HHS Public Access

Author manuscript

Clin Chem. Author manuscript; available in PMC 2017 April 01.

Published in final edited form as:

Clin Chem. 2016 April ; 62(4): 647–654. doi:10.1373/clinchem.2015.249623.

Systematic Evaluation of Sanger Validation of NextGen Sequencing Variants

Tyler F. Beck[1], **James C. Mullikin**[1,2] **on behalf of the NISC Comparative Sequencing Program**, and **Leslie G. Biesecker**[1]

[1]National Human Genome Research Institute, NIH, Bethesda, MD

[2]NIH Intramural Sequencing Center, Rockville, MD

Abstract

BACKGROUND—Next-generation sequencing (NGS) data are used for both clinical care and clinical research. DNA sequence variants identified using NGS are often returned to patients/participants as part of clinical or research protocols. The current standard of care is to validate NGS variants using Sanger sequencing, which is costly and time-consuming.

METHODS—We performed a large-scale, systematic evaluation of Sanger-based validation of NGS variants using data from the ClinSeq® project. We first used NGS data from 19 genes in five participants, comparing them to high-throughput Sanger sequencing results on the same samples, and found no discrepancies among 234 NGS variants. We then compared NGS variants in five genes from 684 participants against data from Sanger sequencing.

RESULTS—Of over 5,800 NGS-derived variants, 19 were not validated by Sanger data. Using newly-designed sequencing primers, Sanger sequencing confirmed 17 of the NGS variants, and the remaining two variants had low quality scores from exome sequencing. Overall, we measured a validation rate of 99.965% for NGS variants using Sanger sequencing, which was higher than many existing medical tests that do not necessitate orthogonal validation.

CONCLUSIONS—A single round of Sanger sequencing is more likely to incorrectly refute a true positive variant from NGS than to correctly identify a false positive variant from NGS. Validation of NGS-derived variants using Sanger sequencing has limited utility, and best practice standards should not include routine orthogonal Sanger validation of NGS variants.

Keywords

exome sequencing; genome sequencing; high-throughput DNA sequencing; molecular diagnostic testing

Corresponding Author: Leslie G. Biesecker, Address: 49 Convent Dr., 4A56, Bethesda, MD 20830, Phone: 301-402-2041, Fax: 301-402-2170, ; Email: lesb@mail.nih.gov

INTRODUCTION

Massively parallel sequencing technologies have revolutionized medical genetics. More than 200,000 genomes and even more exomes have been sequenced to date (1). However, it is still widely accepted that variants found using next-generation sequencing (NGS) should be validated with the current "gold standard" for DNA sequencing, Sanger dideoxy terminator sequencing (2), before returning or publishing results. There have been several reports suggesting that NGS data used in clinical and research settings are at least as accurate—or in some cases more accurate—than Sanger sequencing (3–6). However, several of these studies used small sample sets (168, 37, and 110 variants sequenced, respectively) of secondary data from applied clinical research. Another recent larger-scale study included two separate comparisons of NGS variants with Sanger sequencing results (443 variants and 762 variants, respectively), but this study was performed using target-capture gene panels, which are not representative of the overall genomic landscape, and the authors did not specify if any variants were included in both comparisons (6).

In the study we report here, we set out to determine the utility of Sanger validation using a subset of data from 684 exomes and 2,793,321 Sanger sequencing reads from the ClinSeq® cohort (7). The ClinSeq® project was initiated in 2006, well before NGS was widely available, and began with semiautomated high-throughput Sanger sequencing. By the time that millions of Sanger reads had been generated for the ClinSeq® project, NGS displaced Sanger sequencing as a less expensive, higher throughput tool, which was then applied to the same samples that had already been Sanger sequenced, thus making this dataset ideal for evaluating the utility of orthogonal Sanger validation of NGS variants.

MATERIALS AND METHODS

DNA Isolation

DNA was isolated from whole blood using the salting-out method (Qiagen, Valencia, CA, USA) followed by phenol-chloroform extraction using a Manual Phase Lock Gel extraction kit (5Prime, Gaithersburg, MD, USA) and rehydration with DNA Hydration Solution (Qiagen).

Next-Generation Sequencing

Solution-hybridization exome capture was performed with the SureSelect All Exon System, the SureSelect ICGC System (Agilent Technologies, Santa Clara, CA, USA) or the TruSeq system, V1 or V2 (Ilumina, San Diego, CA, USA). Flow-cell preparation and paired-end read sequencing were performed with either the GAIIx or HiSeq 2000 sequencer (Illumina) as previously described (8). Image analyses and base calling were performed as described (9). Reads were aligned to hg19 (NCBI build 37) using NovoAlign (Novocraft Technologies, Selangor, Malaysia). For exome sequencing, samples were sequenced to sufficient coverage such that 85% of all targeted bases were called with a minimum Most Probable Genotype (MPG) score of at least 10 (10). The MPG genotype caller uses a Bayesian model that calculates the posterior probability of all possible genotypes at a position and reports the most likely genotype with a corresponding score calculated as the natural log of the quotient

of the relative probability of the most likely genotype to the next most likely genotype (10). This means that an MPG score of 10 estimates the probability that the next most likely genotype is correct at e^{-10} or 4.54×10^{-5}. The MPG score is dependent upon both the high-quality sequencing read depth and the zygosity at that base, and is correlated linearly to the overall sequencing read depth (Supplemental Figure 1). Where a base was covered by more than 200 reads, MPG was applied to a random subset of 200 reads. We evaluated a given variant position in 684 exomes if there was at least one variant "sentinel" call at that position with an MPG score of ≥10. If at least one sample met that threshold, then all calls at that position were considered, irrespective of their MPG score, so long as there were at minimum 10 reads covering that position. Structural variants were excluded from analysis.

Sanger Sequencing

Candidate genes for Sanger sequencing were selected based on evidence for association with development of coronary artery calcification and/or atherosclerosis (7). That list was expanded to include genes associated with heart disease identified through the use of mouse models, gene family analyses, and pathway analyses, among others. This resulted in a list of 308 genes sequenced using 16,371 pairs of sequencing primers (Supplemental Table 1). PCR and sequencing primers were generated using PrimerTile, an automated primer design program which utilized the most recent version of the dbSNP database (version 130) to omit common variants from designed primers (11). Amplicons were sequenced as described (7) with a mean amplicon length of 648.8 bp. A subset of five genes found in the Genetic Testing Registry (http://www.ncbi.nlm.nih.gov/gtr/) were chosen for analysis in 684 ClinSeq® samples that had been exome sequenced. These genes were chosen to be representative of the genome based on criteria that included coding DNA sequence (CDS) length, number of exons (minimum 4, maximum 20), GC content, and the presence or absence of pseudogenes in the genome (Table 1, Figure 1). A larger subset of 19 genes, chosen similarly, were interrogated in five ClinSeq® samples that had undergone exome, genome and Sanger sequencing. Validation of variants identified by NGS was simulated using large-scale Sanger sequencing data, which were generated as part of the ClinSeq® project (7). All bases with a Phred quality score of Q20 or greater within covered regions were aligned and interrogated using the Consed graphical sequence editor and genotypes were verified by manual observation of fluorescence peaks (12). Only variants with Sanger data for both forward and reverse read alignments were used in the analysis.

Variants from the exome data that were not validated by the Sanger data were resequenced using the original primers and manually-optimized primers designed using Primer3 software (13,14). Resequencing was performed on a 3130x sequencer using the BigDye 3.1 sequencing kit (Applied Biosystems, Carlsbad, CA, USA). Reads generated by resequencing were aligned to genome build hg19 (NCBI build 37) using Sequencher (Gene Codes Corporation, Ann Arbor, MI, USA) (see Supplemental Table 2 for primer sequences).

Statistical Analysis

The replication failure rates in the 19-gene and five-gene sets from NGS and Sanger sequencing data were compared using Fisher's exact test (15). The extracted data were then subjected to the Jaccard sameness test (16), calculated using all data points with a given

minimal MPG score threshold or higher, and that threshold was iterated across the entire range of MPG scores (Figure 2). Box-and-whisker plots of GC content, CDS length, and exon count for all genes used in this study were plotted using R (Figure 1). The 95% confidence interval of the accuracy estimate was calculated using the Jeffreys interval calculation (17).

Data Access

Exome sequencing data from ClinSeq® participants are available from NCBI's dbGAP database, accession number phs000971.v1.p1. Sanger traces are available from the NCBI SRA database (18) (see Supplemental Table 1 for associated accession numbers and query strings).

RESULTS

We simulated Sanger validation of NGS data by selecting two representative subsets of genes from our dataset. Our objective was to include a range of gene attributes, using GC content, CDS length, and exon count as selection criteria (Table 1, Figure 1). We also included several genes with known pseudogenes to address the challenges of sequence alignment (Table 1).

The mean number of variants with exome coverage of ten reads or more per kb of interrogated DNA in the exome data was 0.8041 (14,258/17,732 kb). Of these, 5,660 non-reference variants were covered bi-directionally by the Sanger sequencing data. For all interrogated variants, the Sanger reads were evaluated manually to emulate techniques typically used in a clinical setting. Among these 5,660 variants, 19 were identified by NGS but not by Sanger sequencing, representing 13 unique single nucleotide variants.

We next set out to address the possibility that this set of five genes was in some way not representative of the wider universe of gene attributes by evaluating a larger set of genes, which necessarily had to be performed on a smaller set of samples. We examined 19 genes (including the five used in our initial analysis) from five samples using both exome and genome sequencing (Table 1). Within these 19 genes, we identified 714 non-reference variants with coverage of at least 10 reads, with a mean variant per kb rate of 0.1256 (714/5,686 kb). There was a strong linear correlation of MPG score and read depth coverage of these variant positions (r^2=0.8978, Supplemental Figure 1). Of these variants, 234 variants were covered bi-directionally by Sanger sequencing data, and all of those variants were present in exome, genome and Sanger sequencing data. The replication failure rate in the 19-gene dataset (0/234) is not significantly different from the five-gene dataset (19/5,660, p=1.000 X^2).

We further evaluated the 19 discrepant results by performing another round of Sanger sequencing using both the original primers designed through automated primer design and new, optimized sets of sequencing primers designed using primer3 software (13). In four cases, sequencing with the original primers yielded a reference (non-variant) genotype, while the newly designed primers validated the variant found via NGS. The original discrepant Sanger results for these cases could be due to polymorphisms within the sample

DNA sequence complementary to the sequence of one of the original primers. In two other cases, resequencing with the original primers confirmed the variant, but the newly designed primers yielded no usable sequence. Finally, in 11 cases resequencing with both sets of primers confirmed the original NGS variant call.

Resequencing reconciled all but two differences between the NGS and Sanger sequencing data. The remaining two discrepancies were found in non-coding regions of the genes *APOA5* and *PDGFRB*, and had MPG quality scores of 4 and 10, respectively (10).

That only two of 5,660 variants were truly discrepant represents an agreement rate of 0.99965 (95% CI 0.99887–0.99993). Jaccard sameness scores were plotted against each possible minimum MPG score threshold from the NGS data and the resulting index ranged from 0.99965 to 1.00000, corresponding to a minimum of 99.965% accuracy ratio for NGS compared to Sanger sequencing (Figure 2).

DISCUSSION

The power and utility of NGS is based on its massively parallel interrogation of nucleic acids. The ability to simultaneously evaluate millions of base pairs allows clinicians and researchers to ask and answer novel and important questions. However, requiring relatively low-throughput dideoxy sequencing as a validation of high-throughput NGS interrogation severely limits the utility of NGS. With the consistently decreasing costs of NGS, the expense and time required to validate variants found in NGS data using Sanger sequencing can quickly outpace the cost of generating the initial NGS data.

Previous studies have provided preliminary evidence that Sanger sequencing validation may not represent the best practice for clinical NGS validation, however these studies were relatively small in scale and used secondary data from clinical diagnostic laboratories (3–5).

In 2013, Sikkema-Raddatz and colleagues (3) evaluated NGS variants in 84 individuals using a targeted panel including 48 genes, validating 168 novel variants using Sanger sequencing, including seven indels. They reported nearly 100% Sanger validation of variants identified through their NGS panel. Notably, the single variant that was not initially validated using Sanger sequencing was validated by a subsequent Sanger sequencing run. They concluded that targeted NGS could be reliably implemented as a stand-alone test, with no orthogonal validation required.

McCourt and colleagues (4) then used a combination of NGS technologies to interrogate variants in a host of cancer-related genes. Of the identified NGS variants, 37 were confirmed by Sanger sequencing validation, leading the authors to conclude that existing NGS technologies perform well in detecting known clinically-relevant mutations.

In 2014, Strom and colleagues (5) addressed the question of Sanger validation using data from 144 clinical exomes, from which they attempted to Sanger-validate 110 total single nucleotide variants. Of these 110 variants, 109 were validated by Sanger sequencing, and the one variant which was not validated had an exome quality score below their quality threshold.

More recently, Baudhuin and colleagues (6) performed a larger-scale study in which data from targeted NGS panels were compared to either Sanger sequence data or data from the 1000 Genomes Project. Sanger sequencing verified 100% of 919 variants identified from the targeted panels.

Combining the data from these four studies yields a total of 1,234 variants, only one of which was not validated by Sanger sequencing. These data, while compelling, are not sufficient to conclude that routine Sanger validation is unnecessary, partially because the largest study included only data from targeted panels with 100x coverage in >99.7% of captured bases (6), which is markedly higher coverage than can be expected from current exome sequencing technologies.

To address the need for systematic and large-scale evaluation of orthogonal Sanger validation of NGS, we used a dataset of 684 exomes comprising approximately 21 TB of sequence and matching Sanger data comprising 2.9 million reads from the same samples. We began with the detection of variants from NGS data generated with well-known exome capture kits (Agilent and TruSeq) and Illumina sequencing, coupled with our well-established variant calling process described in a number of prior successful genetic analysis efforts (19–26). We endeavored to select a range of genes that had attributes that were similar to, or were in some aspects more challenging sequencing targets than, a typical gene. This sample provided us with 5,660 variants that we could validate with our Sanger data, a sample set much larger than prior analyses (3–6). While our Sanger data set included millions of reads that could potentially be compared with the NGS variants, we limited our analysis to model a clinical orthogonal testing scenario as closely as possible. To that end, all variants that met our criteria for interrogation were manually evaluated from the Sanger traces. We also limited our analysis to germline variants from leukocyte DNA, as NGS-based discovery of somatic variants or from formalin-fixed, paraffin-embedded tissue would likely require a separate validation process.

Using this approach, we found only two variants among 5,660 that were not validated by Sanger sequencing. Those two variants had relatively low quality scores for their NGS calls, with one of the variant quality scores being exactly at our standard NGS base-calling quality threshold (MPG=10), and the other being well below that threshold (MPG=4). This suggested that, even without setting a minimum quality threshold for accepting NGS variant calls, 99.965% of those calls would be true positives, based on the lower limit of a 95% confidence interval for this large sample. In addition, our data suggested that, through application of a conservative score threshold, a single high-quality (in our case, MPG ≥ 10) “sentinel” call in any sample leads to the same variant being more likely correctly called in other samples, irrespective of quality score for the variant in that sample. Our sentinel call approach resulted in 583 variant calls with an MPG score of less than 10 being validated by Sanger sequencing. If a flat quality score threshold was applied to all of the data, these variants would have been missed through NGS screening, which could lead to variants that might impact a patient’s health being undetected. Furthermore, if a minimum quality threshold of MPG score 7 was applied to this data (which represents a very conservative threshold approximating to a GATK Q30 score) there would have been only a single non-confirmation, leading to a confirmation rate of 99.9823%. Though MPG is not the most

widely-used quality metric for NGS data, our results can easily be extended to NGS data using any quality metric by using this “sentinel” call approach with equivalent score thresholds.

The more striking conclusion came from re-sequencing the 19 originally discrepant variants, in that the majority of these orthogonal Sanger validations were themselves incorrect. Seventeen of the NGS variants would have been considered false positives if a single round of Sanger sequencing were used as a validation criteria. Our results suggest that if such practice were used in a clinical setting, more positive NGS variants would be discarded as (incorrectly designated) false positives, as compared to using the NGS data directly. Jaccard index analysis supported this assertion, showing no appreciable difference between NGS and Sanger sequencing with respect to variants within our data set, and complete agreement of the two sequencing methods at an MPG score threshold higher than 10.

Our measured validation rate of at least 99.965% for NGS data across a large dataset with no established minimum quality threshold represents higher accuracy than many medical tests currently used by clinicians. Results of such tests are routinely used to determine the course of treatment for a patient without any expectation of orthogonal validation. Given these data, we conclude that Sanger validation of NGS variants that are associated with robust quality scores should not be performed routinely. At the same time, we recognize that some variants detected using NGS technology can have serious medical implications for the tested proband and their family members. In such cases, performing a second orthogonal validation may be appropriate. One can envision a future in which such determinations are made by the ordering clinician, based on the presenting findings and the intended or anticipated clinical use of the genetic testing result. While this assertion will be controversial, we have been unable to find reference to a clinical laboratory test that boasts a 99.965% or higher analytical confirmation rate for which orthogonal confirmatory testing is routinely mandated, and we suggest that leaving the question of confirmatory testing in the hands of the ordering physician is most appropriate. This would align clinical genomics with the practices across many fields of medicine, reduce overall costs of genomic testing, and potentially reduce the error rate of inappropriately labeling a NGS variant as a false positive due to failure of the orthogonal assay. The fact that some CLIA-approved laboratories are already returning NGS variant results for clinically-relevant variants in certain subsets of genes, such as the ACMG-established list of genes for return of secondary findings (27), suggests that the field is already moving toward this practice. We therefore recommend that NGS testing results should be treated as many other clinical tests are treated: imperfect, but highly reliable.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by the Intramural Research Program of the National Human Genome Research Institute. The authors acknowledge Dr. Niraj Trivedi of NHGRI for statistical analyses, and Nancy Hansen of the

Comparative Genomic Analysis Unit, NHGRI for lending her bioinformatics expertise. The authors thank Dr. Steve Lincoln of Invitae for helpful discussions.

Abbreviations

NGS	Next-generation sequencing
MPG	Most Probable Genotype
CDS	coding DNA sequence

Human genes

ACTA2	actin, alpha 2, smooth muscle, aorta
APOC3	apolipoprotein C-III
APOA5	apolipoprotein A-V
CAV1	caveolin 1, caveolae protein, 22kDa
CD40	CD40 molecule, TNF receptor superfamily member 5
CETP	cholesteryl ester transfer protein, plasma
CIITA	class II, major histocompatibility complex, transactivator
FGG	fibrinogen gamma chain
GPX1	glutathione peroxidase 1
LDLRAP1	low density lipoprotein receptor adaptor protein 1
LPL	lipoprotein lipase
MBL2	mannose-binding lectin [protein C] 2, soluble
MMP9	matrix metallopeptidase 9
MVK	mevalonate kinase
PDGFRB	platelet-derived growth factor receptor, beta polypeptide
PITX2	paired-like homeodomain 2
TNFRSF1A	tumor necrosis factor receptor superfamily, member 1A
UCP2	uncoupling protein 2 [mitochondrial, protein carrier]
VEGFA	vascular endothelial growth factor A

References

1. de Souza, F. MIT Technology Review EmTech [Internet]. 2014. [cited 2015 Jul 9]. Available from: http://www.technologyreview.com/news/531091/emtech-illumina-says-228000-human-genomes-will-be-sequenced-this-year/
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977; 74:5463–7. [PubMed: 271968]

3. Sikkema-Raddatz B, Johansson LF, de Boer EN, Almomani R, Boven LG, van den Berg MP, et al. Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. Hum Mutat. 2013; 34:1035–42. [PubMed: 23568810]
4. McCourt, CM.; McArt, DG.; Mills, K.; Catherwood, MA.; Maxwell, P.; Waugh, DJ., et al. Validation of Next Generation Sequencing Technologies in Comparison to Current Diagnostic Gold Standards for BRAF, EGFR and KRAS Mutational Analysis. PLoS ONE [Internet]. 2013. [cited 2015 Jul 9];8. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3724913/
5. Strom SP, Lee H, Das K, Vilain E, Nelson SF, Grody WW, et al. Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. Genet Med Off J Am Coll Med Genet. 2014; 16:510–5.
6. Baudhuin LM, Lagerstedt SA, Klee EW, Fadra N, Oglesbee D, Ferber MJ. Confirming Variants in Next-Generation Sequencing Panel Testing by Sanger Sequencing. J Mol Diagn. 2015; 17:456–61. [PubMed: 25960255]
7. Biesecker LG, Mullikin JC, Facio FM, Turner C, Cherukuri PF, Blakesley RW, et al. The ClinSeq Project: Piloting large-scale genome sequencing for research in genomic medicine. Genome Res. 2009; 19:1665–74. [PubMed: 19602640]
8. Johnston JJ, Teer JK, Cherukuri PF, Hansen NF, Loftus SK, et al. NIH Intramural Sequencing Center (NISC). Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. Am J Hum Genet. 2010; 86:743–8. [PubMed: 20451169]
9. Johnston JJ, Rubinstein WS, Facio FM, Ng D, Singh LN, Teer JK, et al. Secondary Variants in Individuals Undergoing Exome Sequencing: Screening of 572 Individuals Identifies High-Penetrance Mutations in Cancer-Susceptibility Genes. Am J Hum Genet. 2012; 91:97–108. [PubMed: 22703879]
10. Teer JK, Bonnycastle LL, Chines PS, Hansen NF, Aoyama N, Swift AJ, et al. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. Genome Res. 2010; 20:1420–31. [PubMed: 20810667]
11. Chines PS, Swift AJ, Bonnycastle LL, Erdos M, Mullikin JC, et al. NIH Intramural Sequencing Center. PrimerTile: Designing overlapping PCR primers for resequencing. Am J Hum Genet. 2005:A1257.
12. Gordon D, Green P. Consed: a graphical editor for next-generation sequencing. Bioinformatics. 2013; 29:2936–7. [PubMed: 23995391]
13. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3--new capabilities and interfaces. Nucleic Acids Res. 2012; 40:e115. [PubMed: 22730293]
14. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinforma Oxf Engl. 2007; 23:1289–91.
15. Fisher, RA. On the Interpretation of x2 from Contingency Tables, and the Calculation of P. 1922. [cited 2015 Dec 28]; Available from: https://digital.library.adelaide.edu.au/dspace/handle/2440/15173
16. Jaccard P. The Distribution of the Flora in the Alpine Zone. 1. New Phytol. 1912; 11:37–50.
17. Brown LD, Cai TT, DasGupta A. Interval Estimation for a Binomial Proportion. Stat Sci. 2001; 16:101–33.
18. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Res. 2011; 39:D19–21. [PubMed: 21062823]
19. Johnston JJ, Lewis KL, Ng D, Singh LN, Wynter J, Brewer C, et al. Individualized Iterative Phenotyping for Genome-wide Analysis of Loss-of-Function Mutations. Am J Hum Genet. 2015; 96:913–25. [PubMed: 26046366]
20. Sen SK, Boelte KC, Barb JJ, Joehanes R, Zhao X, Cheng Q, et al. Integrative DNA, RNA, and protein evidence connects TREML4 to coronary artery calcification. Am J Hum Genet. 2014; 95:66–76. [PubMed: 24975946]
21. Pierson TM, Yuan H, Marsh ED, Fuentes-Fajardo K, Adams DR, Markello T, et al. GRIN2A mutation and early-onset epileptic encephalopathy: personalized therapy with memantine. Ann Clin Transl Neurol. 2014; 1:190–8. [PubMed: 24839611]

22. Yuan H, Hansen KB, Zhang J, Pierson TM, Markello TC, Fajardo KVF, et al. Functional analysis of a de novo GRIN2A missense mutation associated with early-onset epileptic encephalopathy. Nat Commun. 2014; 5:3251. [PubMed: 24504326]

23. Gonsalves SG, Ng D, Johnston JJ, Teer JK, Stenson PD, Cooper DN, et al. Using exome data to identify malignant hyperthermia susceptibility mutations. Anesthesiology. 2013; 119:1043–53. [PubMed: 24195946]

24. Gartner JJ, Parker SCJ, Prickett TD, Dutton-Regester K, Stitzel ML, Lin JC, et al. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. Proc Natl Acad Sci U S A. 2013; 110:13481–6. [PubMed: 23901115]

25. Ng D, Johnston JJ, Teer JK, Singh LN, Peller LC, Wynter JS, et al. Interpreting secondary cardiac disease variants in an exome cohort. Circ Cardiovasc Genet. 2013; 6:337–46. [PubMed: 23861362]

26. Le Gallo M, O'Hara AJ, Rudd ML, Urick ME, Hansen NF, O'Neil NJ, et al. Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. Nat Genet. 2012; 44:1310–5. [PubMed: 23104009]

27. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, et al. ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing. Genet Med Off J Am Coll Med Genet. 2013; 15:565–74.

28. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002; 12:996–1006. [PubMed: 12045153]

29. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007; 35:D61–5. [PubMed: 17130148]

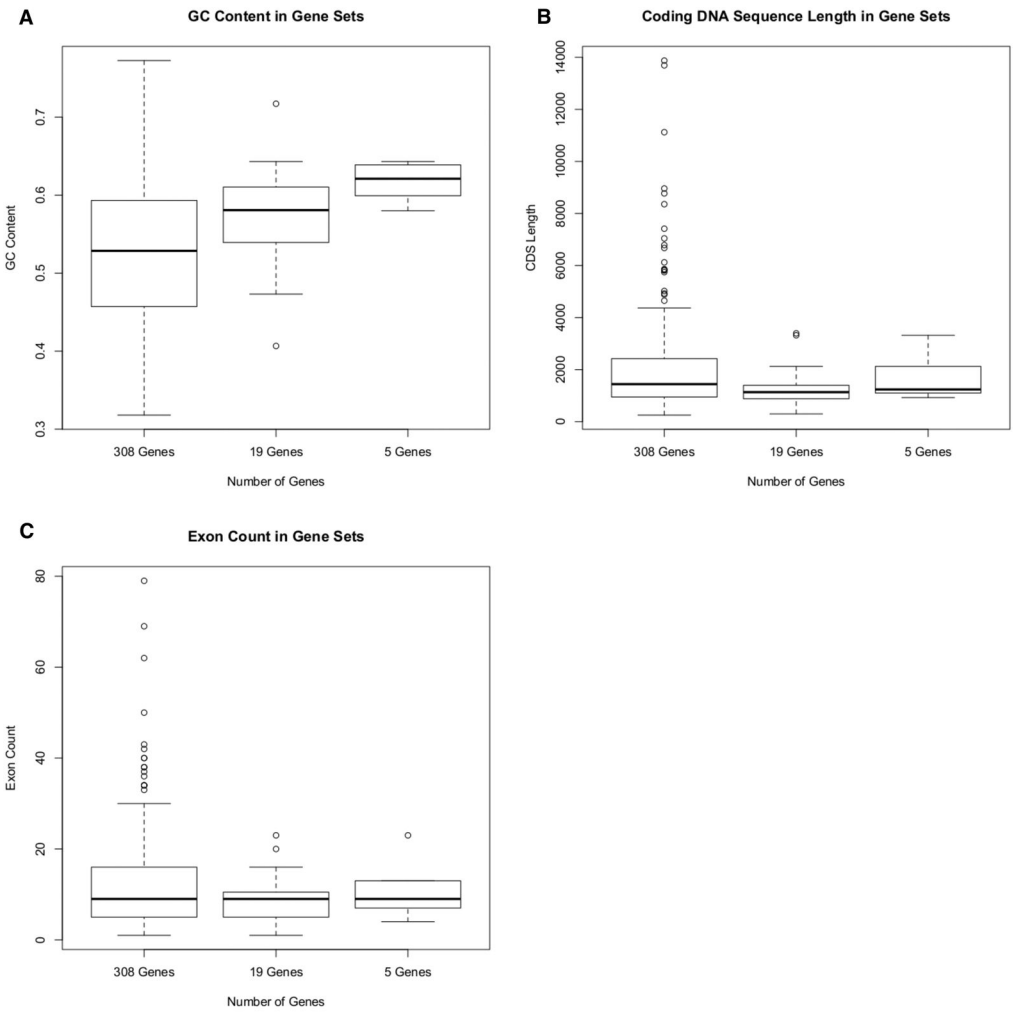

Figure 1. Box-And-Whisker Plots of Gene Statistics

Box-and-whisker plots show the distribution of genes in each candidate set used in this analysis across GC content (Figure 1A), CDS length (Figure 1B), and exon count (Figure 1C). Data on these genes was collected using UCSC Genome Browser (28) or NCBI's Entrez (29). In the case of multiple transcripts, the transcript encoding the longest protein isoform was used.

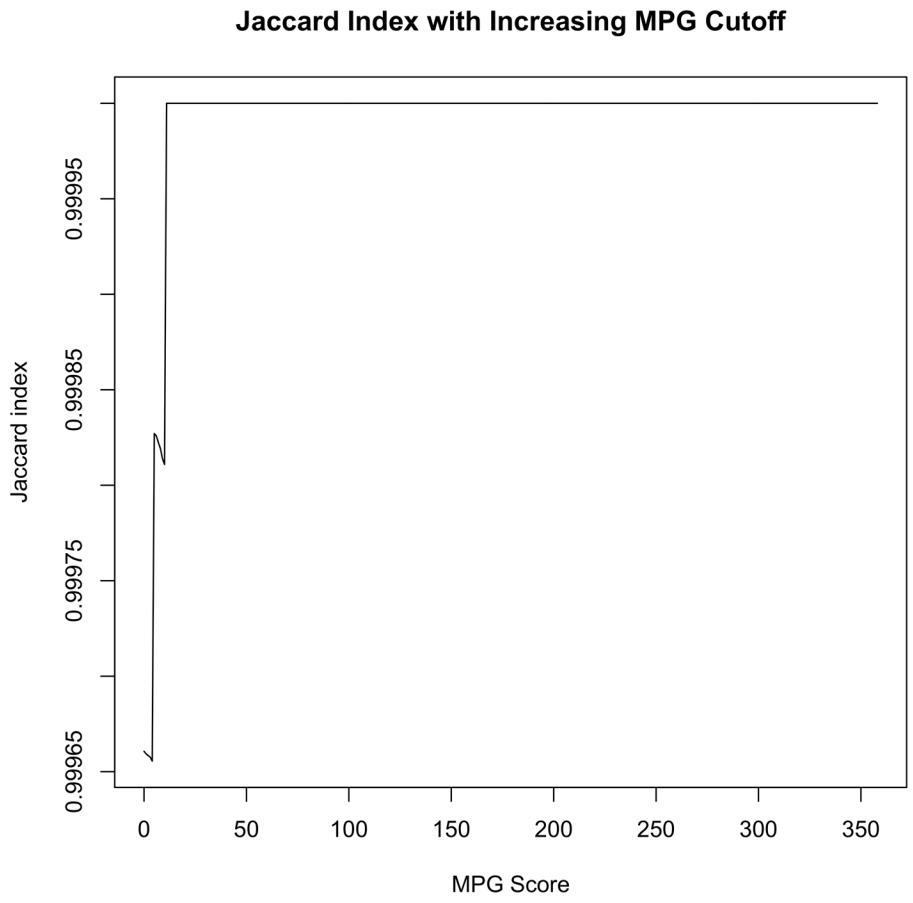

Figure 2. Jaccard Index with Increasing MPG Cutoff

The Jaccard sameness index was used to evaluate the agreement of variants discovered using NGS versus Sanger sequencing, correlated with increasing MPG thresholds, then plotted using R. Data from both sequencing methods were in complete agreement at an MPG higher than 10, resulting in a score of 1.000 at all points thereafter.

Table 1

Genes used in these analyses.[a]

Gene Name	Chromosome	Exon Count	CDS Length	Pseudogene?	GC Content
APOA5	chr11	4	1,101	FALSE	0.64
LDLRAP1	chr1	9	927	FALSE	0.60
MMP9	chr20	13	2,124	FALSE	0.64
PDGFRB	chr5	23	3,321	FALSE	0.58
VEGFA	chr6	7	1,239	FALSE	0.62
ACTA2	chr10	9	1,134	TRUE	0.53
APOC3	chr11	4	300	FALSE	0.59
CAV1	chr7	3	537	FALSE	0.47
CD40	chr20	9	834	FALSE	0.55
CETP	chr16	16	1,482	FALSE	0.54
CIITA	chr16	20	3,396	FALSE	0.62
FGG	chr4	10	1,362	TRUE	0.41
GPX1	chr3	1	612	TRUE	0.72
LPL	chr8	10	1,428	FALSE	0.50
MBL2	chr10	4	747	TRUE	0.53
MVK	chr12	11	1,191	FALSE	0.60
PITX2	chr4	6	954	FALSE	0.59
TNFRSF1A	chr12	10	1,368	FALSE	0.57
UCP2	chr11	8	930	FALSE	0.58

[a]The top five genes listed in this table (above the dark line) were interrogated via exome and Sanger sequencing in 684 samples, while all genes in the table were interrogated in five samples via exome, genome, and Sanger sequencing.