Short Report

Title: Evaluating the genetic diagnostic power of exome sequencing: Identifying missing data.

Running title: Identifying missing data in exome sequencing

Authors: Pascual Lorente-Arencibia BEng, Deyán Guacarán BEng, Antonio Tugores PhD

Affiliation: Unidad de Investigación, Complejo Hospitalario Universitario Insular Materno Infantil, Las Palmas de Gran Canaria, Spain.

Corresponding author:

Antonio Tugores

Unidad de Investigación

Complejo Hospitalario Universitario Insular Materno Infantil

Avda Maritima del Sur, s/n.

35016 Las Palmas de Gran Canaria, SPAIN;

Tel: +34 928 441 531

FAX: +34 928 441 535

email: atugores@yahoo.com

COMPETING INTERESTS

The authors declare no conflicts of interest.
ABSTRACT

A hurdle of exome sequencing is its limited capacity to represent the entire exome. To ascertain the diagnostic power of this approach we determined the extent of coverage per individual sample. Using alignment data (BAM files) from 15 exome samples, sequences of any length that were below a determined sequencing depth coverage (DP) were detected and annotated with the Ensembl exon database using MIST, a novel software tool. Samples sequenced at 50X mean coverage had, on average, up to 50% of the Ensembl annotated exons with at least one nucleotide (L=1) with a DP<20, improving to 35% at 100X mean coverage. In addition, almost 15% of annotated exons were never sequenced (L=50, DP<1) at 50x mean coverage, reaching down to 5% at 100x. The diagnostic utility of this approach was tested for hypertrophic cardiomyopathy, a genetically heterogeneous disease, where exome sequencing covered as much as 80% of all candidate genes exons at DP≥20. This report stresses the value of identifying, precisely, which sequences are below a specific depth in an individual`s exome, and provides a useful tool to assess the potential and pitfalls of exome sequencing in a diagnostic or gene discovery setting.

KEYWORDS

exome / diagnostics / evaluation / coverage / sequence
INTRODUCTION

Genetic screening for specific variants may predict the susceptibility to suffer disease or the capacity to respond adequately to specific medications, providing a framework to make informed decisions about a clinical strategy. Exome sequencing has allowed to search for disease-causing genetic variants in a hypothesis-free fashion on the most understood fraction of the genome (1-2%), where the biological meaning of a variant might be better inferred. This strategy has been able to identify disease-associated variants in, approximately, 25% of the probands.

Both conceptual and technical reasons can be argued to explain this incomplete success rate. First, exome sequencing is a gene-centered approach, so regulatory regions are missed. Secondly, it fails to identify genomic rearrangements, as well as large insertions and deletions (indel). Additionally, even in the best-case scenario, when a variant is a single nucleotide polymorphism or variant or a small indel in an exon, there is still room for failure as (i) not all the exons are represented in the capture kits and (ii) there are sequences that are not sequenced at sufficient depth or at all.

While the overall performance of this analysis has been evaluated at large, a diagnostic approach requires its evaluation on an individual basis, identifying not only the variants, but also the number of protein coding exons that have been missed or sequenced below acceptable quality.

METHODS

Subjects and sample sequencing

Fifteen individuals gave written informed consent to participate in three studies of uncharacterized Mendelian disorders by using whole exome sequencing, approved by the Clinical Research Ethics Committee of our Center. DNA was purified from venous blood (QIAGEN), exon-captured with the SureSelect 50Mb kit (Agilent Technologies), and sequenced with a HiSeq2000 (Illumina) as described. Five samples were sequenced at an average depth of 50 fold (50X) coverage, and the remaining ten samples were re-sequenced to attain a 100 fold (100X) coverage.
Bioinformatic analysis

Sequence FASTQ files (length = 90 nucleotides) were processed following the Genome Analysis Tool Kit (GATK) best-practices\textsuperscript{11}, and aligned to the GRCh37.p13 human reference genome using Burrows-Wheeler Aligner \textsuperscript{12} with default parameters for paired-end sequences to produce raw SAM files. Sequences which partially fall out of the reference genome were soft-clipped, mapping quality values (MAPQ) were set to 0 in unmapped sequences, duplicated PCR fragments were removed, alignments were sorted for each resulting BAM file using Picard Tools \textsuperscript{13}, indel realignment was performed and recalibrated using GATK \textsuperscript{11}.

MIST (MIssing Sequences detection Tool) is a Java-based software that uses two parameters: threshold, which determines the depth (DP) or number of reads to consider for any nucleotide, and length (L), which determines the minimum number of consecutive nucleotides under a given DP. The application defines the intervals that have a DP below the given threshold and intersects them with the Ensembl exon database \textsuperscript{15} including sequences padding the exons (+10 nucleotides) and generates an output file where every line contains a match between an Ensembl exon and a poorly sequenced region (Table 1). MIST source code may be downloaded at: https://github.com/pasculorente/Mist.

RESULTS

Sequence coverage of exome sequencing

Five samples were sequenced at an average depth of 50X, and ten at 100X, obtaining 57 to 64 million reads per sample-lane (50X). Sequences present in autosomal or sex chromosomes, and annotated as protein coding exons, containing at least one nucleotide (L=1) below DP=20, were referred as "incomplete" and were identified by their unique Ensembl ID. Sequencing at 50X rendered, on average, over 192,000 incomplete protein-coding exons, roughly representing 50% of the Ensembl-annotated sequences, improving significantly at 100X, where an average of 35% of the Ensembl
protein-coding exons were below the threshold (Supplementary Table 1; Figure 1A).

The number of sequences that are not represented in the exome capture kits varies depending on the kit, and on the databases used as reference \textsuperscript{6-9}. These sequences should appear as missing in each sample, along with those that are sequenced with poor quality or not at all. A threshold of DP=1 and L= 50 or higher, and the intersection for each group (50X and 100X) identified exons that were consistently missed among all samples from a given group. Almost 15\% of the protein-coding exons (56656/384845) were never sequenced in the 50x coverage group while only 5\% of the protein coding exons (17838/384845) were absent from the 100X group, showing a lower GC content than average (Figure 1B).

**Exome sequencing in genetic diagnosis**

The diagnostic limits of exome sequencing were tested for hypertrophic cardiomyopathy (CMH; MIM: 192600), characterized by great genetic heterogeneity: 22 candidate genes, representing 999 exons present in 96 transcripts \textsuperscript{16}. Out of the list of known candidate genes, exome sequencing was able to detect most exons, revealing that there were, on average, only 169 (17\%) and 121 exons (12\%) with at least one nucleotide below DP=10, and 254 (25\%) and 171 exons (17\%) below DP=20 (Table 2) in the 50x and 100x sample groups respectively.

Furthermore, analysis of the protein-coding exons contained the 27 Megabase region linked to CMH21 (MIM 614676; 115 genes, 1864 exons) \textsuperscript{17}, the analysis revealed that 57\% and 75\% of all annotated exons were fully sequenced at a DP\geq10 in N019 (50X) and S003 (100X) respectively, decreased accordingly to 43\% and 54\% at DP\geq20.

**DISCUSSION**

In a diagnostic sequencing analysis it is important to know the extent of representation of the genes of interest for a particular disease, regardless of the reason why these genes are absent. Exome
capture and sequencing are both physico-chemical processes subjected to varied efficiency, so knowing the un-captured moiety is insufficient to assess missing sequences or below a necessary quality. We found that, as the number of reads increased from 50X to 100X, the number of incomplete sequences decreased. However, a higher number of reads in the 100x sample group did not result in an significantly higher number of fully sequenced exons above the threshold, suggesting that the system could be close to saturation. Moreover, the numbers of commonly shared incomplete or missed protein coding exons were quite close: 118,943 at 50X vs 114,287 at 100X, indicating that missing sequences due to lack of capture and sequencing errors were not significantly rescued by increasing the average coverage at this depth even though it increased the overall quality of the reads. It remains to be determined whether higher average coverage using more sequencing lanes would result in increased detection. We also observed that missing sequences were not due to amplification bias against sequences with high G+C content 18.

Exome sequencing has slowly moved into the clinical diagnostic laboratory: Diseases with great genetic heterogeneity with many candidate genes or when not all the possible candidate genes have been identified may benefit from this approach, and has been recently recommended to be part of an integral genetic diagnostic workflow 3,4,19. Nonetheless, in addition to the challenges of interpretation of the variants that are found, it is also important to identify the relevant sequences that are incomplete or missing for a specific patient, so that the limitations of the test can be fully understood. In this scenario, we ascertained the degree of coverage for a group of genes associated with CMH (MIM: 192600), a genetically heterogeneous disease, finding that exome sequencing was able to detect up to 83% and 88% of the exons at DP>10 and, and as much as 75% and 83% at DP>20 at 50x and 100x average coverage respectively. This example not only highlights the significant improvement resulting from increasing the average coverage from 50x to 100x but, also, identifies exactly the regions where directed Sanger sequencing is needed.
ACKNOWLEDGEMENTS

We thank the patients for their participation. Drs. Sarah Murray and Gary Hardiman provided useful comments on the manuscript. Dr Francisca Quintana’s contribution has been pivotal to create and enable the student exchange program between the Hospital and the School of Informatics and Computer Engineering from the University of Las Palmas de Gran Canaria. Work supported by the Fundación Canaria de Investigación en Salud (FUNCIS), and by the Servicio Canario de Salud.

REFERENCES

1. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. *J Hum Genet* 2014;59:5-15.

2. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *N Engl J Med* 2013;369:1502-1511.

3. Lee H, Deignan JL, Dorrani N, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* 2014;312:1880-1887

4. Yang Y, Muzny DM, Xia F, et al. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* 2014;312:1870-1879.

5. Cooper DN, Chen JM, Ball EV, et al. Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. *Hum Mutat* 2010;31:631-655.

6. Asan, Xu Y, Jiang H, Tyler-Smith C, et al. Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol.* 2011;12:R95.

7. Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR: A comparative analysis of exome capture. *Genome Biol* 2011:12:R97

8. Sulonen AM, Ellonen P, Almusa H, et al. Comparison of solution-based exome capture methods for
next generation sequencing. *Genome Biol* 2011;12:R94

9. Chilamakuri CS, Lorenz S, Madoui MA, et al. Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics* 2014;15:449

10. Meynert AM, Ansari M, FitzPatrick DR, Taylor MS. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics*. 2014;15:247-

11. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;11:11.10.1-11.10.33

12. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;26:589-595

13. Picard Tools, 2014. Available at: http://broadinstitute.github.io/picard/. Accessed December 4, 2015.

14. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078-2079

15. Flicek P, Amode MR, Barrell D, et al. Ensembl 2014. *Nucleic Acids Res* 2014;42:D749-755.

16. McKusick VA. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet*. 2007;80:588-604

17. Song L, DePalma SR, Kharlap M, et al. Novel locus for an inherited cardiomyopathy maps to chromosome 7. *Circulation* 2006;113:2186-2192.

18. Veal CD, Freeman PJ, Jacobs K, et al. A mechanistic basis for amplification differences between samples and between genome regions. *BMC Genomics*. 2012;13:455

19. Xue Y, Ankala A, Wilcox WR, Hegde MR. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. *Genet Med*. 2015;17:444-451
### Table 1. Example of MIST output.

Results shown are from a limited number of lines from a single individual. Every region below a user-defined DP is annotated with the following fields: **chrom** (chromosome), **start exon** (the start position of the exon), **end exon** (the end position of the exon), **start poor** (the start position of the poor region), **end poor** (the end position of the poor region), **gene_id** (the Ensembl gene identifier), **gene name** (the Ensembl gene name), **exon id** (the Ensembl exon identifier), **transcript name** (the Ensembl transcript name), **biotype** (the biotype of the exon), **match** (whether the poor region is left, right, inside or overlapping the exon).

| chrom | start exon | end exon | start poor | end poor | gene_id       | gene name | exon_id       | transcript name | biotype        | match       |
|-------|------------|----------|------------|----------|---------------|-----------|---------------|-----------------|---------------|-------------|
| 2     | 3383446    | 3383610  | 3383436    | 3383620  | ENSG00000171853 | TRAPPC12  | ENSE00001490954 | TRAPPC12-001   | protein_coding | overlap     |
| 2     | 3383457    | 3383647  | 3383447    | 3383657  | ENSG00000171853 | TRAPPC12  | ENSE00001919066 | TRAPPC12-002   | protein_coding | overlap     |
| 2     | 3488491    | 3488865  | 3488848    | 3488848  | ENSG00000171853 | TRAPPC12  | ENSE00001754716 | TRAPPC12-023   | protein_coding | inside      |
| 2     | 3488491    | 3488865  | 3488852    | 348875   | ENSG00000171853 | TRAPPC12  | ENSE00001754716 | TRAPPC12-023   | protein_coding | right       |
| 2     | 3500727    | 3501199  | 3500717    | 3501209  | ENSG00000271868 | RP11-1293J14.1 | ENSE00003699826 | RP11-1293J14.1-001 | lincRNA   | overlap     |
| 2     | 3523139    | 3523507  | 3523422    | 3523517  | ENSG00000182551 | ADI1      | ENSE00001328051 | ADI1-001       | protein_coding | right       |
| 2     | 3501132    | 3502853  | 3501122    | 3502537  | ENSG00000182551 | ADI1      | ENSE00001307713 | ADI1-001       | protein_coding | left        |
| 2     | 3501132    | 3502853  | 3502544    | 3502546  | ENSG00000182551 | ADI1      | ENSE00001307713 | ADI1-001       | protein_coding | inside      |
| 2     | 3518579    | 3521518  | 3518703    | 3521528  | ENSG00000182551 | ADI1      | ENSE00001838794 | ADI1-002       | protein_coding | right       |
| 2     | 3502224    | 3502853  | 3502214    | 3502537  | ENSG00000182551 | ADI1      | ENSE00001490904 | ADI1-002       | protein_coding | left        |
| 2     | 3502224    | 3502853  | 3502544    | 3502546  | ENSG00000182551 | ADI1      | ENSE00001490904 | ADI1-002       | protein_coding | inside      |
| 2     | 3502432    | 3502853  | 3502422    | 3502537  | ENSG00000182551 | ADI1      | ENSE00001799045 | ADI1-003       | protein_coding | left        |
| 2     | 3502432    | 3502853  | 3502544    | 3502546  | ENSG00000182551 | ADI1      | ENSE00001799045 | ADI1-003       | protein_coding | inside      |
| Disease | MIM   | Gene  | Ex.N | N019 10 | N019 20 | N032 10 | N032 20 | N060 10 | N060 20 | W002 10 | W002 20 | S001 10 | S001 20 | S020 10 | S020 20 |
|---------|-------|-------|------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| CMH1    | 192600| CAV3  | 8    | 3       | 4       | 2       | 4       | 3       | 3       | 2       | 2       | 2       | 2       | 2       | 2       |
| CMH1    | 192600| MYH7  | 40   | 3       | 9       | 4       | 14      | 4       | 10      | 6       | 12      | 6       | 12      | 5       | 10      |
| CMH1    | 192600| MYLK2 | 20   | 5       | 10      | 6       | 9       | 5       | 7       | 2       | 3       | 2       | 3       | 1       | 2       |
| CMH2    | 115195| TNNT2 | 81   | 16      | 19      | 15      | 17      | 13      | 17      | 5       | 8       | 5       | 14      | 4       | 8       |
| CMH3    | 115196| TPM1  | 85   | 28      | 32      | 29      | 34      | 28      | 28      | 12      | 16      | 10      | 11      | 9       | 10      |
| CMH4    | 115197| MYBPC3| 39   | 16      | 23      | 19      | 26      | 14      | 23      | 3       | 6       | 3       | 6       | 3       | 3       |
| CMH6    | 600858| PRKAG2| 64   | 15      | 17      | 14      | 17      | 15      | 16      | 9       | 9       | 9       | 9       | 9       | 10      |
| CMH7    | 613690| TNNT3 | 31   | 8       | 11      | 8       | 11      | 8       | 11      | 1       | 4       | 2       | 3       | 0       | 3       |
| CMH8    | 608751| MYL3  | 13   | 5       | 6       | 6       | 7       | 6       | 10      | 4       | 5       | 4       | 5       | 4       | 5       |
| CMH9    | 188840| TTN   | 387  | 17      | 46      | 13      | 38      | 18      | 45      | 33      | 53      | 33      | 48      | 33      | 42      |
| CMH10   | 160781| MYL2  | 14   | 2       | 3       | 2       | 4       | 3       | 3       | 1       | 1       | 1       | 2       | 1       | 1       |
| CMH11   | 612098| ACTC1 | 13   | 2       | 3       | 2       | 3       | 2       | 3       | 2       | 5       | 2       | 4       | 2       | 4       |
| CMH12   | 612124| CSRP3 | 8    | 4       | 6       | 4       | 5       | 4       | 5       | 4       | 4       | 4       | 4       | 4       | 4       |
| CMH13   | 613243| TNNC1 | 11   | 3       | 5       | 3       | 6       | 3       | 5       | 1       | 3       | 2       | 2       | 1       | 2       |
| CMH14   | 613251| MYH6  | 54   | 4       | 11      | 5       | 13      | 4       | 9       | 7       | 12      | 8       | 12      | 6       | 10      |
| CMH15   | 613255| VCL   | 37   | 8       | 12      | 7       | 14      | 9       | 14      | 5       | 9       | 4       | 8       | 5       | 7       |
| CMH16   | 613838| MYOZ2 | 6    | 2       | 3       | 2       | 2       | 2       | 2       | 2       | 2       | 2       | 2       | 2       | 2       |
| CMH17   | 613873| JPH2  | 8    | 7       | 7       | 7       | 7       | 7       | 6       | 7       | 4       | 6       | 5       | 6       | 6       |
| CMH18   | 613874| PLN   | 2    | 2       | 2       | 2       | 2       | 2       | 2       | 2       | 2       | 2       | 2       | 2       | 2       |
| CMH19   | 613875| CALR3 | 14   | 2       | 3       | 3       | 3       | 2       | 3       | 1       | 1       | 0       | 1       | 0       | 1       |
| CMH20   | 613876| NEXN  | 36   | 9       | 10      | 7       | 11      | 7       | 11      | 9       | 9       | 9       | 9       | 9       | 9       |
| CMH22   | 615248| MYPN  | 28   | 8       | 12      | 9       | 12      | 9       | 15      | 8       | 11      | 8       | 11      | 8       | 10      |
| Ensembl total | 22 | 999 | 169 | 254 | 169 | 259 | 168 | 249 | 125 | 184 | 122 | 176 | 115 | 153 |

Table 2. Missing sequences per sample in genes known to be associated with hypertrophic cardiomyopathy. MIST software was run on each individual sample to detect sequences of at least 1 nucleotide in length with thresholds of DP<10 and DP<20. Data was filtered, for each individual sample, to include only unique exons annotated in Ensembl as part of genes that have been associated with hypertrophic cardiomyopathy according to the OMIM database at the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/). Numbers shown for individual samples define exons that have at least one nucleotide with a DP<20.
| Sample | Mean coverage | Reads   | Protein coding | lincRNA | miRNA |
|--------|---------------|---------|----------------|---------|-------|
| N019   | 50x           | 60,589,076 | 193,982        | 27,075 | 2,627 |
| N032   | 50x           | 54,819,342 | 194,511        | 27,093 | 2,654 |
| N060   | 50x           | 56,875,202 | 192,874        | 27,107 | 2,629 |
| N084   | 50x           | 54,881,968 | 195,236        | 27,123 | 2,637 |
| N085   | 50x           | 58,548,886 | 187,687        | 27,056 | 2,540 |
| Average| 50x           | 57,142,895 | 192,858        | 27,091 | 2,617 |
| Common | 50x           | ND      | 118,943        | 26,614 | 1,997 |
| N062   | 100x          | 137,441,026 | 141,303        | 27,129 | 2,310 |
| S001   | 100x          | 109,748,976 | 137,672        | 27,108 | 2,245 |
| S002   | 100x          | 121,246,716 | 134,857        | 27,106 | 2,276 |
| S003   | 100x          | 150,662,618 | 127,732        | 27,033 | 2,212 |
| S020   | 100x          | 161,975,724 | 129,657        | 27,058 | 2,240 |
| S025   | 100x          | 122,775,926 | 142,275        | 27,151 | 2,265 |
| S030   | 100x          | 118,397,668 | 132,031        | 27,103 | 2,221 |
| W001   | 100x          | 113,593,590 | 140,121        | 27,119 | 2,306 |
| W002   | 100x          | 137,984,490 | 142,015        | 27,128 | 2,307 |
| W003   | 100x          | 118,866,548 | 136,375        | 27,115 | 2,262 |
| Average| 100x          | 128,361,284 | 136,404        | 27,105 | 2,264 |
| Common | 100x          | ND      | 114,287        | 26,955 | 2,053 |
| Ensembl| -             | NA     | 384,845        | 31,466 | 3,424 |

**Supplementary Table 1. Sequence reads per sample.** Genomic DNA isolated from 15 individuals were subjected to exon selection with the Agilent SureSelect 50Mb capture kit and sequencing with an Illumina HiSeq2000. Samples were resolved in one (50x average coverage), or 2 lanes (100x average coverage) as indicated. Final number of sequences is shown for each sample, and the average number per group, after removal of low quality sequences and PCR duplicates, as described in the Methods section. MIST software was run on each individual sample with a threshold of DP≤20 read and a length 1 nucleotide or more. Data was filtered, for each individual sample, to include only sequences present in autosomes or sex chromosomes, excluding mitochondrial and other sequences not assigned to specific chromosomes, according to the annotation of protein coding exons, lincRNAs, and miRNAs in Ensembl. Only unique exons are counted, according to their Ensembl exon ID.
Figure 1. Evaluation of missing sequences in exome sequencing projects. (A) Representation of the number of reads versus the number of correctly sequenced protein coding exons. The Y-axis represents the number of complete exons while X-axis represents the number of reads. The 50X and 100X clusters are boxed. The top grey line represents the total number of unique protein coding exons annotated by Ensembl. (B) "Lost" sequences and G+C content per coverage group. Protein coding exons that had, at least, a 50 nucleotide stretch that was never read (DP<1), were
considered as completely lost or missed by the analysis. The graphic shows the total number of "lost" protein coding exons, as determined by the intersection of all samples within the 50x (dark grey shading) and 100x (black shading) groups, in comparison with the total number of unique protein-coding exons annotated by Ensembl (light grey shading). The GC content is shown, as percentage, on the Y axis.