

## ORIGINAL ARTICLE

# HDR: a statistical two-step approach successfully identifies disease genes in autosomal recessive families

Atsuko Imai<sup>1,2,3,4</sup>, Masakazu Kohda<sup>5</sup>, Akihiro Nakaya<sup>2</sup>, Yasushi Sakata<sup>1</sup>, Kei Murayama<sup>6</sup>, Akira Ohtake<sup>7</sup>, Mark Lathrop<sup>4</sup>, Yasushi Okazaki<sup>5,8</sup> and Jurg Ott<sup>3,9</sup>

In the search for sequence variants underlying disease, commonly applied filtering steps usually result in a number of candidate variants that cannot further be narrowed down. In autosomal recessive families, disease usually occurs only in one generation so that genetic linkage analysis is unlikely to help. Because homozygous recessive mutations tend to be inherited together with flanking homozygous variants, we developed a statistical method to detect pathogenic variants in autosomal recessive families: We look for differences in patterns of homozygosity around candidate variants between patients and control individuals and expect that such differences are greater for pathogenic variants than random candidate variants. In six autosomal recessive mitochondrial disease families, in which pathogenic homozygous variants have already been identified, our approach succeeded in prioritizing pathogenic mutations. Our method is applicable to single patients from recessive families with at least a few dozen control individuals from the same population; it is easy to use and is highly effective for detecting causative mutations in autosomal recessive families.

*Journal of Human Genetics* (2016) 61, 959–963; doi:10.1038/jhg.2016.85; published online 30 June 2016

## INTRODUCTION

In genetic linkage analysis, lod score peaks at a true disease variant and surrounding markers tend to be somewhat wider than at a random peak of the same height.<sup>1–5</sup> A special situation exists for cases of recessive disease when parents are related.<sup>6</sup> Markers in extended regions surrounding the disease variant all tend to be homozygous, which has led to the concept of homozygosity mapping, which consists of a search in an individual with a recessive trait for extended regions of homozygosity in adjacent markers.<sup>7</sup> Various forms of homozygosity mapping have been implemented in computer programs.<sup>8</sup> An extension of homozygosity mapping makes use of population haplotype frequencies estimated in unrelated control individuals.<sup>9</sup>

Here we take a novel two-step approach, where step 2 is based on our previous work.<sup>10</sup> We start with a set of candidate variants that were obtained by customary filtering steps (see below). At each candidate variant and surrounding markers, we focus on homozygous genotypes and compare the string of genotypes at these loci between a patient and control individuals. For a random candidate variant, we expect a random similarity of genotypes

between patient and control but for a pathogenic variant, the patient is likely to be much more different than control individuals. Details for our method are presented below.

## MATERIALS AND METHODS

### Sequencing pipeline

For the eight patients used for this study, indexed genomic DNA libraries were prepared from patient genomic DNA, and exomes were captured using TruSeq (Illumina, San Diego, CA, USA) and SureSelect V4 (Agilent Technologies, Santa Clara, CA, USA) exome enrichment kits according to the manufacturers' protocols. Sequencing was performed using 100-bp paired-end reads on a HiSeq2000 or GAIIX (Illumina). Read trimming via base quality was performed using Trimmomatic.<sup>11</sup> Read alignments to the 1000 Genomes Project phase II reference genome (hs37d5.fa) were performed with the Burrows–Wheeler Aligner<sup>12</sup> (BWA, version 0.7.0). PCR duplicate reads were removed using Picard (version 1.89); non-mappable reads were removed using SAMtools<sup>13</sup> (version 0.1.19). After filtering out those reads, we applied the Genome Analysis Toolkit<sup>14</sup> (GATK version 2.4-9-nightly-2013-04-12-g3fc5478) base quality score recalibration and performed SNP and INDEL discovery (UnifiedGenotyper with stand\_call\_conf 50.0 and stand\_emit\_conf 10.0 setting). Detailed protocols of variant calling and prioritization are described in the previous paper.<sup>15</sup>

<sup>1</sup>Department of Cardiovascular Medicine, Osaka University Graduate School of Medicine, Osaka, Japan; <sup>2</sup>Department of Genome Informatics, Osaka University Graduate School of Medicine, Osaka, Japan; <sup>3</sup>Laboratory of Statistical Genetics, Rockefeller University, New York, NY, USA; <sup>4</sup>McGill University and Genome Québec Innovation Centre, Montréal, Québec, Canada; <sup>5</sup>Division of Translational Research, Research Center for Genomic Medicine, Saitama Medical University, Saitama, Japan; <sup>6</sup>Department of Metabolism, Chiba Children's Hospital, Chiba, Japan; <sup>7</sup>Department of Pediatrics, Saitama Medical University, Saitama, Japan; <sup>8</sup>Division of Functional Genomics & Systems Medicine, Research Center for Genomic Medicine, Saitama Medical University, Saitama, Japan and <sup>9</sup>Institute of Psychology, Chinese Academy of Sciences, Beijing, China  
Correspondence: Professor Y Okazaki, Division of Translational Research, Research Center for Genomic Medicine, Saitama Medical University, 1397-1 Yamane, Hidaka City, Saitama 350-1241, Japan.  
E-mail: okazaki@saitama-med.ac.jp

or Professor J Ott, Laboratory of Statistical Genetics, Rockefeller University, 1230 York Avenue, New York, NY 10065, USA.

E-mail: ott@rockefeller.edu

Received 12 February 2016; revised 7 June 2016; accepted 8 June 2016; published online 30 June 2016

**Quality control**

The raw sequence read data passed the quality checks in FASTQC. Variant data used in this study were identified as described above. Raw variants were then filtered for the removal of low-quality variant calls with GATK's Variant Filtration tool, with filtering based on the QualByDepth, ReadPosRankSum, FisherStrand, Depth, Mapping QualityZeroReads, HaplotypeScore, MappingQualityRankSumTest, AlleleBalance, ClusteredSnps and IndelArtifact attributes.

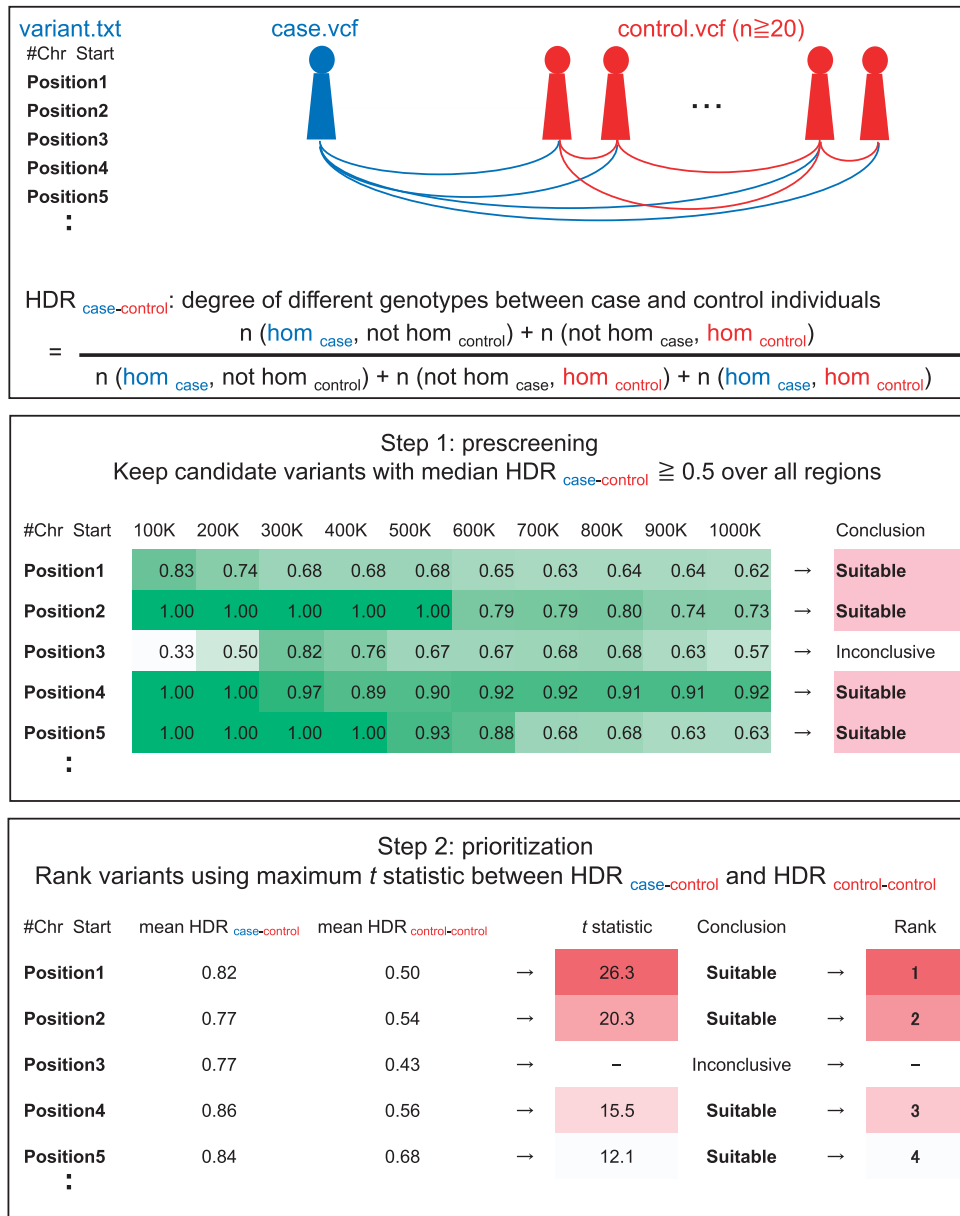
**Case and control individuals**

Among 142 patients with childhood-onset and enzymatically diagnosed mitochondrial respiratory chain complex deficiencies in the original article,<sup>15</sup> we selected eight patients, Pt105, Pt250, Pt268, Pt276, Pt286, Pt314, Pt330 and Pt559 with known pathogenic homozygous mutations for our analysis. For the exome capture kit, SureSelect V4 was used for Pt559 and TruSeq was used for all the remaining seven patients. We had to exclude two patients from our

analysis, Pt652 and Pt711, who also had known pathogenic homozygous mutations,<sup>15</sup> since the exome capture kit used for these two patients was different from TruSeq or SureSelect V4. In the original study,<sup>15</sup> only TruSeq and SureSelect V4 were used for more than 20 control individuals and our method required at least 20 control individuals sequenced by the same platform as the patient. Detailed clinical explanation is available in the original article.<sup>15</sup> As control individuals, we used 52 individuals captured by TruSeq and 39 individuals captured by SureSelect V4, all of whom were sequenced by the same pipeline as the patients without long continuous stretches of homozygosity detected by high-density oligonucleotide array.

**Statistical analysis of dissimilarity**

Consider a candidate variant and flanking markers. At each of these sites we distinguish two genotypes, (a) homozygous for the alternate allele (ALT/ALT) and (b) not homozygous for the alternate allele (REF/ALT or REF/REF). For



**Figure 1** Summary of our procedures. Positions of candidate variants with top five statistics for Pt250 are shown as an example of how our approach works. Among these five positions, position 1 refers to the known pathogenic variant in the *QRSL1* gene. Upper box: How to calculate HDR between two individuals, for example, between case and a control, where n refers to number of variants. Middle box: Step 1 procedure (prescreening). Lower box: Step 2 procedure (prioritization).

two individuals, there are thus four pairs of genotypes at a given site, (a)–(a), (a)–(b), (b)–(a) and (b)–(b), and  $n_1$ ,  $n_2$ ,  $n_3$  and  $n_4$  are the respective numbers of markers with these genotype pairs. However, for each individual we extract from the *vcf* sequencing file only homozygous (ALT/ALT) variants, so in the resulting variant lists for two individuals the last genotype pair is not observed. For the candidate variant and flanking markers, we define as our measure of dissimilarity between the two strings of genotypes of two individuals by the number of markers at which the genotypes of the two individuals are different, that is,  $HD = n_2 + n_3$ , where HD is known as the Hamming distance.<sup>16,17</sup> To accommodate varying numbers of flanking markers, we work with the relative Hamming distance, or Hamming distance ratio,  $HDR = (n_2 + n_3) / (n_1 + n_2 + n_3)$ .<sup>10</sup> The motivation for our approach is that we expect a larger ‘distance’ between case and control individuals for DNA segments containing pathogenic variants than for random DNA regions.

To classify DNA segments containing pathogenic variants more precisely, we developed a new test statistic as outlined below. We work with 10 DNA segments of 100 to 1000 kb (in steps of 100 kb) around a candidate variant since the maximum of our test statistic tends to occur within this interval (see below).

**Step 1.** Initially, we prescreen candidate variants for being suitable for application of our method, that is, we require HDR for case–control pairs to be no smaller than HDR values expected by chance. As outlined in Supplementary Information, for a wide range of variant allele frequencies, random HDR values tend to exceed the value 0.50. Thus, we will only consider variants when at least half of all HDR values between patient and each control individual in each of the 10 DNA segments attain a value of 0.50 or higher. In other words, we reject a candidate variant as ‘unsuitable’ if the median HDR between patient and control individuals is less than 0.50.

**Step 2.** Once a candidate variant passes this prescreening we form all possible pairs of individuals and compute HDR in each pair, contrasting pairs containing the affected individual (group 1) with pairs consisting of two control individuals (group 2).<sup>10</sup> If the candidate variant is pathogenic we expect HDR to be higher in group 1 pairs than in group 2 pairs, that is, members of control–control pairs are expected to be more similar to each other than members of case–control pairs (HDR values in control–control pairs represent our observed random HDR values). At each of the 10 DNA segments, we compute a one-sided *t* statistic with unequal variances to assess the difference, group 1 mean minus group 2 mean. Our final test statistic is  $t_{\max}$ , the maximum *t* obtained over the 10 DNA segments, which is used to prioritize candidate variants.

We evaluate empirical significance levels by choosing each of the *n* control individuals in turn as a pseudo-case and repeat analysis for each such pseudo-dataset. To estimate the *P*-value associated with our observed  $t_{\max}$  statistic, we count the number *k* of pseudo- $t_{\max}$  values that are at least as large as the observed  $t_{\max}$  value. Including the observed data among the pseudo-data, we estimate the empirical significance level as  $p = (k+1)/(n+1)$ . Thus, in the best case scenario ( $k=0$ ), to obtain a *P*-value smaller than 0.05, we need at least  $n=20$  control individuals. Our procedures are summarized in Figure 1.

## Software

Our approach is implemented in a GUI program with pull-down menus, which allows users to compute HDR values.<sup>18</sup> Output from the HDR program will then be used as input to the maxstatRS.FX program<sup>18</sup> for statistical analysis and ranking of candidate variants. Below we briefly describe three input files that users need to prepare for the HDR program.

The *variant.txt* file contains chromosome numbers and start positions of all candidate variants after the usual filtering steps have been exhausted.

For each patient, the user prepares a *case.vcf* file containing only homozygous variants for the alternate allele with quality attribute PASS by GATK filtering from the patient’s *vcf* file as obtained by the sequencing pipeline described above. For each control individual, an analogous *control.vcf* file is generated. Sequencing for control data should have been processed by the same protocol as for case individuals.

**Table 1 Prescreening of candidate variants (step 1) and ranking of known pathogenic variants (step 2)**

| Patient | Gene           | Disease | Step 1 |              | Step 2   |      |                 |
|---------|----------------|---------|--------|--------------|----------|------|-----------------|
|         |                |         | MinHDR | Conclusion   | <i>m</i> | Rank | <i>P</i> -value |
| 105     | <i>SUCLA2</i>  | MC      | 0.56   | Suitable     | 60       | 2    | 0.038           |
| 250     | <i>QRSL1</i>   | LIMD    | 0.62   | Suitable     | 74       | 1    | 0.038           |
| 268     | <i>BOLA3</i>   | LIMD    | 0.75   | Suitable     | 88       | 4    | 0.019           |
| 276     | <i>MRPS23</i>  | HD      | 0.62   | Suitable     | 59       | 2    | 0.038           |
| 286     | <i>BOLA3</i>   | LD      | 0.64   | Suitable     | 69       | 2    | 0.019           |
| 314     | <i>BOLA3</i>   | CM      | 0.60   | Suitable     | 69       | 2    | 0.038           |
| 330     | <i>NDUFAF6</i> | MC      | 0.33   | Inconclusive | —        | —    | —               |
| 559     | <i>TUFM</i>    | NLIMD   | 0.36   | Inconclusive | —        | —    | —               |

Abbreviations: CM, cardiomyopathy; HD, hepatic disease; LD, Leigh’s disease; LIMD, lethal infantile mitochondrial disorder; *m*, number of candidate variants; MC, mitochondrial cytopathy; MinHDR, smallest median HDR in patient–control pairs; NLIMD, non-lethal infantile mitochondrial disorder; rank, order of test statistic among *m* candidate variants (largest  $t_{\max}$  ranked 1); *P*, empirical significance level of test statistic.

**MaxstatRS update.** The user can choose between one of two modes of analysis. (1) In the default setting, candidate variants are ranked by the maximum test statistic described above. (2) If the user specifies a candidate variant as the likely or known disease variant then more detailed results are provided for this variant.

## RESULTS

We applied and validated our method in the eight autosomal recessive mitochondrial disease patients, in which Kohda *et al.*<sup>15</sup> had identified pathogenic homozygous nuclear gene mutations.

Our prescreening step singled out two variants as being unsuitable for our analysis (Table 1).

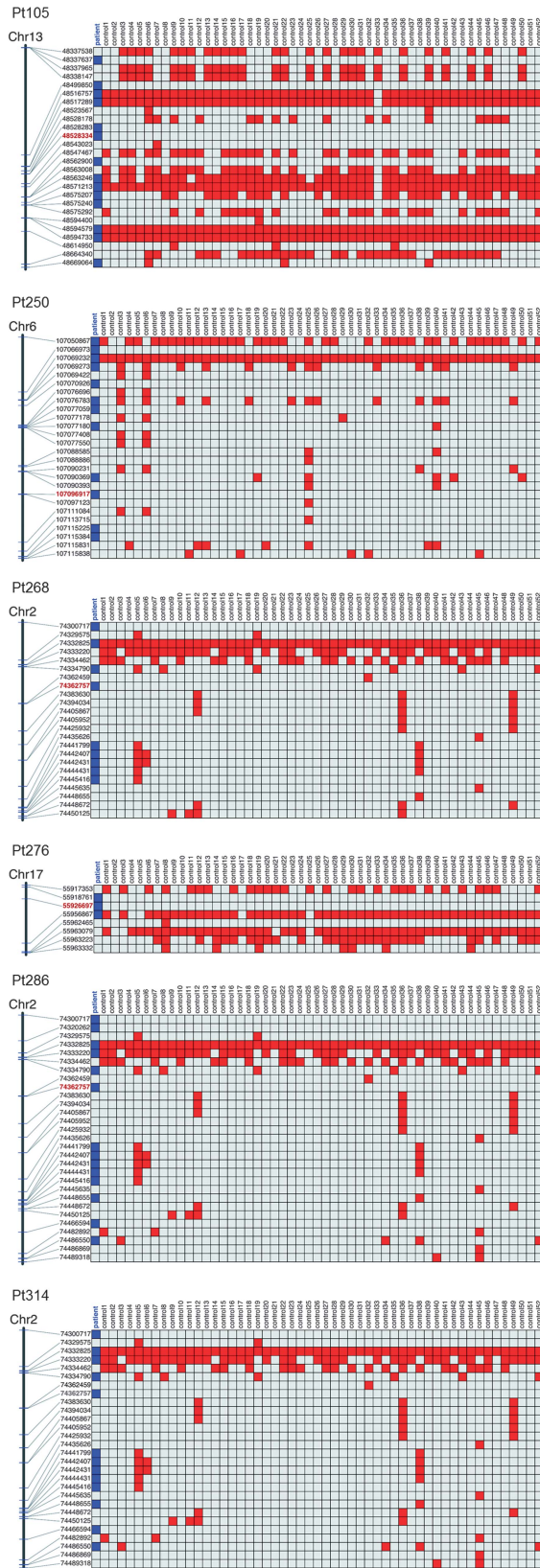
We subsequently applied our previously developed method to the six patients surviving the prescreening step. As shown in Table 1, out of 59–88 candidate variants, our method succeeded in narrowing down pathogenic variants to be ranked 1 through 4. Thus, our method, based on statistical analysis alone, identified six patients with homozygous mutations as being pathogenic.

We further visually represent the data of Table 1 in graphical form. As Figure 2 shows, at the regions with known pathogenic variants, different homozygous patterns between patient and controls are clearly observed.

## DISCUSSION

Here we demonstrate that purely by applying a sophisticated statistical genomics approach we can identify the same genes as those found by Kohda *et al.*<sup>15</sup> Of course, functional analysis<sup>15</sup> is final proof, which our method cannot furnish, but it is comforting to know that our approach can narrow down candidate variants to a very small set of likely pathogenic variants. It is useful both for clinicians who explore causative genes for their patients and also for scientists who try to confirm causative variants by conducting functional analysis.

For their mitochondrial disease patients, Kohda *et al.*<sup>15</sup> completed comprehensive analyses and our result is consistent with their convincing proof. Possible reasons that similar homozygous region patterns are seen between Pt559 and control individuals would be that the *TUFM* gene is located close to the centromere, where homozygous variants tend to accumulate. In comparison, a possible cause why Pt330 has homozygous patterns similar to those in multiple control individuals would be that *NDUFAF6* was mutated in four unrelated patients<sup>15</sup> and only Pt330 harbors a homozygous mutation while the remaining three patients have alleles different from those in Pt330,



**Figure 2** Homozygosity patterns of patient and control individuals. Blue: homozygous variants seen in patient. Red: homozygous variants seen in control individuals. Dark red: Positions of known pathogenic variants.

which suggests that the homozygous regions in *NDUF6* gene seen in Pt330 may not be a typical region originating from common Japanese founders. In comparison, *BOLA3* was mutated in four unrelated patients and all four patients carry the same mutation, c.287A>G:p.H96R, with Pt268, Pt286 and Pt314 all being homozygous and the remaining patient being a compound heterozygote. We can assume that the homozygous regions seen in Pt268, Pt286 and Pt314 are inherited from a common ancestor, which was also suggested by Kohda *et al.*<sup>15</sup> in that p.H96R originated from a single Japanese founder. The remaining variants in the *SUCLA2*, *QRSL1* and *MRPS23* genes are only seen in one unrelated patient each.

Our method is applicable even when genetic data on other family members are unavailable. It is also applicable to prioritizing homozygous mutations in different ethnic populations, for example, in French-Canadians<sup>19–21</sup> and in the Japanese population.

We implemented our program in a GUI version, which is freely available.<sup>18</sup> It is easy to use and does not require specific knowledge in computer science. To our knowledge, no other software comparable to ours is available.

**CONFLICT OF INTEREST**

The authors declare no conflict of interest.

**ACKNOWLEDGEMENTS**

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number JP16K19404 (AI); a grant of Strategic Research Center in Private Universities from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan ([http://www.mext.go.jp/a\\_menu/koutou/shinkou/07021403/002/002/1218299.htm](http://www.mext.go.jp/a_menu/koutou/shinkou/07021403/002/002/1218299.htm)); by the Practical Research Project for Rare/Intractable Diseases from Japan Agency for Medical Research and Development (AMED) (<http://www.amed.go.jp/en/>) (YO); and NSFC grant number 31470070 (JO) from the Chinese Government.

- 1 Terwilliger, J. D., Shannon, W. D., Lathrop, G. M., Nolan, J. P., Goldin, L. R., Chase, G. A. *et al*. True and false positive peaks in genomewide scans: applications of length-biased sampling to linkage mapping. *Am. J. Hum. Genet.* **61**, 430–438 (1997).
- 2 Knapp, M. Discriminating between true and false-positive peaks in a genomewide linkage scan, by use of the peak length. *Am. J. Hum. Genet.* **62**, 1561–1562 (1998).
- 3 Siegmund, D. Is peak height sufficient? *Genet. Epidemiol.* **20**, 403–408 (2001).
- 4 Visscher, P. & Haley, C. True and false positive peaks in genomewide scans: the long and the short of it. *Genet. Epidemiol.* **20**, 409–414 (2001).
- 5 Iyengar, S. K., Klein, B. E., Klein, R., Jun, G., Schick, J. H., Millard, C. *et al*. Identification of a major locus for age-related cortical cataract on chromosome 6p12–q12 in the Beaver Dam Eye Study. *Proc. Natl Acad. Sci. USA* **101**, 14485–14490 (2004).
- 6 Smith, C. A. B. The detection of linkage in human genetics. *J. R. Stat. Soc. B (Methodological)* **15**, 153–192 (1953).
- 7 Lander, E. S. & Botstein, D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567–1570 (1987).
- 8 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.* **77**, 63–72 (2014).
- 9 Zhang, L., Yang, W., Ying, D., Cherny, S. S., Hildebrandt, F., Sham, P. C. *et al*. Homozygosity mapping on a single patient: identification of homozygous regions of recent common ancestry by using population data. *Hum. Mutat.* **32**, 345–353 (2011).
- 10 Imai, A., Nakaya, A., Fahiminiya, S., Tetreault, M., Majewski, J., Sakata, Y. *et al*. Beyond homozygosity mapping: family-control analysis based on Hamming distance for prioritizing variants in exome sequencing. *Sci. Rep.* **5**, 12028 (2015).
- 11 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- 12 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- 13 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
- 14 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

- 15 Kohda, M., Tokuzawa, Y., Kishita, Y., Nyuzuki, H., Moriyama, Y., Mizuno, Y. *et al*. A comprehensive genomic analysis reveals the genetic landscape of mitochondrial respiratory chain complex deficiencies. *PLoS Genet.* **12**, e1005679 (2016).
- 16 Gusfield, D. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology* (Cambridge University Press, New York, NY, USA, 1999).
- 17 Sun, Y., Aljawad, O., Lei, J. & Liu, A. Genome-scale NCRNA homology search using a Hamming distance-based filtration strategy. *BMC Bioinformatics* **13**(suppl. 3): S12 (2012).
- 18 Nakaya, A., Imai, A. & Ott, J. HDR (Hamming Distance Ratio). <http://www.gi.med.osaka-u.ac.jp/software/hdr/>. Accessed 22 March 2016.
- 19 Tetreault, M., Choquet, K., Orcesi, S., Tonduti, D., Balottin, U., Teichmann, M. *et al*. Recessive mutations in POLR3B, encoding the second largest subunit of Pol III, cause a rare hypomyelinating leukodystrophy. *Am. J. Hum. Genet.* **89**, 652–655 (2011).
- 20 Samuels, M. E., Majewski, J., Alirezaie, N., Fernandez, I., Casals, F., Patey, N. *et al*. Exome sequencing identifies mutations in the gene TTC7A in French-Canadian cases with hereditary multiple intestinal atresia. *J. Med. Genet.* **50**, 324–329 (2013).
- 21 Fahiminiya, S., Al-Jallad, H., Majewski, J., Palomo, T., Moffatt, P., Roschger, P. *et al*. A polyadenylation site variant causes transcript-specific BMP1 deficiency and frequent fractures in children. *Hum. Mol. Genet.* **24**, 516–524 (2015).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/4.0/>

© The Author(s) 2016

Supplementary Information accompanies the paper on Journal of Human Genetics website (<http://www.nature.com/jhg>)