of length one and two. The random walk methodology essentially integrates over all interaction paths between seed genes and a candidate gene to generate a similarity score. Although short paths such as those shown in the figure have the most influence on the score, other aspects of the global network structure are also taken into account (Köhler *et al*., 2008)

In contrast for the three cases where ExomeWalker identified the causative gene as the top hit, both eXtasy and Exomiser were unable to achieve this efficient prioritization.

5 CONCLUSION

We have implemented our method in a freely available Web server called ExomeWalker. Users can upload a VCF file and choose one of 243 phenotypic series or enter their own disease-gene family in the form of a list of Entrez Gene identifiers. These genes may already be known to be associated with the disease or be members of a pathway suspected of being disrupted in the disease or just candidates from in-house knowledge. If the VCF file contains multiple samples, ExomeWalker will assume that all samples are from the same family and will ask the user to upload a pedigree (PED) file. It will then perform pedigree filtering on the genes and variants represented in the VCF file using the Jannovar library (Jäger *et al*., 2014). It will subsequently rank the candidate genes and return a list of candidates together with information about the genes. Importantly, it will show all first- and second-degree interactions with the seed genes, allowing users to quickly eyeball candidate lists to determine if there are genes with multiple functional associations with the seed genes that would reward closer inspection. Exome sequencing remains a difficult endeavor, and large-scale exome-sequencing studies for the identification of Mendelian disease causing genes have reported success rates around 20–35% (de Ligt *et al*., 2012; Yang *et al*., 2013). Therefore, it is not realistically to be expected that a prioritization method will place the correct gene in the first place, or first few places, in all cases. An advantage of the methodology presented here is that ExomeWalker quickly shows whether there are candidate genes with both predicted pathogenic variants and multiple functional associations with other genes in the same disease-gene family. If this is not the case, users may wish to explore phenotype-based (Robinson *et al*., 2014) or genomic data fusion (Sifrim *et al*., 2013) prioritization of exome data, or if possible sequence additional family samples to enable linkage filtering (Rödelsperger *et al*., 2011; Smith *et al*., 2011), or sequence additional unrelated individuals for intersection-based (Robinson *et al*., 2011) approaches.

The ExomeWalker server is freely available at http://compbio.charite.de/ExomeWalker/.

Funding: This work was supported by grants of the Bundesministerium für Bildung und Forschung (BMBF project number 0313911), by the European Community's Seventh Framework Programme (Grant Agreement 602300; SYBIL) and by core infrastructure funding from the Wellcome Trust. Additional support was provided by the Director, Office of Science, Office of Basic Energy Sciences, of the US Department of Energy under contract no. DE-AC02-05CH11231.

Conflict of interest: none declared.

REFERENCES

1000 Genomes Project Consortium. et al. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Adzhubei,I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Aerts,S. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, **24**, 537–544.

Amberger,J. *et al.* (2011) A new face and new challenges for Online Mendelian Inheritance in Man (OMIM[textregistered]) *Hum. Mutat.*, **32**, 564–567.

Antonarakis,S.E. and Beckmann,J.S. (2006) Mendelian disorders deserve more attention. *Nat. Rev. Genet.*, **7**, 277–282.

Barabási,A.-L. (2007) Network medicine–from obesity to the "diseasome". *N. Engl. J. Med.*, **357**, 404–407.

Barone,R. *et al.* (2012) Dpm2-cdg: a muscular dystrophy-dystroglycanopathy syndrome with severe epilepsy. *Ann. Neurol.*, **72**, 550–558.

Baxter,K. and Terry,S.F. (2011) International rare disease research consortium commits to aggressive goals. *Genet. Test Mol. Biomarkers*, **15**, 465.

Bochukova,E. *et al.* (2012) A mutation in the thyroid hormone receptor alpha gene. *N. Engl. J. Med.*, **366**, 243–249.

Botstein,D. and Risch,N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.*, **33** (Suppl.), 228–237.

Boycott,K.M. *et al.* (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.*, **14**, 681–691.

Boyden,L.M. *et al.* (2012) Mutations in kelch-like 3 and cullin 3 cause hypertension and electrolyte abnormalities. *Nature*, **482**, 98–102.

Can,T. *et al.* (2005) Analysis of protein-protein interaction networks using random walks. In: *BIOKDD'05: Proceedings of the 5th international workshop on*

Bioinformatics, Association for Computing Machinery, New York, USA, pp. 61-68.
Coutant,S. *et al.* (2012) Eva: exome variation analyzer, an efficient and versatile tool for filtering strategies in medical genomics. *BMC Bioinformatics*, **13** (Suppl. 14), S9.
Danecek,P. *et al.* (2011) The variant call format and vcftools. *Bioinformatics*, **27**, 2156–2158.
De Las Rivas,J. and Fontanillo,C. (2010) Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.*, **6**, e1000807.
de Ligt,J. *et al.* (2012) Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.*, **367**, 1921–1929.
Edvardson,S. *et al.* (2012) Hereditary sensory autonomic neuropathy caused by a mutation in dystonin. *Ann. Neurol.*, **71**, 569–572.
Exome Variant Server. (2013) *Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP)* Seattle, WA. http://evs.gs.washington.edu/EVS/ (8 January 2013 date last accessed).
Falk,M.J. *et al.* (2012) Nmnat1 mutations cause leber congenital amaurosis. *Nat. Genet.*, **44**, 1040–1045.
Fiskerstrand,T. *et al.* (2012) Familial diarrhea syndrome caused by an activating gucy2c mutation. *N. Engl. J. Med.*, **366**, 1586–1595.
Franceschini,A. *et al.* (2013) String v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.
George,R.A. *et al.* (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res.*, **34**, e130.
Gilissen,C. *et al.* (2011) Unlocking mendelian disease using exome sequencing. *Genome Biol.*, **12**, 228.
Glazier,A.M. *et al.* (2002) Finding genes that underlie complex traits. *Science*, **298**, 2345–2349.
Gonzalez,M.W. and Kann,M.G. (2012) Chapter 4: Protein interactions and disease. *PLoS Comput. Biol.*, **8**, e1002819.
Hall,M. *et al.* (2009) The weka data mining software: an update. *SIGKDD Explor.*, **11**, 10–18.
Hussain,M.S. *et al.* (2012) A truncating mutation of cep135 causes primary microcephaly and disturbed centrosomal function. *Am. J. Hum. Genet.*, **90**, 871–878.
Jäger,M. *et al.* (2014) Jannovar: a java library for exome annotation. *Hum. Mut.*, **35**, 548–555.
Jones,M.A. *et al.* (2012) Ddost mutations identified by whole-exome sequencing are implicated in congenital disorders of glycosylation. *Am. J. Hum. Genet.*, **90**, 363–368.
Khurana,E. *et al.* (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*, **342**, 1235587.
Kircher,M. *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
Köhler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.
Koolen,D.A. *et al.* (2012) Mutations in the chromatin modifier gene kansl1 cause the 17q21.31 microdeletion syndrome. *Nat. Genet.*, **44**, 639–641.
Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.
Li,H. (2011) Tabix: fast retrieval of sequence features from generic tab-delimited files. *Bioinformatics*, **27**, 718–719.
Li,M.X. *et al.* (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of mendelian diseases. *Nucleic Acids Res.*, **40**, e53.
Li,M.X. *et al.* (2013) Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet.*, **9**, e1003143.
Lin,Z. *et al.* (2012) Loss-of-function mutations in hoxc13 cause pure hair and nail ectodermal dysplasia. *Am. J. Hum. Genet.*, **91**, 906–911.
Liu,X. *et al.* (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.*, **32**, 894–899.
MacArthur,D.G. *et al.* (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, **335**, 823–828.
Magger,O. *et al.* (2012) Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput. Biol.*, **8**, e1002690.
Mochida,G.H. *et al.* (2012) Chmp1a encodes an essential regulator of bmi1-ink4a in cerebellar development. *Nat. Genet.*, **44**, 1260–1264.
Moreau,Y. and Tranchevent,L.C. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.*, **13**, 523–536.
Navlakha,S. and Kingsford,C. (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, **26**, 1057–1063.
NCBI Resource Coordinators. (2013) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **41**, D8–D20.
Ng,P.C. and Henikoff,S. (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res.*, **12**, 436–446.
O'Roak,B.J. *et al.* (2012a) Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science*, **338**, 1619–1622.
O'Roak,B.J. *et al.* (2012b) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, **485**, 246–250.
Oti,M. *et al.* (2006) Predicting disease genes using protein-protein interactions. *J. Med. Genet.*, **43**, 691–698.
Pelak,K. *et al.* (2010) The characterization of twenty sequenced human genomes. *PLoS Genet.*, **6**, e1001111.
Perez-Iratxeta,C. *et al.* (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, **31**, 316–319.
Rice,G.I. *et al.* (2012) Mutations in adar1 cause aicardi-goutires syndrome associated with a type i interferon signature. *Nat. Genet.*, **44**, 1243–1248.
Robinson,P.N. *et al.* (2011) Strategies for exome and genome sequence data analysis in disease-gene discovery projects. *Clin. Genet.*, **80**, 127–132.
Robinson,P.N. *et al.* (2014) Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.*, **24**, 340–348.
Rödelsperger,C. *et al.* (2011) Identity-by-descent filtering of exome sequence data for disease-gene identification in autosomal recessive disorders. *Bioinformatics*, **27**, 829–836.
Santoni,F.A. *et al.* (2014) Simultaneous identification and prioritization of variants in familial, de novo, and somatic genetic disorders with variantmaster. *Genome Res.*, **24**, 349–355.
Schaefer,M.H. *et al.* (2013) Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLoS Comput. Biol.*, **9**, e1002860.
Schrauwen,I. *et al.* (2012) A mutation in cabp2, expressed in cochlear hair cells, causes autosomal-recessive hearing impairment. *Am. J. Hum. Genet.*, **91**, 636–645.
Schuurs-Hoeijmakers,J.H.M. *et al.* (2012) Recurrent de novo mutations in pacs1 cause defective cranial-neural-crest migration and define a recognizable intellectual-disability syndrome. *Am. J. Hum. Genet.*, **91**, 1122–1127.
Schwarz,J.M. *et al.* (2010) Mutationtaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.
Sifrim,A. *et al.* (2012) Annotate-it: a swiss-knife approach to annotation, analysis and interpretation of single nucleotide variation in human disease. *Genome Med.*, **4**, 73.
Sifrim,A. *et al.* (2013) extasy: variant prioritization by genomic data fusion. *Nat. Methods*, **10**, 1083–1084.
Smith,K.R. *et al.* (2011) Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol.*, **12**, R85.
Tranchevent,L.-C. *et al.* (2011) A guide to web tools to prioritize candidate genes. *Brief. Bioinform.*, **12**, 22–32.
Turner,F.S. *et al.* (2003) Pocus: mining genomic sequence annotation to predict disease genes. *Genome Biol.*, **4**, R75.
Vedrenne,V. *et al.* (2012) Mutation in pnpt1, which encodes a polyribonucleotide nucleotidyltransferase, impairs rna import into mitochondria and causes respiratory-chain deficiency. *Am. J. Hum. Genet.*, **91**, 912–918.
Vuillaumier-Barrot,S. *et al.* (2012) Identification of mutations in tmem5 and ispd as a cause of severe cobblestone lissencephaly. *Am. J. Hum. Genet.*, **91**, 1135–1143.
Wu,C.-H. *et al.* (2012) Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis. *Nature*, **488**, 499–503.
Yandell,M. *et al.* (2011) A probabilistic disease-gene finder for personal genomes. *Genome Res.*, **21**, 1529–1542.
Yang,Y. *et al.* (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.*, **369**, 1502–1511.
Zhang,L. *et al.* (2013) Privar: a toolkit for prioritizing snvs and indels from next-generation sequencing data. *Bioinformatics*, **29**, 124–125.
Zivony-Elboum,Y. *et al.* (2012) A founder mutation in vps37a causes autosomal recessive complex hereditary spastic paraparesis. *J. Med. Genet.*, **49**, 462–472.
Zollino,M. *et al.* (2012) Mutations in kansl1 cause the 17q21.31 microdeletion syndrome phenotype. *Nat. Genet.*, **44**, 636–638.