

# Novel mutations in genes causing hereditary spastic paraplegia and Charcot-Marie-Tooth neuropathy identified by an optimized protocol for homozygosity mapping based on whole-exome sequencing

Daliya Kancheva, MSc<sup>1-3</sup>, Derek Atkinson, MSc<sup>1,2</sup>, Peter De Rijk, PhD<sup>4</sup>, Magdalena Zimon, PhD<sup>1,2,12</sup>, Teodora Chamova, MD, PhD<sup>5</sup>, Vanyo Mitev, MD, DSc<sup>3</sup>, Ahmet Yaramis, MD<sup>6</sup>, Gian Maria Fabrizi, MD, PhD<sup>7</sup>, Haluk Topaloglu, MD, PhD<sup>8</sup>, Ivailo Tournev, MD, DSc<sup>5,9</sup>, Yesim Parma, MD, PhD<sup>10</sup>, Esra Battaloglu, PhD<sup>11</sup>, Alejandro Estrada-Cuzcano, PhD<sup>1,2</sup> and Albena Jordanova, PhD<sup>1-3</sup>

**Purpose:** Homozygosity mapping is an effective approach for detecting molecular defects in consanguineous families by delineating stretches of genomic DNA that are identical by descent. Constant developments in next-generation sequencing created possibilities to combine whole-exome sequencing (WES) and homozygosity mapping in a single step.

**Methods:** Basic optimization of homozygosity mapping parameters was performed in a group of families with autosomal-recessive (AR) mutations for which both single-nucleotide polymorphism (SNP) array and WES data were available. We varied the criteria for SNP extraction and PLINK thresholds to estimate their effect on the accuracy of homozygosity mapping based on WES.

**Results:** Our protocol showed high specificity and sensitivity for homozygosity detection and facilitated the identification of

novel mutations in *GAN*, *GBA2*, and *ZFYVE26* in four families affected by hereditary spastic paraplegia or Charcot-Marie-Tooth disease. Filtering and mapping with optimized parameters was integrated into the HOMWES (homozygosity mapping based on WES analysis) tool in the GenomeComb package for genomic data analysis.

**Conclusion:** We present recommendations for detection of homozygous regions based on WES data and a bioinformatics tool for their identification, which can be widely applied for studying AR disorders.

*Genet Med* advance online publication 22 October 2015

**Key Words:** Charcot-Marie-Tooth neuropathy; hereditary spastic paraplegia; homozygosity mapping; PLINK; whole-exome sequencing

## INTRODUCTION

Gene hunting for autosomal-recessive (AR) disorders in consanguineous families is largely facilitated by a positional cloning technique known as homozygosity mapping.<sup>1</sup> In contrast to classical linkage analysis, in which multiple patients in several generations of a family are needed to obtain conclusive results, homozygosity mapping allows the study of small families with only a few affected individuals. This approach has been applied successfully for the identification of genetic defects in families from isolated geographic areas,<sup>2</sup> ethnic minorities,<sup>3</sup> and even outbred populations.<sup>4</sup> Homozygosity mapping is based on the principle that the affected children of a consanguineous

marriage are likely to have inherited the same mutation from a recent common ancestor, passed on by both parents on the same chromosomal segment, referred to as “identical by descent.”<sup>1,5</sup> Detection of homozygous genomic regions would limit the number of loci that are potentially involved in the etiology of the particular disease, thus facilitating the mapping of the AR causative variant.

Genomic runs of homozygosity (ROH) can be detected by genotyping polymorphic markers, such as short tandem repeats or single-nucleotide polymorphisms (SNPs). While each SNP has usually only two alleles, it is their number, ability to detect heterozygous regions (and hence to exclude linkage), and

The last two authors shared coauthorship.

The first two authors and the last two authors contributed equally to this work.

<sup>1</sup>Molecular Neurogenetics Group, Department of Molecular Genetics, VIB, University of Antwerp, Antwerp, Belgium; <sup>2</sup>Neurogenetics Laboratory, Institute Born-Bunge, University of Antwerp, Antwerp, Belgium; <sup>3</sup>Department of Medical Chemistry and Biochemistry, Molecular Medicine Center, Medical University-Sofia, Sofia, Bulgaria; <sup>4</sup>Department of Molecular Genetics, VIB, University of Antwerp, Antwerp, Belgium; <sup>5</sup>Department of Neurology, Medical University-Sofia, Sofia, Bulgaria; <sup>6</sup>Department of Pediatrics, Dicle University School of Medicine, Diyarbakir, Turkey; <sup>7</sup>Department of Neurological, Neuropsychological, Morphological and Motor Sciences, University of Verona, Verona, Italy; <sup>8</sup>Division of Pediatric Neurology, Department of Pediatrics, Faculty of Medicine, Hacettepe University, Ankara, Turkey; <sup>9</sup>Department of Cognitive Science and Psychology, New Bulgarian University, Sofia, Bulgaria; <sup>10</sup>Department of Neurology, Istanbul Medical Faculty, Istanbul University, Istanbul, Turkey; <sup>11</sup>Department of Molecular Biology and Genetics, Bogazici University, Istanbul, Turkey; <sup>12</sup>Current address: Cell Biology and Biophysics Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.

Correspondence: Albena Jordanova (albena.jordanova@molgen.vib-ua.be)

Submitted 13 May 2015; accepted 25 August 2015; advance online publication 22 October 2015. doi:10.1038/gim.2015.139

amenability to automated scoring that provides their potential in pinpointing homozygous regions. Homozygosity mapping in consanguineous families typically yields ROHs extending up to dozens of centimorgans and containing numerous positional candidate genes. Sanger sequencing of these regions is costly, labor-intensive, and often requires prior functional prioritization of the genes to be analyzed. The introduction of next-generation sequencing (NGS) allows simultaneous, unbiased, and fast analysis of all genes within the regions of interest. Yet NGS poses its own challenges by yielding too many sequence variations, including errors, making it difficult to distinguish the disease-causing mutation among them. Combining homozygosity mapping with NGS provides a powerful tool to narrow down the list of candidate variants and greatly facilitates gene discovery. With rapidly decreasing costs and increasing throughput of NGS technologies, it became attractive to apply homozygosity mapping directly on whole-exome sequencing (WES) data, without the need of prior SNP genotyping. Several computer programs have recently been developed or adapted to assist in ROH identification of WES data, including GERMLINE,<sup>6</sup> AgileGenotyper/AgileVariantMapper,<sup>7</sup> HomSI,<sup>8</sup> HomozygosityMapper,<sup>9</sup> and H<sup>3</sup>M<sup>2</sup>,<sup>10</sup> each of which demonstrates specific limitations.

Lower performance of the GERMLINE algorithm in comparison to PLINK<sup>11</sup> and BEAGLE has been shown for SNP array-based ROH analysis.<sup>12</sup> The GERMLINE software has been adapted to analyze WES data; however, the accuracy of ROH detection remained prohibitively low.<sup>6</sup> Magi *et al.*<sup>10</sup> recently reported an H<sup>3</sup>M<sup>2</sup> method that takes into account the distance between consecutive SNPs when defining ROH regions. H<sup>3</sup>M<sup>2</sup> has higher accuracy than PLINK (with default parameters) and GERMLINE for detection of ROHs smaller than 1.5 Mb. For larger ROHs, both PLINK and H<sup>3</sup>M<sup>2</sup> show comparably good results. Nevertheless, a drawback of the H<sup>3</sup>M<sup>2</sup> method is that it exploits all the 1000 Genome Project SNPs for map construction, resulting in the majority of the SNPs being uninformative and yielding false-positive results.<sup>13</sup>

Carr *et al.*<sup>7</sup> described the programs AgileGenotyper and AgileVariantMapper for extraction of SNP genotypes from WES and graphical visualization of homozygous regions. A similar graphical output is created by HomSI.<sup>8</sup> Both AgileGenotyper/AgileVariantMapper and HomSI were able to retrieve the homozygous regions containing the disease-causing mutations from WES data. Nevertheless, their output data requires visual inspection and subjective interpretation. Seelow *et al.*<sup>9</sup> presented a new version of the Web-based tool HomozygosityMapper, which allows detection of homozygous stretches from NGS data.

A common limitation of the software listed above is that none of them performs quality filtering (QF) for the SNPs extracted from WES data, for example, based on genotype quality or presence in repeated sequences. WES data are known to generate many false-positive variant calls as a result of sequencing artifacts, misalignment, and other factors.<sup>14</sup> Therefore, QF of the data is necessary to minimize the spurious calls that could break up the homozygous regions.

The variety of parameters and methods used for SNP extraction and ROH detection hinders the comparison between different studies and challenges researchers in their choice.<sup>12,13</sup> In this work we aimed to define optimal parameters for homozygosity mapping on WES data using PLINK, one of the state-of-the-art software programs for this type of analysis.<sup>12</sup> We used as a test set SNPs retrieved from microarray data and WES in a group of families with AR forms of Charcot-Marie-Tooth disease (CMT) and hereditary spastic paraplegia (HSP). Both disorders share common clinical features and genetic causes, and result from a length-dependent, dying-back degeneration of peripheral or corticospinal axons, respectively.<sup>15</sup>

## MATERIALS AND METHODS

### Subjects

This study included a total of 15 patients (12 families) who have suggestive AR inheritance, belong to consanguineous families and/or inbred minorities (Roma, Turkish), and are affected with HSP or CMT. All participating individuals underwent standard clinical examination performed by experienced neurologists. Two families with CMT were diagnosed with a demyelinating hereditary motor and sensory polyneuropathy (CMT type I), six with axonal CMT (CMT type II). One family with HSP had a pure HSP phenotype; the remaining ones were classified as having complicated forms (**Table 1**). The study was approved by the local ethics committees, and all individuals provided blood samples after signing an informed consent form. DNA was isolated using standard protocols.

Ten of the 15 patients were analyzed by both SNP array and WES, and five by WES alone (**Table 1**, **Supplementary Figures S1 and S2** online). Optimization was performed on a training set of five individuals (HSP-1.1, CMT-1.1, CMT-1.2, CMT-2.1, CMT-3.1), whereas the rest of the individuals (CMT-4.1, CMT-5.1, CMT-6.1, CMT-7.1, CMT-8.1, HSP-2.1, HSP-2.2, HSP-3.1, HSP-4.1, HSP-4.2) were used to validate the optimal parameters. Two subjects were included as positive controls since their causative mutations had already been identified as residing in the largest homozygous region (CMT-2.1) and in a region only 1.7 Mb in size (CMT-3.1), respectively. In the 13 remaining patients the causative mutations were unknown before this study.

Relatedness was estimated using WES data, as previously described.<sup>16</sup>

### SNP genotyping and WES

Whole-genome SNP genotyping was performed using the Illumina HumanOmniExpress BeadChip. Genotypes were called using the GenomeStudio software, version 2011.1 (Illumina, San Diego, CA). The overall call rate for all individuals was >99%.

Exon sequences were captured by SureSelect Enrichment array (Agilent, Santa Clara, CA) followed by paired-end HiSeq2000 sequencing, yielding ~6 Gb of raw data with 90% of the target region covered  $\geq 20\times$ . Primary analysis was performed using a software pipeline, developed in house, called

Table 1 Description of the individuals and homozygosity mapping results

Family	Individual	Clinical diagnosis	Declared consanguinity	Total ROH size			Total ROH size			SNP array— region size (ranking)	WES— region size (ranking)	Reference
				Mlb	%	Number of ROHs	Average ROH size	Average ROH size				
								Mlb	%			
HSP-1	1	cHSP	1c	252.3	7.9	52	4.9	281.5	8.8	69	4.1	
CMT-1	1	CMT II	Unknown	160.7	5.4	45	3.6	141.3	4.7	38	3.7	
	2			245.4	8.2	38	6.5	269.8	8.9	58	4.7	
Overlap												
CMT-2	1	CMT II	1c	95.0	3.2	17	5.6	99.8	3.3	26	3.8	
CMT-3	1	CMT I	1c	308.3	10.3	41	7.5	306.0	10.2	64	4.8	HSPB1 James <sup>24</sup>
	1			162.9	5.4	37	4.4	150.3	5.0	35	4.3	PRX Auer-Grumbach <sup>25</sup>
CMT-4	1	CMT II	1c	243.6	8.1	52	4.7	275.8	9.2	63	4.4	
CMT-5	1	CMT II	nc <sup>a</sup>	258.9	8.6	41	6.3	260.1	8.7	44	5.9	
CMT-6	1	CMT I	nc <sup>a</sup>	323.6	10.8	66	4.9	332.5	11.1	70	4.8	
CMT-7	1	CMT II	1c	264.0	8.8	42	6.3	283.4	9.4	58	4.9	
CMT-8	1	CMT II	Unknown	243.7	7.6	44	5.5	244.6	7.6	50	4.9	
HSP-2	1	cHSP	Unknown	—	—	—	—	299.3	10.0	126	2.4	GAN 43.7Mb (1) This study
	2			—	—	—	—	434.9	14.5	150	2.9	GBA2 — 2.5 Mb (35) This study
Overlap												
HSP-3 <sup>b</sup>	1	cHSP	Unknown	—	—	—	—	147.2	4.9	56	2.6	GBA2 — 2.5 Mb (48) This study
HSP-4 <sup>b</sup>	1	pHSP	Unknown	—	—	—	—	256.0	12.3	123	2.1	ZFYVE26 — 9.3 Mb (2) This study
	2			—	—	—	—	311.0	10.4	121	2.6	ZFYVE26 — 9 Mb (6) This study
Overlap												
No SNP data available	1	pHSP	Unknown	—	—	—	—	317.5	10.6	147	2.2	ZFYVE26 — 9.3 Mb (3) This study
	2			—	—	—	—	123.9	4.1	49	2.5	ZFYVE26 — 9 Mb (2) This study

1c, first cousins; cHSP, complicated hereditary spastic paraplegia; CMT, Charcot-Marie-Tooth; nc, nonconsanguineous; pHSP, pure hereditary spastic paraplegia; ROH, run of homozygosity; SNP, single-nucleotide polymorphism. <sup>a</sup>Parents are from the same village. <sup>b</sup>HSP-3 and HSP-4 are related.

GenomeComb<sup>17</sup> (under condition of GPL; available at <http://genomecomb.sourceforge.net>), which allows comparison, annotation, and filtering of NGS data. The pipeline used the Burrows-Wheeler algorithm, version 0.7.5a,<sup>18</sup> to align sequencing reads to the reference human genome (hg19). Duplicate reads were marked using Picard, version 1.87 (<http://broadinstitute.github.io/picard/>); GenomeAnalysisToolKit (GATK), version 2.4–9 (ref. 19), was used to perform realignment around insertions/deletions. Variants were called using both GATK and Samtools, version 0.1.19-44428cd,<sup>20</sup> and were consequently combined and annotated with a variety of different databases (in-house exomes, dbSNP, 1000 Genomes Project). GenomeComb was also used for the selection of SNPs from all individuals with available WES data. We extracted the genotypes from all loci that showed a variant call in at least one individual per family.

### Detection of homozygous regions

PLINK, version 1.07, was used to assess the individual ROHs and the ones shared between siblings. Homozygosity mapping based on SNP array data was performed as described previously.<sup>21</sup> To measure the accuracy of ROH detection from WES data, we used the SNP array data as a reference and estimated the specificity and sensitivity of detection as follows:

$$\text{Sensitivity (\%)} = \frac{\text{ROH}_{\text{WES}} \text{ overlapping with ROH}_{\text{array}}}{\text{All ROH}_{\text{array}}}$$

$$\text{Specificity (\%)} = \frac{\text{ROH}_{\text{WES}} \text{ overlapping with ROH}_{\text{array}}}{\text{All ROH}_{\text{WES}}}$$

We defined false-positive ROH as present in WES-derived data, but not in the SNP array-derived data, and false-negative ROHs were the ones found in SNP array-derived and not in WES-derived data. The Z-test<sup>22</sup> was used to assess the differences between the mean sensitivity/specificity estimated in the training set versus the test set of individuals.

For the performance comparison with HomozygosityMapper and H<sup>3</sup>M<sup>2</sup>, default parameters were used for both programs. HomozygosityMapper's settings included a minimum coverage of 10 reads per nucleotide and no lower limit of the block length. Genotypes at positions that varied from RefSeq were extracted, as suggested when no unaffected relatives are available. For H<sup>3</sup>M<sup>2</sup>, we used the recommended parameters DNorm = 100,000, P1 = 0.1, and P2 = 0.1.

### Mutation analysis

All potential pathogenic variants in known HSP or CMT genes identified by WES were validated by bidirectional Sanger sequencing. The polymerase chain reaction products were purified with ExoSAP-IT (USB, Cleveland, OH), directly sequenced with a Big Dye Terminator kit version 3.1 (Applied Biosystems, Foster City, CA), and electrophoretically separated on an ABI3730xl DNA Analyzer (Applied Biosystems). Mutations were described and classified according to the Human Genome Variation Society nomenclature

(<http://www.hgvs.org/mutnomen>) and the American College of Medical Genetics and Genomics guidelines,<sup>23</sup> respectively. Cosegregation analysis of the variants with the disease was performed for all available family members. The novel mutations were screened in 50 Roma or 190 Turkish unrelated control individuals. Computational prediction of the pathogenicity of the missense changes was performed using Condel (<http://bg.upf.edu/fannsdb>) and Meta-SNP (<http://snps.biofold.org/meta-snp>), and NetGene2 (<http://www.cbs.dtu.dk/services/NetGene2>) was used for the splice-site alteration.

## RESULTS

### SNP extraction parameters

To evaluate the effect of QF on the accuracy of ROH detection, we applied six levels of stringency in SNP extraction from the WES data of our training set (individuals HSP-1.1, CMT-1.1, CMT-1.2, CMT-2.1, CMT-3.1). The levels are based on different combinations of the following quality filters: the GATK “pass” status, consensus calling of the SNPs by the GATK and Samtools variant callers, presence of the SNPs within simple repeats or microsatellites, and genotype quality >40 (**Supplementary Figure S3** online). For testing the QF, ROHs were detected using PLINK with 10 SNPs defining a ROH, 20 SNPs defining a sliding window, minimum density of one SNP in 10,000 kb, maximum gap of 10,000 kb. For each condition we estimated the sensitivity and specificity of ROH detection, allowing one, two, or three heterozygous SNPs per window. Although varying the number of heterozygous SNPs resulted in a reasonable range of sensitivity (82–94%), the specificity for two or three heterozygous SNPs was prohibitively low (43–73%). The highest QF stringency combining all filters provided the best results in terms of equilibrium between sensitivity (86%) and specificity (82%), underscoring the importance of extracting high-quality SNPs. Dropping the consensus filter retained identical specificity but lowered sensitivity by 1%. All subsequent analyses were performed using the highest degree of QF stringency.

### Optimization of homozygosity mapping on WES data

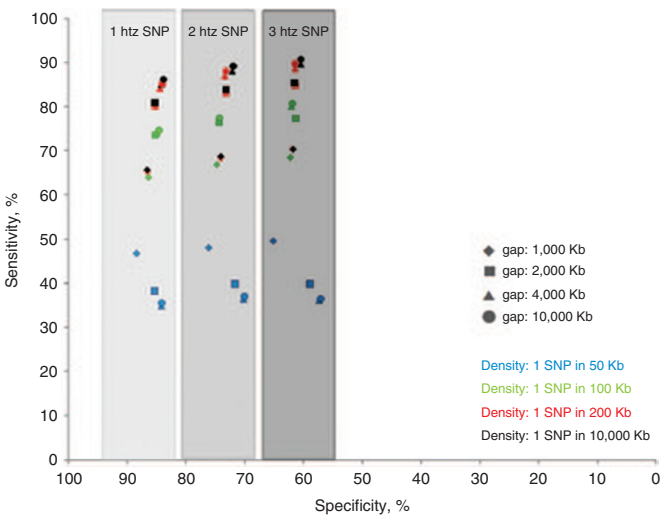
To determine the optimal parameters for homozygosity mapping using PLINK, multiple combinations of parameters were tested in our training set (**Table 2**). The number of SNPs defining an ROH (10 SNPs) and a sliding window (20 SNPs) were decreased compared with the SNP array analysis in order to keep a ratio similar to that between the number of SNPs extracted from WES data (~150,000 SNPs) and the ones present in the SNP array (715,000 SNPs). The other parameters were varied (minimum density, number of heterozygous SNPs per window, and maximum allowed gap). Using a minimum SNP density of 1 in 50, 100, 200, and 10,000 kb, we saw that a lower-density threshold resulted in an increase of the sensitivity, whereas the specificity was comparable (**Figure 1**). This means that more ROHs were found, without including too many false-positive regions. Also, there was no obvious difference between the results with a SNP density of 1 in 200 kb and 1 in 10,000 kb.



**Table 2** Parameters of runs of homozygosity detection in PLINK

Parameter	Command	SNP array	WES data	HOMWES optimal parameters
Size threshold (kb) to call an ROH	Homozyg-kb	1,000 kb	1,000 kb	1,000 kb
SNP number threshold to call an ROH	Homozyg-snp	100 SNPs	10 SNPs	10 SNPs
Sliding window size in SNPs	Homozyg-window-snp	50 SNPs	20 SNPs	20 SNPs
Allowed missing SNPs	Homozyg-window-missing	5 SNPs	10 SNPs	10 SNPs
Proportion of homozygous windows threshold	Homozyg-window-threshold	0.05	0.05	0.05
Minimum SNP density to call an ROH	Homozyg-density	1 SNP in 50 kb	1 SNP in 50, 100, 200, or 10,000 kb	1 SNP in 200 kb
Maximum allowed gap	Homozyg-gap	1,000 kb	1,000, 2,000, 4,000, or 10,000 kb	4,000 kb
Allowed heterozygous SNPs	Homozyg-window-het	3 SNPs	1, 2, or 3 SNPs	1 SNP

An overview of the parameters used for single-nucleotide polymorphism (SNP) array run of homozygosity (ROH) detection, homozygosity mapping based on whole-exome sequencing analysis (HOMWES) optimization, and the final HOMWES parameters.  
kb, kilobase; WES, whole-exome sequencing.



**Figure 1** Scatterplot showing sensitivity and specificity of the parameters for homozygosity mapping based on WES analysis optimization. Comparison of homozygous regions derived from single-nucleotide polymorphism (SNP) array and whole-exome sequencing data using a range of parameters: density of 1 SNP in 50, 100, 200, or 10,000 kb; allowing 1, 2, or 3 heterozygous SNPs per window (outlined in grey); and a gap of 1,000, 2,000, 4,000, or 10,000 kb. Calculations are based on data from individuals HSP-1.1, CMT-1.1, CMT-1.2, CMT-2.1, and CMT-3.1. The mean values of specificity and sensitivity are shown.

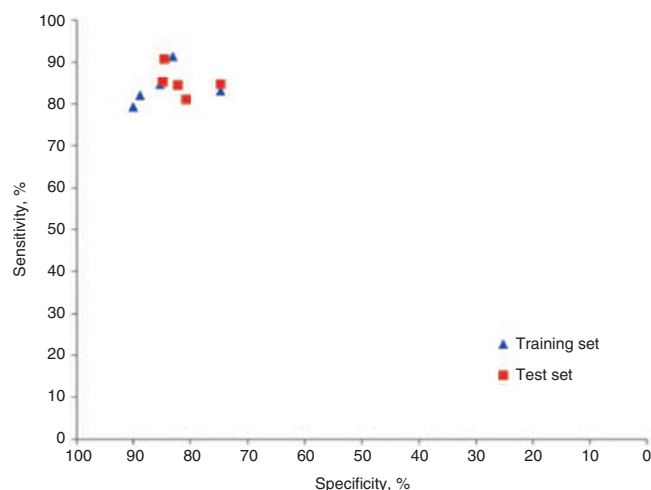
Allowing a higher degree of heterozygous calls per window (one to three) decreased the number of false-negative regions, but at the same time the false-positive ones were enriched. This is illustrated by the slight increase in sensitivity and significant decrease of specificity (Figure 1).

PLINK splits a homozygous region when two SNPs are located farther apart than the given threshold for the allowable gap. Although increasing the maximum gap size (1,000 to 10,000 kb) had no effect on specificity, it did cause an increase in sensitivity. This implies that more homozygous regions were included with a comparable percentage of false-positives results. In addition, we did not detect a substantial difference in the ROH selection between an allowed gap of 4,000 and 10,000 kb.

Evaluation of all combinations of parameters revealed six conditions with both high sensitivity and specificity (~80%). These conditions included a gap of 2,000, 4,000, or 10,000 kb, allowing only 1 heterozygous SNP per window, and a density of 1 in 200 or 10,000 kb. There is a small increase in sensitivity using a gap of 4,000 compared with 2,000 kb. For the extreme parameters (10,000-kb gap, 1 SNP/10,000 kb), we observed no difference compared with the more stringent setting. Therefore, we selected as optimal the following parameters: gap of 4,000 kb, 1 heterozygous SNP per window, and density of 1 in 200 kb.

To check the reproducibility of the results, five individuals who were not included in the optimization (CMT-4.1, CMT-5.1, CMT-6.1, CMT-7.1, CMT-8.1) and who had both SNP and WES data were used as a test set; they were analyzed with the determined HOMWES parameters (Table 2, Figure 2). Specificity and sensitivity were not significantly different between the subjects of the training set versus the test set ( $P = 0.81$  and  $P = 0.58$  for sensitivity and specificity, respectively). This implies that our parameters, and therefore our analysis, provide reproducible results. Overall, comparison of WES- and SNP array-derived homozygous regions demonstrated a high specificity ( $83.0 \pm 5.1\%$ ) and sensitivity ( $84.7 \pm 3.9\%$ ) of the method for all individuals. The optimized parameters were integrated into the HOMWES tool in the GenomeComb package (version 0.11.0) for genomic data analysis.

Although we used two different variant callers (GATK and Samtools) in our filtering pipeline, variant call format files are routinely generated using one variant caller only. The initial filtering test did not show large differences between high degrees of QF stringency using one or two variant calling software (Supplementary Figure S3 online). For that reason, we performed ROH detection using the HOMWES tool with only one variant caller for the 10 individuals with both SNP and WES data. The results when applying GATK, Samtools, or both showed similar specificity ( $82.5 \pm 5.1$ ,  $83.7 \pm 9.2$ , and  $83.0 \pm 5.1\%$ , respectively) and sensitivity ( $84.6 \pm 4.2$ ,  $79.8 \pm 11.8$ , and  $84.7 \pm 3.8\%$ , respectively). HOMWES can thus be applied to typical variant call format files from a single variant caller without a significant loss of performance.

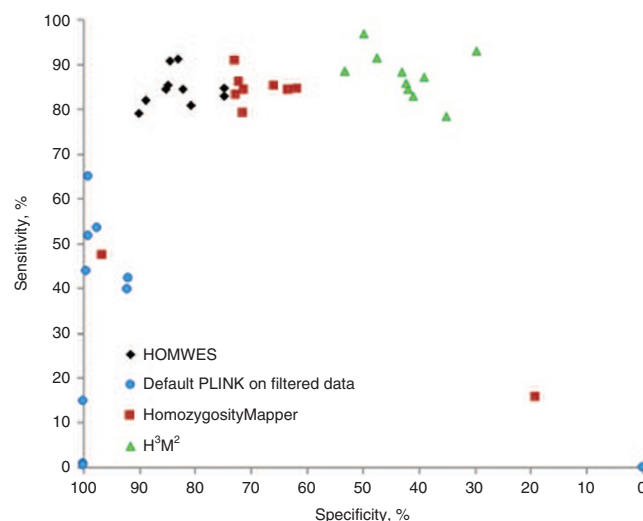


**Figure 2** Scatterplot showing the validation of the optimal parameters for homozygosity mapping based on whole-exome sequencing analysis (gap of 4,000 kb, 1 heterozygous single-nucleotide polymorphism per window, and a density of 1 in 200 kb).

### Mutation identification

Before this study, the causative mutations were already identified in two families (CMT-2: *HSPB1* c.250G>A, p.G84R<sup>24</sup>; CMT-3: *PRX* c.2098delG, p.A700Pfs\*17 (refs. 21, 25); **Supplementary Table S1** online) by SNP array-based homozygosity mapping and Sanger sequencing. These patients served as positive controls for the chosen parameters. The *HSPB1* mutation was located in the largest homozygous region found in CMT-2 (32.9 Mb) based on the array data. Using our optimal HOMWES conditions, the pathogenic variant was located in a 9.9-Mb region, ranked as the 11th largest homozygous region because of the fragmentation affecting larger regions (**Table 1**; **Supplementary Table S2** online; see Discussion). In CMT-3, the causative mutation in *PRX* was found in a region only 1.7 Mb in size; the largest ROH was 19 Mb. This region and the associated mutation were successfully retrieved with our HOMWES analysis.

In addition, in four families (CMT-8, HSP-2, HSP-3, HSP-4) novel pathogenic variants were identified in known CMT or HSP genes (**Table 1**; **Supplementary Table S1** online). The clinical features of the patients with novel mutations are summarized in **Supplementary Tables S3 and S4** online. For one family (CMT-8), for which both SNP and WES data were available, the mutation was found in one of the largest homozygous regions defined by both approaches (*GAN* c.140T>C, p.I47T). In three additional families, we successfully identified the disease-causing mutation in the ROHs generated exclusively from WES data (HSP-2: *GBA2* c.451+2T>C; HSP-3 and HSP-4: *ZFYVE26* c.2639T>C, p.L880P). The *ZFYVE26* mutation is located in one of the largest homozygous regions (HSP-3: 9.8 Mb, HSP-4: 9 Mb), whereas the mutation in *GBA2* was identified in a region of only 2.5 Mb (rank: 41). The *ZFYVE26* variant was found to reside on a shared haplotype between the two Roma families, who were known to be distantly related (**Supplementary Figure S2b** online). Our relatedness calculations revealed a proportion of identical by descent alleles between HSP-3.1 and



**Figure 3** Comparison of the accuracy of runs of homozygosity detection using homozygosity mapping based on WES analysis (HOMWES), HomozygosityMapper, and H<sup>3</sup>M<sup>2</sup> methods. Calculations were performed using the optimized protocol based on PLINK (HOMWES), the default PLINK parameters for homozygosity mapping using quality filtered data extracted from whole-exome sequencing (default PLINK on filtered data), and default parameters for HomozygosityMapper and H<sup>3</sup>M<sup>2</sup>.

HSP-4.1, corresponding to an approximately first-cousin relationship (**Supplementary Table S5** online).

All mutations segregated in additional family members and were not present in 1000 Genomes Project; National Heart, Lung, and Blood Institute Exome Variant Server; or population-matched controls (**Supplementary Figures S1 and S2** online; **Supplementary Table S1** online). The *GBA2* c.451+2T>C variant is situated in the canonical donor splice site of the second exon–intron junction of *GBA2*, and skipping exon 2 would result in a frameshift and stop codon 26 amino acids after exon 1. Alternatively, intron 2 retention would also create a stop codon 140 nucleotides after the splice site. No material from the patients was available for testing these predictions.

All variants are predicted to be deleterious, affect evolutionary conserved nucleotides and amino acid residues, and are classified as “pathogenic” or “likely pathogenic” according to the American College of Medical Genetics and Genomics guidelines (**Supplementary Table S1** online).

### HOMWES performance compared to HomozygosityMapper and H<sup>3</sup>M<sup>2</sup>

The accuracy of ROH detection was compared between HOMWES and HomozygosityMapper or H<sup>3</sup>M<sup>2</sup> using the individuals for whom both SNP array and WES data were available. As shown in **Figure 3**, we observed comparable sensitivity for all three methods, with mean of  $84.20 \pm 3.44$  and  $86.91 \pm 5.28\%$  for HomozygosityMapper and H<sup>3</sup>M<sup>2</sup>, respectively, compared with  $84.69 \pm 3.69\%$  for our protocol. Importantly, our optimized protocol showed a significantly higher specificity ( $82.94 \pm 5.12\%$ ) juxtaposed with  $67.52 \pm 5.96\%$  for HomozygosityMapper and  $42.19 \pm 7.15\%$  for H<sup>3</sup>M<sup>2</sup>. An analysis of quality filtered SNPs

with the default PLINK parameters also was included for completeness (Figure 3). It showed lower performance with a huge decrease in sensitivity (mean of  $31.22 \pm 24.75\%$ ), despite a slightly higher specificity ( $87.92 \pm 31.05\%$ ).

## DISCUSSION

Recent literature and our own experience suggest that gene hunting is most successful in families with AR conditions in whom homozygosity mapping can be applied, as this positional cloning approach dramatically reduces the number of variants requiring genetic and functional validation.<sup>21,26,27</sup> Here we evaluated the performance of different SNP extraction strategies and PLINK parameters for homozygosity mapping, using WES data of patients with AR HSP and AR CMT. We determined the parameters for detection of ROHs from WES data with high sensitivity (84.4%) and specificity (82.4%), using high-density SNP arrays as a reference, and implemented them in the publicly available HOMWES tool. In the first step of HOMWES analysis, high-quality data are selected for further processing by including only SNPs with GATK “pass” status, genotype quality >40, not situated within microsatellites or simple repeats, and having consensus calls by GATK and Samtools. Next, homozygosity mapping is performed with PLINK using the optimized parameters summarized in Table 2. In our data set, this reduced the regions to be examined by more than 85%. Inclusion of additional affected family members (when available) further decreased the candidate regions, as illustrated in Table 1. Limiting validation and analysis of potential disease-causing mutations to those located within ROHs resulted in a three- to sixfold decrease in candidate variants per individual (Supplementary Table S6 online).

We were able to retrieve all homozygous regions in the control families with a mutation in known CMT genes, based on the WES data. In family CMT-3, the mutation in *PRX* was identified in a region only 1.7 Mb in size, which is below the standard homozygosity mapping detection thresholds.<sup>28,29</sup> Moreover, in four families, our approach led to the identification of novel pathogenic variants in known HSP and CMT genes.

The p.I47T in *GAN* was found in two pairs of siblings in the same family of Turkish origin who were diagnosed with an axonal form of CMT. Mutations in *GAN* have been associated with giant axonal neuropathy type 1, a chronic polyneuropathy of childhood that affects both the peripheral and central nervous systems. The phenotype of our patients is compatible with the clinical presentation attributed to mutations in this gene.

The c.451+2T>C variant in *GBA2* was identified in two siblings of Roma origin who were affected by complicated HSP accompanied by cerebellar ataxia and mild axonal polyneuropathy. Mutations in *GBA2* have been associated with SPG46 (complicated HSP with ataxia, mental impairment, and cataracts)<sup>30</sup> and a recessive form of cerebellar ataxia with neuropathy and spasticity.<sup>31</sup> The majority of the identified mutations, similar to the variant described here, cause loss of function of the protein.

The p.L880P substitution in *ZFYVE26* was identified in two related Roma pedigrees. The two younger affected individuals

had pure HSP at the time of the diagnosis (4–6 years after clinical onset), whereas the oldest patient developed a complex phenotype with axonal polyneuropathy, a thin corpus callosum, and white matter lesions on neuroimaging. *ZFYVE26* defects are the second most common cause (after *SPG11*) of complicated HSP with a thin corpus callosum.<sup>32</sup> Only one missense mutation in *ZFYVE26* (p.Ile508Asn) has been reported previously. All the remaining described variants cause truncation of the protein, pointing to a predominant loss-of-function mechanism.<sup>33</sup> Notably, our study is the first report of *GBA2* and *ZFYVE26* mutations causing HSP in the Roma population.

Several homozygosity mapping methods based on WES have been developed. We compared the performance of our protocol with only HomozygosityMapper and H<sup>3</sup>M<sup>2</sup>, since the others (AgileGenotyper/AgileVariantMapper and HomSI) have only a graphical output that cannot be used for objective comparisons. All three algorithms achieved comparably high sensitivity, but regarding the specificity, our method outperformed the other two. HomozygosityMapper failed to detect the 1.7-Mb ROH containing the *PRX* mutation in CMT-3. H<sup>3</sup>M<sup>2</sup> found all the regions with the disease-causing mutation in the studied families and showed the highest sensitivity, but at the price of an almost twofold reduction in specificity. One explanation for this low performance might be the enrichment of the H<sup>3</sup>M<sup>2</sup>-generated map with uninformative homozygous reference SNPs. Our results apparently contradict the observation by Magi et al.<sup>10</sup> of better accuracy of H<sup>3</sup>M<sup>2</sup> compared with PLINK, but these authors did not allow lower stringency thresholds in PLINK and used a different SNP extraction method. Furthermore, neither HomozygosityMapper nor H<sup>3</sup>M<sup>2</sup> allow QF of the extracted SNPs.

Despite our efforts to select optimal ROH detection parameters, the established protocol still imposes certain limitations. One obvious problem concerns the SNP distribution. WES provides SNPs located exclusively in exonic regions, whereas the SNP arrays yield markers evenly distributed across the genome. This creates difficulties in delineating ROHs encompassing fewer genes. A second drawback concerns the number and size of the homozygous regions. We observed an increase in the number of ROHs based on WES compared with SNP arrays (~21%), whereas the total size of the homozygous regions remains similar (Table 1). These differences reflect the tendency of large ROHs to be fragmented into several smaller ones by PLINK (Supplementary Table S2 online), which is largely a result of the presence of errors in WES data. Nevertheless, in our experience, the breaks between the fragmented regions are usually less than 1 Mb, allowing the recognition of the large regions upon manual examination.

Another consideration is related to the false-negative and false-positive results associated with smaller ROHs (<5 Mb, average size 1.57 Mb). The larger regions are found in a consistent way by both WES- and array-based analysis. So we sought to determine whether false-positive and false-negative regions result in loss or gain of homozygous variants. To test this, we selected two affected individuals from CMT-1 and filtered the variants in all



false-positive and false-negative regions (47.30 and 66.86Mb for CMT-1.1 and CMT-1.2, respectively). Homozygous variants within these regions were selected followed by excluding common polymorphisms with minor allele frequency >5% in dbSNP, 1000 Genomes Project, and the NHLBI Exome Variant Server. After this selection no variants remained, indicating noninformative regions. Therefore, we consider the false-negative regions as less likely to contain mutations derived from recent ancestors. On the other hand, the false-positive regions do not significantly increase the list of candidate variants, and therefore do not considerably affect the quality of the analysis.

In conclusion, we propose a protocol for homozygosity mapping using HOMWES implemented in the GenomeComb package, which is applicable for gene hunting in AR diseases. Our method showed increased specificity in comparison to HomozygosityMapper and H<sup>2</sup>M<sup>2</sup>, highlighting the importance of establishing the most appropriate parameters for this type of analysis. The usefulness of the protocol is demonstrated by the identification of three novel mutations in genes that are causative for HSP and CMT.

## SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at <http://www.nature.com/gim>

## ACKNOWLEDGMENTS

The authors are grateful to all study participants and their families for their cooperation and thank Luba Kalaydjieva for helpful discussions and Banu Anlar for assisting in patient descriptions. This work was funded in part by the University of Antwerp (TOP BOF 29069 to A.J.); the Fund for Scientific Research–Flanders (to A.J.), Tom Wahlig Foundation, Jena, Germany (to A.J. and I.T.); and the Bogazici University Research Fund (11B01P7 to E.B.). D.A. and A.E.C. are supported by fellowships from the Fund for Scientific Research–Flanders. D.K. received a travel grant from the Boehringer Ingelheim Fonds, Germany.

## DISCLOSURE

The authors declare no conflict of interest.

## REFERENCES

- Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* 1987;236:1567–1570.
- Pereiro I, et al. New mutations in BBS genes in small consanguineous families with Bardet-Biedl syndrome. Detection of candidate regions by homozygosity mapping. *Mol Vis* 2010;16:137–143.
- Gooding R. A novel Gypsy founder mutation, p.Arg1109X in the CMT4C gene, causes variable peripheral neuropathy phenotypes. *J Med Genet* 2005;42:e69.
- Hildebrandt F, Heeringa SF, Rüschendorf F, et al. A systematic approach to mapping recessive disease genes in individuals from outbred populations. *PLoS Genet* 2009;5:e1000353.
- Sheffield VC, Stone EM, Carmi R. Use of isolated inbred human populations for identification of disease genes. *Trends Genet* 1998;14:391–396.
- Zhuang Z, Gusev A, Cho J, Pe'er I. Detecting identity by descent and homozygosity mapping in whole-exome sequencing data. *PLoS One* 2012;7:e47618.
- Carr IM, Bhaskar S, O'Sullivan J, et al. Autozygosity mapping with exome sequence data. *Hum Mutat* 2013;34:50–56.
- Görmez Z, Bakir-Gungor B, Sagioglu MS. HomSI: a homozygous stretch identifier from next-generation sequencing data. *Bioinformatics* 2014;30:445–447.
- Seelow D, Schuelke M. HomozygosityMapper2012—ridging the gap between homozygosity mapping and deep sequencing. *Nucleic Acids Res* 2012;40(Web Server issue):W516–W20.
- Magi A, Tattini L, Palombo F, et al. H3M2: detection of runs of homozygosity from whole-exome sequencing data. *Bioinformatics* 2014;30:2852–2859.
- Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559–575.
- Howrigan DP, Simonson MA, Keller MC. Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. *BMC Genomics* 2011;12:460.
- Pippucci T, Magi A, Gialluisi A, Romeo G. Detection of runs of homozygosity from whole exome sequencing data: state of the art and perspectives for clinical, population and epidemiological studies. *Hum Hered* 2014;77:63–72.
- Fuentes Fajardo KV, Adams D, Mason CE, et al.; NISC Comparative Sequencing Program. Detecting false-positive signals in exome sequencing. *Hum Mutat* 2012;33:609–613.
- Timmerman V, Clowes VE, Reid E. Overlapping molecular pathological themes link Charcot-Marie-Tooth neuropathies and hereditary spastic paraplegias. *Exp Neurol* 2013;246:14–25.
- Kancheva D, Chamova T, Guergueltcheva V, et al. Mosaic dominant TUBB4A mutation in an inbred family with complicated hereditary spastic paraplegia. *Mov Disord* 2015;30:854–858.
- Reumers J, et al. Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat Biotechnol* 2011;30:61–68.
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
- McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.
- Li H, Handsaker B, Wysoker A, et al.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
- Zimon M, et al. Unraveling the genetic landscape of autosomal recessive Charcot-Marie-Tooth neuropathies using a homozygosity mapping approach. *Neurogenetics* 2015;16:33–42.
- Sheskin D. *Handbook of Parametric and Nonparametric Statistical Procedures*. 3rd edn. Chapman & Hall/CRC: Boca Raton, FL, 2004.
- Richards S, Aziz N, Bale S, et al.; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405–424.
- James PA, Rankin J, Talbot K. Asymmetrical late onset motor neuropathy associated with a novel mutation in the small heat shock protein HSPB1 (HSP27). *J Neurol Neurosurg Psychiatry* 2008;79:461–463.
- Auer-Grumbach M, Fischer C, Papić L, et al. Two novel mutations in the GDAP1 and PRX genes in early onset Charcot-Marie-Tooth syndrome. *Neuropediatrics* 2008;39:33–38.
- Alkuraya FS. The application of next-generation sequencing in the autozygosity mapping of human recessive diseases. *Hum Genet* 2013;132:1197–1211.
- Zimón M, Baets J, Almeida-Souza L, et al. Loss-of-function mutations in HINT1 cause axonal neuropathy with neuromyotonia. *Nat Genet* 2012;44:1080–1083.
- Woods CG, Cox J, Springell K, et al. Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. *Am J Hum Genet* 2006;78:889–896.
- Pippucci T, Benelli M, Magi A, et al. EX-HOM (EXome HOMozygosity): a proof of principle. *Hum Hered* 2011;72:45–53.
- Martin E, Schüle R, Smets K, et al. Loss of function of glucocerebrosidase GBA2 is responsible for motor neuron defects in hereditary spastic paraplegia. *Am J Hum Genet* 2013;92:238–244.
- Hammer MB, Eleuch-Fayache G, Schottlaender LV, et al. Mutations in GBA2 cause autosomal-recessive cerebellar ataxia with spasticity. *Am J Hum Genet* 2013;92:245–251.
- Goizet C, Boukhris A, Maltete D, et al. SPG15 is the second most common cause of hereditary spastic paraplegia with thin corpus callosum. *Neurology* 2009;73:1111–1119.
- Pensato V, Castellotti B, Gellera C, et al. Overlapping phenotypes in complex spastic paraplegias SPG11, SPG15, SPG35 and SPG48. *Brain* 2014;137(Pt 7):1907–1920.