Genetics inMedicine | ORIGINAL RESEARCH ARTICLE

Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing

Diana Mandelker, MD, PhD[1,7], Ryan J. Schmidt, MD, PhD[1], Arunkanth Ankala, PhD[2], Kristin McDonald Gibson, PhD[3,7], Mark Bowser, MS, MPH[4], Himanshu Sharma, MS[4], Elizabeth Duffy, BS[4], Madhuri Hegde, PhD, FACMG[5], Avni Santani, PhD[3], Matthew Lebo, PhD, FACMG[1,4] and Birgit Funke, PhD, FACMG[4,6]

Purpose: Next-generation sequencing (NGS) is now routinely used to interrogate large sets of genes in a diagnostic setting. Regions of high sequence homology continue to be a major challenge for short-read technologies and can lead to false-positive and false-negative diagnostic errors. At the scale of whole-exome sequencing (WES), laboratories may be limited in their knowledge of genes and regions that pose technical hurdles due to high homology. We have created an exome-wide resource that catalogs highly homologous regions that is tailored toward diagnostic applications.

Methods: This resource was developed using a mappability-based approach tailored to current Sanger and NGS protocols.

Results: Gene-level and exon-level lists delineate regions that are difficult or impossible to analyze via standard NGS. These regions are ranked by degree of affectedness, annotated for medical relevance, and classified by the type of homology (within-gene, different functional gene, known pseudogene, uncharacterized noncoding region). Additionally, we provide a list of exons that cannot be analyzed by short-amplicon Sanger sequencing.

Conclusion: This resource can help guide clinical test design, supplemental assay implementation, and results interpretation in the context of high homology.

Genet Med advance online publication 26 May 2016

Key Words: homology; mappability; next-generation sequencing; pseudogene; whole-exome sequencing

INTRODUCTION

Short-read next-generation sequencing (NGS) has become the diagnostic assay of choice in many clinical laboratories given its potential for efficient large-scale analysis. For genetically heterogeneous disorders, especially those presenting with clinical heterogeneity, clinical laboratories are beginning to transition from targeted gene panels to whole-exome sequencing (WES). However, with the expansion to exome-wide analysis, the limitations of NGS in regions of high homology will become increasingly apparent. Diagnostic laboratories have historically had deep gene-specific knowledge regarding the presence of homologous sequences. Such expertise does not scale easily, and unless this knowledge is precurated, a clinical laboratory may risk reporting false-positive and false-negative variant calls resulting from inaccurate mapping of short reads to highly homologous regions, including pseudogenes. Although most bioinformatic NGS data analysis pipelines are homology-aware, adequate resources and guidance tailored toward use at early stages of test development are lacking. Awareness of problematic regions is critical at the test design stage as well as the reporting stage to guide decision making regarding whether to exclude regions from the test and to decide whether alternative assays must be used for critical genes. This is particularly the case for WES, in which sets of genes to be analyzed can be configured ad hoc based on a specific clinical scenario.

With the GENCODE project identifying 11,216 unique pseudogenes,[1] the potential for homology to interfere with clinical sequencing is widespread and must be examined before launching a clinical test. Homology is especially concerning for genes with high detection rates for the disease of interest or for genes where professional societies suggest return of results regardless of the patient's diagnosis. The American College of Medical Genetics and Genomics (ACMG) recognizes the inherent difficulty in interrogating regions of high homology and says in its most recent guidelines that "the laboratory must develop a strategy for detecting disease-causing variants within regions with known homology."[2] Likewise, the College of American Pathologists states that laboratories testing highly homologous

The first two authors contributed equally to this work.
[1]Department of Pathology, Harvard Medical School/Brigham and Women's Hospital, Boston, Massachusetts, USA; [2]Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia, USA; [3]Division of Genomic Diagnostics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA; [4]Partners HealthCare Personalized Medicine, Laboratory for Molecular Medicine, Cambridge, Massachusetts, USA; [5]Emory Genetics Lab, Emory University School of Medicine, Atlanta, Georgia, USA; [6]Department of Pathology, Harvard Medical School/Massachusetts General Hospital, Boston, Massachusetts, USA; [7]Current affiliation: Department of Pathology, Memorial Sloan Kettering Cancer Center, New York City, New York, USA (D.M.); Medical Genetics, Invitae Corporation, San Francisco, California, USA (K.M.G.). Correspondence: Birgit Funke (bfunke@partners.org)

Submitted 13 February 2016; accepted 24 March 2016; advance online publication 26 May 2016. doi:10.1038/gim.2016.58

genes must devise methodology that can distinguish between gene and pseudogene and document the accuracy of their testing.[3] However, translating these recommendations into clinical practice can be challenging.

Multiple resources have been developed computationally to describe pseudogenes on a genomic level.[4–7] Although these resources are extremely useful to define pseudogene structure, location, and extent of homology to the parent gene, they have not been created with diagnostic applications in mind. Some features of high interest to molecular diagnostic laboratories include the extent to which a homologous gene is problematic for NGS and/or Sanger sequencing, information about the medical importance of the gene, as well as information on the genomic context of the homology. Annotation of homologous genes and exons can facilitate decisions regarding whether to exclude such regions from the NGS assay or to develop ancillary assays to ensure accuracy and optimize clinical sensitivity.

An oftentimes underappreciated portion of the NGS bioinformatic pipeline is the alignment of the short reads to the reference genome, which is challenged by reads deriving from regions with high homology. Mapping quality (MQ), which is an output of alignment algorithms that are associated with each read,[8,9] provides an estimate of the probability that the read was aligned to the wrong location in the reference genome. This metric is derived from a complex calculation that takes into account multiple factors, including the probability of the read arising from other areas of the genome and base quality (the probability of a sequencing error) at bases that differ from the reference sequence. MQ may be misleading in specific instances, especially in cases where genuine variants are present in the read. An alternative approach to measuring homology is mappability.[10] Mappability is easy to compute and can be performed in silico using only the sequence of the reference genome. Consequently, this approach does not incur the time and materials cost of running actual samples and can be run before finalizing the design of an NGS assay. Importantly, mappability is not affected by variations across experiment runs or different sequencing platforms, making it a good candidate to serve as the foundation of a universal homology resource.

We adjusted the mappability algorithm[10] to mimic the sizes of typical Sanger sequencing amplicons and NGS library fragments and searched for sites of either ≥98% or 100% homology elsewhere in the genome. From this analysis, we generated exon and gene level lists relevant to current clinical testing that include regions that are difficult or impossible to analyze by standard Sanger or NGS approaches. These data confirm the widespread nature of high homology throughout the human exome, which is well understood in general but poses risk as laboratories expand their analyses beyond small panels of familiar genes.

MATERIALS AND METHODS

Mappability analysis

Whole-genome sequence (hg19, GRCh37) was downloaded from http://genome.ucsc.edu/ and indexed using gem-indexer (http://algorithms.cnag.cat/wiki/The_GEM_library). The gem-mappability[10] was run using the following arguments: (i) -l 1000 -m 0 -t 2, (ii) -l 250 -m 0 -t 2, and (iii) -l 250 -m 5 -t 7. The mappability outputs were converted to .wig files using gem-2-wig and then subsequently to .bed files using BEDOPS wig2bed.[11] These .bed files were adjusted to 1-based coordinates using a custom Python script. Genomic coordinates for each exon in the human exome (hg19) were downloaded from http://genome.ucsc.edu and extended 65 bp upstream and downstream of each exon to include flanking sequences that are typically analyzed in hybrid capture NGS approaches and contain clinically important regulatory sequences for RNA splicing. These coordinates are referred to as "start_minus_65bp" and "end_plus_65bp" positions for each exon in the **Supplementary Tables** online and constitute the regions analyzed in our analysis.

A pipeline of custom Python scripts was run that extracts the mappability score for each position within an exon and calculates homology metrics at the exon level (**Supplementary Tables S1–S4** online). These tables were sorted to identify exons that contain high homology to sequences elsewhere in the genome using the criteria outlined in the Results section. Scripts used for this analysis are available upon request.

Generating lists of homologous exons and genes

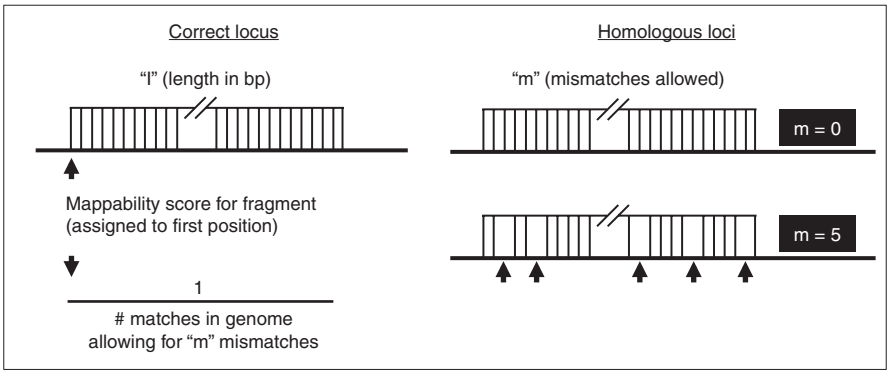
A mappability score[10] was assigned to each genomic position that reflects the degree of homology associated with the local sequence of length l. We varied settings on the mappability algorithm to derive exon-level lists of particular relevance to clinical genetics laboratories (see **Figure 1** for details).

Homology-type annotation

Sequences of the 11,557 exons along with flanking ±65 bp from the "NGS Problem List–Low Stringency" were obtained and a local alignment for each region was performed against the genome using the BLAT algorithm[12] at 90% minimum identity threshold. The target coordinates of the BLAT hits were recorded and annotated as follows: (i) "same gene"—target coordinates do not overlap with the query coordinates but the target coordinates overlap with the query gene coordinates; (ii) "different gene"—target coordinates overlap with another gene coding sequence (CDS) in the exome and do not overlap with the query gene coordinates; (iii) "pseudogene"—target coordinates overlap with psiDRv0 pseudogene regions[6] but do not overlap with any CDS in the exome; and (iv) "non-CDS"—target coordinates do not overlap with psiDRv0 pseudogene regions or with any CDS in the exome.

Medical relevance gene filter and annotation

To focus our analysis on genes with suspected or established medical relevance, our lists of genes containing problematic exons were intersected with a list of 4,773 genes obtained from OMIM (http://omim.org, accessed November 2012), HGMD (http://www.hgmd.cf.ac.uk, accessed November 2012), and ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/, accessed December 2012) (**Supplementary Table S13** online).



List (Parameters)	NGS dead zone (l = 250, m = 0)	NGS problem list - high stringency (l = 250, m = 5)	NGS problem list - low stringency (l = 250, m = 5)	Sanger dead zone (l = 1,000, m = 0)
Exons (% of total)	4,264 (2.2%)	7,691 (3.9%)	11,557 (5.9%)	3,078 (1.6%)
Genes	619	1,168	2,512	467
Medically relevant genes	73	193	464	54
Description	Short reads cannot be unambiguously mapped to any positions in an exon or to a large stretch within a larger exon. The probability of mapping to an identical sequence elsewhere is the same for these regions.	Difficult or unable to be analyzed regions. A large fraction of the reads obtained for these regions likely have a high risk of misalignment.	Contains at least some potentially difficult or unable to be analyzed positions. There is a risk of read misalignment due to high homology at these sites.	Obtaining a unique Sanger amplicon for this region is difficult if not impossible. There is at least one position in the exon that is part of a 1 kb region with an identical match elsewhere in the genome.
Exon-Level criteria	100% of positions with mappability scores < 1 Or Large contiguous region (≥ 250 bp) with mappability scores < 1	≥ 90% of positions with mappability scores < 1 Or Large contiguous region (≥ 250 bp) with mappability scores < 1	At least one position in the exon with mappability score < 1	At least one position in the exon with mappability score < 1

Figure 1 Mappability-based generation of homologous exon and gene lists. A sequence of length "l" is scanned across the genome to find homologous sequences allowing for "m" mismatches. Matched sequences at homologous loci with zero (m = 0, 100% match) or five (m = 5, up to 2% mismatch) mismatches are shown. The mappability score is defined as the reciprocal of the number of matches in the genome and is assigned to the first position in the sequence. A unique sequence will be present only once in the genome and will have a mappability score of 1. By contrast, any sequence that matches to more than one location will have a mappability score <1. Mappability scores were calculated for each position in the exome. Four lists were generated using the indicated mappability settings and criteria for assembling exon-level and gene-level lists. For all NGS-relevant mappability analyses, the k-mer length was set to 250 bp (l = 250), because this approximates a commonly used library fragment size. To conservatively mimic the amplicon size used in standard Sanger sequencing approaches, the k-mer length was increased to 1,000 bp (l = 1,000). Distinct exon-level criteria were used to generate four lists. The total number of exons, genes, and genes with medical relevance is presented for each list.

Medically relevant genes identified on the "NGS Problem List–High Stringency" (193 genes; **Figure 1**) were further annotated to provide basic information useful in a molecular diagnostic setting through searches of OMIM, HGMD, ClinVar, GeneReviews (http://www.ncbi.nlm.nih.gov/books/NBK1116/), and PubMed (http://www.ncbi.nlm.nih.gov/pubmed). Evidence from the published literature was recorded with corresponding PubMed IDs when available. Additional annotations include age of onset, prevalence, and categorization (Mendelian, association, somatic, pharmacogenetic). The evidence for each gene–disease association was graded using specific criteria (**Supplementary Figure S1** online) from evidence level 0 (undetermined association) through evidence level 3 (definitive association).

MQ analysis

MQ scores were obtained from 30 WES experiments for each read using the BWA-MEM algorithm[13] performed on the Illumina HiSeq 2500 platform. The scores were averaged at each base within the exome (±15 bp surrounding each exon) from the

reads aligned to this position. The percentage of positions with low MQ scores (<17) was obtained for each gene and reported in the gene-level Tables (**Supplementary Tables S5–S12** online). This threshold was chosen because it is consistent with settings commonly used in NGS bioinformatic pipelines[14] to exclude low-quality reads from the variant calling process.

RESULTS

Generation of lists of homologous regions in the exome for clinical diagnostic use

To identify regions of the exome that may be problematic for diagnostic testing due to high homology, we used the gem-mappability algorithm[10] and tailored it to reflect standard NGS and Sanger sequencing approaches (**Figure 1**).

We compiled four exon-resolution lists with increasing degrees of sequence homology, each containing quantifiable measures of the degree of involvement by high homology, such as the percentage of affected positions and the largest stretch of contiguously affected positions. "Dead zone" lists indicate that identical matches exist elsewhere in the genome, making it impossible to unambiguously map an NGS read or generate a standard-length Sanger amplicon in these affected regions. "NGS problem" lists allow for up to 2% mismatch. In these regions, a read that matches the reference sequence has the potential to align correctly. However, there is a significant risk that a read containing a sequencing error or variant may misalign to a highly homologous region elsewhere. Data from these exon-level lists were aggregated by gene name and served as the basis for building gene-level lists expressing the degree to which a gene is affected by high sequence homology. These lists include any gene with at least one affected exon (**Figure 1**). Summary metrics are provided for each gene detailing the percentages of affected exons and positions. Gene-level lists were subdivided by intersection with a list of 4,773 genes with known or suspected medical relevance to highlight clinically important genes.

List 1. "NGS Dead Zone" (2.2% of exons; 619 genes): These regions are entire exons or large contiguous portions of exons that have 100% identity to other loci. A contiguous region of 250 or more affected positions indicates a stretch of 100% homology that extends 499 or more base pairs. When attempting to sequence the central positions of a homologous sequence of this length, a read's mate pair cannot assist with unambiguous alignment of the read because it also falls inside the region of 100% homology, given a library fragment size of 250 bp.

Lists 2 and 3. "NGS Problem Lists–High Stringency" (3.9% of exons; 1,168 genes) and "Low Stringency" (5.9% of exons; 2,512 genes): After defining regions that are definitively problematic for standard NGS approaches, we adjusted the mappability settings to allow for up to 2% mismatches (**Figure 1**) to capture additional regions that may pose problems due to high homology but do not share 100% identity with other loci. The purpose of the NGS problem lists is to warn users about regions that may pose problems for standard NGS, especially under certain conditions, but are not altogether unanalyzable. A "high stringency" exon-based list included exons with ≥90% of their positions affected by high homology or a contiguous stretch of 250 or more affected positions to account for significant regions of high homology within large exons (**Figure 1**). In these cases, the overwhelming majority of reads generated from this region will possess high homology to other loci, which may lead to misalignment to other portions of the genome. Variant calls based on reads mapped to these affected regions should be viewed with extreme caution. The "low stringency" list captures any exon with at least one position whose associated 250-bp sequence fragment maps to more than one place in the genome with 2% or less mismatches (**Figure 1**). This analysis aims to detect all regions at risk for homology interference, and thus a substantial fraction of exons on this list may not cause analytical problems in practice.

List 4. "Sanger Dead Zone" (1.6% of exons; 467 genes): Some clinical laboratories routinely use Sanger sequencing to confirm variants detected by NGS. A homologous region that is difficult to analyze by NGS may still be interrogated accurately by Sanger sequencing provided that it is possible to design amplicons larger than the region affected by sequence homology. We identified regions in the exome that cannot be Sanger sequenced using standard short amplicon protocols, assuming a maximum amplicon size of 1,000 bp.

To test whether our mappability-based approach adequately flags regions with empirically demonstrated poor MQ, we compared our data with MQ scores derived from 30 WES data sets (see Methods for details). The majority of genes on our lists contained positions with low average MQ (**Figure 2**, **Supplementary Tables S5–S12** online). Additionally, there was a positive correlation between the percentage of positions with low MQ and the percentage of affected positions by our analyses. The "NGS Problem List–Low Stringency" identified 1,557 of 1,676 (92.9%) of genes in the exome with at least one position with a low average MQ score. This comparison serves as an empirical validation of our method and suggests that our resources can help identify regions that are at risk for problems due to high homology before implementing and running an NGS assay.

High-priority clinically relevant genes affected by homology

It is particularly important to be aware of genes that are both affected by homology and strongly associated with a clinical phenotype because these are the most likely to be tested, and variants found in these genes are most likely to be causally linked to a patient's phenotype.

To focus attention on "cannot miss" genes affected by high homology that should be considered when designing and interpreting clinical assays, we selected the 193 medically relevant genes from the "NGS Problem List–High Stringency" analysis (**Supplementary Table S8** online) for further manual annotation

NGS dead zone

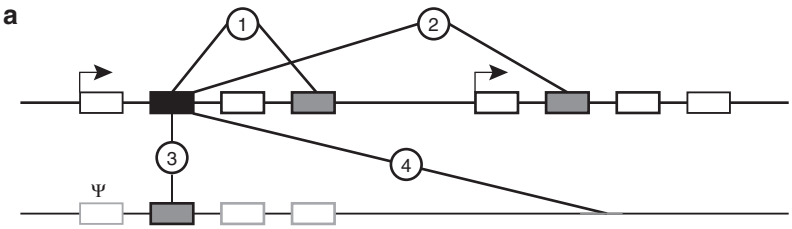
Gene	Affected exons (%)	Affected positions (%)	% Observed low MQ	Homology type	Disease(s)	Category
SMN1	14/16 (87.5)	3,488/3,850 (90.6)	92.7	Different gene | non-CDS | pseudogene	Spinal muscular atrophy	M
RPS17	8/10 (80)	1,850/2,116 (87.4)	76.4	Same gene | non-CDS | pseudogene	Diamond-blackfan anemia	M
SMN2	14/18 (77.8)	3,488/4,140 (84.3)	93.0	Different gene | non-CDS | pseudogene	Spinal muscular atrophy	M
IKBKG	7/10 (70)	1,921/2,764 (69.5)	63.6	Pseudogene	Incontinentia pigmenti	M
CFC1	5/6 (83.3)	837/1,471 (56.9)	76.7	Different gene	Congenital heart defects	M
ADAMTSL2	9/18 (50)	2,738/5,196 (52.7)	60.0	Non-CDS	Geleophysic dysplasia	M
OPN1MW	7/12 (58.3)	1,915/3,750 (51.1)	67.2	Same gene | different gene	Colorblindness, deutan; blue cone monochromacy	M
STRC	10/29 (34.5)	3,987/9,098 (43.8)	80.9	Different gene | non-CDS | pseudogene	Sensorineural hearing loss	M
KRT86	3/9 (33.3)	904/2,631 (34.4)	37.9	Different gene | non-CDS | pseudogene	Monilethrix	M
TUBB2B	1/4 (25)	528/1,858 (28.4)	72.0	Different gene | non-CDS | pseudogene	Polymicrogyria	M
LPA	10/39 (25.6)	3,003/11,193 (26.8)	39.5	Same gene | different gene | non-CDS | pseudogene	Coronary artery disease	A
CHRNA7	2/10 (20)	668/2,896 (23.1)	59.6	Different gene	15q13.3 microdeletion syndrome	M
KRT81	2/9 (22.2)	474/2,688 (17.6)	36.0	Different gene | non-CDS | pseudogene	Monilethrix	M
NCF1	2/11 (18.2)	454/2,603 (17.4)	22.3	Pseudogene	Chronic granulomatous disease	M
OTOA	4/30 (13.3)	1,124/7,358 (15.3)	28.5	Non-CDS	Sensorineural hearing loss	M
KIR3DL1	1/9 (11.1)	346/2,505 (13.8)	41.9	Different gene | non-CDS | pseudogene	HIV disease progression	A
TNXB	10/56 (17.9)	2,890/21,942 (13.2)	25.5	Same gene	Ehlers-danlos syndrome	M
OPN1LW	1/6 (16.7)	241/1,875 (12.9)	10.8	Different gene	Blue cone monochromacy	M
NEB	16/181 (8.8)	4,786/49,213 (9.7)	15.3	Same gene	Nemaline myopathy	M
CORO1A	1/10 (10)	235/2,686 (8.7)	7.9	Non-CDS	Immunodeficiency	M
OCLN	1/8 (12.5)	172/2,609 (6.6)	36.0	Pseudogene	Band-like calcification with simplified gyration and polymicrogyria	M
FLG	1/2 (50)	802/12,446 (6.4)	20.0	Same gene	Ichthyosis vulgaris	M
HYDIN	6/86 (7)	1,701/26,643 (6.4)	67.4	Non-CDS | pseudogene	Primary ciliary dyskinesia	M
RHCE	1/10 (10)	157/2,554 (6.1)	17.4	Different gene	Rh blood group antigens	M
PMS2	1/15 (6.7)	274/4,539 (6)	20.4	Non-CDS | pseudogene	HNPCC	M, S
STAT5B	1/18 (5.6)	266/4,704 (5.7)	15.0	Different gene	Growth hormone insensitivity with immunodeficiency	M
TTN	7/363 (1.9)	1,308/161,621 (0.8)	2.2	Same gene	Dilated cardiomyopathy	M

Sanger dead zone

Gene	Affected exons (%)	Affected positions (%)	% Observed low MQ	Homology type	Disease(s)	Category
RPS17	6/10 (60)	1,286/2,116 (60.8)	76.4	Same gene | non-CDS | pseudogene	Diamond-blackfan anemia	M
SMN1	12/16 (75)	2,310/3,850 (60)	92.7	Different gene | non-CDS | pseudogene	Spinal muscular atrophy	M
SMN2	12/18 (66.7)	2,310/4,140 (55.8)	93.0	Different gene | non-CDS | pseudogene	Spinal muscular atrophy	M
ADAMTSL2	9/18 (50)	2,736/5,196 (52.7)	60.0	Non-CDS	Geleophysic dysplasia	M
IKBKG	6/10 (60)	1,447/2,764 (52.4)	63.6	Pseudogene	Incontinentia pigmenti	M
OPN1MW	5/12 (41.7)	1,431/3,750 (38.2)	67.2	Same gene | different gene	Colorblindness, deutan; blue cone monochromacy	M
CFC1	3/6 (50)	387/1,471 (26.3)	76.7	Different gene	Congenital heart defects	M
OPN1LW	2/6 (33.3)	346/1,875 (18.5)	10.8	Different gene	Blue cone monochromacy	M
STRC	4/29 (13.8)	1,484/9,098 (16.3)	80.9	Different gene | non-CDS | pseudogene	Sensorineural gearing loss	M
CORO1A	2/10 (20)	344/2,686 (12.8)	7.9	Non-CDS	Immunodeficiency	M
GRK1	2/7 (28.6)	300/2,602 (11.5)	0.0	Different gene	Oguchi disease	M
LPA	6/39 (15.4)	1,220/11,193 (10.9)	39.5	Same gene | different gene | non-CDS | pseudogene	Coronary artery disease	A
OCLN	2/8 (25)	213/2,609 (8.2)	36.0	Pseudogene	Band-like calcification with simplified gyration and polymicrogyria	M
NCF1	1/11 (9.1)	202/2,603 (7.8)	22.3	Pseudogene	Chronic granulomatous disease	M
RHCE	1/10 (10)	157/2,554 (6.1)	17.4	Different gene	Rh blood group antigens	M
NEB	11/181 (6.1)	2,560/49,213 (5.2)	15.3	Same gene	Nemaline myopathy	M
OTOA	2/30 (6.7)	375/7,358 (5.1)	28.5	Non-CDS	Sensorineural hearing loss	M
TNXB	4/56 (7.1)	406/21,942 (1.9)	25.5	Same gene	Ehlers-danlos syndrome	M
PMS2	1/15 (6.7)	67/4,539 (1.5)	20.4	Non-CDS | pseudogene	HNPCC	M, S

Figure 2 Gene-level lists include genes with high medical relevance. Medically relevant genes with definitive disease association (evidence level 3, **Supplementary Figure S1** online) from the "NGS Dead Zone" and "Sanger Dead Zone" lists are shown with selected annotations sorted by percentage of affected positions. % Observed low mapping quality (MQ) indicates the percentage of positions with low average MQ scores within the gene calculated from whole-exome sequencing and is presented for comparison with our mappability-based metrics. Homology type classifies the genomic context of the homologous sequences for each gene. Categories of mutant alleles are abbreviated as follows: M, Mendelian; A, risk association; S = somatic. Taking stereocillin (*STRC*) as an example, one can see that it has been associated with sensorineural hearing loss. Homologous loci elsewhere in the genome include an annotated pseudogene. Ten of its exons fall within the "NGS Dead Zone" and cannot be reliably analyzed by standard next-generation sequencing (NGS). A candidate orthogonal approach to "fill in" these homologous regions after NGS might be Sanger sequencing. However, four of the *STRC* exons are also found on the "Sanger Dead Zone" list and an alternative method, such as long-range polymerase chain reaction, should be used to avoid pseudogene interference.[18]

a



	Homology type	%Match	Query region	Query coordinates	Target region	Target coordinates	psiDR pseudogene
1	Same gene	99.52%	TTN exon 194	chr2:179518479-179518689	TTN exon 176	chr2:179527000-179527210	–
2	Different gene	98.33%	MYH6 exon 33	chr14:23853586-23853991	MYH7 exon 34	chr14:23884199-23884502	–
3	Pseudogene	99.59%	STRC exon 17	chr15:43899996-43900238	–	chr15:43999470-43999712	ENST00000509801.1
4	Non-CDS	99.55%	OTOA exon 21	chr16:21742093-21742316	–	chr16:22558219-22558442	–

b

Homology type	NGS problem list - low stringency genes affected (%)
Same gene	348/2,731 (12.7)
Different gene	1,626/2,731 (59.5)
Pseudogene	1,530/2,731 (56.0)
Non-CDS	1,406/2,731 (51.5)

Figure 3 Annotation of homology type. (**a**) Schematic overview of classification into four homology types—same gene, different gene, pseudogene, and non-CDS (1–4). An example of a local alignment search is shown using a homologous exon (black) as the query that returns a match to each of the four genomic contexts (gray, 1–4). ψ indicates an annotated pseudogene. An actual match from our analysis is shown for each homology type in the table. Note that exon numbering is arbitrary and does not correspond to any specific transcript. (**b**) Table listing the percentage of genes with each homology type annotation in the "NGS Problem List–Low Stringency." Note that individual genes may be affected by multiple types of homology.

to highlight well-known genes with established roles in disease. Efforts to fully and deeply curate all gene–disease relationships are underway by the Clinical Genome Resource (https://clinicalgenome.org/), but this level of curation is beyond the scope for our work. However, we curated this set of genes as a first step for understanding the clinical importance of the highly homologous genes identified in our analysis, with the understanding that this curation will need to be frequently updated as new knowledge on gene–disease relationships emerges. Of the 193 genes, 85 (44.0%) were well-established disease genes (evidence level 3, **Supplementary Figure S1** online). Additionally, genes with definitive gene–disease associations from the "NGS Dead Zone" and the "Sanger Dead Zone" lists are presented in **Figure 2** and sorted by the percentage of affected positions. These genes are significant contributors to prevalent and severe genetic disorders, including *PMS2* (colon cancer), *STRC* (hearing loss), and *TTN* (dilated cardiomyopathy). Three genes on the "NGS Problem List–High Stringency" (**Supplementary Table S8** online)—*MYH7*, *PMS2*, and *SDHC*—appear on the ACMG incidental findings list,[15] with the recommendation that they be analyzed during exome or genome sequencing. Other genes with high homology frequently tested in clinical practice include *PKD1* (autosomal dominant polycystic kidney disease), *CYP2D6* (pharmacogenetics), and *CHEK2* (cancer predisposition).

Categorization by type of homology

Although all highly homologous sequences pose the same challenges when reads are aligned to the reference genome, the specific genomic context of the affected regions can have significantly different implications for assays that are chosen to supplement NGS and allow for comprehensive interrogation. The subsequent sections describe four common scenarios, their implications for clinical testing, and possible solutions. These solutions may not fit every laboratory's needs because operational implementation of NGS assays can differ, but they provide a template for the clinical testing of genes with these specific difficult genomic contexts. We identified the locations of the sequences throughout the genome homologous to each gene on our lists and classified these regions to provide a homology-type annotation (**Figure 3**). It is important to note that a gene may be affected by multiple types of homology due to the presence of multiple highly homologous sequences or due to a single highly homologous sequence that closely matches multiple regions of differing homology types.

Intragenic homology (same gene). In this case, the genomic context for homology resides within the CDS of the same gene (**Figure 3**). An example is the Nebulin (*NEB*) gene, for which an 8-exon segment (exons 82–105) is triplicated with nearly 100% sequence identity between the repeated blocks. For this type of intragenic large tandem repeat, commonly used approaches to discriminate homologous sequences (e.g., long-range polymerase chain reaction) are inappropriate. Solutions include gene-aware bioinformatic filtering that applies a different allelic fraction threshold for the gene that recognizes the potential for a heterozygous genotype in any one of the repeated regions to result in a reduced percentage of variant reads due to read misalignment. Other solutions include visual analysis of NGS read alignments or Sanger sequencing traces to identify the presence of variant alleles at low-read fractions, although any

variant detected through this approach could not be specifically assigned to any one of the triplicated regions.

Homology to functional genes (different gene). In the case of high homology between paralogous functional genes, both may need to be analyzed if they are relevant for the same disorder (**Figure 3**). One example is the cardiac muscle alpha and beta heavy chain genes, *MYH6* and *MYH7*. Both are associated with hypertrophic cardiomyopathy, are located adjacent to each other on chromosome 14, and share significant homology of exon 26 such that NGS read mapping is problematic. However, the high homology between the two genes is restricted to the coding sequences and does not extend into the introns. In such cases where homology is restricted to coding sequences, it may be possible to design a unique Sanger sequencing assay to resolve variant calls post-NGS.

Homology to nonfunctional pseudogenes (pseudogene). Pseudogene sequence can be seen as a contaminant and possible source of assay interference because unique analysis of the functional gene is desired (**Figure 3**).[16] One approach to enable clinical testing of genes in this category is to use long-range polymerase chain reaction sequencing assays to uniquely amplify the gene of interest. This approach has been used successfully for the analysis of *PKD1* in the diagnosis of autosomal dominant polycystic kidney disease[17] and *STRC* in the diagnosis of autosomal recessive sensorineural hearing loss.[18]

Homology to other sequences not annotated above (non-CDS). These regions of homology fall outside the exonic CDS and annotated pseudogenes (**Figure 3**). High homology to these sequences can be potentially overcome using a long-range polymerase chain reaction as described for annotated pseudogenes.

Resource availability

Our exon-level and gene-level lists are freely accessible on the precision FDA (https://precision.fda.gov/) and NCBI GeT-RM (http://www.ncbi.nlm.nih.gov/variation/tools/get-rm/) websites.

DISCUSSION

Exome sequencing allows for the large-scale analysis of genes without the intimate knowledge that comes with testing single genes or small gene panels. Highly homologous sequences are prevalent within the exome and have the potential to confound molecular diagnostic testing. We present a resource to empower clinical laboratories to recognize and overcome challenges presented by homology. This resource is primarily geared toward laboratory directors and probably has the greatest utility during the test design phase, but it can be used at all steps in the diagnostic testing process, including bioinformatic filtering to recognize homologous sequences that may result in a false-positive or false-negative variant call.

Our gene-level lists provide a quick reference to gauge possible challenges, design test content, and determine whether ancillary tests need to be developed for certain genes. These lists can be sorted by medical relevance and the degree of affectedness at the exon or base pair levels. Alternatively, specific genes can be queried by name. Use of the exon-level lists provides additional detail such as the number of positions affected within the exons as well as their precise location. The genomic coordinates can be interfaced with bioinformatic pipelines or used to generate tracks that will flag regions with high homology.

We compared our mappability-based approach to MQ, which is an output of commonly used read alignment algorithms such as BWA.[8,9] Mappability exclusively measures homology within the reference genome, whereas MQ is calculated empirically for each read within an NGS experiment and is influenced by multiple variables in addition to the presence of homologous sequences. It is critical to recognize that MQ scores are not designed to handle sequence variants and can be misleading in the presence of variants that increase the percent identity of a read with a homologous region. Although our mappability-based approach and MQ should be considered distinct metrics that are not redundant or interchangeable, there is substantial overlap in a global sense due to the shared effect of homology. The agreement that we observed between these measures validated that our method could identify problematic portions of the exome that are at risk for homology interference before performing an NGS assay. We believe that our mappability-based approach and MQ provide complementary measures that should be used together in NGS assay development and implementation.

The latest versions of many variant callers attempt to remedy homology issues by not calling variants in regions of low MQ. Although this approach minimizes but does not completely eliminate false-positive variant calls in regions of high homology, it also raises the possibility of false-negative results because actual variants in homologous regions may be filtered out. Moreover, if a clinical laboratory is not aware that reads in these regions are excluded due to low MQ scores, then there can be an illusion of comprehensive analysis, when in reality pertinent genomic regions remain ineffectively analyzed. Thus, laboratories should closely scrutinize homologous regions to avoid missing important variants and promising coverage of regions that cannot be accurately assessed. This is especially important for genes with high medical relevance, including those on the ACMG incidental findings list.[15] For other regions of lesser importance whose analysis is impaired by homology, a laboratory may choose to recognize the technical limitations of the assay and forego analysis of these loci.

Limitations

The gene curation performed as part of this resource is limited in scope and focused on a set of homologous, medically relevant genes that are likely to pose problems for NGS assays. Knowledge of medical relevance is continually expanding and quickly outdated, and information presented in this resource

should not be relied upon solely. Using the databases that were the basis for the data in this article, the list of (potentially) disease-associated genes has now grown to more than 7,000 genes; therefore, future curation will undoubtedly add critical disease genes to the analysis presented here. The data presented here should be regarded as a starting point for clinical laboratories, keeping in mind that that they will continue to evolve.

Future directions

We anticipate that future developments will bring new challenges and opportunities with regard to the problem of homology interference in NGS. The inclusion of additional sequences in the GRCh38 genome assembly may increase the number of sites affected by high homology and exacerbate the problems encountered in currently affected regions. Additionally, polymorphic pseudogenes have been identified that may compromise clinical testing in a small percentage of the population. For example, a processed pseudogene that is homologous to SMAD4 has been found in ~0.26% of the population.[19] Further studies are necessary to continue to elucidate the variability in homologous sequences among individuals. Long-read sequencing technologies, such as those developed by Pacific Biosciences and Oxford Nanopore Technologies, can largely overcome the problem of inaccurate alignment of homologous reads, although these platforms have not yet been implemented in many diagnostic laboratories. Additionally, synthetic long-read approaches can be used with current-generation sequencing instruments to try to circumvent problems posed by homology. Although these approaches and technologies hold promise for more accurate sequencing read alignments, the resource described herein should help laboratories using standard NGS and Sanger sequencing approaches to clinical testing minimize analytical errors due to homology interference.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

The authors acknowledge Partners HealthCare Research Information Services & Computing for providing and maintaining the infrastructure on which our computational analysis was performed.

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

1. Pei B, Sisu C, Frankish A, et al. The GENCODE pseudogene resource. *Genome Biol* 2012;13:R51.
2. Rehm HL, Bale SJ, Bayrak-Toydemir P, et al.; Working Group of the American College of Medical Genetics and Genomics Laboratory Quality Assurance Commitee. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med* 2013;15:733–747.
3. Aziz N, Zhao Q, Bry L, et al. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch Pathol Lab Med* 2015;139:481–493.
4. Bischof JM, Chiang AP, Scheetz TE, et al. Genome-wide identification of pseudogenes capable of disease-causing gene conversion. *Hum Mutat* 2006;27:545–552.
5. Chan WL, Yang WK, Huang HD, Chang JG. pseudoMap: an innovative and comprehensive resource for identification of siRNA-mediated mechanisms in human transcribed pseudogenes. *Database (Oxford)* 2013;2013: bat001.
6. Karro JE, Yan Y, Zheng D, et al. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res* 2007;35(Database issue):D55–D60.
7. Lam HY, Khurana E, Fang G, et al. Pseudofam: the pseudogene families database. *Nucleic Acids Res* 2009;37(Database issue):D738–D743.
8. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
9. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 2008;18:1851–1858.
10. Derrien T, Estellé J, Marco Sola S, et al. Fast computation and applications of genome mappability. *PLoS One* 2012;7:e30377.
11. Neph S, Kuehn MS, Reynolds AP, et al. BEDOPS: high-performance genomic feature operations. *Bioinformatics* 2012;28:1919–1920.
12. Kent WJ. BLAT–the BLAST-like alignment tool. *Genome Res* 2002;12:656–664.
13. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 2013;1303.3997v1([q-bio.GN]).
14. 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;43:11.10.1–11.1033.
15. Green RC, Berg JS, Grody WW, et al.; American College of Medical Genetics and Genomics. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* 2013;15:565–574.
16. De Vos M, Hayward BE, Picton S, Sheridan E, Bonthron DT. Novel PMS2 pseudogenes can conceal recessive mutations causing a distinctive childhood cancer syndrome. *Am J Hum Genet* 2004;74:954–964.
17. Tan YC, Michaeel A, Blumenfeld J, et al. A novel long-range PCR sequencing method for genetic analysis of the entire PKD1 gene. *J Mol Diagn* 2012;14: 305–313.
18. Mandelker D, Amr SS, Pugh T, et al. Comprehensive diagnostic testing for stereocilin: an approach for analyzing medically important genes with high homology. *J Mol Diagn* 2014;16:639–647.
19. Millson A, Lewis T, Pesaran T, et al. Processed pseudogene confounding deletion/duplication assays for SMAD4. *J Mol Diagn* 2015;17:576–582.