

Filtering for Compound Heterozygous Sequence Variants in Non-Consanguineous Pedigrees

Tom Kamphans¹, Peggy Sabri², Na Zhu², Verena Heinrich², Stefan Mundlos^{2,3}, Peter N. Robinson², Dmitri Parkhomchuk², Peter M. Krawitz^{2,3*}

1 Smart Algos, Berlin, Germany, **2** Institute for Medical Genetics and Human Genetics, Charité Universitätsmedizin, Berlin, Germany, **3** Max Planck Institute for Molecular Genetics, Berlin, Germany

Abstract

The identification of disease-causing mutations in next-generation sequencing (NGS) data requires efficient filtering techniques. In patients with rare recessive diseases, compound heterozygosity of pathogenic mutations is the most likely inheritance model if the parents are non-consanguineous. We developed a web-based compound heterozygous filter that is suited for data from NGS projects and that is easy to use for non-bioinformaticians. We analyzed the power of compound heterozygous mutation filtering by deriving background distributions for healthy individuals from different ethnicities and studied the effectiveness in trios as well as more complex pedigree structures. While usually more than 30 genes harbor potential compound heterozygotes in single exomes, this number can be markedly reduced with every additional member of the pedigree that is included in the analysis. In a real data set with exomes of four family members, two sisters affected by Mabry syndrome and their healthy parents, the disease-causing gene *PIGO*, which harbors the pathogenic compound heterozygous variants, could be readily identified. Compound heterozygous filtering is an efficient means to reduce the number of candidate mutations in studies aiming at identifying recessive disease genes in non-consanguineous families. A web-server is provided to make this filtering strategy available at www.gene-talk.de.

Citation: Kamphans T, Sabri P, Zhu N, Heinrich V, Mundlos S, et al. (2013) Filtering for Compound Heterozygous Sequence Variants in Non-Consanguineous Pedigrees. PLoS ONE 8(8): e70151. doi:10.1371/journal.pone.0070151

Editor: Kai Wang, University of Southern California, United States of America

Received: April 20, 2013; **Accepted:** June 20, 2013; **Published:** August 5, 2013

Copyright: © 2013 Kamphans et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by a grant from the Deutsche Forschungsgemeinschaft grant to P.M.K. (DFG KR 3985/1-1). GeneTalk is supported by a grant from the Bundesministerium für Wirtschaft und Technologie (BMWT; 03EGSBB082). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Tom Kamphans, who is affiliated to SmartAlgos, is a self-employed software developer and consultant (SmartAlgos). This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: peter.krawitz@gmail.com

Background

In recessive genetic disorders, both copies of a certain gene are defective. For autosomal recessive genes, this means that the maternally as well as the paternally transmitted copy of a gene harbors a pathogenic mutation. The occurrence of a pathogenic mutation can be viewed as a random process, and many different pathogenic mutations have arisen for recessive genes in the human population over time. This also means the lower the kinship of the parents the higher is the chance that two different mutant alleles of the disease gene are present in a patient affected by a recessive disease, whereas in a closely related, consanguineous partnership it is more likely that an affected child will inherit the same pathogenic allele from both parents and thus be homozygous for the disease causing mutation. This translates to a simple rule of thumb: If the parents are non-consanguineous, the most likely explanation for a recessive disease is compound heterozygosity for two different pathogenic mutations. Exceptions from this rule of thumb may be founder mutations in certain populations and specific gain of function mutations in certain genes such as e.g. *FGFR2*.

A challenge in filtering sequence variants for compound heterozygotes is that one has to figure out whether the two heterozygous variants affect different copies of a gene or the same copy. Usually, that cannot be determined from a single DNA

sequence, if the read length is less than the distance between the variants or if it is not possible to phase the haplotypes by any other means. However, when sequence variants of more than one family member are available, one can exclude certain variants based on rules of Mendelian inheritance. We will describe a set of rules that we used for compound heterozygous filtering and analyze how effectively the sequence variants can be reduced in certain case scenarios.

Methods

We implemented a compound heterozygous filter in Ruby inside the GeneTalk framework [1]. We assume that the phenotype is fully penetrant and that all sequenced individuals are either affected or not affected. The first two rules work on a single variant level:

- 1) A variant has to be in a heterozygous state in all affected individuals.
- 2) A variant must not occur in a homozygous state in any of the unaffected individuals.

If a variant were homozygous in an unaffected individual it could not be disease causing, otherwise the individual would have

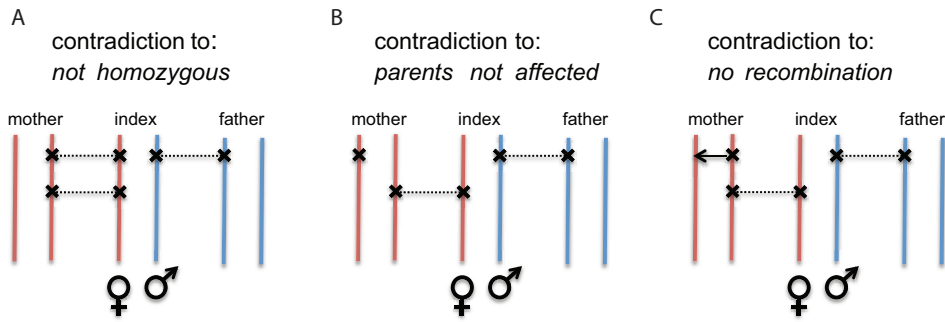


Figure 1. Compound Heterozygote Filtering Rules. If both parents of the index patient are unaffected it is not possible that one of the heterozygous disease causing mutations is present in a heterozygous state in both parents unless a recombination occurred between this variant and the second compound heterozygous mutation.
doi:10.1371/journal.pone.0070151.g001

to be affected, as both copies of the gene harbor the same mutation.

If the genotypes of both parents of an affected child are available and they are both unaffected there is a third rule that is very powerful in reducing the variants:

- 3) A variant that is heterozygous in an affected child must be heterozygous in exactly one of the parents.

This rule is a compact version of:

- 3a) The variant must not be heterozygous in both parents.
- 3b) The variant must be present in at least one of the parents.

Rule 3a is applicable only if no recombination occurred between the tested loci in any of the parents. However, most genes have an extension considerably less than one megabase in the genome and thus the probability of a recombination is usually far below one per cent and the assumption of no recombination is well justified. In Figure 1, we illustrate why a variant that does not fulfill 3a may be removed as not pathogenic. If we keep such a variant it will result in a violation of one of prerequisites for a compound heterozygous mode of inheritance. Without loss of generality we may consider two heterozygous variants in an affected individual. One of them is in a heterozygous state in both unaffected parents. If the variant, for which mother and father are heterozygous, is transmitted by both of them to the index patient,

then the variant would be homozygous in the child and would be therefore removed based on the first rule (Figure 1 A). The index patient could be heterozygous for both variants if both variants occurred on different copies of the gene in one of the parents. However, in this case this particular parent has to be affected as well and the third rule does not apply, as its assumption is not fulfilled (Figure 1B). The third case describes a scenario in which one of the parents has the two heterozygous mutations on the same copy of the gene, while the other parent is heterozygous for only one of them. In this scenario the child could be compound heterozygous only if a recombination happened in the germline of its ancestor (Figure 1C). As already mentioned, this case is so unlikely that we exclude it.

Rule 3b is applicable only if we assume that no *de novo* mutations occurred. The number of *de novo* mutations is estimated to be below five per exome per generation [2,3], thus, the likelihood that an individual is compound heterozygous and at least one of these mutations arose *de novo* is low. If more than one family member is affected, *de novo* mutations are even orders of magnitudes less likely as a recessive disease cause. On the other hand, excluding these variants from the further analysis helps to remove many sequencing artifacts. In linkage analysis for example it is common practice of data cleaning to exclude variants as Mendelian errors if they cannot be explained by Mendelian inheritance.

When the filtering rules 1–3 have been applied on a single variant level, the fourth and fifth rule test on a gene level, whether

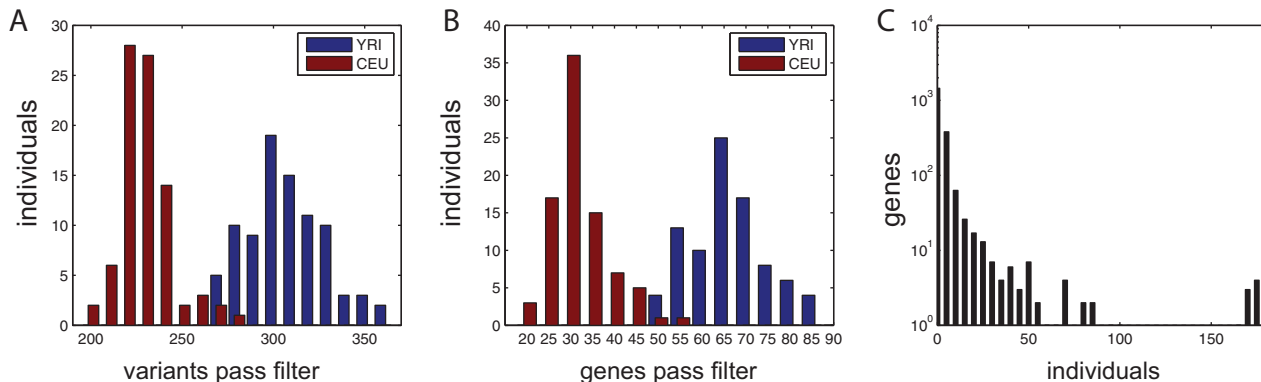


Figure 2. Exomes of 85 European individuals (CEU) as well as 88 African individuals (YRI) were filtered for rare compound heterozygous candidate variants. A) In average around 230 variants pass the filter in CEU exomes and 309 in YRI exomes. B) The potential compound heterozygotes are distributed over 31 genes in CEU individuals and 67 genes in YRI individuals. C) Altogether 1998 genes harbored potential compound heterozygous variants in the tested individuals and compound heterozygotes in 1066 genes occurred only in singular cases.
doi:10.1371/journal.pone.0070151.g002

Chrom	Pos	dbSNP ID	Gene	Ref	Genotype 464 (Affected)	Genotype 465 (Affected)	Genotype 466 (Unaffected)	Genotype 467 (Unaffected)	Effect							
1	145296451	rs140256918	NBPF10	G	C/G	135/93	C/G	170/63	C/G	170/70	G/G	237/0	missense	Annotate	more	
1	145368518	rs61813437	NBPF10	C	C/T	192/44	C/T	203/37	C/C	212/30	C/T	199/38	missense	Annotate	more	
9	35090263	rs142164373	PIGO	G	A/G	28/33	A/G	30/34	A/G	26/32	G/G	67/0	missense	Annotate	Show	more
9	35091529	.	PIGO	G	G/GG	35/42	G/GG	59/49	G/G	101/0	G/GG	38/41	frameshift-insertion	Annotate	more	
19	8999449	rs76798407	MUC16	G	G/T	176/29	G/T	179/40	G/T	185/37	G/G	188/25	missense	Annotate	more	
19	9002597	.	MUC16	C	C/T	62/10	C/T	63/13	C/C	69/3	C/T	74/14	missense	Annotate	more	

Gene Info Exons

Name	PIGO	Entrez	84720
Location	9 : 35088685 - 35096591	CCDS	CCDS6575.1
ID Type	Gene symbol	GeneCards	PIGO
max. CDS	3270		
MRHC	0.0895		
Comment			

Source: CCDS.2011-09-07

Annotations

User	Chrom	Pos	Genotype	Gene	Comment	OMIM-ID	Actions
petkraw	9	35090058	C/T	PIGO	this mutation impairs the correct spl...		
petkraw	9	35090263	A/G	PIGO	This variant impairs the function of ...		
petkraw	9	35091522	T/TG	PIGO	This insertion impairs the function o...		

Figure 3. Filtering results for compound heterozygotes in a case study. With the filter settings for genotype frequency <0.01 , effect on protein level (functional filter: missense, nonsense, stop loss, splice site, insertions or deletions), and compound heterozygotes yields six variants in three genes. *MUC16* and *NBPF10* are both genes from large gene families known for their high variability and detection artifacts due to pseudogenes. The heterozygotes in *PIGO* remain as the likeliest candidates. The *Show* icon at the right end of the line links to an expert curated annotation database that indicates that the mutation in *PIGO* is causing Hyperphosphatasia with mental retardation syndrome and has been published in [9]. The gene view for *PIGO* lists all variant annotations for this gene and links to further knowledge bases. The length of the coding sequence of the longest transcript (max. CDS) and the mean number of rare heterozygous variant calls per exome (MRHC) are important parameters for the assessment of candidate genes.

doi:10.1371/journal.pone.0070151.g003

enough variants remained to fulfill the requirements for a compound heterozygous mode of transmission and whether they were transmitted biparentally:

- 4) A gene must have two or more heterozygous variants in each of the affected individuals.
- 5) There must be at least one variant transmitted from the paternal side and one transmitted from the maternal side.

We did not use two as an upper bound in these rules as not necessarily all of the variants that pass these rules have to be pathogenic. However, as we will discuss later, a gene with many variants passing all five rules, is less likely to be a disease gene. The fifth rule makes sure that only genes pass the filter for which there are at least two heterozygous variants in the affected individuals that are transmitted in a biparental mode. Another way of phrasing this rule is: There must not be two identical haplotypes around the disease gene in an unaffected and an affected individual. If all heterozygous mutations of a gene in an affected individual match all the heterozygous mutations of the same gene in an unaffected individual of the family, then we exclude this gene. Imagine a scenario where sequence variants of only one grandfather or grandmother are available as unaffected control but not directly the sequence variants of the parents of the affected child. Let us assume that this index patient has exactly two heterozygous variants in a gene that match the genotypes of this gene in the grandmother. Excluding recombination this means

that one of the parents of the child was either carrier of these two heterozygous variants or not. However, this means that one of these variants cannot be disease causing or rule three would be broken. The intervals for which we counted and compared the heterozygous genotypes were defined by the gene start and end points that we derived from Ensembl/BioMart [4]. All filters that we applied prior to the compound heterozygote filter are available through the GeneTalk platform. The allele frequency filter is based on genetic variation data from the 1000 genomes project [5] as well as the 5000 exomes project [6]. The effect of the variants on the protein level which is subject to the functional filter was predicted by ANNOVAR [7].

The length of a gene, as well as the variability of its sequence in a healthy reference population affects the probability to observe rare, heterozygous variants in a test individual. We derived the length of the coding sequence of the longest transcript per gene (max. CDS) and determined the mean number of rare, heterozygous variant calls in healthy individuals (MRHC) based on the 5000 exomes data [8] to assist in the interpretation of candidate genes after filtering. The MRHC was computed by adding the frequencies of heterozygous variant calls for all positions in a gene that were below 0.01:

$$MRHC = \sum_{i \in \text{gene}} \frac{r_i}{c_i}, \text{ with } \frac{r_i}{c} \leq 0.01$$

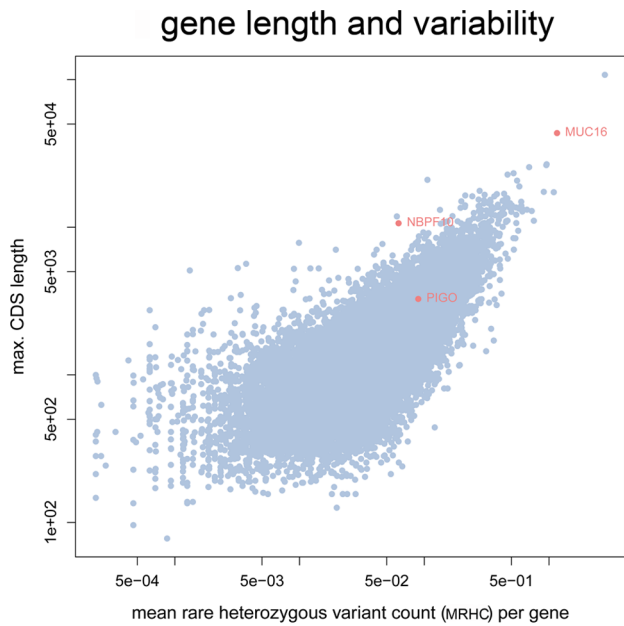


Figure 4. The length of the coding sequence and the mean number of rare alleles per gene. In an average healthy individual from the 5000 exomes project there is more than one rare heterozygous variant in *MUC16* that has an allele frequency below 0.01 in the reference population. In contrast, the coding sequence of *PIGO* is much shorter and rare heterozygous variants occur in less than 8 out of 1000 exomes.
doi:10.1371/journal.pone.0070151.g004

where c_i is the total genotype count at position i and r_i is the heterozygous genotype count at position i .

Results

We tested the effectiveness of the compound heterozygotes filter in GeneTalk for single samples as well as in more complex pedigree structures. If single individuals are analyzed only rules 1 and 4 can be used. Rules 2 and 5 require data of at least one additional unaffected family member and rule 3 is applicable only if sequence data of both parents is available. Any additional filter settings should be used prior to rule 4 as this rule is more effective

in reducing the number of candidate genes the lower the expected number of heterozygous variants is per gene. We used an allele frequency cutoff of 1% for heterozygous variants, removed all synonymous variants and known sequencing artifacts before applying the compound heterozygous filter. We choose the frequency cutoff of 1% as an upper bound, as this is above the allele frequency of the most common pathogenic alleles in cystic fibrosis (MIM 219700), one of the more common autosomal recessive disorders.

With these parameter settings, we filtered 85 European exomes (CEU) as well as 88 African exomes (YRI) available from the 1000 genomes project [5]. All these individuals are healthy and the numbers of variants as well as genes passing our filter settings serve as a background distribution that one has to expect when filtering single exome data. In the CEU exomes in average 230 variants distributed over 31 genes passed the filter, whereas in average 309 variants distributed over 67 genes passed the filter in the Yoruban exomes (Figure 2A and 2B). In the 173 tested individuals we identified variants as possible compound heterozygotes in altogether 1998 genes, and 1066 of these genes were unique to only one of the tested samples (Figure 2C).

As all these exomes were of healthy individuals it would be difficult to identify a single additional gene that passes the filters in a patient with a rare recessive disorder due to the true disease causing compound heterozygotes.

We then analyzed the effectiveness of the compound heterozygous filter for cases in which more than only a single exome of a family is available. For this purpose we used two parent-child trios of European (NA12878, NA12891, NA12892) as well as Yoruban (NA19238, NA19239, NA19240) descent from the 1000 genomes project. We assigned the status of the affected index to the offspring (NA12878, NA19238), so that for both trios all five rules were applicable. As before, in advance to the compound heterozygous filter, we reduced the exome variants of the trios to rare variants with a presumable effect on the protein level and removed known calling artifacts. This prior filtering step yielded 1668 variants in the European trio and 2653 variants in the Yoruban trio. The compound heterozygous filter reduced this number down to 48 variants in 17 genes in the European individual NA12878, while 68 variants in 29 genes passed in the Yoruban individual NA19238. If we remove genes for which compound heterozygous variants also occurred in other unrelated individuals (Figure 1C and Table S1), only twelve candidate genes remained in NA12878 (*CACHD1*, *DPP4*, *CACNA1D*, *PLXNA1*,

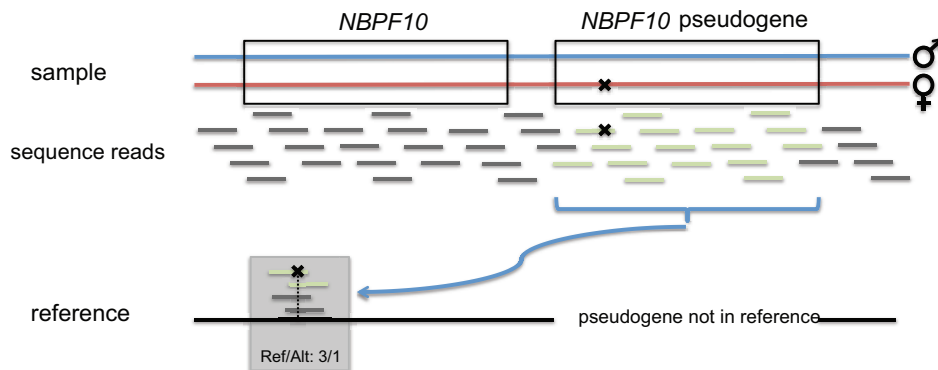


Figure 5. Illustration of mapping artifacts resulting in false positive variant detection. The illustrated sample carries a mutation in the maternal copy of a pseudogene of *NBPF10*. If the pseudogene is not included in the reference sequence, the reads originating from this pseudogene are mismapped. This may result in a false variant call. Indicative for false genotype calls are proportions of reads supporting the alternate allele that strongly deviate from 0.5 or 1.
doi:10.1371/journal.pone.0070151.g005

MECOM, *PSMD2*, *TBC1D4*, *DCAF5*, *ZNFX614*, *MYH9*, *PKDREJ*, *MAGEC1*) and in NA19238 (*CD180*, *ERAP2*, *EPB41L2*, *COL1A2*, *MUC16*, *AASS*, *PZP*, *ZMYM2*, *ZNFX423*, *LRRC48*, *DMC1*, *GPR64*). Thus, in comparison to filtering a single individual the inclusion of the parental genotypes reduces the number of variants by a factor of around five. The removal of highly variable genes as observed by the analysis of single individuals cut the remaining variants in half.

We continued our analysis with exome data from a family with two daughters affected by Mabry syndrome also known as hyperphosphatasia with mental retardation syndrome (MIM 614749) [9]. The parents were of European descent and unrelated. Altogether 27584 distinct variants were detected in the coding regions of all four analyzed family members. Applying the genotype frequency filter and removing known sequencing artifacts reduced this number to 2208 variants. 1446 of these variants had a predicted effect on the protein level. The compound heterozygous filter reduces this set to six variants in three distinct genes, *NBPF10*, *MUC16*, and *PIGO* (see Figure 3).

MUC16 is a member of a large family of mucin coding genes and *NBPF10* is from the neuroblastoma breakpoint family, which consists of 22 genes and pseudogenes that arose by gene duplication [10,11]. Both genes are highly variable and harbor many low frequency variants [8]. Genes with a long coding sequence (CDS) and many rare heterozygous variants are more likely to appear as candidate genes after compound heterozygous filtering. In Figure 4 a scatterplot is shown for the mean number of rare heterozygous variants and the CDS length of the longest transcript for each gene. *MUC16* is not only an extraordinarily large gene with over 40kb coding sequence, in average there were also 1.16 rare heterozygous variant calls per healthy individual with an allele frequency below 0.01 in the reference population of the 5000 exomes project. In e.g. NA19238, there were also two such rare heterozygous calls that passed the filter. Genes with pseudogenes, such as *NBPF10* are also prone to genotyping artifacts in reference guided resequencing due to mismatched reads: A variant that is classified as a heterozygous genotype is likely to be a false positive call, if the coverage of this position is high, however the proportion of reads supporting the alternate allele deviates strongly from the value of 0.5 that is expected for a heterozygous genotype. Figure 5 illustrates the mismatching of reads originating from a pseudogene of *NBPF10* that result in such genotyping artifacts. For *NBPF10* many such calling artifacts have been reported in GeneTalk and were automatically excluded. Also the variants detected in *NBPF10* and *MUC16* are highly suggestive for such false calls, for instance in the heterozygously called variant with the dbSNP ID rs61813437 the alternate allele is only supported by 44 out of altogether 236 sequence reads in the first individual and likewise in the others (alternate allele/reference allele: 203/37, 212/30, 199/38) and in dbSNP this variant is also listed as “not validated”. After this quick assessment of the trustworthiness of the variant calls, *PIGO* remains as the most promising disease candidate gene in this case and the mutations were indeed confirmed as the causative mutations of the disorder [9].

With the additional second affected individual in this case scenario the number of candidate genes was markedly lower than in the two trios. In autosomal recessive disorders all affected family members share both haplotypes around the disease locus and are

therefore identical by descent for both copies of the gene ($IBD = 2$) [12]. Two siblings are in average $IBD = 2$ in only one quarter of their genome which reduces the number of candidate genes likewise with every additional affected sibling. If we analyzed in this family only one of the affected sisters at a time in a trio approach, we would have seen the additional candidate genes *HRNR*, *PLCD1*, and *MUC5B*. KGGseq [13], a statistical framework for analyzing exome data of multiple individuals has only implemented rule 2 for compound heterozygous filtering and reduced to 223 rare candidate variants.

Conclusions

In this work we developed a filter for identifying compound heterozygotes in exome data of one or more individuals of a family. The rule set on which our filter is based, is comprehensive for analyzing multiple samples and advances the prioritization of compound heterozygous candidate variants beyond existing tools for analyzing data of exome sequencing studies [14,15]. We showed that filtering for compound heterozygous mutations is an effective means in identifying disease candidate genes especially if multiple family members are available for the analysis. In a trio analysis with exome data of the parents and one affected child, typically, mutations in only about a dozen candidate genes remain. This manageable number of remaining genes can then be assessed based on the expertise of the investigator or further prioritized by suitable tools [16–18]. We implemented the compound heterozygous filter as an intuitively usable web service that allows a quick reduction of the exome variants to such a candidate set. The filter as well as the European and Yoruban trios are accessible via the demo account at www.gene-talk.de [1].

Web Resources

The URL for data presented herein are as follows:

1000 genomes project website. Available: <http://www.1000genomes.org>. Accessed 2013 May 2.

NHLBI Exome Sequencing Project (ESP) website. Available: <http://evs.gs.washington.edu/EVS/>. Accessed 2013 May 2.

GeneTalk website. Available: www.gene-talk.de. Accessed 2013 May 2.

Supporting Information

Table S1.
(PDF)

Acknowledgments

The authors would like to thank the NHLBI GO Exome Sequencing Project and its ongoing studies which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010).

Author Contributions

Conceived and designed the experiments: TK PMK. Performed the experiments: PS NZ VH DP. Analyzed the data: PMK VH DP PNR SM. Wrote the paper: PMK PNR.

References

1. Kamphans T, Krawitz PM (2012) GeneTalk: an expert exchange platform for assessing rare sequence variants in personal genomes. *Bioinformatics* 28: 2515–2516.
2. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485: 242–245.

3. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328: 636–639.
4. Haider S, Ballester B, Smedley D, Zhang J, Rice P, et al. (2009) BioMart Central Portal—unified access to biological data. *Nucleic Acids Res* 37: W23–27.
5. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.
6. (ESP) NGENSP (2013) Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP). Seattle, WA.
7. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38: e164.
8. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337: 64–69.
9. Krawitz PM, Murakami Y, Hecht J, Kruger U, Holder SE, et al. (2012) Mutations in PIGO, a member of the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation. *Am J Hum Genet* 91: 146–151.
10. Karro JE, Yan Y, Zheng D, Zhang Z, Carriero N, et al. (2007) Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res* 35: D55–60.
11. Pei B, Sisu C, Frankish A, Howald C, Habegger L, et al. (2012) The GENCODE pseudogene resource. *Genome Biol* 13: R51.
12. Rodelsperger C, Krawitz P, Bauer S, Hecht J, Bigham AW, et al. (2011) Identity-by-descent filtering of exome sequence data for disease-gene identification in autosomal recessive disorders. *Bioinformatics* 27: 829–836.
13. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27: 2987–2993.
14. Li MX, Gui HS, Kwan JS, Bao SY, Sham PC (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res* 40: e53.
15. Sifrim A, Van Houdt JK, Tranchevent LC, Nowakowska B, Sakai R, 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.
16. Bornigen D, Tranchevent LC, Bonachela-Capdevila F, Devriendt K, De Moor B, et al. (2012) An unbiased evaluation of gene prioritization tools. *Bioinformatics* 28: 3081–3088.
17. Tiffin N, Andrade-Navarro MA, Perez-Iratxeta C (2009) Linking genes to diseases: it's all in the data. *Genome Med* 1: 77.
18. Moreau Y, Tranchevent LC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet* 13: 523–536.