RESEARCH NOTE

Coverage analysis of lists of genes involved in heterogeneous genetic diseases following benchtop exome sequencing using the ion proton

CAROLINE LACOSTE[1,2], JEAN-PIERRE DESVIGNES[1], DAVID SALGADO[1], CHRISTOPHE PECHEUX[2], LAURENT VILLARD[1,2], MARC BARTOLI[1,2], CHRISTOPHE BEROUD[1,2], NICOLAS LEVY[1,2], CATHERINE BADENS[1,2] and MARTIN KRAHN[1,2]*

[1]*Aix Marseille Université, INSERM UMR_S 910, GMGF, 13385, Marseille, France*
[2]*APHM, Département de Génétique Médicale, Hôpital Timone Enfants, 13385, Marseille, France*

[Lacoste C., Desvignes J.-P., Salgado D., Pecheux C., Villard L., Bartoli M., Beroud C., Levy N., Badens C. and Krahn M. 2016 Coverage analysis of lists of genes involved in heterogeneous genetic diseases following benchtop exome sequencing using the ion proton. *J. Genet*. **95**, 203–208]

Introduction

Exome sequencing (ES) has been proven efficient for clinical applications in heterogeneous genetic diseases (Neveling *et al*. 2013; Biesecker and Green 2014), including ES associated with data filtering for selected genes (Dias *et al.* 2012; Bartoli *et al.* 2014). However, previous reports have been based mainly on the use of large-scale next-generation sequencing (NGS) platforms, difficult to implement in small to medium scale genetic diagnosis laboratories. Recent technological developments include the Ion Proton™ (Rothberg *et al.* 2011) (Life Technologies, Carlsbad, USA), allowing ES through an easy setup in 6 h runs on a benchtop apparatus. The Ion Proton™ has been the first commercialized platform specifically adapted to the environment and diagnostic setting of small to medium scale laboratories, in particular through the simplified workflow and considerably shortened sequencing run times, while the production of high quality exome data has been demonstrated, including comparable average depth and read length, and accuracy for SNP variant calls between the Ion Proton™ semiconductor technology and the Illumina technology (Boland *et al.* 2013; Motoike *et al.* 2014).

The rationale of our study was to specifically evaluate sequence coverage using Ion AmpliSeq™ exome enrichment and Ion Proton™ sequencing, for genes implicated in diverse genetically heterogeneous diseases, to determine which genes may be retained using this procedure as first-tier mutation screening.

We selected the Ion Proton™ platform to implement ES associated with data filtering for selected genes, to allow for standardization of mutation screening procedures for the analysis of diverse groups of heterogeneous diseases, through a common workflow, not feasible through the use of disease-group-specific gene-panel sequencing. While available sequence coverage data for different platforms mainly focus on average exome values, the selection of genes which can be analysed for diagnostic purposes through first-tier ES warrants targeted coverage analysis of genes of interest. Therefore, as a requirement for diagnostic mutation screening using Ion Proton™ ES, we evaluated sequence coverage data for six groups of genetically heterogeneous diseases, of interest for genetic diagnostic laboratories (table 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/): myopathies (82 genes), hereditary motor and sensory neuropathies (55 genes), early onset epileptic encephalopathies (30 genes), isolated and combined dystonia (12 genes), non-syndromic deafness and hereditary hearing loss (60 genes), and intellectual disability (107 genes for X-linked transmission; 39 genes for autosomal recessive transmission, 37 genes for autosomal dominant transmission). In addition, we evaluated sequence coverage for a list of 57 genes for which the American College of Medical Genetics and Genomics (ACMG) recommends to report incidental findings (Green *et al*. 2013).

Methods

We generated and analysed sequence coverage data of 45 exome runs (14 duplex runs following barcoding using Ion Xpress™ Barcode Adapters (Life Technologies) and 31 simplex runs) from DNA samples from our Biological Resources Center (Timone Hospital, Marseille). This study was conducted according to legal institutional and national regulations, and according to the Declaration of Helsinki

*For correspondence. E-mail: martin.krahn@univ-amu.fr.
Catherine Badens and Martin Krahn contributed equally to this work.

Keywords. exome; coverage; diagnosis; sequencing; mutation.

on ethical principles for medical research involving human subjects. All patients signed a consent form before the procedure. ES was performed using Ion AmpliSeq™ exome enrichment and library preparation (Life Technologies), template preparation using the Ion PI™ Template OT2-200-Kitv2 on the Ion OneTouch™ 2 System (Life Technologies), and sequencing using the Ion PI™ Chip-Kitv2 and Ion PI™ Sequencing200-Kitv2 on the Ion Proton™ Sequencer (Life Technologies) with sequencing data processing using the Torrent Suite™ Software (ver. 4.0.2) on a Torrent Server (Life Technologies) (figure 1). The Torrent Suite™ has been used to handle the bioinformatics data analysis from the acquisition through the generation of bam files. The suite chains several applications; we used the default parameters for all of them. The base calling step has been accomplished using the Ion Torrent's base-caller algorithm. Quality trimming has been performed using the per-base quality scores. The window size is set to 30 bases, and the threshold below which the trimming occurs is a quality score of 15. The alignment was performed using the TMAP aligner with map4 parameter, which is the default option to balance rapidity and maintain a high degree of sensitivity and specificity, on human reference genome 19.

The Ion AmpliSeq™ exome enrichment kit targets ~33 Mb of coding exons (97% of coding regions as described by consensus coding sequences (CCDS) annotation), corresponding to a total design coverage including padding and flanking regions (mean number of five flanking base pairs 5′ and 3′ of coding exons) of ~58 Mb. To comprehensively determine the coverage for the gene lists of interest, we selected for each gene in all lists the longest CCDS transcript isoforms (as recommended by the Human Genome Variation

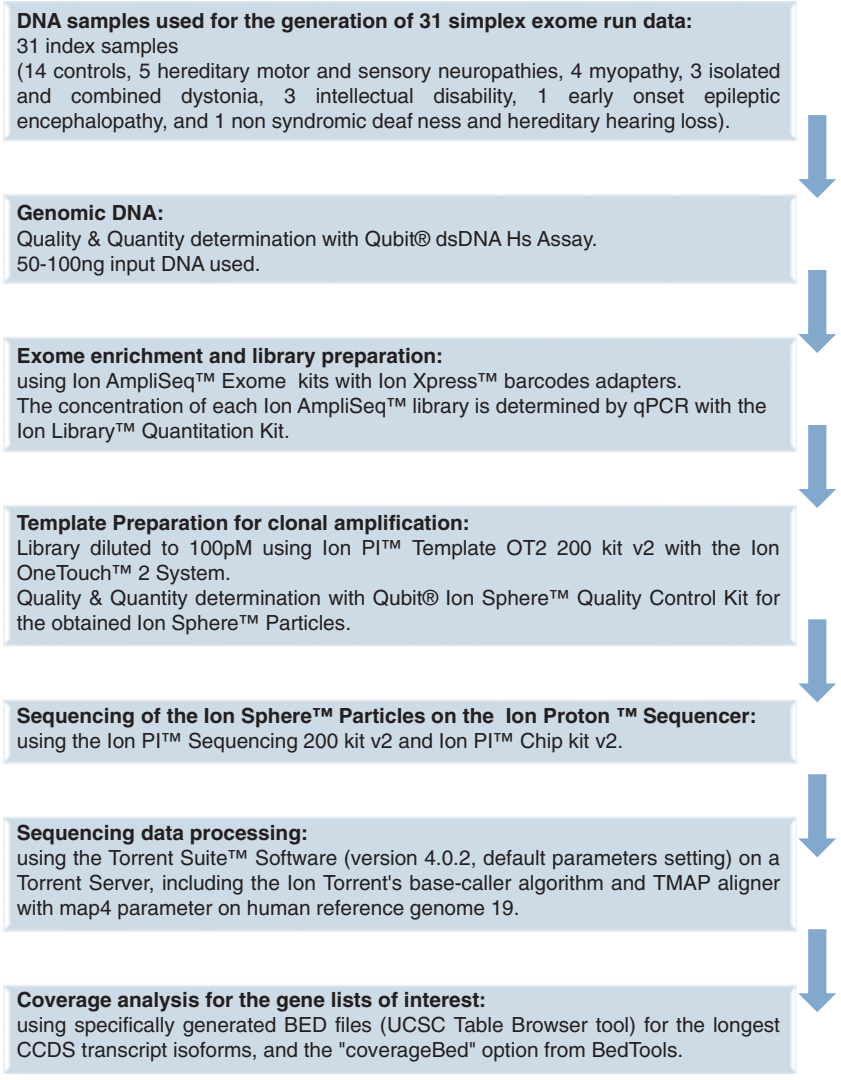

Figure 1. Workflow used to determine the mean sequence coverage at ≥20× for simplex exome runs analysed in the present study. All reagents are from Life Technologies and were used according to the manufacturer's recommendations.

Society guidelines (http://www.hgvs.org/mutnomen) for mutational analyses). We then used the UCSC Table Browser tool (http://genome.ucsc.edu/cgi-bin/hgTables) to retrieve a single BED file containing all coding exons. Effective coverage calculation based on obtained sequencing data from 31 simplex runs was based on the generated BED files specifically for the respective transcripts, using the 'coverageBed' option from BedTools (Quinlan and Hall 2010), to report the depth of each position for each sample processed. Finally, we implemented a perl script to group the coverage values at exon and transcript levels. This perl script summarized the per cent of base pairs covered at different depth levels (1×, 5×, 10×, 20× and 30×). In our study, genes were considered as sufficiently covered for diagnostic mutation screening at a ≥90% mean sequence coverage at ≥20×, with mean exome sequencing depths ≥100×, based on the guidelines from the American College of Medical Genetics and Genomics (Rehm *et al.* 2013).

Results

The sequence data generated for 45 exome runs yielded an average of 11.2 Gb per run (ranging from 6.6 to 15.8 Gb; standard deviation (SD) 2.38), including an average of 7.68 Gb of AQ20 (ranging from 4.3 to 12.5 Gb; SD: 2.08), corresponding to a mean of 70765587 reads (ranging from 40981018 to 90139803; SD: 12264334) with a mean length of 157.62 bases (range from 129 to 178; SD: 12.48). The exome sequence depth obtained for simplex runs ranged from 103× to 223×, with a mean value of 162× (SD: 36) for the 31 analysed simplex exome runs. This corresponds to mean sequence coverage of 96.26% at 1× (SD: 0.33), 92.64% at 10× (SD: 1.02) and 89.85% at 20× (SD: 1.98).

Using duplex runs, the exome sequence quality decreased to a mean value of 93× (ranging from 42× to 143×, SD: 25), corresponding to a mean sequence coverage of 95% at 1× (SD: 2.86), 88.82% at 10× (SD: 5.35) and 82.48% at 20× (SD: 7.08). Therefore, further detailed coverage analysis focussing on the different gene lists was performed for the 31 simplex runs. To identify genes suitable or not for diagnostic mutation screening, we determined mean sequence coverage at a depth of at least 20× (recommended for the detection of heterozygous constitutive germline mutations) as detailed in tables 1 and 2 in electronic supplementary material.

Among the different gene lists of interest, genes considered as sufficiently covered for diagnostic mutation screening (≥90% mean sequence coverage at ≥20×) distributed as: 58 (71.9%) among the 82 genes involved in myopathies; 45 (81.8%) among the 55 genes involved in hereditary motor and sensory neuropathies; 18 (60%) among the 30 genes involved in early onset epileptic encephalopathy; six (50%) among the 12 genes involved in isolated and combined dystonia; 45 (75%) among the 60 genes involved in nonsyndromic deafness and hereditary hearing; 77 (71.9%) among the 107 genes involved in X-linked intellectual disability; 29 (74.3%) among the 39 genes involved in autosomal recessive intellectual disability; 24 (64.8%) among the 37 genes involved in autosomal dominant intellectual disability; and 38 (66.6%) among the 57 genes for which the ACMG recommends to report incidental findings to patients. Among the different experiments, mean sequence coverage values are consistent for the respective lists of interest (figure 2).

Discussion

First-tier mutation screening using ES associated with data filtering offers promising perspectives for molecular diagnosis of genetically heterogeneous diseases. While most initial diagnostic applications of NGS relied on the use of disease-specific gene panel enrichment procedures (Rehm 2013), the feasibility of molecular diagnosis using ES has been demonstrated in the past years through the pioneering work of different laboratories having access to large-scale NGS platforms (Choi *et al*. 2009; Ng *et al*. 2010; Neveling *et al*. 2013; Yang *et al*. 2013; Biesecker and Green 2014). As compared with disease-specific gene panels, ES associated with data filtering for selected genes has the advantage to allow for regular updating of gene lists of interest, with regard to the always growing number of genes in several diseases groups, such as intellectual disability and neuromuscular disorders. Effectively, updating of disease-specific gene panel enrichment procedures through adding novel targets readily leads to disequilibrium in amplification and/or sequence capture performances, and therefore warrants customized revalidation of the respective gene panels at each update. On the contrary, for ES associated with data filtering for selected genes, updating of gene lists simply requires modifying the bioinformatics filtering settings. Moreover, ES offers the second-tier possibility of analysing the whole exome, including genes not previously implicated in the respective disease-groups of interest. In particular, in small to medium scale diagnostic laboratories implicated in the diagnosis of genetically heterogeneous diseases, but with limited numbers of proceeded samples for the respective diseases, the aforementioned flexibility constitutes in our opinion an important element of choice for using ES associated with data filtering for selected genes as an alternative to disease-specific gene panel enrichment procedures.

Importantly, as ES is being more widely used in genetic diagnosis, its implementation in small to medium scale diagnostic laboratories should not be limited by technical hurdles relating to large-scale platforms, to allow for large access and availability in prospective diagnosis. The simple workflow of the Ion Proton™ platform (Rothberg *et al.* 2011) is adapted to this purpose, with an easy setup and technical handling for ES in a benchtop format while producing high quality exome data (Boland *et al.* 2013).

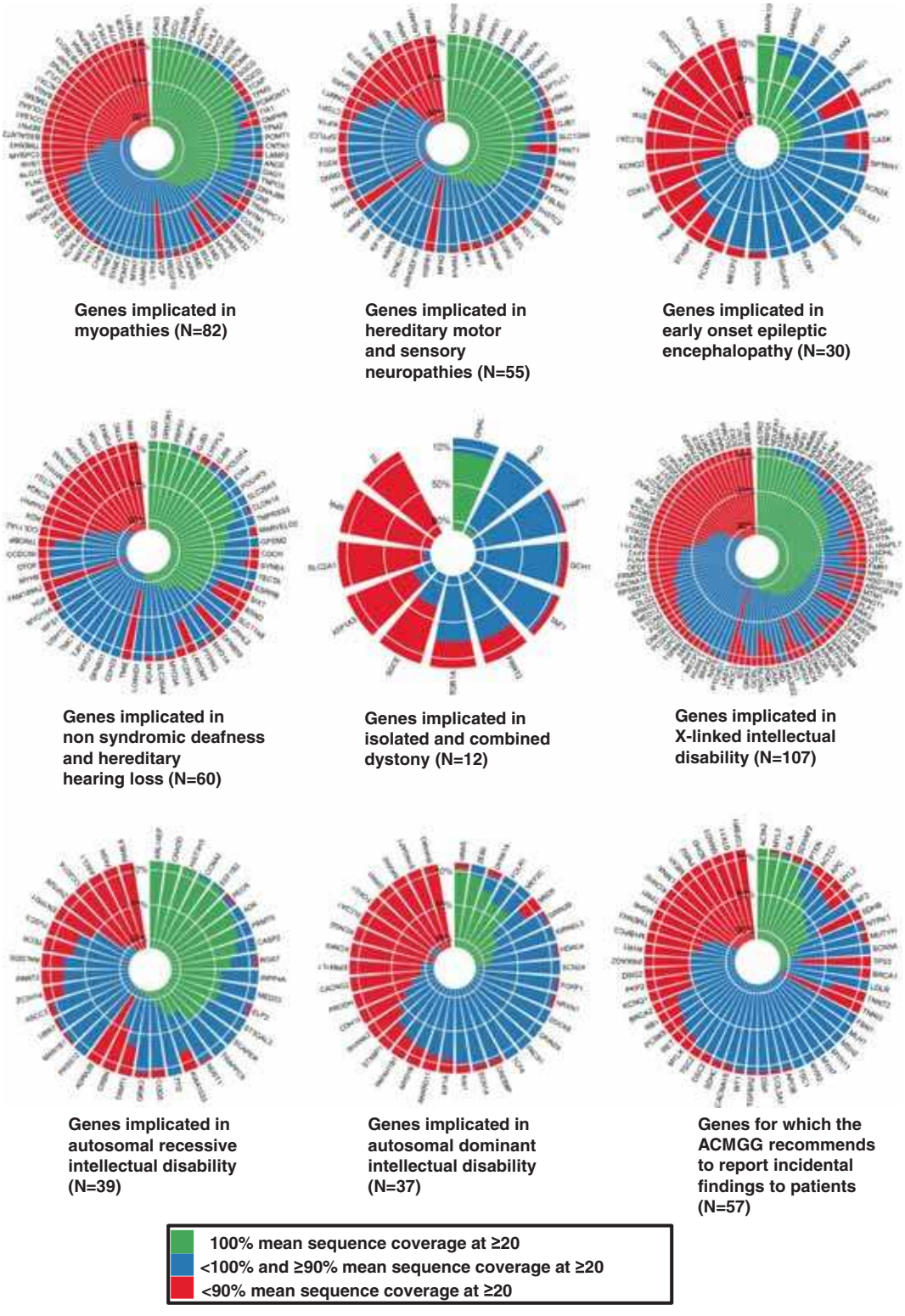

Figure 2. Sequence coverage at a depth of ≥20× among 31 simplex exome runs for the analysed gene lists. The polar histogram shows the distribution of sequence coverage at a depth of ≥20× among the 31 simplex exome runs analysed (representing 100% on the radial axis), for each gene of the respective gene lists of interest. Polar histograms were plotted using R software scripts and a specific library developed by C. Ladroue (http://chrisladroue.com/2012/02/polar-histogram-pretty-and-useful/).

Our results determine for the first time the specific genes implicated in diverse genetically heterogeneous diseases, for which the Ion Proton™ platform in its current configuration is suitable for first-tier mutation screening using ES associated with data filtering.

While mean sequence coverage values are consistent among the different experiments, variability was observed for individual genes of the respective lists of interest (figure 2), and has to be taken into account. In fact, it constitutes a criteria of optimization (including through automation), that may be used for the definition of thresholds such as total number of generated AQ20 reads before variant calling steps. Importantly, exception made for the list of 12 genes involved in isolated and combined dystonia, approximately two-third of all genes among the different groups are efficiently sequenced at >90% mean sequence coverage at $\geq 20\times$.

The sequence coverage achieved in the current configuration is adapted for efficient first-tier mutation screening using ES associated with data filtering for the majority of genes, and anticipated increase in diagnostic yield with regard to the large genetic heterogeneity of the diseases addressed here. More specifically, the information on the mean sequence coverage values determined in our study gene by gene among the analysed lists (table 2 in electronic supplementary material), allows for retaining or excluding selected genes for diagnostic applications using the first-tier Ion AmpliSeq™ exome enrichment and Ion Proton™ sequencing workflow, while mutational analysis of genes currently insufficiently covered using this procedure can rely on supplementation with second-tier Sanger sequencing for full coverage, as previously demonstrated (Dias *et al.* 2012). Genomic regions to be targeted with second-tier Sanger sequencing can be identified precisely for any sample of interest using the same coverage analysis workflow as used in our study for the determination of mean sequence coverage values. Importantly, to avoid biases relating to variability among exome sequencing runs, this should optimally rely on individual coverage determination on a sample by sample basis, to ensure full coverage, as exemplified for one selected sample in tables 3 and 4 in electronic supplementary material, respectively for the 82 genes involved in myopathies, and the 57 genes for which the ACMG recommends to report incidental findings to patients. In particular, supplementation with second-tier Sanger sequencing has to be considered for cases in which clinical data strongly orientate towards the implication of specific genes among the respective lists (i.e. suspected specific single gene disorders), or if comprehensive gene list screening is required.

For laboratories implementing the Ion Proton™ benchtop platform for first-tier mutation screening using ES associated with data filtering, the choice of providing either a mutation screening approach (with precisely determined associated limitations especially regarding coverage of the coding regions of interest), or a comprehensive gene list analysis, will further determine the diagnostic strategy and inclusion of second-tier sequencing supplementation, while the future increase in sequencing capacities of the platform should allow for correlated increase in sequence coverage.

Acknowledgements

We sincerely thank the patients for their contribution to this study. We thank Nicole Philip, Emmanuelle Salort-Campana, Jean Pouget, Shahram Attarian, Véronique Blanck-Labelle, Annachiara DeSandre-Giovannolli, Rafaëlle Bernard, Nathalie Bonello-Palot, Perrine Malzac, Patrice Bourgeois, Mathieu Milh, Gwenaëlle Collod-Béroud, Karine Bertaux, Cécile Mouradian and Valérie Delague for their contribution to this work. We also thank the Association Française contre les Myopathies, the Jain Foundation, Assistance Publique Hôpitaux de Marseille, Inserm, Aix-Marseille Université, ARS-PACA and Agence de la Biomédecine for supporting this work. The researchers implicated in this project also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 2012-305121 'Integrated European –omics research project for diagnosis and therapy in rare neuromuscular and neurodegenerative diseases (NEUROMICS).

References

Bartoli M., Desvignes J. P., Levy N. and Krahn M. 2014 Exome sequencing as a second-tier diagnostic approach for clinically suspected dysferlinopathy patients. *Muscle Nerve* **50**, 1007–1010.

Biesecker L. G. and Green R. C. 2014 Diagnostic clinical genome and exome sequencing. *N. Engl. J. Med.* **371**, 1169–1170.

Boland J. F., Chung C. C., Roberson D., Mitchell J., Zhang X., Im K. M. *et al.* 2013 The new sequencer on the block: comparison of life technology's proton sequencer to an illumina hiseq for whole-exome sequencing. *Hum. Genet.* **132**, 1153–1163.

Choi M., Scholl U. I., Ji W., Liu T., Tikhonova I. R., Zumbo P. *et al.* 2009 Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. USA* **106**, 19096–19101.

Dias C., Sincan M., Cherukuri P. F., Rupps R., Huang Y., Briemberg H. *et al.* 2012 An analysis of exome sequencing for diagnostic testing of the genes associated with muscle disease and spastic paraplegia. *Hum. Mutat.* **33**, 614–626.

Green R. C., Berg J. S., Grody W. W., Kalia S. S., Korf B. R., Martin C. L. *et al.* 2013 ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15**, 565–574.

Motoike I. N., Matsumoto M., Danjoh I., Katsuoka F., Kojima K., Nariai N. *et al.* 2014 Validation of multiple single nucleotide variation calls by additional exome analysis with a semiconductor sequencer to supplement data of whole-genome sequencing of a human population. *BMC Genomics* **15**, 673–686.

Neveling K., Feenstra I., Gilissen C., Hoefsloot L. H., Kamsteeg E. J., Mensenkamp A. R. *et al.* 2013 A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. *Hum. Mutat.* **34**, 1721–1726.

Ng S. B., Buckingham K. J., Lee C., Bigham A. W., Tabor H. K., Dent K. M. *et al.* 2010 Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35.

Quinlan A. R. and Hall I. M. 2010 BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842.

Rehm H. L. 2013 Disease-targeted sequencing: a cornerstone in the clinic. *Nat. Rev. Genet.* **14**, 295–300.

Rehm H. L., Bale S. J., Bayrak-Toydemir P., Berg J. S., Brown K. K., Deignan J. L. *et al.* 2013 ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **15**, 733–747.

Rothberg J. M., Hinz W., Rearick T. M., Schultz J., Mileski W., Davey M. *et al.* 2011 An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352.

Yang Y., Muzny D. M., Reid J. G., Bainbridge M. N., Willis A., Ward P. A. *et al.* 2013 Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* **369**, 1502–1511.

Received 15 April 2015, in revised form 2 July 2015; accepted 11 August 2015
Unedited version published online: 14 August 2015
Final version published online: 24 February 2016