



Published in final edited form as:

Genet Med. 2018 August ; 20(8): 855–866. doi:10.1038/gim.2017.192.

Characterizing reduced coverage regions through comparison of exome and genome sequencing data across ten centers

Rashesh V. Sanghvi*,

Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX

Christian J. Buhay*,

Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX

Bradford C. Powell,

Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC

Ellen A. Tsai,

Laboratory for Molecular Medicine, Partners HealthCare Personalized Medicine, Cambridge, MA.

Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School

Michael O. Dorschner,

University of Washington, UW Medicine Center for Precision Diagnostics, and Department of Pathology

Celine S. Hong,

Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, MD, USA, Current affiliation: National Institute of Allergy and Infectious Diseases, NIH, Bethesda, MD, USA

Matthew S. Lebo,

Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Cambridge, MA.

Department of Pathology, Brigham & Women's Hospital and Harvard Medical School, Boston, Massachusetts

Ariella Sasson,

Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

Corresponding Authors: Donna Muzny, M.Sc, Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston TX 77030, Nikhil Wagle, MD, Dana-Farber Cancer Institute, 450 Brookline Avenue, Dana 820A, Boston, MA 02215.

*co-first author

**co-senior author

Financial Disclosures/Conflicts of Interest: L.G.B. is an uncompensated consultant for Illumina, receives royalties from Genentech, and honoraria from Wiley Blackwell. L.A.G. is a consultant for Foundation Medicine, Novartis, Boehringer Ingelheim, Third Rock; an equity holder in Foundation Medicine; and a member of the Scientific Advisory Board at Warp Drive. L.A.G. receives sponsored research support from Novartis, Astellas, BMS, and Merck. N.W. is a consultant for Novartis; is an equity holder in Foundation Medicine; and receives sponsored research support from Novartis, Genentech, and Merck.

SUPPLEMENTAL DATA

Supplemental Data include seven tables, three figures, and additional methods.

David S. Hanna,

University of Washington, UW Medicine Center for Precision Diagnostics, and Department of Pathology

Sean McGee,

Department of Genome Sciences, University of Washington

Kevin M. Bowling,

HudsonAlpha Institute for Biotechnology, Huntsville, AL

Gregory M. Cooper,

HudsonAlpha Institute for Biotechnology, Huntsville, AL

David E. Gray,

HudsonAlpha Institute for Biotechnology, Huntsville, AL

Robert J. Lonigro,

Department of Pathology, University of Michigan

Andrew Dunford,

Broad Institute of MIT and Harvard, Cambridge, MA

Christine A. Brennan,

Michigan Center for Translational Pathology, University of Michigan Medical School, Ann Arbor, MI 48109, USA

Carrie Cibulskis,

Broad Institute of MIT and Harvard, Cambridge, MA

Kimberly Walker,

Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX

Mauricio O. Carneiro,

Broad Institute of MIT and Harvard, Cambridge, MA

Joshua Sailsbery,

University of North Carolina Department of Genetics and Neurology

Lucia A. Hindorff,

Division of Genomic Medicine, National Human Genome Research Institute, National Institutes of Health

Dan R. Robinson,

Department of Pathology, University of Michigan

Avni Santani,

Perelman School of Medicine, University of Pennsylvania

Department of Path and Lab Medicine, Children's Hospital of Philadelphia

Mahdi Sarmady,

Division of Genomic Diagnostics, Department of Pathology & Lab Medicine, The Children's Hospital of Philadelphia

Heidi L. Rehm,

Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Cambridge, MA
Department of Pathology, Brigham & Women's Hospital and Harvard Medical School, Boston,
Massachusetts

Broad Institute of MIT and Harvard, Cambridge, MA

Leslie G. Biesecker,

Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute,
NIH, Bethesda, MD, USA

Deborah A. Nickerson,

Department of Genome Sciences, University of Washington

Carolyn M. Hutter,**

Division of Genomic Medicine, National Human Genome Research Institute, National Institutes of
Health

Levi Garraway,**

Center for Cancer Precision Medicine and Department of Medical Oncology, Dana-Farber Cancer
Institute, Boston, MA

Broad Institute of MIT and Harvard, Cambridge, MA

Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston,
MA

Howard Hughes Medical Institute, Chevy Chase, MD

Donna M. Muzny , and**

Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza Houston, TX

Nikhil Wagle**

Center for Cancer Precision Medicine and Department of Medical Oncology, Dana-Farber Cancer
Institute, Boston, MA

Broad Institute of MIT and Harvard, Cambridge, MA

Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston,
MA

on behalf of the NHGRI Clinical Sequencing Exploratory Research (CSER) Consortium

Abstract

PURPOSE: As massively parallel sequencing is increasingly being used for clinical decision-making, it has become critical to understand parameters that affect sequencing quality and to establish methods for measuring and reporting clinical sequencing standards. In this report, we propose a definition for *reduced coverage regions* and have established a set of standards for variant calling in clinical sequencing applications.

METHODS: To enable sequencing centers to assess the regions of poor sequencing quality in their own data, we optimized and used a tool (*ExCid*) to identify reduced coverage loci within

genes or regions of particular interest. We used this framework to examine sequencing data from 500 patients generated in ten projects from sequencing centers in the NHGRI/NCI Clinical Sequencing Exploratory Research (CSER) Consortium.

RESULTS: This approach identified reduced coverage regions in clinically relevant genes, including known clinically relevant loci that were uniquely missed at individual centers, in multiple centers, and in all centers.

CONCLUSIONS: This report provides a process roadmap for clinical sequencing centers looking to perform similar analyses on their data.

Keywords

Sequencing standards; clinical sequencing; exome; genome

INTRODUCTION

Exome and genome sequencing (ES and GS) using massively parallel sequencing (also known as next generation sequencing, NGS) is increasingly being implemented in clinical settings¹. At present there is the potential for wide variation in sequencing metrics between institutions, between samples sequenced at the same institution, and even within a single sample. Thus, as sequencing moves increasingly into the clinical arena, the application of these methods needs to be accompanied by the development of performance metrics and by an understanding of potential technical limitations of GS and ES as a clinical test. Indeed, the misnomers of “whole” exome and “whole” genome sequencing demonstrate that our field is communicating a confusing message to end-users – neither are truly whole. To this end, there has been increased focus on clinical sequencing standards, including the development of professional standards and guidelines for the use of NGS in clinical laboratories^{2–7}.

Under current recommendations, putative clinically relevant variants identified through NGS should be validated using Sanger sequencing or other orthogonal methods^{3,4}, although this practice has been challenged^{6,8,9}. In addition to knowing that positive results are accurate, clinicians and patients need information to accurately interpret a “negative” clinical sequencing result. This includes distinguishing when negative findings may be attributable to incomplete sequencing results. A key contributor to incomplete sequencing is reduced coverage in regions lacking sufficient high quality aligned bases for variant calling^{3,10}. Understanding the effects of reduced coverage requires a number of steps, including: a) setting definitions for “reduced coverage” regions that are not well represented in NGS results; b) establishing methods for measuring and reporting reduced coverage regions as part of clinical sequencing quality; and c) examining the potential impact of reduced coverage in the interpretation of clinically relevant regions of the genome.

The Clinical Sequencing Exploratory Research (CSER) Consortium, funded by the National Human Genome Research Institute and the National Cancer Institute, supports both the methods development needed to integrate sequencing into the clinic and the ethical, legal, and psychosocial research required to responsibly apply personal genomic sequence data to

medical care¹¹. The CSER Sequencing Standards Working Group (SSWG) worked to collectively establish a framework for identifying reduced coverage regions in the clinical sequencing setting. This report provides a summary of, and rationale for, the definitions and methods used in this framework. As a demonstration we examined clinical sequencing data on 500 patients generated between 2011 and 2015 from ten CSER centers (eight performing germline ES and two performing germline GS) and identified reduced coverage regions within and across projects. To provide clinical context, we examined reduced coverage regions in an exemplar gene list: 4,656 genes taken from the GeneTests database, a collection of genes for which a clinical test is available in a diagnostic lab (as of February 2015).

By presenting a framework for identifying reduced coverage bases, this report provides a process roadmap for other clinical sequencing centers looking to perform similar analyses on their data. This work summarizes factors, such as capture methods and GC content, that contribute to reduced coverage. In addition, the demonstrative analysis on 500 samples identifies regions of clinically relevant genes that appear to be universally difficult to sequence using Illumina-based sequencing technology. This highlights the importance of communicating sequencing standards in clinical reports and suggests that orthogonal or advanced methods may be needed to identify variants in some clinically relevant regions.

METHODS

Sequencing at Each Center

All human subjects provided informed consent to participate in these studies. Institutional Review Boards at each center approved their respective studies. Data sets analyzed in this study have been deposited in dbGaP, corresponding to the CSER studies at each center.

Generation of GeneTests Target Files

The February 2015 version of GeneTests, was obtained February 24 2015. Genomic coordinates for coding exons associated with transcripts of coding genes and genomic coordinates for all exons associated with non-coding genes were compiled.

Analysis of Reduced Coverage Regions

The Exome Coverage and Identification (*ExCID*) Report is a software tool to assess sequence depth in user-defined regions from read data (BAM file), annotate regions with gene, transcript and exon information and report intervals below a user-defined coverage threshold. For this study we consider a base to have reduced coverage if the base is covered <20X in at least 90% of the samples within each center. ExCID Version 2.1 was used for all analyses, and is available on GitHub '<https://github.com/cbuhay/ExCID>'.

Regions of Clinical Interest

We used two curated databases, the May-04-2015 release of ClinVar² and the 2015.1 release of Human Gene Mutation Database (HGMD)³, to assess if any of these reduced coverage regions in the GeneTests list contained clinically actionable variants. Variants in each database were separately intersected with reduced coverage regions identified in GeneTests

across eight ES centers, across two GS centers and across all ten centers using BEDTools¹. We also used the Aug-03-2015 release of Online Mendelian Inheritance in Man (OMIM) to relate reduced coverage regions identified in GeneTests to clinical phenotypes.

RESULTS

Clinical sequencing cohorts across ten institutions

To compare reduced coverage regions across multiple clinical sequencing projects, we collected data from ten centers conducting clinical sequencing research projects as part of the NHGRI CSER consortium. Each center provided sequencing data for a cohort of 50 patient samples sequenced in their respective projects. An overview of the sequencing approach taken in each project at the time of data collection is shown in Table 1. Eight of the ten centers used germline ES for their projects, while two centers used germline GS. All ten projects used Illumina sequencing, though the read length, targets captured, depth of coverage, and other parameters varied between projects (Table 1). Standard sequencing metrics for each cohort using these respective approaches is shown in Supplementary Table 1.

An Approach for Identifying Reduced Coverage Regions in Sequencing Data

To identify the reduced coverage regions in these projects, we first needed to establish definitions for regions that had insufficient coverage for variant calling, and therefore clinical utility. Although raw coverage counts (the number of reads at any given locus) is frequently used as a marker for usability, coverage is not the sole determinant of the ability to call variants. In addition to coverage, the assessment of a region requires base pair level inspection of the mapping quality (MQ) of the reads placed in this region and base quality (BQ) of the individual base pairs in each read. Therefore, we established the concept of *usable bases*: a high quality base (BQ \geq 20) that comes in a read with high mapping quality (MQ \geq 20) from a properly paired read. Setting quality thresholds at 20 assures there is a 99% probability that each base in each read are correctly called and uniquely mapped^{12,13}. We defined any site (locus) to be well covered if it contains at least 20 *usable bases* (\geq 20X coverage with BQ \geq 20 and MQ \geq 20). It is important to note that although raw coverage may exceed 20X, requiring 20 usable bases increases confidence that most germline variants are detected¹⁴ and that stochastic sequencing errors are not detected as false positive variants¹⁵. To focus on loci that were consistently unusable across multiple samples at any given sequencing center, we defined the *reduced coverage loci* as those that had less than 20 usable bases in at least 90% of the samples in that center's cohort.

To identify the reduced coverage loci in any collection of sequencing data, we used a tool called Exome Coverage and Identification (*ExCID*), which assesses sequence depth in user-defined targeted regions from read data (BAM file), annotates targets with gene, transcript, and exon information and reports intervals below a user defined coverage threshold. Input parameters (see Methods) include the input BAM files, the targets to interrogate, and the definitions of reduced coverage loci (i.e., more than 90% of samples in the cohort with less than 20x usable base coverage). Targets may be specified to include any regions of particular interest. The information reported includes coverage metrics and bases covered below the

coverage threshold for each sample in the cohort, exact bases that are reduced coverage, the length of the reduced coverage region, gene, transcript and exon information, and percentage of genes covered greater than the threshold. We further determined the GC content of each interval, position in the original target and mapability over the reduced coverage regions using standard approaches (see Methods).

Defining Critical Genes

While these tools could be applied to all sequencing data, a more practical approach might be to apply these sequencing standards to a list of critical genes for any given clinical application. Such a list might include the genes of interest in a particular clinical condition or a set of genes selected for interpretation by a clinical sequencing lab. To demonstrate the approach of identifying critical regions, we selected an exemplar gene list that might be representative of the types of genes a clinical sequencing lab would find of interest. For this exemplar gene list, we used a publicly available curated list of 4,656 genes for which a clinical test is available in a diagnostic lab, as registered in GeneTests (<https://www.genetests.org>) as of February 2015 (Supplementary Table 2).

To examine reduced coverage loci within this gene list, we first needed to convert the list of gene names to specific genomic coordinates that corresponded to coding regions of interest. We compiled the coordinates for the coding exons for the canonical isoform of each gene, using the RefSeq transcript annotated in the Human Gene Mutation Database (HGMD)¹⁶ (<http://www.hgmd.cf.ac.uk/ac/index.php>). The RefSeq and HGMD nomenclatures and genomic coordinates continued a standard that was already being used at the participating institutions; it adheres to the nomenclature guidelines in HUGO Gene Nomenclature Committee (HGNC; <http://www.genenames.org/guidelines.html>). We chose the CDS coordinates to include each coding gene. For the GeneTests gene list, the total size of the target regions was 9 Mbp. Sequencing metrics across the GeneTests genes for each cohort are shown in Supplementary Table 3.

Reduced Coverage Regions in GeneTests genes across ten centers

Using the *ExCID* tool and the GeneTests list, we surveyed the clinical sequencing data from each of the ten project cohorts. The goal of this analysis was to determine the reduced coverage regions—characterized as bases below 20x usable base coverage in 90% of the 50 samples in each cohort—specifically within the GeneTests genes.

The survey of reduced coverage regions in the 4,656 genes in the GeneTests list demonstrated some variability among the ten projects (Figure 1). The total reduced coverage bases in each of the ten projects ranged from 107 kb to 817 kb, comprising between 1.2% and 9.1% of the total coding bases contained within the GeneTests genes (Figure 1A). The number of reduced coverage exons (defined as exons containing at least one reduced coverage base) across the projects varied from 1,237 to 6,519, comprising between 1.8% and 9.6% of the total number of exons within the GeneTests list (Figure 1B). There were 533 exons (0.8%) that had reduced coverage at all ten centers. There was wider variation in the number of reduced coverage genes at each site (defined as genes containing at least one reduced coverage base), with a range of 526 to 2,816, comprising between 11.3% and 60.5%

of the 4,656 GeneTests genes (Figure 1C). 146 genes (3.1%) were affected by bases that had reduced coverage at all centers, totaling up to 66.4 kb (Supplementary Table 4). To test the robustness of results, we also used the *GATK DepthOfCoverage* tool on the same data sets, which yielded similar results (Supplementary Table 5)

This 66.4 kb region that had reduced coverage in all ten centers comprises 0.74% of the coding bases in the GeneTests lists. These loci represent the regions that had reduced coverage in 90% of the 50 samples from each of the eight ES projects as well as the two GS projects. Repeat analyses using 100 samples from nine centers demonstrated similar results, suggesting that this is not a function of sample size (Supplementary Table 6). Figure 1D illustrates a pairwise comparison of the centers, demonstrating both the percent overlap and absolute overlap in reduced coverage regions between any two centers. Greater correlation was seen between the two GS centers (I and J) and between the ES centers using similar capture designs.

Amongst the centers performing ES, 201,011 bp had reduced coverage at all eight centers (Figure 2A). Additional bases with reduced coverage uniquely at each ES center ranged from 171 bp to 205,293 kb. Figure 2B illustrates the number of bases, complete exons, and complete genes in the GeneTests list successfully covered at one or more centers using ES. 94% of the genes in GeneTests (4,370) were successfully covered in their entirety by at least one of the eight ES centers. Only 17% of the genes (814) were completely covered by all eight centers. 286 genes from GeneTests contained at least one region that had reduced coverage in all eight ES centers. The specific genes involved in these reduced coverage regions are listed in Supplementary Table 4. There were different regions that had uniquely reduced coverage by the ES centers and the GS centers. Of the 201kb that had reduced coverage by all eight ES centers, 134.7 kb were successfully sequenced in the two GS centers (Supplementary Figure 1). Conversely, of the 105 kb that had reduced coverage in both GS centers, 39.4 kb were successfully sequenced in at least one of the exome centers.

To assess potential clinical relevance, these reduced coverage regions were cross referenced with two curated databases of relationships between genetic variants and clinical phenotypes, ClinVar¹⁷ and HGMD¹⁶. The goal of this analysis was to determine if any of the reduced coverage regions contained any known disease-associated loci. Amongst 30,861 disease-associated variants in the ClinVar database, the reduced coverage bases at each sequencing center ranged from 93 to 1,323 (Figure 1E). Of the 146 GeneTest genes that had reduced coverage by all ten centers, 22 genes had a reduced coverage region that overlapped with a clinical variant in the HGMD or Clinvar databases (Table 2). Similarly, of the 286 GeneTest genes that had reduced coverage by the eight ES centers, 28 genes had a reduced coverage region that overlapped with a clinical variant in the HGMD or Clinvar databases (Table 2). Details of the phenotypes associated with these specific clinically relevant positions can be found in Table 2.

Characteristics of Reduced Coverage Regions and Exons

The reduced coverage bases across all centers were made up of 735 distinct contiguous intervals. 42% of these intervals (309) had lengths ranging from 1–5 bp (Figure 3A), and these short fragments were predominantly GC-rich (>70%). The remaining reduced

coverage intervals ranged in length from 5–490 bp and though they only made up 58% of the reduced coverage intervals, they accounted for 65.7 kb (97%) of the total bases for the reduced coverage regions in this analysis. A similar analysis for just the exome centers is shown in Supplementary Figure 2.

The 4,656 genes in the GeneTests list are made up of 67,759 exons. Of those, 533 exons had reduced coverage loci at all ten centers. The reduced coverage loci fell into two distinct groups. In the first group, less than 20% (and most often less than 10%) of the entire exon had reduced coverage bases. In the second group, most of the exon (>90%) had reduced coverage (Figure 3B). Exons with missing intervals spanning >90% of the entire coding sequence had lengths ranging from about 17 bp to over 2,112 bp, with the median length around 105 bp. 36% (194 of 533) of the problematic exons were in either the first or last exon of the gene, much higher than expected by chance alone, given that only 17% of the 67,759 exons are either first or last exons (Supplementary Table 7). The GC content also contributed to a large number of the reduced coverage loci, with 29% of the reduced coverage regions with a GC content >80%, as compared to 0.15% of the bases in GeneTests overall. (Figure 3C).

Potential Clinical Implications of Reduced Coverage Regions

One goal of identifying the reduced coverage regions in clinical sequencing data was to understand the clinical implications of incomplete sequencing of potentially relevant genes. Several reduced coverage regions were identified that could affect the molecular diagnosis of patients for a variety of phenotypes (Table 2; Supplementary Table 4). Many of these appeared to be due to issues with high similarity with other parts of the genome leading to an inability of reads to be uniquely aligned. For instance, the *STRC* gene has recently been revealed to be a major contributor to congenital sensorineural hearing loss; however, single nucleotide variants and small indels cannot be reliably detected via NGS due to *STRC* having 99.6% identity with the coding region of a nearby pseudogene^{18–20}. Similarly, current ES and GS approaches have a difficult time detecting pathogenic variants for adult onset polycystic kidney disease, a disease with a prevalence of 1/400 – 1/1000 and for which 85% of pathogenic variants identified are in *PKD1*²¹. *PKD1* is part of a genomic region that has been duplicated six times on chromosome 16, leading to an inability to accurately map reads over a large portion of the gene^{22–24}. For both *STRC* and *PKD1*, specific targeted sequencing strategies can be deployed to improve the coverage of these genes to address the presence of reduced coverage bases resulting from ES or GS.

Other clinically relevant genes that have reduced coverage likely have simpler solutions to accurately detect variation. *SHOX*, a gene associated with short stature, is part of the pseudo-autosomal regions (PAR) that occurs at the ends of both the X and Y chromosome. Since ES and GS typically align against the default human reference, this region is indicated as occurring on two different chromosomes (X and Y), thus mimicking a region with high similarity. Defining this region as its own chromosome (XY) would mitigate the mapability issues associated with this region.

DISCUSSION

Massively parallel sequencing data are increasingly being used for clinical decision-making. In these clinical contexts, it is critical to understand parameters that affect sequencing quality and to establish methods for measuring and reporting clinical sequencing results. In addition to knowing that data are accurate, clinicians and patients using clinical sequencing data need reassurance that the lack of clinically relevant findings are true negatives, and not due to inconclusive sequencing results. Moreover, it is important for everyone to understand that no test is 100% sensitive and that clinical decisions need to be made in light of a valid metric of sensitivity. Coverage metrics are a crucial component of the overall sensitivity of NGS.

In this report, we have proposed a definition for *reduced coverage regions* and have established a set of standards for variant calling in clinical sequencing applications. To enable sequencing centers to assess the regions of poor sequencing quality in their own data, we optimized a tool, *ExCID* (now publically available for use), which provides a list of reduced coverage loci within genes or regions of particular interest. To demonstrate an approach for examining reduced coverage regions in clinical sequencing data, we used these tools on clinical data generated in ten projects from different sequencing centers. This approach identified reduced coverage regions in clinically relevant genes, including known clinically relevant loci that were uniquely missed at individual centers, in multiple centers, or in all centers.

Comparing the reduced coverage regions across the various centers allowed categorization of the problematic regions and suggests possible solutions to improve standards. Reduced coverage regions that were unique to an individual center conducting ES were likely due to specific methodology at that center (Table 1). Choices about capture method, mean target coverage, or analysis pipeline may account for differences between centers. Since these parameters can be changed, regions that have reduced coverage only at a particular center can likely be salvaged by modifying the sequencing approach. For example, deeper sequencing is likely to reduce the number of reduced coverage regions that are unique to any single center (Supplementary Figure 3).

In contrast, loci that had reduced coverage in all or most of the ES centers but not by GS were likely due to difficulty with hybridization capture of the region. This might include difficulty generating appropriate baits for the relevant regions or difficulty with capture itself. Alternative bait design or orthogonal methods for sequencing these regions might help salvage these specific regions. Finally, there were certain regions in the genome that had reduced coverage at all ten centers, regardless of sequencing strategy. Although the total number of regions in this category were few, they did contain some potentially clinically relevant sites (Table 2). While it may be difficult to develop additional methods to salvage these regions, it is important for clinical sequencing centers to note the inability to provide conclusive sequencing information about these regions, especially those that may have clinical implications.

Centers conducting clinical sequencing can use the tools and approach described in this study to analyze their own clinical sequencing data, which may be particularly important for clinical laboratories conducting test validation. Once a list of genes or regions of interest have been identified, these can be converted to a list of coordinates (see Methods). Using these coordinates and a collection of representative sequencing data, centers can run *ExCID* on this data to identify those regions that have reduced coverage in 90% of the samples. By comparing these reduced coverage regions to the data presented in this study, centers may be able to better understand the reasons for poor coverage in their own data and develop potential salvage approaches, as detailed above.

While this study provides a generalized approach for analyzing clinical sequencing data for reduced coverage regions, there are a number of caveats. First, the data from each center represented specific cohorts from each project. These were intended to be exemplars to demonstrate the approach to analyzing data at an individual center and between multiple centers; they do not represent a comprehensive survey of human sequencing data. Second, although eight centers using ES data were represented, only two of our centers used GS data. Therefore, conclusions regarding the differences between GS and ES must be considered with that limitation in mind. Third, this analysis assumes that centers are performing germline sequencing to find heterozygous SNPs. For labs seeking to identify variants with a lower allelic fraction (for example, cases of germline mosaicism or somatic mutations), additional considerations will be relevant and standards proposed here are unlikely to be sufficient. Finally, all ten centers used Illumina-based sequencing approaches. At present, there are several other platforms available for clinical sequencing, which may provide different results. Moreover, the rapid evolution of platforms suggests that the specific data here may not be exactly reproduced using the most current technologies. Indeed, the technology used at most of the centers in this study has already evolved since the generation of data for this study. Nevertheless, though the data on reduced coverage regions at ten centers may not be representative of the breadth of technologies used today, the approach to analyzing reduced coverage regions remains platform agnostic, and can be applied to any clinical sequencing data.

When communicating results to clinicians and patients, sequencing centers need to be able to recognize and address reduced coverage regions. The tools and framework defined in this study can provide information about regions that have systematically reduced coverage at the center, and regions that have reduced coverage in any individual sample, which may also include additional regions that have sporadically reduced coverage. Both are important for quality control and communication of clinical sequencing data test results. It is important for centers to understand *systematic* reduced coverage regions to understand limitations of their clinical sequencing testing, and, when appropriate, to modify approaches to improve sequencing quality. At the individual level, regions that have reduced coverage in any particular sample may have direct clinical implications. Therefore, it is important that these regions be accurately communicated to the clinicians using the reports, so that they can correctly factor this into their decision-making. Communicating clinical sequencing data to clinicians and patients remains challenging, and including issues of sequencing standards adds further complexity to this—but is important. Further research into specific approaches to documenting and communicating these standards will be required.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Financial Support:

U01HG006500 and U41HG006834 (M.L. and H.L.R.)

NHGRI/NCI: U01 HG006507 and U01 HG007307 (M.O.D.)

NHGRI U01 HG006487 (B.C.P.)

NHGRI U01 HG006500 (E.A.T.)

5R01DA030976–04, 5U01HG006487–03, 1U19HD077632–02 (K.C.W.)

HG006507. HG007292 (D.A.N.)

NHGRI U01 HG006492 (L.A.G. and N.W.)

NHGRI/NCI 1U01HG00648, NHGRI U54-HG003273, and NIH/NHGRI 1UM1HG008898–01(C.J.B, R.V.S., K.W., and D.M.M.)

ZIA HG200359–08 (L.G.B. and C.S.H.)

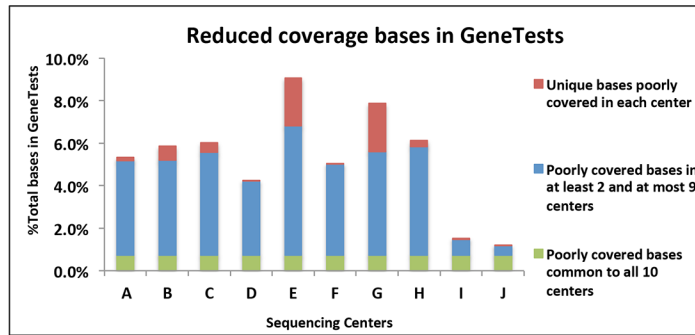
L.A.H. and C.M.H. are members of the NIH CSER staff team, responsible for scientific management of the CSER program.

REFERENCES

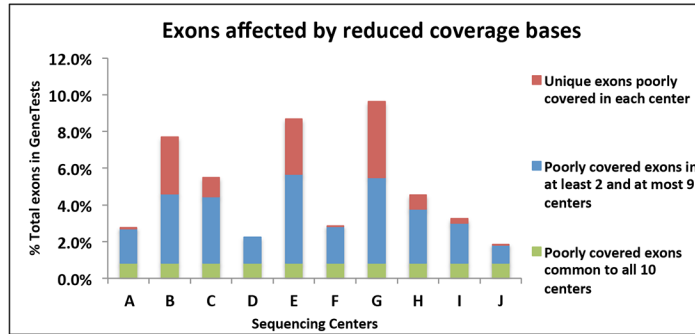
1. Biesecker LG & Green RC Diagnostic clinical genome and exome sequencing. *The New England journal of medicine* 371, 1170 (2014).
2. Brownstein CA, et al. An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. *Genome biology* 15, R53 (2014). [PubMed: 24667040]
3. Gargis AS, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nature biotechnology* 30, 1033–1036 (2012).
4. Rehm HL, et al. ACMG clinical laboratory standards for next-generation sequencing. *Genetics in medicine : official journal of the American College of Medical Genetics* 15, 733–747 (2013). [PubMed: 23887774]
5. Weiss MM, et al. Best practice guidelines for the use of next-generation sequencing applications in genome diagnostics: a national collaborative study of Dutch genome diagnostic laboratories. *Human mutation* 34, 1313–1321 (2013). [PubMed: 23776008]
6. Sikkema-Raddatz B, et al. Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. *Human mutation* 34, 1035–1042 (2013). [PubMed: 23568810]
7. Matthijs G, et al. Guidelines for diagnostic next-generation sequencing. *European journal of human genetics : EJHG* 24, 2–5 (2016). [PubMed: 26508566]
8. Beck TF, Mullikin JC, Program NCS & Biesecker LG Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. *Clinical chemistry* 62, 647–654 (2016). [PubMed: 26847218]
9. Strom SP, et al. Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. *Genetics in medicine : official journal of the American College of Medical Genetics* 16, 510–515 (2014). [PubMed: 24406459]
10. Ajay SS, Parker SC, Abaan HO, Fajardo KV & Margulies EH Accurate and comprehensive sequencing of personal genomes. *Genome research* 21, 1498–1505 (2011). [PubMed: 21771779]

11. Green RC, et al. Clinical Sequencing Exploratory Research Consortium: Accelerating Evidence-Based Practice of Genomic Medicine. *American journal of human genetics* 98, 1051–1066 (2016). [PubMed: 27181682]
12. Li H, Ruan J & Durbin R Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* 18, 1851–1858 (2008). [PubMed: 18714091]
13. Ewing B, Hillier L, Wendl MC & Green P Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research* 8, 175–185 (1998). [PubMed: 9521921]
14. Bainbridge MN, et al. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome biology* 12, R68 (2011). [PubMed: 21787409]
15. Parla JS, et al. A comparative analysis of exome capture. *Genome biology* 12, R97 (2011). [PubMed: 21958622]
16. Stenson PD, et al. Human Gene Mutation Database (HGMD): 2003 update. *Human mutation* 21, 577–581 (2003). [PubMed: 12754702]
17. Landrum MJ, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* 42, D980–985 (2014). [PubMed: 24234437]
18. Francey LJ, et al. Genome-wide SNP genotyping identifies the Stereocilin (STRC) gene as a major contributor to pediatric bilateral sensorineural hearing impairment. *American journal of medical genetics. Part A* 158A, 298–308 (2012).
19. Mandelker D, et al. Comprehensive diagnostic testing for stereocilin: an approach for analyzing medically important genes with high homology. *The Journal of molecular diagnostics : JMD* 16, 639–647 (2014). [PubMed: 25157971]
20. Vona B, et al. DFNB16 is a frequent cause of congenital hearing impairment: implementation of STRC mutation analysis in routine diagnostics. *Clinical genetics* 87, 49–55 (2015). [PubMed: 26011646]
21. Harris PC & Torres VE Polycystic Kidney Disease, Autosomal Dominant in GeneReviews(R) (eds. Pagon RA, et al.) (Seattle (WA), 1993).
22. Qi XP, et al. Genetic diagnosis of autosomal dominant polycystic kidney disease by targeted capture and next-generation sequencing: utility and limitations. *Gene* 516, 93–100 (2013). [PubMed: 23266634]
23. Loftus BJ, et al. Genome duplications and other features in 12 Mb of DNA sequence from human chromosome 16p and 16q. *Genomics* 60, 295–308 (1999). [PubMed: 10493829]
24. The polycystic kidney disease 1 gene encodes a 14 kb transcript and lies within a duplicated region on chromosome 16. The European Polycystic Kidney Disease Consortium. *Cell* 77, 881–894 (1994). [PubMed: 8004675]

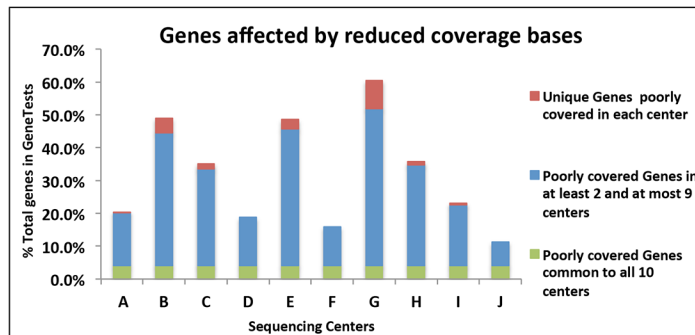
A



B

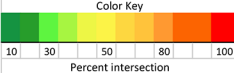


C



D

	A	B	C	D	E	F	G	H	I	J
A	100%	367,264	351,905	335,741	443,910	331,876	327,026	398,117	97,681	87,242
B	57%	100%	416,748	312,293	408,388	380,228	285,350	373,607	109,635	95,212
C	52%	64%	100%	302,147	381,694	450,918	266,054	351,960	116,012	100,421
D	64%	52%	49%	100%	360,041	287,300	286,028	379,122	76,977	70,721
E	52%	44%	39%	43%	100%	348,835	439,425	476,585	116,588	97,382
F	55%	63%	83%	53%	38%	100%	237,651	326,404	107,490	96,066
G	38%	30%	27%	36%	40%	26%	100%	398,314	105,864	90,696
H	63%	53%	47%	69%	53%	48%	19%	100%	109,728	94,118
I	19%	20%	21%	17%	14%	22%	14%	19%	100%	105,705
J	17%	18%	18%	17%	12%	21%	12%	17%	76%	100%

Color Key

 10 30 50 80 100
 Percent intersection

E

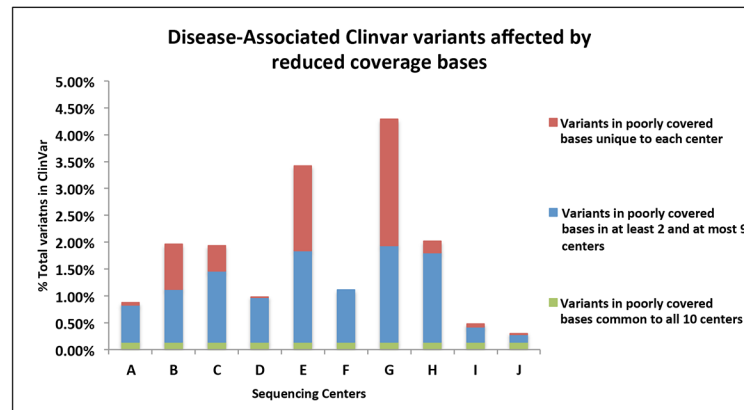


Figure 1. Reduced coverage regions in the GeneTests List.

(A) Comparison of reduced coverage bases among all centers. (B) Comparison of GeneTests exons affected by the reduced coverage bases among all centers. (C) Comparison of GeneTests exons affected by the reduced coverage bases among all centers. (D) Pairwise comparisons of reduced coverage regions between any two centers. Absolute values represent the reduced coverage bases common to two centers. Percentages represent the overlap in reduced coverage bases between two centers as compared to the union of reduced coverage bases at the two centers. High correlation existed between the two GS centers (I and J) and among the ES centers using same capture design. (E) Disease-associated ClinVar variants overlapping the reduced coverage bases in each center.

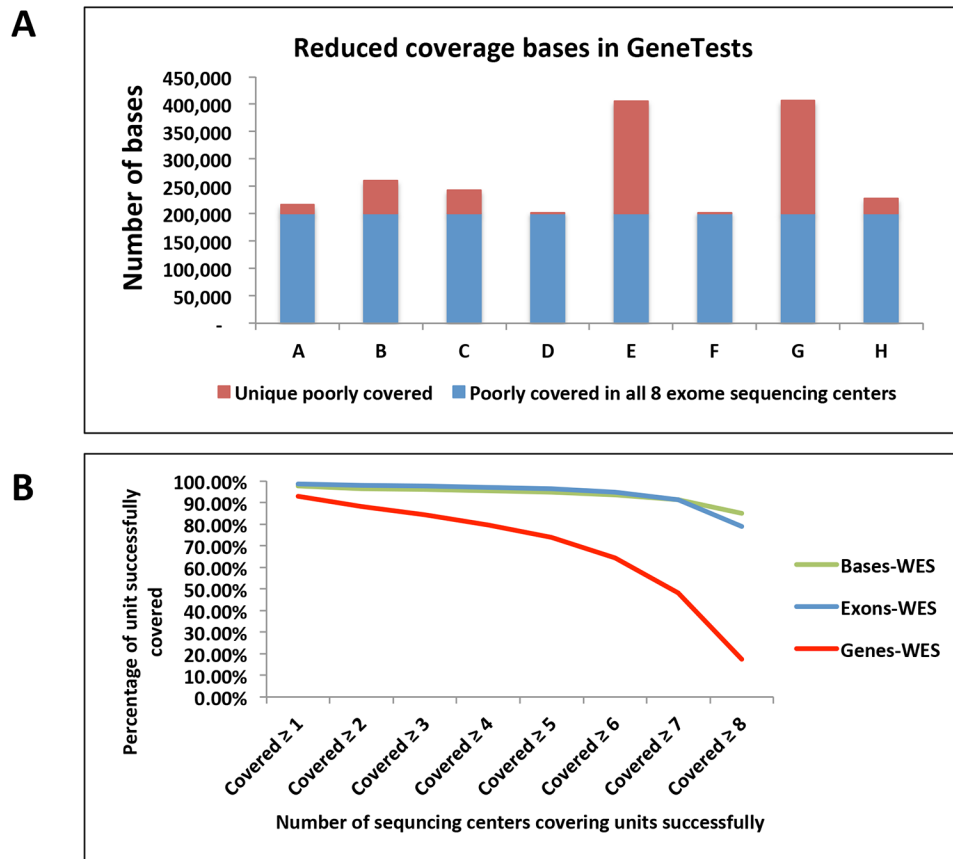
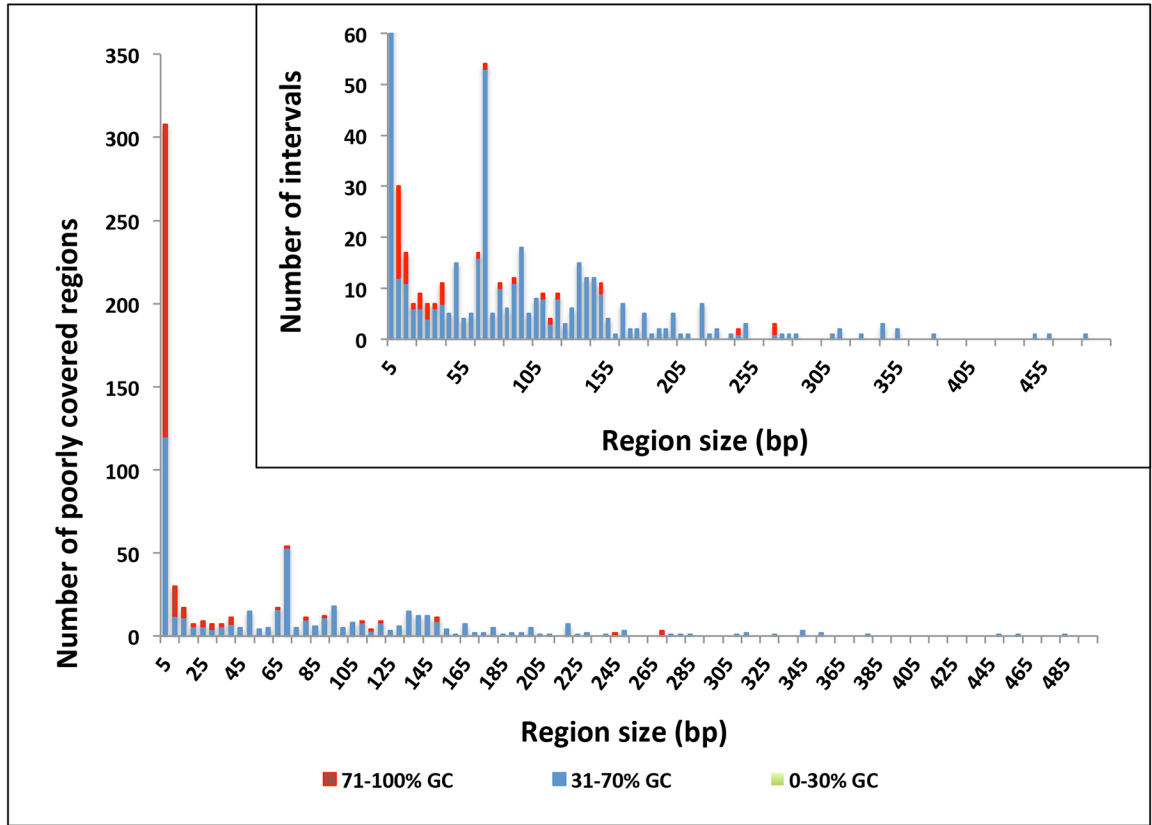


Figure 2. Reduced coverage bases in GeneTests.

(A) Comparison of reduced coverage regions among eight ES centers. (B) The percent of total bases, total whole exons, and total whole genes amongst the 4,656 GeneTests list successfully covered at one or more centers. To be included, every base in an exon or gene must have been a usable base (coverage $\geq 20X$, mapping quality ≥ 20 , base quality ≥ 20).

A



Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

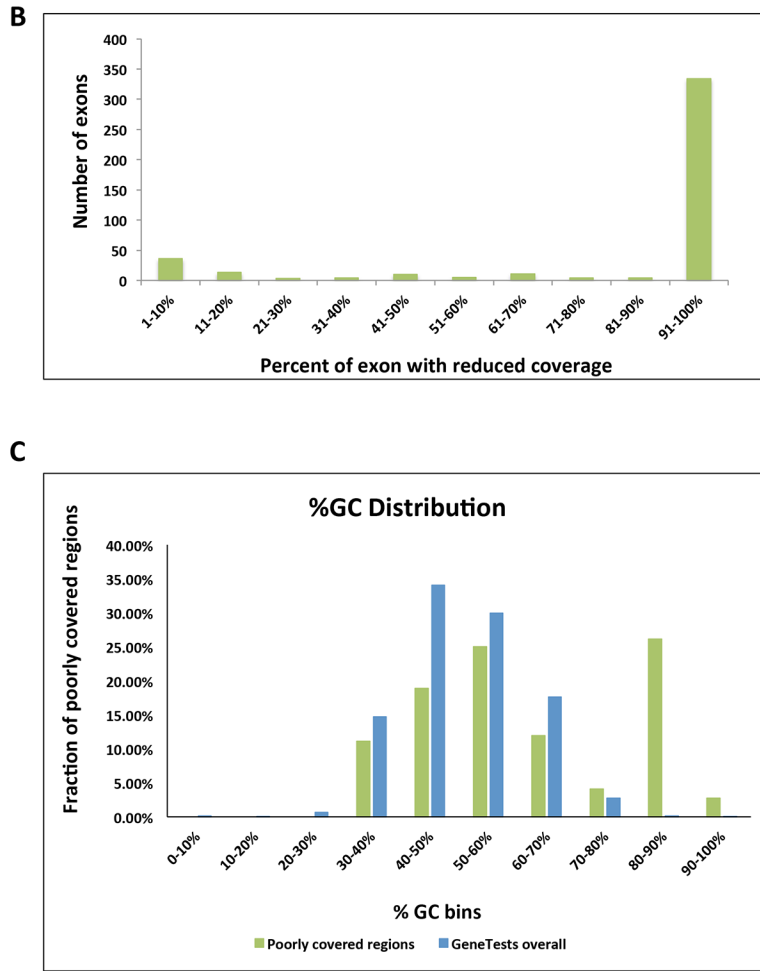


Figure 3. Analysis of Reduced Coverage Regions common to all Centers

(A) Overall, there were 735 missing intervals totaling 66.4 Kbp in the intersection of exon and genome centers. Forty-two percent of all missing intervals had lengths that were 5 bp or shorter. The remainder of missing intervals lengths ranged widely and occurred with less frequency, but they accounted for 65.7 Kbp or 97.7% of the total length of all missing intervals combined. (B) Of the >67K exons accounting for 4,656 genes in GeneTests, 533 had reduced coverage regions. These regions fell into two distinct groups—either a small part (<20%, with the vast majority less than 10%) of the entire exon had reduced coverage, or most of an exon (>90%) had reduced coverage. (C) Comparison of GC% distribution between the GeneTests baseline and the reduced coverage regions in all centers.

Overview of the sequencing approach used at each center. The sample sizes used for this analysis is 50 patients per center. As shown, eight of the ten centers used exome sequencing (ES) for their projects, while two centers used genome sequencing (GS).

Table 1

	SITE A	SITE B	SITE C	SITE D	SITE E	SITE F	SITE G	SITE H	SITE I	SITE J
Approach	ES	ES	ES	ES	ES	ES	ES	ES	GS (PCR-free)	GS (PCR-free)
Source	Blood	Blood	Blood	Blood	Blood	Blood	Blood	Blood	Blood	Blood
Targets / Capture Set	HGSC VCRome2.1 custom exome chip	Agilent SureSelect Human All Exon V5	Agilent SureSelect Human All Exon V4	Agilent SureSelect Human All Exon V4	Agilent SureSelect Human All Exon V2	Agilent SureSelect Human All Exon V4	Illumina TruSeq	NimbleGen SeqCap EZ Exome v3.0	N/A	N/A
Sequencing Platform	Illumina	Illumina	Illumina	Illumina	Illumina	Illumina	Illumina	Illumina	Illumina	Illumina
Read Length	2 × 101 bp	2 × 101 bp	2 × 101 bp	2 × 101 bp	2 × 101 bp	2 × 125 bp	2 × 101 bp	2 × 101 bp	2 × 101 bp	2 × 101 bp
Mb Target	35.4	50.6	51.3	64	33	51	61.8	64	N/A	N/A
Approximate Goal Mean Target Coverage	250X	40X	100X	100X	75X	150X	100X	65X	30X	30x
Input DNA / Sample	1 microgram	1 microgram	6 micrograms	1 microgram	0.1 micrograms	3 micrograms	1 microgram	2.5 micrograms	1.1 micrograms	1 microgram
Libraries / Sample	1	1	1	1	1	1	1	1	1	1
Captures / Sample	1	1	1	1	1	1	1	0.33	N/A	N/A
Lanes / Sample	0.33	0.25	0.25	1	0.3	0.4	0.33	0.25	3	3
Aligner	BWA	BWA	NovoAlign	BWA	BWA	NovoAlign	Novoalign	BWA	BWA	BWA
Variant Caller	ATLAS	GATK	GATK	GATK	GATK	VarScan	MPG	GATK	GATK	GATK

Table 2

Genes containing reduced coverage bases found in all ten centers or all eight ES centers that overlap with known disease-associated variants in ClinVar or HGMD.

Genes	Genes with at least one reduced coverage base common to all ten centers	Genes with at least one reduced coverage base common to all eight ES centers	Clinvar or HGMD variant overlapping with reduced coverage bases	Present in OMIM	OMIM phenotype
ACAN	✓	✓	✓	✓	Spondyloepiphyseal dysplasia, Kimberley type, 608361 (3); Spondyloepimetaphyseal dysplasia, aggrecan type, 612813 (3); Osteochondritis dissecans, short stature, and early-onset osteoarthritis, 165800 (3)
ARX	✓	✓	✓	✓	Epileptic encephalopathy, early infantile, 1, 308350 (3); Lissencephaly, X-linked 2, 300215 (3); Mental retardation, X-linked 29 and others, 300419 (3); Proud syndrome, 300004 (3); Partington syndrome, 309510 (3); Hydranencephaly with abnormal genitalia, 300215 (3)
B3GALT6	✓	✓	✓	✓	Spondyloepimetaphyseal dysplasia with joint laxity, type 1, with or without fractures, 271640 (3); Ehlers-Danlos syndrome, progeroid type, 2, 615349 (3)
CFC1	✓	✓	✓	✓	Heterotaxy, visceral, 2, autosomal, 605376 (3); Double-outlet right ventricle, 217095 (3); Transposition of the great arteries, dextro-looped 2, 613853 (3)
COL5A1		✓	✓		No Reported Phenotype
CRI	✓	✓	✓	✓	CR1 deficiency (1); {?SLE susceptibility} (1); [Blood group, Knops system], 607486 (3); {Malaria, severe, resistance to}, 611162 (3)
CTF1	✓	✓	✓	✓	No Reported Phenotype
EVC	✓	✓	✓	✓	Ellis-van Creveld syndrome, 225500 (3); Weyers acrorenal dysostosis, 193530 (3)
GALNT12	✓	✓	✓	✓	{Colorectal cancer, susceptibility to, 1}, 608812 (3)
HBA2	✓	✓	✓	✓	Thalassemia, alpha-, 604131 (3); Heinz body anemia, 140700 (3); Erythrocytosis (3); Hypochromic microcytic anemia (3); Hemoglobin H disease, nondeletional, 613978 (3)
HSD11B2		✓	✓		No Reported Phenotype
IKBKG	✓	✓	✓	✓	Incontinentia pigmenti, 308300 (3); Ectodermal dysplasia, hypohidrotic, with immune deficiency, 300291 (3); Ectodermal, dysplasia, anhidrotic, lymphedema and immunodeficiency, 300301 (3); Immunodeficiency, isolated, 300584 (3); Immunodeficiency 33, 300636 (3); Invasive pneumococcal disease, recurrent isolated, 2, 300640 (3)
KCNQ1		✓	✓		No Reported Phenotype
LRP5	✓	✓	✓	✓	Osteoporosis-pseudoglioma syndrome, 259770 (3); [Bone mineral density variability 1], 601884 (3); Hyperostosis, endosteal, 144750 (3); van Buchem disease, type 2, 607636 (3); Osteosclerosis, 144750 (3); {Osteoporosis}, 166710 (3); Exudative vitreoretinopathy 4, 601813 (3); Osteopetrosis, autosomal dominant 1, 607634 (3)

Genes	Genes with at least one reduced coverage base common to all ten centers	Genes with at least one reduced coverage base common to all eight ES centers	Clinvar or HGMD variant overlapping with reduced coverage bases	Present in OMIM	OMIM phenotype
MNX1	✓	✓	✓	✓	Currarino syndrome, 176450 (3)
NAGLU		✓	✓		No Reported Phenotype
OPN1LW	✓	✓	✓	✓	Colorblindness, protan, 303900 (3); Blue cone monochromacy, 303700 (3)
OPN1MW	✓	✓	✓	✓	Colorblindness, deutan, 303800 (3); Blue cone monochromacy, 303700 (3)
OPN1MW2	✓	✓	✓		No Reported Phenotype
PKD1	✓	✓	✓	✓	Polycystic kidney disease, adult type I, 173900 (3)
RPS17	✓	✓	✓	✓	Diamond-Blackfan anemia 4, 612527 (3)
SEPN1	✓	✓	✓	✓	Muscular dystrophy, rigid spine, 1, 602771 (3); Myopathy, congenital, with fiber-type disproportion, 255310 (3)
SGCB		✓	✓		No Reported Phenotype
SHOX	✓	✓	✓	✓	Short stature, idiopathic familial, 300582 (3); Leri-Weill dyschondrosteosis, 127300 (3); Langer mesomelic dysplasia, 249700 (3)
SMN1	✓	✓	✓	✓	Spinal muscular atrophy-1, 253300 (3); Spinal muscular atrophy-2, 253550 (3); Spinal muscular atrophy-3, 253400 (3); Spinal muscular atrophy-4, 271150 (3)
SMN2	✓	✓	✓	✓	{Spinal muscular atrophy, type III, modifier of}, 253400 (3)
STRC	✓	✓	✓	✓	Deafness, autosomal recessive 16, 603720 (3)
TNFRSF11A		✓	✓		No Reported Phenotype