European Journal of Human Genetics (2016) 24, 581–586

www.nature.com/ejhg

ARTICLE

Can whole-exome sequencing data be used for linkage analysis?

Steven Gazal*[,1,2], Simon Gosset[1,3], Edgard Verdura[4,5], Françoise Bergametti[4,5], Stéphanie Guey[4,5], Marie-Claude Babron[6,7] and Elisabeth Tournier-Lasserve[4,5,8]

Whole-exome sequencing (WES) has become the strategy of choice to identify causal variants in monogenic disorders. However, the list of candidate variants can be quite large, including false positives generated by sequencing errors. To reduce this list of candidate variants to the most relevant ones, a cost-effective strategy would be to focus on regions of linkage identified through linkage analysis conducted with common polymorphisms present in WES data. However, the non-uniform exon coverage of the genome and the lack of knowledge on the power of this strategy have largely precluded its use so far. To compare the performance of linkage analysis conducted with WES and SNP chip data in different situations, we performed simulations on two pedigree structures with, respectively, a dominant and a recessive trait segregating. We found that the performance of the two sets of markers at excluding regions of the genome were very similar, and there was no real gain at using SNP chip data compared with using the common SNPs extracted from WES data. When analyzing the real WES data available for these two pedigrees, we found that the linkage information derived from the WES common polymorphisms was able to reduce by half the list of candidate variants identified by a simple filtering approach. Conducting linkage analysis with WES data available on pedigrees and excluding among the candidate variants those that fall in excluded linkage regions is thus a powerful and cost-effective strategy to reduce the number of false-positive candidate variants.
European Journal of Human Genetics (2016) **24,** 581–586; doi:10.1038/ejhg.2015.143; published online 15 July 2015

INTRODUCTION

Identification of causal variants in Mendelian disorders was previously performed by positional cloning whose first step was gene mapping through linkage analysis conducted with DNA chips. The progress of high-throughput sequencing led geneticists to directly perform whole-exome sequencing (WES) for the identification of candidate variants, especially for families with limited linkage informativity. Indeed, the limited number of rare variants with a predicted effect on the protein (estimated between 130 and 400 non-synonymous sites per individual[1]) allows to reduce the number of rare candidate variants using a simple filtering approach based on variant frequency in reference samples, predicted effect on the protein and inheritance of the disease.[2]

Nevertheless, this number can be quite large and include a number of false-positive candidate variants. First, numerous false calls are generated during the variant calling step, especially for rare variants[3] and near insertions or deletions,[4] and some of them segregate with the disease by chance. Second, variants reported as absent or rare in reference samples can be common in the population of the studied family; when present in all the affected members of a family in which founders are unavailable, such variants would be retained although they might be inherited from distinct ancestors and not consistent with the inheritance pattern. Third, it is sometimes necessary to keep variants with incomplete genotype information for some individuals due to coverage or quality limitations. Finally, when studying a family with unavailable parents under a recessive compound heterozygous model, WES filtering can select two variants that are on the same haplotype and that are thus not consistent with the recessive model.

It is essential to minimize the number of candidate variants to reduce the heaviness of the Sanger sequencing validation step. For this reason, many studies have combined WES filtering with a multipoint linkage analysis performed with microsatellites or SNP chip data. Multipoint linkage analysis, under a model with complete penetrance and no phenocopy, can identify regions of the genome where all cases share one haplotype identical-by-descent (IBD) for a dominant disease or two haplotypes IBD for a recessive disease. When individuals affected by a recessive disease are inbred, homozygosity mapping[5] will identify homozygous regions of the genome where all cases share twice the same haplotype IBD. The use of linkage analysis LOD scores, which summarize the concordance between IBD information and a specified disease model, can thus help to eliminate false-positive candidate variants described above. In addition, when the causal variant is not captured, linkage data suggest regions of interest that can be targeted in follow-up studies.

Although this approach combining WES and DNA chip linkage analysis is therefore of great interest to minimize the number of candidate variants, it has, however, a significant cost burden. Performing linkage analysis directly with polymorphisms extracted from WES data is an attractive strategy to reduce this cost. However, because of the small proportion of the genome sequenced in WES, the size of the gaps between sequenced regions, and the high amount of

[1]INSERM, IAME, UMR 1137, Paris, France; [2]Plateforme de Génomique Constitutionnelle du GHU Nord, Assistance Publique des Hôpitaux de Paris (APHP), Hôpital Bichat, Paris, France; [3]Univ Paris Diderot, IAME, UMR 1137, Sorbonne Paris Cité, Paris, France; [4]INSERM, UMR 1161, Paris, France; [5]Univ Paris-Diderot, Génétique et Physiopathologie des Maladies Cérébro-Vasculaires, UMR 1161, Sorbonne Paris Cité, Paris, France; [6]INSERM, Genetic Variability and Human Diseases, UMR 946, Paris, France; [7]Univ Paris-Diderot, UMR 946, Sorbonne Paris Cité, Paris, France; [8]Assistance Publique des Hôpitaux de Paris, Hôpital Lariboisière, Paris, France
*Correspondence: Dr S Gazal, INSERM, IAME, UMR 1137, Inserm, UFR de Médecine—site Bichat, 16 rue Henri Huchard, 75018 Paris, France. Tel: +33 (0) 1 57 27 77 68; Fax: +33 (0) 1 57 27 75 21; E-mail: steven.gazal@inserm.fr
Received 13 March 2015; revised 7 May 2015; accepted 26 May 2015; published online 15 July 2015

linkage disequilibrium (LD) observed within exons, these data might not be suitable for the Lander–Green multipoint algorithm[6] implemented in the present linkage analysis software.

We aimed to compare the performance of linkage analysis conducted with WES and DNA chips genotyping data and then test the benefit of this combined linkage/WES filtering approach. We compared first by simulation and then application to real data, linkage analyses conducted with WES genotype data and with SNP chip data on two families segregating an autosomal-dominant and an autosomal-recessive disorder. We then used Sanger sequencing to test the benefit of this combined approach in reducing the number of false-positive variants.

MATERIALS AND METHODS

Simulation study

Performance of linkage analysis conducted with various WES genotype and/or SNP chip data sets was first evaluated by simulation. Two families were studied (Figure 1): one with a dominant disease (Family A), and one inbred with a recessive disease (Family B). Simulation of genotype data was conducted by randomly assigning one of European (EUR) haplotypes from 1000 Genomes (1000G) panel[1,7] to each family founder and by simulating Mendelian inheritance along the pedigree. A reference LOD score was then computed and compared with the ones obtained by linkage analyses performed with the markers of the Affymetrix 250K SNP chip and with the common polymorphisms present in WES data. The simulation study workflow is summarized in Supplementary Figure S1.

SNP chip and WES genotype data sets used for simulations. To have realistic genome patterns, we downloaded 758 EUR 1000G autosomal haplotypes obtained by phasing the whole chromosomes of 379 unrelated EUR individuals with SHAPEIT2.[8] We then extracted subsets of markers corresponding to different map designs. First, to reproduce SNP chip genotyping, we kept the 248 290 markers present in the Affymetrix 250K chip. Second, to mimic the common polymorphisms present in WES data, we selected the 71 206 variants referenced in the exome variant server (EVS) database that have a minor allele frequency (MAF) $\geq 5\%$ in EUR of 1000G. We thus used haplotypes with a total of 318 609 markers.

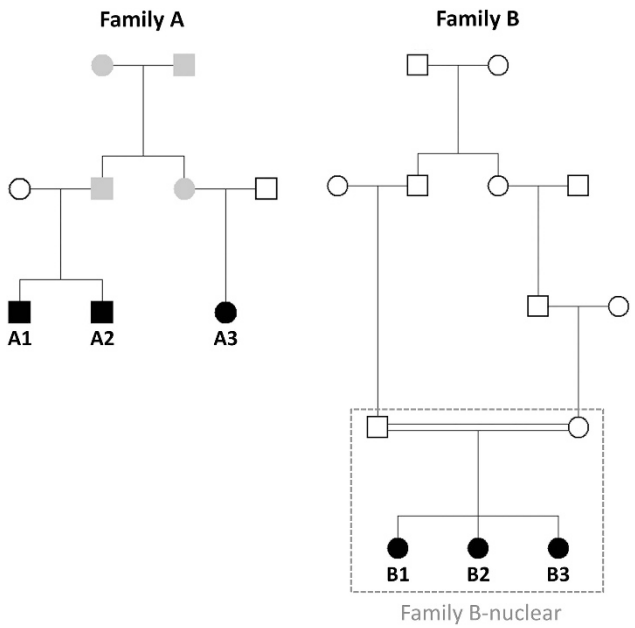

Figure 1 Pedigrees of Families A and B. In both families, the three labeled individuals have been genotyped on Affymetrix 250K SNP chip and sequenced by WES. Linkage analysis on Family B was performed with and without (Family B-nuclear) the inbreeding loop. Black, gray and white symbols represent affected, unknown and unaffected status, respectively.

To simulate a family replicate, these haplotypes were randomly drawn without replacement for each chromosome and were assigned to pedigree founders. Then the recombination process along each chromosome and Mendelian inheritance were simulated with Genedrop program of MORGAN version 2.9 (http://www.stat.washington.edu/thompson/Genepi/MORGAN). As our goal was to study the genome-wide linkage accuracy and not the power to find the causal variant, Mendelian inheritance was simulated unconditionally to a trait. For each family, 100 replicates were simulated, each replicate consisting for three data sets. The first one, labeled reference data set, was constituted of Genedrop founder labels at all the 318 609 markers (Supplementary Figure S1) for all the pedigree founders and the 3 affected individuals. These founder labels allowed us to exactly know the inheritance vector inside the pedigree and to create markers fully informative for linkage analysis. The SNP chip data set consisted for the genotype data for the 248 290 markers present in the Affymetrix 250K chip for the three affected individuals. Finally, the WES genotype data set consisted for the genotype data for the 71 206 common polymorphisms presented in WES data for the 3 affected individuals.

To mimic realistic WES genotype data, markers with homozygous genotypes for the reference alleles for all three individuals of a family were removed from the WES genotype data set, as variant calling algorithms only report genetic positions with at least one variant different from the reference allele. For Family A, monomorphic markers were also removed from SNP chip and WES genotype data sets, as they are non-informative (NI) for the study of dominant diseases. The remaining numbers of markers in SNP chip and WES simulated data sets are available in Supplementary Table S1, which also confirms the non-uniform coverage of the genome by WES data.

Linkage analysis performance using simulated datasets. Linkage analyses were performed with Merlin software version 1.1.2[9] using allele frequencies from the EUR of 1000G. For Family A, the genetic model was set to autosomal dominant with complete penetrance and no phenocopy, and the disease allele frequency was set to 0.001. For Family B, the genetic model was set to autosomal recessive with complete penetrance and no phenocopy, and the disease allele frequency was set to 0.01. Two linkage analyses were performed on this family. First, Merlin was run on the pedigree with the inbreeding loop in order to perform homozygosity mapping. Second, in order to allow for allelic heterogeneity for the disease, only the nuclear pedigree (Family B-nuclear, Figure 1) was specified to Merlin. Note that breaking the inbreeding loop can be necessary when the homozygous-by-descent region is small, that is, when two crossovers occurred close to the causal variant.

For each replicate, linkage analysis was performed on the three different data sets described above. LOD score computed on the reference data set was designated as the reference (REF) LOD score, as all the markers of this data set are fully informative. After each linkage analysis, the genome was categorized into either linked, excluded or NI region. A region was arbitrarily considered as linked if the LOD scores were higher than a linkage threshold, which we rounded as ELOD − 0.1, where ELOD is the expected LOD score at the disease locus. As ELOD is equal to 0.903, 2.709 and 1.204 for Families A, B and B-nuclear, respectively, the linkage thresholds were set to 0.8, 2.6 and 1.1, respectively. A region was considered as excluded if the LOD scores of its markers were < -2, as originally recommended.[10] Finally, a region was considered as NI if the LOD scores of its markers lied between -2 and the linkage threshold.

To evaluate the accuracy of linkage analysis performed with SNP chip and WES genotype data sets, we created a 3x3 cross table comparing the cumulated lengths in cM of their three different linkage regions (linked, excluded and NI) with those obtained with the REF LOD score. A region was labeled as false positive if it was excluded in the reference data set but linked in SNP chip or WES genotype data sets.

Minimizing false-positive signals due to LD. In order to minimize LD present in the data while retaining a set of markers dense enough to be informative, we compared linkage analysis performance when selecting one SNP every 10, 25, 50, 100 and 250 kb with the FSuite software.[11]

Application study

To analyze the putative benefits of adding a linkage analysis performed with WES genotyping data to the WES filtering step to reduce the number of bona fide candidate variants, we used 'real' SNP chip and WES data obtained in Families A and B combined with Sanger sequencing of all candidate variants identified at the filtering step.

Genotyping data. Genomic DNA from Families A and B and consenting relatives was extracted from peripheral blood leukocytes according to standard procedures.

Genotyping was performed with the Affymetrix GeneChip Human Mapping EA 250 K Array (Affymetrix, Santa Clara, CA, USA), as previously described.[12] Exome sequencing was also performed as previously described.[12] Briefly, Agilent SureSelect Human All Exon Kit V4+UTRs were used for Family A and V4 for Family B. Each genomic DNA fragment was then sequenced on a sequencer as 75-bp paired-end reads (Illumina HISEQ, Illumina, San Diego, CA, USA). Image analysis and base calling were performed with Real Time Analysis Pipeline v.1.8 with default parameters (Illumina). The bioinformatic analysis of sequencing data was based on the CASAVA v.1.8 Illumina pipeline. CASAVA performs alignment against the human reference genome (hg19, UCSC Genome Browser), calls the SNPs on the basis of the allele calls and read depth and detects variants (SNPs and indels). Genetic-variation annotation was performed by an in-house pipeline (IntegraGen, Evry, France), and results were provided per sample in tabulated text files.

Linkage analysis. All data handling was performed with the graphical user interface Alohomora.[13] We verified sample genders with the CheckGender procedure of the package. Alohomora was used to remove rare variants from WES data (MAF <0.05), to select markers according to their physical positions to minimize LD between markers, and to format data for the Merlin software. Genetic models were the same as those of our simulations. Merlin was run twice on each data set: first to detect unlikely genotypes and to remove them through the program Pedwipe (Merlin package), and second to compute the LOD scores. Linkage analysis with WES data was performed only with variants with a high variant calling quality in each individual. Linkage analyses were computed with European allele frequencies and DeCode genetic map furnished by Affymetrix.

Candidate variant detection. Candidate variant detection was performed with a Perl script and the genetic-variation-annotation files. Variant filtering was based on annotation information (missense, nonsense, splice-site and indel frame-shifts), consistency with the inheritance of the disease model, and reference allele frequencies of three reference samples: EUR of 1000G, European American of EVS database, and 96 control individuals of IntegraGen database (Supplementary Figures S2 and S3). For Family B, we considered homozygous variants, and compound heterozygotes (ie, genes with at least two heterozygous variants for each affected individuals). Note that, to be less sensitive to sequencing coverage and alignment accuracy, we allowed genotypes with a low alignment confidence or not covered by the sequencing. All candidate variants have been re-sequenced by PCR and traditional Sanger method.[14]

RESULTS

Performance of linkage analyses conducted with simulated SNP chip and WES genotype data sets

We first evaluated the impact of LD on false-positive evidence of linkage, using sets of markers separated by variable physical distances (Figure 2). Cross tables obtained with all informative markers and after removing markers are shown in Supplementary Table S2 and Table 1, respectively.

The top row of Figure 2 shows the amount of NI genome, that is, the genetic length of genome with a LOD score between -2 and the linkage threshold, and the second row shows the amount of false-positive genome, that is, the genetic length of genome with a LOD score higher than the linkage threshold but with a REF LOD score <-2. Whatever the genotyping strategy and the family, selecting one marker every 25 kb significantly decreased the amount of false-positive signals while retaining a set of markers dense enough to be informative. Bins longer than 25 kb did not further decrease significantly the amount of false-positive signals but increased the proportion of NI genome. More detailed results in terms of number of markers, false negative, true positive and true negative amounts of genome are presented in Supplementary Figure S4. Interestingly, we observed that increasing marker density also increased the proportion

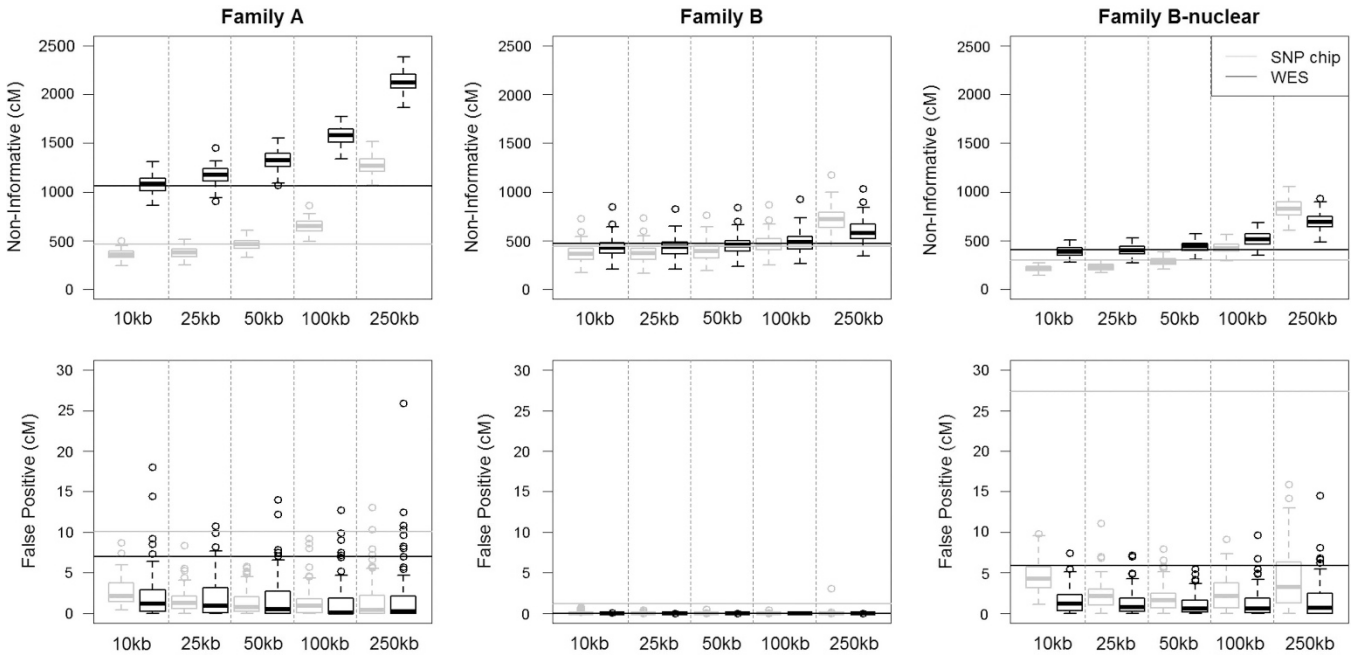

Figure 2 Marker selection to minimize false-positive signals in simulated data sets. Each boxplot shows genetic lengths computed on 100 replicates, with a marker selection according to different physical length bins. Black and gray lines represent the median of values computed with all the markers (ie, without marker selection).

of false-negative signals in Family A, that is, regions with a LOD score < -2 while the REF LOD score is higher than the linkage threshold. We looked at several of these regions with a false-negative signal and observed that they were due to a double recombination on the haplotype of the individual A3 (Supplementary Figure S5). Indeed, as haplotype frequencies are poorly estimated in the presence of LD (because they are estimated by multiplying allele frequencies together), Merlin gave a higher likelihood to observe a haplotype coming from another ancestor rather than a double recombinant.

Linkage analysis performed using the WES genotype data set, even though genome coverage was not uniform, excluded a high proportion of the genome (around 2/3 and more than 3/4 for Families A and B, respectively), while keeping a low amount of false-positive and false-negative signals (Table 1). Linkage analysis was more informative with the SNP chip data set, especially for Family A (375.46 cM NI cumulative distance with SNP chip data *versus* 1169.67 cM with WES genotype data). For Family B, informativity of the two data sets was similar (364.75 *versus* 493.94 cM for Family B and 229.56 *versus* 402.87 cM for Family B-nuclear). However, WES genotype data sets were less powerful to detect homozygous-by-descent regions in Family B: 6.56 cM are detected with SNP chip data set, against 4.36 cM with WES genotype data sets.

Table 1 Linkage analysis performance on simulated data sets after marker selection

	SNP chip[a]			*WES*[b]		
REF[c]	*Exclusion*	*NI*	*Linkage*	*Exclusion*	*NI*	*Linkage*
Family A						
Exclusion	2930.10[d]	226.05	1.54	2230.55	931.72	2.04
NI	0.62	7.77	0.33	0.78	9.2	0.19
Linkage[e]	4.83	141.65	289.97	2.75	228.75	205.23
	2935.55	375.46	291.84	2234.08	1169.67	207.46
Family B						
Exclusion	3247.91	143.53	0.01	3170.49	213.13	0.00
NI	0.11	220.32	0.04	0.66	218.76	3.52
Linkage[f]	0.00	0.90	6.56	0.00	3.05	4.36
	3248.02	364.75	6.61	3171.15	434.94	7.88
Family B-nuclear						
Exclusion	3165.64	214.67	2.28	3013.93	359.58	1.29
NI	0.02	6.48	0.43	0.11	6.32	0.27
Linkage[g]	0.04	8.41	221.42	0.27	36.97	191.72
	3165.70	229.56	224.13	3014.31	402.87	193.28

[a]The average number of selected informative markers was 44 532, 60 145 and 60 136 for Families A, B and B-nuclear, respectively.
[b]The average number of selected informative markers was 14 227, 14 201 and 14 198 for Families A, B and B-nuclear, respectively.
[c]Reference LOD score.
[d]Genetic length in cM.
[e]Linkage threshold = 0.8.
[f]Linkage threshold = 2.6.
[g]Linkage threshold = 1.1.

Linkage analysis conducted with real SNP chip and WES data in Families A and B

We then used 250K SNP chip data and Agilent SureSelect WES data available for Families A and B to compare linkage performance of these two data sets. For Family A, selection of one marker every 25 kb identified 48 057 and 15 902 informative markers for SNP chip data and WES genotype data, respectively; for Family B, 65 231 and 12 961 markers were selected. Linkage data obtained with these two data sets were very similar (Table 2), and few regions showed discordant results, that is, linked with one type of data and excluded with the other. There was no discordant region for Family B. For Families A and B-nuclear, cumulated lengths of 2.41 and 3.22 cM were, respectively, discordant, that is, $<0.1\%$ of the genome. Finally, note that the genetic lengths of linked, excluded and NI regions were in the same range as those observed in the simulated data sets (Table 1).

Combination of linkage and WES data reduced significantly the number of candidate variants in Families A and B

WES data filtering (Supplementary Figures S2 and S3) identified a total of 25 candidate variants, including 9, 7 and 9 variants for Families A, B and B-nuclear, respectively (Table 3 and Supplementary Table S3). Variants of Family B-nuclear included the seven homozygous candidate variants of Family B, and two variants (B8 and B9) of the same gene with heterozygous genotypes.

LOD scores around the 25 candidate variants with both genotyping strategies are shown in Supplementary Table S3. Linkage analysis conducted with WES genotypes excluded 13 of the 25 candidate variants. Linkage analysis conducted with SNP chip data excluded 15 variants, including the 13 previous ones plus variants A7 of Family A and B2 of Family B-nuclear.

Sanger sequencing of all candidate variants allowed us to confirm 8 of the candidate variants and to exclude 14 of them (Table 3 and Supplementary Table S3). For three INDELs, the quality of sequence was not good enough to be conclusive. None of the variants located in excluded regions using either SNP chip or WES genotype data was validated by Sanger sequencing, except for the two heterozygous B8 and B9 variants of Family B-nuclear that were shown to be on the same haplotype. Indeed, as Merlin proposes outputs of the haplotype reconstruction, we have been able to observe that the three affected

Table 2 Comparison of linkage data obtained with real WES genotype and SNP chip data

	Family A				*Family B*				*Family B-nuclear*			
	SNP chip				*SNP chip*				*SNP chip*			
	Exclusion	*NI*	*Linkage*[a]		*Exclusion*	*NI*	*Linkage*[b]		*Exclusion*	*NI*	*Linkage*[c]	
WES												
Exclusion	2144.38[d]	128.75	1.48	2274.61	2940.42	92.01	0.00	3032.43	2694.55	147.13	1.01	2842.68
NI	674.29	219.97	132.67	1026.93	213.47	245.04	9.36	467.87	369.19	69.67	42.00	480.86
Linkage	0.93	29.37	179.59	209.90	0.00	0.07	14.26	14.33	2.21	1.21	187.67	191.09
	2819.61	378.09	313.74		3153.89	337.11	23.62		3065.94	218.01	230.67	

[a]Linkage threshold = 0.8.
[b]Linkage threshold = 2.6.
[c]Linkage threshold = 1.1.
[d]Genetic length in cM.

Table 3 Number of candidate variants of Families A, B and B-nuclear with and without linkage filtering

	WES filtering	*WES filtering+linkage analysis with WES data*	*WES filtering+linkage analysis with SNP chip data*
Family A	9 (5)[a]	6 (5)	5 (5)
Family B	7 (1)	2 (1)	2 (1)
Family B-nuclear	9 (3)	4 (1)	3 (1)

[a]Numbers between parentheses are the number of candidate variants validated by Sanger sequencing.

individuals only shared one haplotype IBD at this locus (Supplementary Figure S6). This confirmed that these two variants were on the same haplotype, which was not consistent with a recessive model. Finally, the two variants A7 and B2, which were only excluded by linkage analysis conducted with SNP chip data, were not validated by Sanger sequencing.

DISCUSSION

Herein we used simulated and real genotyping data obtained in two small families to investigate the performances of SNP chip and WES genotype data for linkage analysis. We then analyzed the power of combined linkage and WES to reduce the number of candidate variants in these two families. We showed that performance of linkage analyses conducted with either SNP chip or WES genotype data sets was very similar, providing that one marker per 25 kb was selected to minimize LD while retaining a set of markers dense enough to be informative. Using these guidelines with real WES data allowed us to roughly decrease by half the number of candidate variants in the two families. Sanger sequencing proved that candidate variants located in excluded regions were sequencing errors or inconsistent with the disease model. Altogether, these data strongly suggest that linkage analysis conducted with WES genotypes is an accurate and cost-effective strategy to reduce the number of candidate variants in small family WES studies. Finally, it is important to emphasize that excluding candidate variants located in regions showing a $LOD < -2$ is a safer strategy than keeping the ones in linkage peaks.

Linkage analysis can be performed independently on each variant (two-point linkage analysis) or by taking into account all markers at the same time to reconstruct haplotypes of each individual (multi-point linkage analysis). As numerous false rare variants are generated during the WES data calling step,[3] we advice to avoid two-point linkage analysis that will give a high LOD score to a false rare variant segregating with the disease by chance. One of the advantages of our strategy is to use multipoint linkage analysis with common polymorphisms surrounding rare variants to remove these false positives. However, LD present in dense data is known to lead multipoint linkage analysis to false-positive evidence of linkage. This result is due to an incorrect estimation of haplotype frequencies by the linkage software that estimate them by multiplying allele frequencies together. A large number of studies have thus proposed different strategies to remove LD of the data based on r^2 or D' calculation[15–20] or to model LD by clustering markers to estimate haplotype frequencies.[21] When dealing with small families (as in the present case), a reference population sample is thus needed to compute these LD statistics. As such a sample might be unavailable or difficult to obtain, we proposed to remove markers based on physical length bins. Our simulation study highlighted a 25-kb bin. This result was quite surprising at first glance as this bin still exhibits LD between markers, and other studies often use a bin size between 250 and 500 kb (or 0.25 and 0.5 cM) to remove LD. However, a 25-kb bin appeared to be sufficient to significantly decrease the amount of false-positive signals while keeping the data as informative as possible. Note that the choice of bins in physical distance, rather than genetic distance, was driven by the fact that the Alohomora software, which is widely used and that we used in our application, only permits to select markers according to their physical distances. Nevertheless, we observed similar results when we selected markers according to their genetic positions (data not shown).

It is now accepted that linkage analysis is emerging again as an important tool for the identification of causal variants using sequencing data.[22,23] However, to our knowledge, our paper is the first methodological one proving the robustness of linkage analysis performed on WES genotype data. Although application of linkage analysis, or IBD detection, directly on WES genotype data has already successfully reduced the search space of the causative disease gene in a few studies,[24–27] the conditions of validity of this strategy still remained uncertain. Smith *et al.*[28] showed that there was good graphical agreement of linkage peaks obtained with SNP chip and WES genotype data for three small families with only one inbred or two outbred individuals sequenced, but they did not look at the agreement of excluded regions and were not able to quantify the amounts of false-positive and falsenegative regions of both genotyping strategies. Here, our simulation process enabled us to show that these two values were close to 0 and prove the robustness of the linkage results.

Altogether, we showed that linkage conducted with WES genotype data is accurate and cost effective. This genome-wide approach combining linkage with WES genotype data and WES variant filtering linkage would also be of major interest when analyzing multiple small pedigrees in which genetic heterogeneity is suspected. Finally, note that this approach can also be used with whole-genome sequencing data.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank Emmanuelle Génin for her very helpful comments. E Verdura was supported by Programme Hospitalier de Recherche Clinique AOM06037 (grant to ET-L). S Guey was a recipient of a fellowship from the Fondation pour la Recherche Médicale.

1 1000 Genomes Project Consortium: An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**: 56–65.
2 Ng SB, Buckingham KJ, Lee C *et al*: Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 2010; **42**: 30–35.
3 Browning BL, Browning SR: Detecting identity by descent and estimating genotype error rates in sequence data. *Am J Hum Genet* 2013; **93**: 840–851.
4 Li H, Homer N: A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform* 2010; **11**: 473–483.
5 Lander ES, Botstein D: Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* 1987; **236**: 1567–1570.
6 Lander ES, Green P: Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA* 1987; **84**: 2363–2367.

7 1000 Genomes Project Consortium: A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1061–1073.
8 Delaneau O, Zagury J-F, Marchini J: Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 2013; **10**: 5–6.
9 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002; **30**: 97–101.
10 Morton NE: Sequential tests for the detection of linkage. *Am J Hum Genet* 1955; **7**: 277–318.
11 Gazal S, Sahbatou M, Babron M-C, Génin E, Leutenegger A-L: FSuite: exploiting inbreeding in dense SNP chip and exome data. *Bioinformatics* 2014; **30**: 1940–1941.
12 Hervé D, Philippi A, Belbouab R *et al*: Loss of $\alpha 1\beta 1$ soluble guanylate cyclase, the major nitric oxide receptor, leads to moyamoya and achalasia. *Am J Hum Genet* 2014; **94**: 385–394.
13 Rüschendorf F, Nürnberg P: ALOHOMORA: a tool for linkage analysis using 10 K SNP array data. *Bioinformatics* 2005; **21**: 2123–2125.
14 Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977; **74**: 5463–5467.
15 Huang Q, Shete S, Amos CI: Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *Am J Hum Genet* 2004; **75**: 1106–1112.
16 Boyles AL, Scott WK, Martin ER *et al*: Linkage disequilibrium inflates type I error rates in multipoint linkage analysis when parental genotypes are missing. *Hum Hered* 2005; **59**: 220–227.
17 Goode EL, Badzioch MD, Jarvik GP: Bias of allele-sharing linkage statistics in the presence of intermarker linkage disequilibrium. *BMC Genet* 2005; **6** (Suppl 1): S82.
18 Levinson DF, Holmans P: The effect of linkage disequilibrium on linkage analysis of incomplete pedigrees. *BMC Genet* 2005; **6** (Suppl 1): S6.
19 Cho K, Yang Q, Dupuis J: Handling linkage disequilibrium in linkage analysis using dense single-nucleotide polymorphisms. *BMC Proc* 2007; **1** (Suppl 1): S161.
20 Bellenguez C, Ober C, Bourgain C: Linkage analysis with dense SNP maps in isolated populations. *Hum Hered* 2009; **68**: 87–97.
21 Abecasis GR, Wigginton JE: Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am J Hum Genet* 2005; **77**: 754–767.
22 Ott J, Wang J, Leal SM: Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* 2015; **16**: 275–284.
23 Wang GT, Zhang D, Li B, Dai H, Leal SM: Collapsed haplotype pattern method for linkage analysis of next-generation sequence data. *Eur J Hum Genet* 2015 **23**: 1739–1743.
24 Norton N, Li D, Rampersaud E *et al*: Exome sequencing and genome-wide linkage analysis in 17 families illustrate the complex contribution of TTN truncating variants to dilated cardiomyopathy. *Circ Cardiovasc Genet* 2013; **6**: 144–153.
25 Krawitz PM, Schweiger MR, Rödelsperger C *et al*: Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet* 2010; **42**: 827–829.
26 Guergueltcheva V, Azmanov DN, Angelicheva D *et al*: Autosomal-recessive congenital cerebellar ataxia is caused by mutations in metabotropic glutamate receptor 1. *Am J Hum Genet* 2012; **91**: 553–564.
27 Eggers S, Smith KR, Bahlo M *et al*: Whole exome sequencing combined with linkage analysis identifies a novel 3 bp deletion in NR5A1. *Eur J Hum Genet* 2014; **23**: 486–493.
28 Smith KR, Bromhead CJ, Hildebrand MS *et al*: Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol* 2011; **12**: R85.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)