Good Laboratory Practice for Clinical Next-Generation Sequencing Informatics Pipelines

Supplementary Principles and Recommendations

Authors: Amy S. Gargis[1,2*], Lisa Kalman[1], David P. Bick[3], Cristina da Silva[4], David P. Dimmock[3], Birgit H. Funke[5,6], Sivakumar Gowrisankar[5,6,7*], Madhuri R. Hegde[4], Shashikant Kulkarni[8,9,10], Christopher E. Mason[11], Rakesh Nagarajan[10], Karl V. Voelkerding[12,13], Elizabeth A. Worthey[3], Nazneen Aziz[14,15*], John Barnes[16], Sarah F. Bennett[17], Himani Bisht[18], Deanna M. Church[19,20*], Zoya Dimitrova[21], Shaw R. Gargis[22], Nabil Hafez[23,24*], Tina Hambuch[25], Fiona C.L. Hyland[26], Ruth Ann Luna[27], Duncan MacCannell[28], Tobias Mann[29,30*], Megan R. McCluskey[31], Timothy K. McDaniel[32], Lilia M. Ganova-Raeva[21], Heidi L. Rehm[5,6], Jeffrey Reid[33.34*], David S. Campo[21], Richard B. Resnick[23], Perry G. Ridge[12,35*], Marc L. Salit[36], Pavel Skums[21], Lee-Jun C. Wong[33], Barbara A. Zehnbauer[1], Justin M. Zook[36], Ira M. Lubin[1]

[1]Division of Laboratory Systems, Centers for Disease Control and Prevention, Atlanta GA, USA. [2]Division of Preparedness and Emerging Infections, Centers for Disease Control and Prevention, Atlanta, GA, USA. [3]Department of Pediatrics, Medical College of Wisconsin, Milwaukee, Wisconsin, USA. [4]Department of Human Genetics, Emory University School of Medicine, Atlanta GA, USA. [5]Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Cambridge, Massachusetts, USA. [6]Department of Pathology, Harvard Medical School, Boston, Massachusetts, USA. [7]Novartis Institutes for Biomedical Research, Cambridge, Massachusetts, USA. [8]Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA. [9]Department of Pediatrics, Washington University School of Medicine, St. Louis, Missouri, USA. [10]Department of Pathology and Immunology, Washington University School of Medicine, USA. [11]Department of Physiology and Biophysics, Cornell University,

Nature Biotechnology: doi:10.1038/nbt.3237

New York, New York, USA. [12]Department of Pathology, University of Utah, Salt Lake City, Utah, USA. [13]Institute for Clinical and Experimental Pathology, Associated Regional and University Pathologists (ARUP) Laboratories, Salt Lake City, Utah, USA. [14] College of American Pathologists, Northfield, Illinois, USA. [15]Phoenix Children's Hospital, Phoenix, Arizona, USA. [16]National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA. [17]Division of Laboratory Services, Centers for Medicare and Medicaid Services, Baltimore, Maryland, USA. [18]Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, Maryland, [19]National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland, USA. [20]Personalis, Menlo Park, California, USA. [21]Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, Georgia, USA. [22]Division of Select Agents and Toxins, Centers for Disease Control and Prevention, Atlanta, Georgia, USA. [23]GenomeQuest, Westborough, Massachusetts, USA. [24]Neurology, Quest Diagnostics, Marlborough, Massachusetts, USA. [25]Clinical Services, Illumina, San Diego, California, USA. [26]Thermo Fisher Scientific, South San Francisco, California, USA. [27] Texas Children's Microbiome Center, Texas Children's Hospital and Department of Pathology & Immunology, Baylor College of Medicine, Houston, Texas, USA. [28]National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA USA. [29]Illumina, San Diego, California, USA. [30]Progenity, Ann Arbor, Michigan, USA. [31]SoftGenetics, State College, Pennsylvania, USA. [32]Oncology, Illumina, San Diego, California, USA. [33]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA. [34]Regeneron Pharmaceuticals, Tarrytown, New York, USA. [35]Department of Biology, Brigham Young University, Provo,

Nature Biotechnology: doi:10.1038/nbt.3237

Utah, USA. [36]Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland, USA.

*The following author affiliations have changed during the course of this work: Division of Preparedness and Emerging Infections, Center for Disease Control and Prevention, Atlanta, Georgia, USA (A.S.G.); Novartis Institutes for Biomedical Research, Cambridge, Massachusetts, USA (S.G.); Phoenix Children's Hospital, Phoenix, Arizona, USA (N.A.); Personalis, Menlo Park, California, USA (D.M.C.); Quest Diagnostics, Marlborough, Massachusetts, USA (N.H.); Progenity, Ann Arbor, Michigan, USA (T.M.); Regeneron Pharmaceuticals, Tarrytown, New York, USA (J.R.); and Brigham Young University, Provo, Utah, USA (P.G.R.).

Disclaimers: The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the Centers for Disease Control and Prevention , the Agency for Toxic Substances and Disease Registry, or the Food and Drug Administration. Certain commercial equipment, instruments, or materials are identified in this document. Such identification does not imply recommendation or endorsement by the Centers for Disease Control and Prevention the Agency for Toxic Substances and Disease Registry, or the Food and Drug Administration nor does it imply that the products identified are necessarily the best available for the purpose. The identification of certain commercial equipment, instruments or materials in this document does not imply recommendation or endorsement by the US National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose.

Nature Biotechnology: doi:10.1038/nbt.3237

Table of Contents

Nature Biotechnology: doi:10.1038/nbt.3237

1. Introduction:

Next-Generation Sequencing (NGS) technologies produce data that require substantial computational infrastructure for data storage, analysis, and interpretation. These rapidly evolving technologies are utilized in the clinical setting by a growing number of laboratories to analyze gene panels, exomes, and whole genomes by a variety of genetic disorders. NGS assays utilize complex testing algorithms, which require laboratory-based sequencing and computational processes to generate a final result.

Data generated by NGS technologies are analyzed in a series of steps. Millions to billions of sequence reads are initially generated by the sequencing platform (primary analysis). The combination of NGS data analysis tools, referred to as the "informatics pipeline," processes and analyzes the raw data generated by the sequencing instrument to produce a report. A variety of software tools have been developed to analyze NGS data for specific contexts or applications. As of 2015, most laboratories use custom-built combinations of commercial or publicly available tools in addition to in-house developed analytic procedures to develop their informatics pipelines.

In the next step of data analysis (secondary analysis), NGS reads are aligned to a reference that is needed to identify where sequence variants exist. If multiple patient samples are pooled (multiplexed) prior to sequencing, the reads associated with each patient are separated before alignment to the reference assembly and analyzed independently. Read alignments are then systematically examined and often re-aligned in local regions of the genome to remove artifacts and ensure accurate genotype calls. Finally, these alignments are used to identify differences between the patient's sequence and the reference assembly in a process referred to as variant calling. The end product of this "secondary analysis" is the identification of sequence variants

Nature Biotechnology: doi:10.1038/nbt.3237

that are stored in a variant call file. Between 3 and 4 million sequence variants are typically detected when the entire human genome is sequenced and aligned to a reference sequence[1]. These include single nucleotide variants (SNVs), small insertions, deletions, and their combination (indels). The detection of other variants that include copy number variations, complex variants, and structural rearrangements can be problematic and often requires specialized analysis

Additional analysis of the variant found is performed to identify those that are relevant to the patient's clinical condition (tertiary analysis). Variants are first annotated with predicted molecular consequences (such as the creation of a premature stop codon by a nonsense or frameshift variant, or the substitution of an amino acid by a missense variant). This permits filtering of those not expected to be clinically relevant. Additional annotations are added that includes what is known about disease association, population prevalence, and other factors. A clinical assessment is next performed to identify variants relevant to the patient's medical condition. As of 2015, this evaluation is a labor-intensive process, which often includes both automated analysis and manual review by an expert able to evaluate the available literature and other data sources. The understanding of the spectrum of pathogenic variants that exists in the human population is limited; however, when clinical information is available, it is used during the analysis to prioritize variants that are potentially disease-associated. The final step is the development of a test result report, which integrates the findings from the sequencing analysis with the patient's clinical data to determine whether any or a combination of the variants detected can explain the patient's disease. Much of primary, secondary and tertiary analyses involve substantial automated informatics components, which is a significant change in operations for many clinical molecular testing laboratories.

Nature Biotechnology: doi:10.1038/nbt.3237

Currently, most clinical NGS tests are offered as laboratory developed tests (LDTs), which are tests designed, manufactured and used within a single laboratory. These tests are carried out using commercially available sequencing platforms to generate raw sequence data that is subsequently analyzed using software algorithms (informatics pipeline). In the US, LDTs are subject to the Clinical Laboratory Improvement Amendments (CLIA) regulations, which require that laboratories introducing a test system not cleared or approved by the US Food and Drug Administration (FDA), establish analytical performance specifications of the assay for accuracy, precision, analytic sensitivity and specificity, and other measures, as relevant. In 2013, FDA cleared the Illumina MiSeqDX as a Class II Exempt device along with its associated reagent kit[2], and in 2014, two additional sequencing platforms (the Life Technologies Ion PGM Dx sequencers and the Vela Sentosa SQ301) were registered, listed, and can be marketed under the same regulation (http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm375742.htm, http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfRL/rl.cfm?lid=427645&lpcd=PFF, and http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfRL/rl.cfm?lid=430009&lpcd=PFF). However, laboratories using these instruments must still establish an informatics pipeline for the intended clinical application(s). The clinical test is therefore an LDT and requires validation to establish performance specifications under CLIA[3-7], even though the FDA has cleared the sequencing platform. The FDA also cleared two tests for diagnosis of cystic fibrosis using NGS. In these instances, no component of either test is laboratory developed and as such clinical laboratories need not validate these tests but they do need to verify that they can achieve the performance specifications established by the manufacturer. Development and optimization of

Nature Biotechnology: doi:10.1038/nbt.3237

these processes, especially the analytical methods, is non-trivial and requires specialized informatics expertise relevant to NGS.

The design of an informatics pipeline is more complex for exome and genomic analysis than for gene panels. This is because the analysis of gene panels is confined to a discreet set of selected genes. For exome and genome analysis, the pipeline must assess which genes are likely to be relevant to the indication for testing and this guides which variants will be considered as potentially clinically relevant, and more data must be analyzed.

The optimization of an informatics pipeline does not necessarily provide an ideal analysis pathway for each patient sample. In some instances, the test may not identify a clinically relevant variant. In this situation, the laboratory may reanalyze variant data using different filters or software settings within the confines of the test validation. This approach has been used to identify clinically relevant sequence variants that otherwise would have been missed.

Although optimization is a prerequisite for test validation, there is limited available guidance for the establishment and optimization of clinical NGS informatics pipelines. To address this, the Centers for Disease Control and Prevention (CDC) established a national workgroup to identify principles and develop recommendations (Table A, Letter of Correspondence) for the establishment and optimization of an NGS informatics pipeline that can be validated for clinical applications. This workgroup was developed as a result of a high priority recommendation from a previous CDC-facilitated national workgroup [Next Generation Sequencing- Standardization of Clinical Testing (Nex-StoCT) April, 2011] which formulated general quality guidance for the integration of NGS into clinical laboratory settings[3]. The current manuscript describes the guidelines and recommendations developed by the Next Generation Sequencing- Standardization of Clinical Testing II (Nex-StoCT II) Informatics

Nature Biotechnology: doi:10.1038/nbt.3237

Workgroup (see Letter of Correspondence, Table 1 and Supplementary Appendix). These guidelines and recommendations are intended to be used by clinical genetic testing laboratories when developing and optimizing a NGS informatics pipeline. This is especially important as many clinical laboratories do not have bioinformatics staff or expertise and need basic guidance.

2. Methods:

The CDC's Division of Laboratory Programs, Standards and Services (DLPSS) convened invited stakeholders for the Nex-StoCT II Informatics Workgroup meeting in October, 2012 in Atlanta, Georgia. The primary focus was the use of NGS for the detection of germline sequence variants; however, the workgroup also discussed some aspects unique to NGS applications for cancer and infectious disease testing. The workgroup addressed the analytic processes of NGS (e.g., demultiplexing, alignment, variant calling, annotation, etc.) and not issues associated with data representation or messaging, with respect to the electronic health record or the laboratory information system.

The topics discussed included:

- de-multiplexing
- sequence mapping and alignment
- variant calling
- variant annotation
- downstream processes for clinical interpretation, such as variant/gene prioritization (ranking), classification, and clinical integration into the test result report
- applicable metrics and controls for each of these topics

Nature Biotechnology: doi:10.1038/nbt.3237

Meeting participants, who were selected based on their expertise in areas included in the scope of the workgroup, included informatics experts, clinical and research laboratory professionals, test platform and software developers, and physicians with experience using NGS for patient care. The workgroup also included participants from several federal agencies tasked with promoting the quality of clinical NGS applications (Centers for Medicare & Medicaid Services, CMS; National Institute of Standards and Technology, NIST; National Institutes of Health, NIH; FDA, and CDC), members from an accrediting body (College of American Pathologists, CAP) and others holding leadership positions in professional organizations (e.g., American College of Medical Genetics and Genomics, ACMG; Association for Molecular Pathology, AMP; Clinical and Laboratory Standards Institute, CLSI; the Global Alliance for Genomics and Health (GA4GH) and the Genomes in a Bottle (GIAB) Consortium). The two-day meeting consisted of plenary, roundtable, and workgroup sessions designed to facilitate discussion, foster collaboration, and build agreement among participants. Following the meeting, participants were engaged in teleconference calls to complete and augment the discussions that began at the in-person meeting. The group focused on recommendations for SNV identification, insertions, deletions, and indels. The detection of copy number changes, structural variants, mosaicism, mitochondrial heteroplasmy, methylation, somatic variants in cancer and microbial genome sequencing were not extensively considered.

3. Analytical Workflow Overview: Primary, Secondary, and Tertiary Analysis

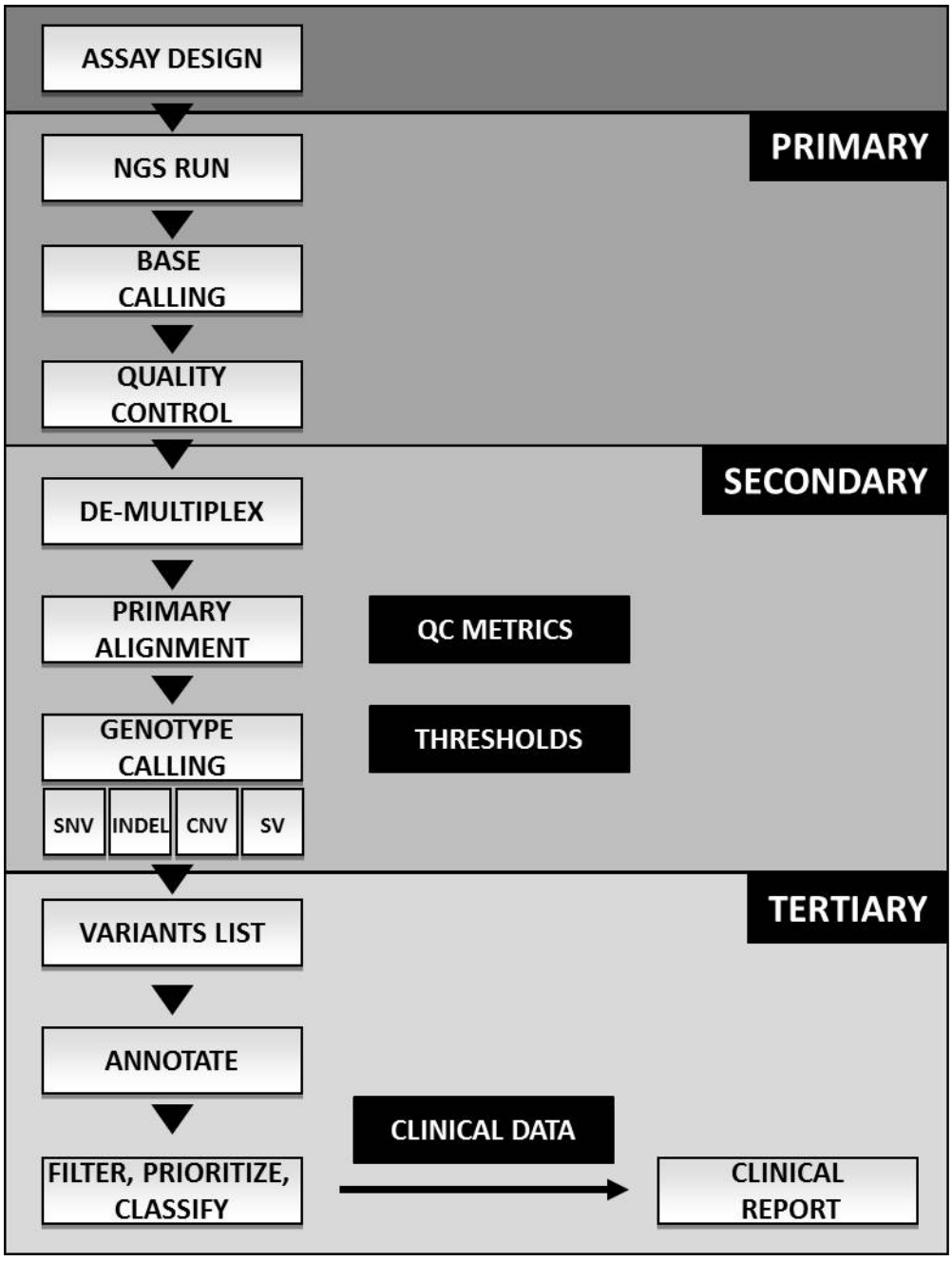
3.1 Primary Analysis

NGS testing is often divided into three distinct phases: primary, secondary, and tertiary analyses (Supplementary Figure 1). The design and optimization of an informatics pipeline

Nature Biotechnology: doi:10.1038/nbt.3237

primarily involves the secondary and tertiary analysis of the data derived after DNA sequencing is performed. The primary phase is largely developed by the vendor and is platform specific. Nonetheless, there are some features of primary analysis that should be considered when optimizing an informatics pipeline. Primary analysis takes place within the sequencing instrument and typically involves base calling to generate a file containing the set of sequence "reads" and associated base quality scores. Two common file formats are often used: FASTQ can store reads with per base quality scores [8, 9], and BAM, a binary alignment format file [10], can contain mapped and/or unmapped reads. BAM are binary files that are based upon the Sequence Alignment/Map (SAM) format and were developed to provide a common alignment format that supports aligners and downstream analyses (e.g., variant detection). The data undergo a quality control step, in which reads are filtered to remove those with base calls that do not meet vendor or laboratory-established quality criteria before alignment to the reference assembly. During this process, the 5' and/or 3' ends of reads are trimmed automatically by the sequencing instrument software or by manual adjustment according to established criteria because the base quality scores are often lower at the termini of each read. Sequencing instrument software does not routinely remove PCR primers. For targeted sequencing, primer sequences must be removed or soft-clipped (i.e. the primer sequence is retained for use in alignment, but is masked during variant calling) to ensure that SNPs within primer regions are called with the correct variant allele frequency. Additional recommendations and standards for primary analysis have been extensively described elsewhere[3, 5].

Nature Biotechnology: doi:10.1038/nbt.3237



Supplementary Figure 1: NGS workflow. NGS can be divided into three phases: primary, secondary, and tertiary analysis. Primary analysis includes sequencing to produce a set of reads with quality scores which are deposited into an electronic file. Quality control procedures are established to remove low quality reads from downstream analysis. During secondary analysis,

Nature Biotechnology: doi:10.1038/nbt.3237

pooled (multiplexed) reads are separated or "de-multiplexed" and associated with their respective patient samples. Reads are then mapped and aligned to a reference assembly. Secondary analysis results in the identification of sequence variants (e.g. SNV: single nucleotide variants, INDEL: insertions and deletions, CNV: copy number variants, and SV: structural variants, including duplications, deletions, inversions, and translocations of large >100 (nucleotides) blocks of DNA sequence), and positions that are the same as the reference sequence, and in some cases positions that cannot be assigned a genotype call with confidence, due to insufficient coverage or other reasons. Primary and secondary analyses are typically automated. Tertiary analysis involves annotation and clinical assessment of the sequence variants to identify those that are relevant to the patient and the clinical indication for testing. Tertiary analysis includes both automated and manual processes. Classification and interpretation of variants usually requires manual review, especially for exome and genome analysis.

The time it takes to perform a clinical NGS test varies greatly. Timing depends on the extent of the genome interrogated (e.g., panel, exome, or genome), the level of automation and procedures used for library preparation, sequencing method, alignment, variant calling, filtering, and clinical assessment. The number of skilled genomic variant analysts and the laboratory's computational capacity are also a factor. The clinical assessment can be the most time consuming component of the process. Most cases contain a number of variants that are not well described in the literature as disease associated, but may be associated with the patient's phenotype. Therefore, the time required for a clinical assessment of these variants can range from about 30 minutes to hours for each variant that requires manual review to make an educated assertion to the classification(benign, likely benign, variant of unknown significance (VUS), likely pathogenic,

Nature Biotechnology: doi:10.1038/nbt.3237

pathogenic). For heritable conditions, collection of phenotypic information and additional testing of family members are sometimes warranted. As of 2015, the time it takes to complete an NGS test can range from hours to weeks.

3.2 Secondary Analysis

Secondary analysis is the process of aligning reads to a reference sequence and generating variant calls. If multiple patient samples are pooled (each individual sample is separately tagged for identification; see Section 4.1), or multiplexed during the sequencing process, the resulting data must be de-multiplexed to separate patient sequence prior to analysis. This is followed by mapping and subsequent alignment of each read to a reference assembly. Reads may not align uniquely to the reference assembly for many reasons. For example, reads may align identically to multiple locations if they are derived from highly homologous regions of the genome[5]. Reads derived from loci not represented in the reference assembly, for example certain HLA alleles, may map to the HLA locus but fail to align. Alternatively, they may align to one or more off-target locations that can lead to the false positives. Regions that exhibit extreme allelic diversity, such as HLA, can also complicate alignment if the read derived from the sample is from a different haplotype than the sequence used in the reference assembly. For this reason, the GRC (http://genomereference.org, accessed August 18, 2014) includes alternate sequences paths in the assembly for regions with extreme diversity when there is data sufficiency. This approach does introduce allelic diversity, and currently most of the commonly used analysis pipelines cannot distinguish allelic duplication from paralogous duplication. However, it is clear that a more complete reference assembly can improve read alignment. The robust identification of small insertion and deletions variants (indels) is also quite challenging. Inconsistent alignments can lead to false SNV calls, or mis-calling of the indel within the initial

Nature Biotechnology: doi:10.1038/nbt.3237

read alignment. Many commonly used tools, e.g. the Genome Analysis Toolkit (http//www.broadinstitute.org/gatk/, accessed August 18, 2014), perform a post-processing step to do local realignment in regions containing potential indels. Also, *de-novo* assembly of a patient's reads may be helpful when standard alignment protocols do not detect large insertions, deletions, or indels not represented in available assemblies [11]. Detection of these copy number changes is challenging and requires normalization against single-copy regions of a reference genome sequence.

Differences between the reference assembly and patient reads are identified following alignment. Several software programs which use a variety of algorithms are available to assess the likelihood that a variant is present or absent These algorithms typically use several methods that can include counting the number of reads associated with each allele after appropriate thresholds and mapping qualities are set, previous information about variants, allele frequencies, properties of the sequencing platform, and linkage disequilibrium data. These approaches take advantage of the diploid nature of individuals therefore their successful use for single copy sex chromosomes may be problematic. These and other approaches use Bayesian models for calling variants [8]. Often, settings for variant calling, such as minimum variant frequency and minimum coverage requirements can be adjusted by the user. These settings should be adjusted for the test and the type of variant(s) being tested.

A list of the identified variants is often represented in a digital file format. In 2015, the variant call format (VCF) file specification is the most broadly used by the clinical community, but its implementation and the annotation information it contains vary among laboratories. The VCF, which was developed by the 1000 Genomes project, represents a generic format for storing DNA sequence variants, including single nucleotide variants (SNVs), indels, and structural

variants, together with annotations for [12]. This format contains a header and data (sequence variants) section. The header permits standardized information tailored to the sequence data. The VCF format is widely used in the clinical setting to support downstream analyses (e.g., variant filtration). Other formats are also used, including the Genome Variant Format (GVF)[13]. VCF and GVF allow for genotype representation and provide additional flexibility for describing additional attributes about a variant that are essential for downstream analysis. It is important to note that for a given file specification (e.g., VCF), content representation such as the use of different methods for sequence alignment and variant calling, or the methods to describe the variant and its type within the file can vary significantly among laboratories. The lack of standardized VCF requirements is a significant obstacle to comparing and exchanging variant call files among laboratories. This is of growing importance as laboratories outsource parts of the NGS testing process and need to seamlessly input external VCF files into their downstream pipeline for tertiary analysis.

3.3 Tertiary Analysis

Tertiary analysis uses the product of the secondary analysis to filter, prioritize, and classify variants to identify those that are meaningful to the patient's clinical condition. This process begins with variant annotation, which is the process of collecting and linking all available information about a particular variant. Annotated variants can then be sorted, filtered, and prioritized using custom rules to determine the variant(s) that are relevant to the indication for testing. Although some of the annotations come directly from the secondary analysis phase (e.g. the variant quality and depth of coverage) the majority are derived from external information that is not readily apparent from the derived sequence (e.g., the known or predicted

functional consequences of a variant, its population frequency, and/or relationship of a gene to disease or phenotype). These annotations can often be obtained through automated processes using commercial and open source tools.

The prioritized list of putative genomic variants, their associated genes, and predicted functional effects undergo a clinical assessment to identify the relevant variants to be included in the laboratory result report. These steps generally require the manual review of candidate variants by an expert who is able to offer professional judgment about their relevance to the patient and to make the final decision about which ones to report, their implications, and the limitations of the test and its interpretation. The time required to review candidate variants can vary significantly. A variant that is well characterized in the literature as disease associated and relevant to the indication for testing can be readily interpreted within the context of the patient's phenotype. Variant(s) with limited representation in the literature, but with data consistent with a clinical association (e.g. segregation data, structural/functional correlates, or presence in a disease-associated gene) may take significantly more time to evaluate and interpret. Variant filtration and prioritization may be more automated through machine learning and other algorithmic approaches in the future, but the in-depth review of variants and genes, including literature assessment, is expected to remain largely manual for quite some time.

The variants in genes with an established disease association are prioritized based on predicted pathogenicity in the context of a patient's clinical presentation during manual assessment and classification. Many clinical laboratories classify variants into five discrete categories (benign, likely benign, a variant of uncertain significance, likely pathogenic or pathogenic) [14, 15]. It is important to note that clinical classification is not completely separate from the filtration and prioritization steps, because it is not always a linear process. For

example, a set of variants may be filtered early in the analysis because of their association with a gene known to be unrelated to the medical condition(s) in question.

The final step in the analysis is clinical result reporting. Clinical result reports for NGS should contain sufficient information to communicate the test result and its limitations. Significant guidance and commentary has been offered for the effective reporting of molecular genetic test results[5, 14, 16-18]. The challenge in reporting results for NGS is to balance the useful and necessary information that a clinician requires with the appropriate level of detail so the uses and limitations of the analysis are clearly understood and/or the user can be directed to useful resources and other assistance. Examples of NGS reports can be found in the ACMG Guidelines [5] for clinical laboratory standards for next-generation sequencing.

4. Workgroup Recommendations: Secondary Analysis

4.1 Multiplexing/De-multiplexing

Multiplexing is the physical pooling and simultaneous sequencing of multiple patient samples in a single NGS reaction. The number of samples that can be multiplexed for sequence analysis is dependent on the type of analysis (e.g., gene panel, exome, or genome), the depth of sequencing required, the technical limitations of the manual and automated procedures for the timely and accurate preparation of multiplexed samples, and the throughput of the sequencing platform. Currently, it is estimated that clinical laboratories pool 10-12 patient samples per run[19-26]. This number will likely change over time and is highly dependent on the application, platform, and capacity of the laboratory to handle multiple samples.

Samples are multiplexed by "tagging" each fragment with an index (a short sequence that is added to fragmented genomic DNA, also known as a barcode, or multiplex identifier) to

Nature Biotechnology: doi:10.1038/nbt.3237

identify the patient from whom the sequence is derived prior to pooling. De-multiplexing refers to the use of computational tools to separate and associate each sequence read with the correct patient sample using the barcode indexes once sequencing is complete. It is important to prevent an incorrect assignment of reads to a patient's sample during de-multiplexing. The reliability of multiplexing is greatly influenced by several factors associated with the index design and the process by which they are added to the fragments prior to sequence analysis. The optimization and validation of multiplexing and de-multiplexing takes into consideration:

1) Index design (length and diversity of base composition)

2) Processes for addition of the indexes to each fragment to be sequenced

3) Selection and use of software tools for de-multiplexing sequence reads

Commercially available indexes and software programs to analyze these indexes are typically used in the clinical laboratory. There are a variety of methods that can be used to add the index to the patient sample during library construction. The ligation-based method (also known as in-line barcoding) involves incorporation of the index into the library adaptors that are ligated to the fragmented sample DNA[27-31]. Indexes are typically embedded within one or both of the forward or reverse sequencing library adaptors[27, 32, 33], via PCR primers[32, 34, 35], or by a combination of adaptors and PCR primers [36]. Platform vendors have adapted and optimized these procedures for use with their instruments and provide the indexes and protocols to end users. **The Nex-StoCT workgroup recommended that laboratories use the commercially available indexes and protocols recommended by the platform manufacturers if they can be optimized and validated for the intended clinical application.**

Clinical laboratories without the relevant expertise are discouraged from designing their own indexes because custom-developed indexes require extensive validation to ensure that they

Nature Biotechnology: doi:10.1038/nbt.3237

can be discriminated from each other during de-multiplexing[37]. Indexes (either commercial or custom designed) should be validated before clinical use. This can be done using known samples to assess the fidelity and combined error rate (includes error rate of all steps of sequencing, experimental handling, analytical, de-multiplexing, variant annotation, etc.).

Clinical laboratories that design indexes need to consider the "edit distance" (the number of substitutions, insertions, and deletions necessary to transform one index sequence into another) between any two index sequences. The edit distance should be optimized to assure sequence identity and prevent errors that may lead to mis-assignment (sometimes called cross-contamination) of reads from different patient samples[37]. To avoid these errors, **the workgroup recommended that laboratories use indexes that differ by more than a single base in the same reaction/lane.** The balance of the index base composition (mixtures of A,C,G,T at each base position) should be optimized to assist with cluster detection for some technologies[36]. Another feature to consider is the length of the index. Indexes of six or more bases in length allow more accurate sample assignment[27], whereas problems with sample identification have been reported when shorter sequences were used (e.g. 4 base indexes) (workgroup observation). Shorter indexes have a higher likelihood of replication errors when PCR is used during library generation. Non-specific priming during PCR can also lead to the conversion of one index to another. This can cause reads to be associated with the incorrect patient sample[37]. Multiplexing/de-multiplexing errors can result in a low- level presence of unexpected (mismatched) indexed reads in a patient sample, and these are typically represented at a low allelic fraction. A low-level of mismatched indexes is typically not a significant problem for germ-line autosomal diseases, but can be an important issue for other applications, including testing for somatic variation, mosaicism, and mitochondrial heteroplasmy. Allowing

Nature Biotechnology: doi:10.1038/nbt.3237

mismatches during the test optimization process may be informative for understanding the likelihood for loss of discrimination among indexes as a consequence of replication and other types of errors during sequencing.

The workgroup recommended that clinical laboratories discard reads with mismatched indexes. There may be a reduction in the percentage of reads that contribute to coverage when no mismatches are allowed, but this can be overcome by increasing the depth of coverage[19, 21, 22].

The workgroup considered workflow issues related to multiplexing and **recommended to index samples as soon as possible and prior to targeted capture;** however, this may not apply to PCR-based target enrichment methods that require a PCR step prior to indexing. This practice also enables the sample to be pooled before capture, thereby minimizing reagent cost, human error, and sample switch. Different sets of barcodes can be used for adjacent samples in the sequencing instrument to avoid mis-assignment of reads in the event that physical cross-contamination occurs.

The fidelity of de-multiplexing should be assessed and validated to assure the correct assignment of sequence reads to their respective patient samples. This can be achieved using a "concordance and contamination screen", in which the patient's sample is split and both a SNP array analysis (or orthogonal method) as well as multiplexed NGS are performed[38]. Concordance between the two analysis methods occurs when sequence reads are assigned correctly to the indexed sample. Lack of concordance is often caused by sample mix-up or a poorly designed index. Cross-contamination errors can also occur during sequencing, (e.g. nucleotides are mis-incorporated, a nucleotide is read incorrectly) and lead to mis-assignment of sequence reads. The concordance and contamination screen is dependent on having a second procedure, which allows for the comparison of results from the split sample. Alternate methods

Nature Biotechnology: doi:10.1038/nbt.3237

(e.g., sequencing samples individually versus in pools) can also be used to determine the fidelity of sample pooling and de-multiplexing. Laboratories should ensure that the barcodes detected are only from those included in the assay. This approach not only assesses error, but also monitors the efficiency of barcoding for each index used, and may be used to remove un-barcoded, artifactual sequences. Deviations in the expected allelic fraction or mitochondrial heteroplasmy may indicate errors in multiplexing/de-multiplexing or other steps of the sequence analysis. Control procedures provide assessment of the error rate for each patient sample represented within the multiplex pool. Detailed procedures for the inclusion of internal and external QC samples to help assess the concordance and cross-contamination for each clinical sample in a clinical laboratory setting have been published[20, 21, 39].

4.2 Mapping and Alignment of NGS reads

De novo assembly of human genomes from NGS reads is currently neither reliable nor computationally favorable for routine use, thus variant analysis generally depends on mapping and alignment of sequence data to the reference assembly. In theory this approach involves mapping a read in order to identify the locus from which it is derived, then aligning it to determine whether there are any differences in the read when compared to the reference assembly. In practice, the mapping of a read to the locus from which it was derived may be complicated by biological duplication (pseudogenes and segmental duplication[40] , assembly error, and genomic regions with significant allelic diversity, such as the HLA locus. The human genome reference assembly is produced by the Genome Reference Consortium (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/, accessed August 18, 2014)[41]. While other groups, such as University of California, Santa Cruz (UCSC; https://genome.ucsc.edu/, accessed August 18, 2014) and Ensembl (http://www.ensembl.org,

accessed August 18, 2014) distribute the assembly, it is recommended that the full assembly be obtained directly from the GenBank FTP site (http://www.ncbi.nlm.nih.gov/genbank/ftp/, accessed August 18, 2014) to obtain the proper sequence identifiers and alignments of alternate sequence representation to the primary assembly, including the sequence or assembly accession and versions. Using a single source for a reference assembly is helpful for minimizing problems in communicating results or data sharing when identifying variants based on different reference assemblies, which may use different annotations or coordinate systems. The last major update to the assembly was GRCh37 and a new assembly, GRCh38, was released in December 2013 (see http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/, accessed August 18, 2014). It is also important to note that naming conventions among different groups have not always been synonymous. For example, UCSC hosts a commonly used web browser and annotation dataset that refers to GRCh37 as hg19 (see https://genome.ucsc.edu/cgi-bin/hgGateway, accessed August 18, 2014).

The GRC releases patches (otherwise known as version changes) to update the assembly quarterly. These patches do not change the primary assembly genomic coordinates. Patch updates are provided to correct errors in existing sequence and/or introduce new sequence data. For the latter, this is offered as a scaffold to the primary assembly. Such a patch is assigned its own coordinates so as not to disrupt the genomic coordinate structure of the current accessioned assembly. These sequences will be integrated and will likely change the genomic coordinate structure when the reference assembly is updated and assigned a new accession number. The patch updates are available on the GenBank FTP site (http://www.ncbi.nlm.nih.gov/genbank/ftp/, accessed August 18, 2014). An NGS alignment to a particular region may change based on the version of the assembly used. **As a consequence, the workgroup recommended that the**

Nature Biotechnology: doi:10.1038/nbt.3237

assembly accession and version number (available at http://www.ncbi.nlm.nih.gov/assembly, accessed August 18, 2014) should be documented for each alignment for traceability in both informatics workflows and result reporting. The full assembly has an assembly accession.version as do all assembly units. Additionally, all sequences within the assembly use GenBank sequence accession.versions to track sequence changes. When reporting data on an assembly, it is important to note the accession.version of the full assembly or specific assembly units used for analysis.

The human reference assembly is complex and multi-allelic[41]. The full assembly (GRCh37, http://www.ncbi.nlm.nih.gov/assembly/2758/, accessed August 18, 2014) is composed of a Primary Assembly Unit, which is meant to be a non-redundant representation of a single haplotype. The Primary Assembly Unit contains the chromosome sequences as well as all unplaced and unlocalized sequence that cannot be ordered and oriented on the chromosome, but are not likely allelic. Additional assembly units are created with sequence representations for regions with extreme allelic diversity (e.g. HLA). Most commonly used NGS data analysis tools are not designed to work with a multi-allelic reference because they cannot distinguish allelic duplication from other types of biological duplication within the assembly. To utilize these sequences, existing software tools will either need to be modified or new software tools will need to be developed. Only 3 regions in GRCh37 contained alternate loci, another 60 regions were added as patches (in GRCh37.p14). While many of these regions have only one alternate, some complex loci such as the HLA, LRC and KIR loci have many alternate representations.

Most mapping and alignment algorithms try to find a balance between sensitivity, specificity and speed. Therefore many aligners may place a read at the correct locus in the assembly, but will produce a sub-optimal alignment. This may be corrected by local re-alignment, which is a

Nature Biotechnology: doi:10.1038/nbt.3237

common feature of many informatics pipelines. An important metric that describes the fidelity of the read mapping is the mapping quality score. This score estimates the probability that a read is misplaced. It is based on several parameters including the number of distinct regions within the reference assembly to which the software could map the read together with the number of base differences between the read and the candidate locations in the reference sequence. This mapping score is typically stored in a BAM file and is recognized by variant calling algorithms. Mapping scores from different algorithms are not comparable and software varies with regard to how low quality mapping scores are handled. For example, if a read maps equally well to multiple locations in the reference assembly, some mappers will discard it, some will place it randomly, some will place it at multiple locations, and some will map to the location with the highest mapping quality score. This behavior can significantly affect downstream variant calling and is typically accounted for during the optimization and clinical validation of the test. Errors in alignment may also be measured by evaluation of the accuracy of the variant calls. It may be useful to use variant callers that use mapping quality information: for example, variant callers that have a threshold below which reads with low mapping scores are not used in variant calling, or are down-weighted as evidence for a variant.

A variety of mapping and alignment algorithms have been developed and incorporated into NGS software packages. These alignment tools can be adjusted to different levels of sensitivity for the detection of different variant types. The relevant algorithms and software packages have been described and compared elsewhere[41-47].

Two fundamental alignment techniques are employed in the tools commonly used today: hash table–based implementations and Burrows–Wheeler transform (BWT)-based methods [46]. Hash table-based algorithms index and scan the sequence data to facilitate rapid searching and

placement of reads on the reference genome sequence[46]. These tools work by building a data structure (or hash table) that is usually an index of short oligomers (also called seeds) that are present in either the reads or the reference genome sequence (e.g. MAQ[48]). This table identifies candidate mapping positions by finding locations in the reference and in the reads to be aligned that share these short seed sequences. The candidate mapping positions are then evaluated to determine the final alignments [e.g. BFAST [49], NovoAlign (http://www.novocraft.com, accessed August 18, 2014), MOSAIK (https://wiki.gacrc.uga.edu/wiki/MOSAIK, accessed August 18, 2014), and Isaac (http://bioinformatics.oxfordjournals.org/content/early/2013/06/04/bioinformatics.btt314, accessed August 18, 2014). The software tools that utilize hash table-based algorithms differ based on the following: length of the seed, number of mismatches allowed in the initial mapping, weight of the initial mapping, type of seed extension, memory requirements, speed, and accuracy.

Instead of a table of short oligomers to align the reads, the Burrows–Wheeler Transform (BWT)-based methods[50] use a string matching approach to create a space-efficient index of the reference genome to facilitate rapid searching[46, 48]. This method rapidly identifies genomic locations as good matches for a read and then, similar to the hash table-based methods, fully evaluates these candidates to place reads to specific locations[51]. BWT-based methods take less time to execute and are more memory efficient than most methods based on hash tables[46, 48]. Examples of short read alignment programs that are based on BWT are Bowtie 2[52], BWA[53], TMap[47] and SOAP2[54].

The selection of an alignment algorithm and strategy should consider the NGS application (e.g., class of variants to be detected [short insertion, SV, SNV etc.], the next-generation

platform used, whether the analysis covers the whole genome or targeted regions, etc.), as well as the laboratory's computational capacity (e.g. is there a high performance cluster environment, etc.). Each aligner will require unique settings for the optimal detection of different classes of variants. Some aligners are optimized for use with a particular platform, for example TMAP[47] is specifically designed for mapping Ion Torrent data to a reference assembly. Other aligners, e.g., BLASR[55], are capable of addressing platform-specific errors, such as increased indel sequencing errors that may occur. Certain aligners are optimized to detect specific types of sequence variants, such as short insertions or deletions, SNVs or CNVs , or provide more precise local alignment but are too resource intensive for read mapping as an initial step in the workflow. By combining aligners (i.e. in parallel or in series) with different detection capabilities into an informatics pipeline, laboratories can design assays that can detect a wider variety of variants than is currently achievable using a single aligner[56]. As such, the workgroup **recommended that clinical laboratories evaluate a combination of aligners or the same aligner with different settings to effectively identify the types of variants targeted (e.g. for SNV and CNV detection).** Alternatively, users may choose software packages containing complete workflows that are pre-optimized for particular panels and applications. For example, a more stringent quality threshold related to acceptable mapping quality scores might be used for alignment in order to call SNVs as compared to the threshold used to align for detection of indels. It is important to understand that many tools use the same underlying algorithm, and therefore different software packages may have the same fundamental strengths and weaknesses[43, 45, 56]. Optimization of sensitivity and specificity in NGS assays have been previously described[3, 5, 16]. Laboratories are encouraged to optimize assays to minimize the number of false negative calls, while avoiding excessive numbers of false positive calls. This can

be accomplished by developing and optimizing assays using characterized reference materials that contain a wide variety of variant types (eg. SNVs, large and small indels) located throughout the genomic regions targeted by the test. NIST is currently developing highly characterized genomic DNA reference materials that can be used for this purpose, and they are also creating additional standards that will address false negative calls at low allele frequencies[67].

NGS software is still evolving and new versions frequently become available. Different versions of the same software may vary in performance for a particular application. For example, a recent version of BWA, BWA-MEM (http://bio-bwa.sourceforge.net/, accessed August 18, 2014), works better on longer reads and can tolerate higher error rates[57] than the earlier version, BWA-SW [53]. As of 2015, the majority of software tools used for clinical NGS applications were developed for research activities; although a number of these have been adapted for clinical applications, for a listing and description of these tools see CLSI MM09 A2 [16]. Vendors generally design their software and establish default settings for optimal performance for general and specific applications. **The work group recommended that laboratories initially use the software's default settings and only modify them (with validation) when appropriate for their clinical application(s).** Examples of software settings that may be modified include: low quality sequence trimming, number of allowed mismatches, the allowed gap opening and gap extension, and minimum mappability for reads. Changes require re-validation to assure that the desired outcomes are achieved and other elements of the alignment process are not compromised. Therefore, **the work group recommended that changes to default settings and subsequent evaluation be performed in consultation with an informatician with the requisite expertise.** These steps are performed prior to test validation,

Nature Biotechnology: doi:10.1038/nbt.3237

which serves to document the final settings of the software that will be used during patient testing.

Several factors should be considered when selecting and optimizing alignment software to assure the quality of the alignment. These include the different error profiles of the various NGS platforms, the general and variant-type specific error rate (e.g. mismatches vs. gaps, repetitive region errors), and the average read length of the instrument. Alignment errors may be detected by evaluation of quality metrics that include the mapping quality score, the transition/transversion ratio (Ti/Tv) of subsequent variant calls, and ratios of synonymous to non-synonymous changes in subsequent variant calls. Each alignment algorithm generates a unique mapping quality score (which are not comparable between algorithms), and assigns reads in different ways (e.g. some attempt to map, while some bin reads) as described above. The Ti/Tv ratio serves as a general quality indicator because it is approximately constant for a particular targeted region (e.g., gene panel, exome, or genome) [58]. In addition, there are non-computational sources of error. For example, the quality of the sample and the fidelity of the target enrichment process used for gene panel and exome sequencing may influence the quality of the sequence generated and the subsequent alignment[3], which may affect the ability to map reads. This is due to the presence of low quality base calls that alter the total percent of sequences suitable for alignment, as well as the ratio of on-target to off-target alignments. Consequently, the on-target to off-target ratio is a measure of the quality of the capture across different runs of the same assay, with high on-target rates directly related to the fidelity of the final sequence calls.

Homologous sequences present a challenge to optimal alignment[3, 5], especially if they are longer than the average length of the sequence read generated by NGS. Paired end sequencing

or long reads helps but is no longer effective when the region of homology is much larger than the library fragment size. Coverage at the target locus will be reduced when long stretches with100% identity to the region of interest are present elsewhere in the genome. This can lead to missed variants. Sequences that are less than 100% identical but exceed a level of homology such that aligners cannot unambiguously map reads will be prone to false positive variant calls (i.e. a variant call that reflects the difference between gene and pseudogene rather than an actual change at the gene of interest). For exome and genome sequencing and most gene panels, **the workgroup recommended that reads be aligned to the reference assembly, and not just the target region, to minimize the potential for off-target or forced alignments, unless methods to ensure the optimal alignment to targeted regions of the genome have been developed.** This may not be necessary for some gene panel or other sub-exome/genomic analysis when PCR captured methods are used, stipulating that the fidelity of the alignment must be optimized and validated to minimize the potential for mis-alignment of reads. Comprehensive testing for some disorders must include certain genes with high sequence homology to other loci. An example is the stereocilin (STRC) gene, a major gene in nonsyndromic hearing loss[59]. While NGS is potentially a universal technology platform that can be used to interrogate many different variant types and consolidate diagnostic testing assays, homologous genes will remain a complicated issue for the foreseeable future. In some cases it may be necessary to use an alternative method, such as Sanger, that enables amplification or sequencing of only the targeted region using location-specific primers or probes to confirm variant calls.

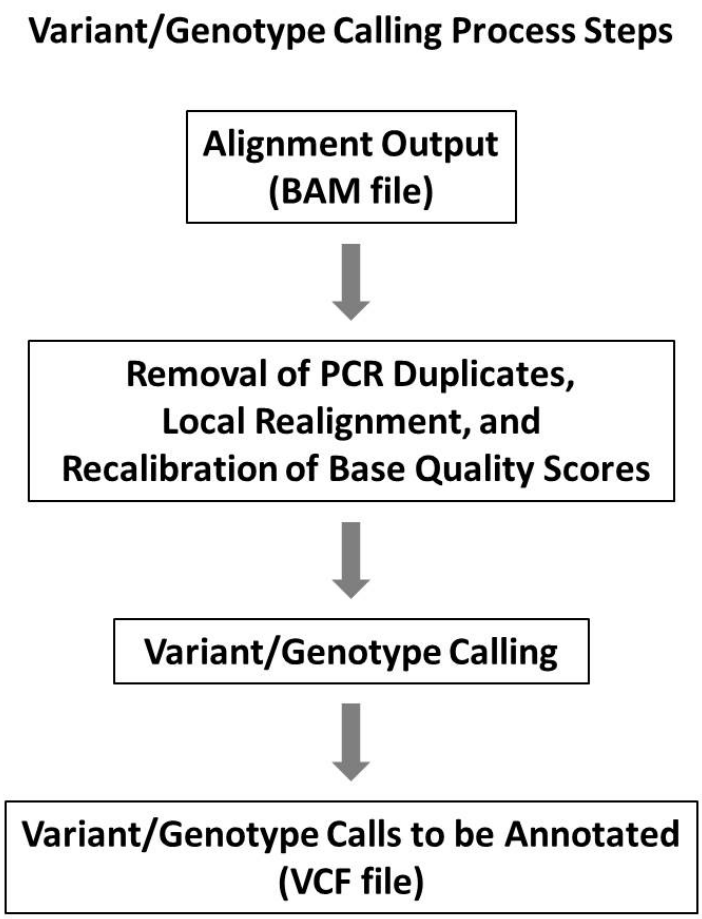
.

4.3 Post-alignment processing, Genotype / Variant, Calling, Sequence Annotation, and Filtration

Additional processing is often performed prior to variant calling to optimize the alignment and account for errors in the initial mapping and alignment. These steps include local realignment, and for some variant callers and enrichment methods, removal of polymerase chain reaction (PCR) duplicates , and the recalibration of base call quality scores[8, 60, 61]. It is important to note that duplicate removal is generally not performed with amplification-based enrichment protocols and some variant callers don't require base call recalibration.

Variant calling is the process in which nucleotides (and their position) that differ between a patient's sample and the reference sequence used for the alignment are identified. This process is shown in Supplementary Figure 2.

Nature Biotechnology: doi:10.1038/nbt.3237



Supplementary Figure 2. A patient sequence is aligned to the reference assembly and/or other characterized sequences prior to downstream analysis for variant and genotype calling. The aligned sequences are typically contained a BAM or similar file type. These files are amenable to analysis that identifies a set of variants or sequences that differ from the reference. These sequence variants are deposited into a variant file that permits additional feature annotation and downstream analysis. The BAM file comprises mapped data on a "per-fragment" basis whereas the VCF file represents the variant calls on a "per sample" basis.

High rates of duplicate reads can occur during amplification from the same initial sample molecule. This can result in an overrepresentation of certain reads that can result in an increased

Nature Biotechnology: doi:10.1038/nbt.3237

false positive rate and incorrect assignment of zygosity[8] Programs such as PICARD (http://picard.sourceforge.net, accessed August 18, 2014) and SAMtools[53] can be used to remove duplicate reads from FASTQ files using algorithms that detect excess reads with the same start and stop coordinates. Duplicate removal is typically not performed in assays using amplicon-based capture reagents, because all amplification products will have the same start and stop positions. In such systems duplicates can be minimized by increasing input material or reducing PCR cycles. The level at which duplicates influence accuracy can be determined by measuring false positive and negative rates across an input DNA titration series.

4.3.1 *Local realignment*

During analysis, the set of mapped reads at a locus is locally re-aligned to the primary reference assembly, (see http://www.broadinstitute.org/gatk/events/2038/GATKwh0-BP-2-Realignment.pdf , accessed August 18, 2014). This process can improve the quality of the alignment and the sensitivity of variant calling while reducing false positives, especially for small indels[62] (see http://www.broadinstitute.org/gatk/events/2038/GATKwh0-BP-2-Realignment.pdf, accessed August 18, 2014) . There is evidence that misalignment around indels is an important source of error[61, 63]. Although reads may be placed in the correct genomic location during the primary alignment, their placement may be shifted if an insertion or deletion is present and this shift can introduce false-positive variant calls in the region flanking the variant [8]. For example, mapping software will not have sufficient evidence to introduce an indel into the alignment when there is an insertion or deletion near the end of a read. This usually results in a small number of bases placed into the alignment rather than the correct identification of an indel. However, when a number of reads align to a locus, particularly when the indel is

closer to the interior of the read, realignment algorithms may be used to correctly identify the indel and remove the spurious substitutions. Confirmatory testing using an alternate method can effectively identify these false positives or identify indels that otherwise may be missed. The Genome Analysis Toolkit (GATK)[63, 64] , Torrent Suite and Ion Reporter software[47] and NextGENe (http://www.softgenetics.com/NextGENe.html, accessed August 18, 2014), are examples of commonly used data analysis pipelines that perform local realignment, quality-score recalibration, and variant/genotype calling. Several recent publications describe evaluations of NGS software for calling short indels and associated challenges[44, 61, 65].

4.3.2 *De novo assembly*

Some programs, including the GATK2 HaplotypeCaller and Torrent Suite/ Ion Reporter software, perform local de novo assembly, which is also useful for accurate alignment and detection of variants [66]. This approach gathers all of the reads that map to a region of interest and assembles them without the use of a reference sequence to avoid propagation of errors associated with alignment to a reference assembly. These programs use heuristic algorithms that calculate the likely order of overlapping reads, and it has been shown that this approach can improve the accuracy of variant calling or identify long insertions and deletions that may be missed by alignment based approaches due to soft clipping which is an unmatched fragment in a partially mapped read.[67]

4.3.3 Quality score recalibration

Quality scores are associated with each primary base call in the aligned reads. Base quality score recalibration generates higher quality variant calls than those derived from the raw

Nature Biotechnology: doi:10.1038/nbt.3237

per-base quality scores that were originally assigned by the platform-specific base caller[8, 60]. The raw Phred-scaled quality scores do not always accurately reflect the true base-calling error rate[60]. To recalibrate quality scores, NGS data analysis pipelines, for example GATK, utilize alignment-based algorithms that employ a set of established variants, e.g., SNVs of both types (transitions and transversions), for accurate alignment to the reference sequence. It has been proposed that performing base quality score recalibration using well characterized spike-in DNA sequences can also improve recalibration accuracy[68]. Variant callers that use other types of sequence data, such as flow-space[69] information, do not typically require base quality score recalibration.

4.3.4 **Variant callers**

The variant calling process generates between 20,000 and 100,000 variants per exome, and approximately 3-4 million variants for whole genome sequencing. At least 90% of these variants are typically SNVs[1, 70]. Genotype callers output results to a variant call file[12]. At a minimum, these files record information and annotations about the sequence variants identified, such as their type (e.g., SNV, indel, etc.). Many of these files support the inclusion of additional information such as structural and functional consequences of the variant, and the capacity to record reference sequence or no-calls. Many software tools are used for variant identification and have been described and compared elsewhere[8, 60, 61].

NGS bioinformatics pipelines may call SNVs, small insertions and deletions, as well as larger structural variations (e.g. larger insertions, interchromosomal translocations, and copy number variants). In 2015, no single software tool identifies all of these variant classes with equal accuracy. As a consequence, **the workgroup and others[42, 71, 72] recommended that more**

than one variant caller should be evaluated to identify the combination of settings and/or software able to detect the spectrum of variants targeted by the intended clinical application and the suitability of a variant caller for the platform should be considered. For example, some laboratories analyze the data using variant calling software optimized for SNV detection, followed by a separate analysis using software optimized to call indels. The results from these separate analyses can then be combined. However, there is a high bioinformatics/IT overhead that comes with constructing combinatorial pipelines, and discordance between the multiple approaches must be resolved[73]. In some cases, alternate methods using different but established technologies, such as quantitative PCR or Sanger sequencing, may be required to resolve discrepancies between the different analyses.

Errors can occur during both alignment and variant calling. The two processes are tightly linked; therefore variant calling software should be optimized in conjunction with the alignment strategy. Software packages that provide optimized mapping, aligning, and variant calling for particular applications may be appropriate. In the clinical laboratory, this assessment should focus on the sensitivity and specificity of variant calling for the type of variants that are under investigation (e.g. SNVs, indels, etc.). **The workgroup recommended the use of both real and simulated data for optimization of variant or genotype calling[74]. However, the workgroup recommended that simulated data not be used in the absence of data derived from patient samples for optimization and validation of the informatics pipeline. When possible, well-characterized human genome reference materials such as those developed by the GeT-RM, the Genome in a Bottle Consortium, and similar efforts should be used for test development, optimization, and validation[67] (GeT-RM browser:**

Nature Biotechnology: doi:10.1038/nbt.3237

http://www.ncbi.nlm.nih.gov/variation/tools/get-rm/, www.genomeinabottle.org, accessed August 18, 2014).

A variety of systematic errors can occur during sequencing, mapping, and alignment, thus many variant callers evaluate the metrics associated with these processes and filter out variants that do not meet set criteria[75]. These filters generally use the base call quality score, mapping quality score, coverage, an estimation of strand bias, and allelic read percentage (allelic fraction or zygosity), among other parameters for the identification of variants with a high degree of confidence. The acceptance criteria for many of these metrics are often application specific but general principles apply. For example, the allelic fraction for a heterozygote call is expected to center around 50% and therefore acceptance criteria can be set to include a reasonable range (e.g. some laboratories use cut-offs that range from 30% - 70%). Different settings may be necessary for different variant types (SNVs vs. indels). If the settings and metrics are too stringent, data may be lost or filtered, resulting in false negatives or, if less stringent, too many false positives. Additionally, when variant detection is performed in genomic regions with low coverage, a large number of sequencing errors may be called as variants. For example, a single read supporting a heterozygous variant call in a region covered by only 5 or 6 reads is difficult to differentiate from a sequence or mapping error. Some laboratories apply a strict threshold for the particular number of quality reads needed for calling variants, while others have no hard threshold but instead require confirmation of these variants via an orthogonal method, or by comparison to data from family members to exclude errors. However, an overabundance of coverage can also be indicative of an error, since reads from regions of low complexity or those with shared sequence identity can erroneously map to a single region causing higher than expected

coverage[40]. Higher than expected coverage may be a consequence of mapping errors that can be caused by over amplification as can occur in the EGFR gene.

4.3.5 Considerations for diseases with atypical allelic fraction ranges and challenging variant profiles

Allelic fractions for diseases such as somatic cancer, mitochondrial diseases, and Mendelian disorders in which mosaicism is observed can be significantly lower than those observed for a typical germline variant. This requires a different pipeline (or different allele ratio threshold setting) to adequately detect variant calls that would normally be filtered out. In addition, some aligners and/or variant callers are optimized to identify particular variant types, such as insertions and deletions and structural rearrangements. Therefore, more than one aligner or variant caller may be required to analyze the expected variant types that are more prominent in cancer. The detection of structural variants for cancer applications can also be problematic because of the large number of chromosomal rearrangements and other aberrations such aneuploidy present in the sequence data. The significant tissue heterogeneity in some cancer samples also confounds variant calling and may require an alternative or refined pipeline. Sequencing of the mitochondrial genome, which may also show significant heterogeneity, also dictates careful consideration of the pipeline and settings used[20, 21]. Additional details in addressing these sequencing issues are beyond the scope of this manuscript.

5. Workgroup Recommendations: Tertiary Analysis

5.1 Variant and Gene Annotation, Filtration and Prioritization

During tertiary analysis, a variety of data sources and algorithms are used to evaluate the identified variants to determine which are relevant to the indication for testing and should therefore be included in the clinical report. Tertiary analysis includes variant annotation, followed by automated filtration and prioritizations using these annotations and finally an in-depth clinical assessment of the most relevant variants.

5.1.1 Annotation (in addition to that described during secondary analysis)

Annotations used for filtering and prioritizing variants typically include structured information such as the population frequency, biochemical properties, computational pathogenicity prediction, variant type and predicted impact of the variant on the protein (missense, loss of function, etc). Databases such as those developed by the 1000 genome project (http://www.1000genomes.org/ accessed August 18, 2014) and Exome Variant Server (EVS, http://evs.gs.washington.edu/EVS/, accessed August 18, 2014) are useful sources for population frequencies. Also typically included in high throughput annotations is structured information regarding whether a variant is present in variant databases such as the Human Gene Mutation Database (HGMD, http://www.hgmd.org/, accessed August 18, 2014), Online Mendelian Inheritance in Man, (OMIM, http://www.omim.org/, accessed August 18, 2014) or ClinVar (http://www.clinvar.com/, accessed August 18, 2014) and whether or not it has been previously labeled as clinically significant. ClinVar contains all of the OMIM variants in a format that can be mapped back to a sequence, which is a notable advantage over other databases. Many of the variants found in OMIM are not formatted to permit this.

Unstructured information such as evidence embedded within published literature or information about segregation of a variant, the pattern of inheritance, and the patient's phenotype

are difficult to incorporate into this upfront, high throughput process but is critical in the subsequent in-depth assessment of variants that are considered potentially relevant for the patient's clinical presentation[5, 76, 77]. Additional detail about the type of information that is collected during clinical variant annotation are described elsewhere[5, 76].

Both public and commercial annotation software tools are available. Some annotation tools rely on information contained in external data sets (e.g. UCSC transcript information to determine splice sites [https://genome.ucsc.edu/, accessed August 18, 2014], species sequence comparison programs, like PhastCons[78], as well as functional variant pathogenicity prediction programs like Mutalyzer (https://mutalyzer.nl/, accessed August 18, 2014), Provean (http://provean.jcvi.org/index.php, accessed August 18, 2014), Mutation Assessor (http://mutationassessor.org/, accessed August 18, 2014), SIFT[79] (http://sift.jcvi.org/, accessed August 18, 2014), and PolyPhen2 (http://genetics.bwh.harvard.edu/pph2, accessed August 18, 2014), may be run on specific variants to hypothesize the functional consequence of amino acid changes[72]. Other tools, including ANNOVAR[80], VAAST[81], Carpe Novo (http://www.hmgc.mcw.edu/BIR/carpe_novo/carpe_novo.htm, accessed August 18, 2014), Variant Effect Predictor[82], SNFeff (http://snpeff.sourceforge.net/, accessed August 18, 2014), Ion Reporter[47], and Mutation Taster (http://www.mutationtaster.org/, accessed August 18, 2014), can annotate the variants and support their subsequent filtration and prioritization based on user defined rules, nucleotide and amino acid level evolutionary conservation, and predicted protein impact (e.g. a variant that causes a premature stop, affects canonical splice site or changes the start codon to another amino acid). A survey of variant annotation tools, along with their respective input/output formats and variant identification capabilities was described[42]. **The choice of annotation tools should be based on the types of sequence variants that are to be**

detected by the clinical test as well as the strengths and limitations of the software to detect particular variant types. Annotation and filtration tools need to be integrated into the entire informatics pipeline to enable seamless automation that must be validated.

Some annotation data are drawn from external databases including genome wide databases (e.g. [ClinVar[83] http://www.ncbi.nlm.nih.gov/clinvar/, accessed August 18, 2014, HGMD, OMIM, dbVar [http://www.ncbi.nlm.nih.gov/dbvar/], accessed August 18, 2014) and locus specific databases (e.g. Leiden Open Variation Database [LOVD, www.lovd.nl/, accessed August 18, 2014], Universal Mutation Database [UMD, http://www.umd.be/, accessed August 18, 2014]). These databases are not well-curated for medical applications, making it essential that the user understands the quality of evidence obtained from these resources when making a clinical assessment[70, 84]. The incorrect retention or filtering of sequence data as a consequence of an incorrect annotation in an external database may cause errors. For example, a variant may be erroneously listed in a database as pathogenic with insufficient evidence provided by a single study, or may be an artifact due to an error in the data analysis process[70]. Conversely, databases, such as dbSNP (http://www.ncbi.nlm.nih.gov/SNP/, accessed August 18, 2014), that may contain information about benign polymorphisms also contain rare and pathogenic variants. NCBI provides files that allow users to distinguish variants that were submitted with pathogenicity assertions; however, other pathogenic variants can be missed. These limitations necessitate manual review by a clinical laboratory professional with the necessary expertise. The laboratory must cross-reference findings derived from databases with other resources and the primary literature when making a determination about the likely pathologic association of a variant. To streamline the workflow, many laboratories develop internal databases of customized annotation information based on the results from their manual review processes which are used as a curated,

reliable data source for variants that may be identified in future analyses[76]. **The workgroup recommended that until reliable, medically-curated databases are available, data used to annotate variants for making a clinical assessment should be carefully evaluated to assure that it is supported by sufficient evidence.**

5.1.2 Variant filtration and prioritization

Filtering of variants uses computer algorithms designed to identify variants that are clinically relevant to the patient and the reason for testing. False- positives can be removed from consideration using software tools for automated variant filtration that base their assessment on quality criteria, knowledge of the region targeted, and variant type[70, 72]. A substantial number of variants can be filtered or prioritized based on criteria such as the allele frequency in the population, predicted effect on protein function, and the presence of the variant in clinically relevant genes. Variants with high frequencies in the population can also be eliminated prior to analysis as they are usually not expected to cause disease.

Heuristic and probabilistic ranking are the two major approaches for filtration. Heuristic ranking utilizes a series of logic steps that are applied to the variant calls (e.g. ANNOVAR[80], SIFT[79]). The current iterations of heuristic filtering assume that causal variants alter protein sequence in most instances (unless there is evidence otherwise); therefore, synonymous variants may be removed to reduce the number of variants in a given exome data set from approximately 15,000-20,000 coding SNVs to approximately 7,000 – 10,000 nonsynonymous coding SNVs[1]. Importantly, synonymous (and non-synonymous) changes can affect gene function through disruption of other mechanisms such as splicing, mRNA processing, and transcriptional regulation [85]. This can be a limitation of heuristic ranking approaches.

Probabilistic ranking tools used for filtration (e.g.VAAST[81]) prioritize variants using a likelihood prediction that involves analysis of allele and variant frequencies in the sequence under analysis, compared to a control population database, which is then combined with an amino acid impact score to generate a ranked list of variants. The probabilistic ranking approach uses larger control databases from the same population as the case to increase the likelihood that the relative rarity of the variant is properly assessed. Probabilistic ranking methods can score nonsynonymous variants and variants in noncoding regions and incorporate other approaches that increase the statistical power and accuracy of prioritization. For example, analysis methods may use amino acid substitution data combined with pedigree, phased data sets, and disease inheritance models[81, 86]. Some methods use a combination of both probabilistic and heuristic ranking to provide a more accurate list of prioritized candidate genes and potential causal variants[8]. **The workgroup recommended that laboratory professionals recognize differences in approaches to variant filtration and consider these in the design of the informatics pipeline.**

Heuristic and probabilistic methods primarily assess the likelihood of changes to the structure and function of proteins as a consequence of the detected variant in the ~22,000 human protein-coding genes. Information about non coding regions is available from the GENCODE database (http://www.gencodegenes.org/, accessed August 18, 2014), which annotates over 25,000 noncoding RNA genes (non-protein-coding genes) and thousands of expressed pseudogenes. Large-scale functional genomics screens or datasets such as ENCODE (http://www.genome.gov/10005107, accessed August 18, 2014) which provides information about functional elements, such as regulatory elements that control gene transcription in the human genome, will be useful for the annotation of noncoding regions. Non-protein-coding

genes now outnumber coding genes and an increasing number of regulatory elements are being defined, therefore more activity in this area will be needed in the future. Clinical laboratories are usually not incorporating assessments of non-coding regions due to the lower probability of disease impact and substantial limitations in predicting the effects of variation.

Population frequency is often used to filter data when evaluating the molecular basis of rare disease. The global, maximum population and race-specific allele frequencies can be applied as separate variables, when data of acceptable quality are available. Variants are typically filtered if they have a higher prevalence than that of the disease in the patient's ethnic/racial group. This is challenging because limited data are available for most ethnic/racial groups and usually only for medical conditions in which a single highly penetrant variant is the likely cause of disease. In some disorders, such as dilated cardiomyopathy, pathogenic variants are common in the general population and known to cause late onset but a milder form of the disease. Similarly, certain recessive variants have high carrier rates (e.g. *GJB2* c.35delG and *CFTR* p.F508del), such variants may be incorrectly filtered from the analysis. As a general rule, variants at or above approximately 1% are filtered. While this may be appropriate for the majority of persons, caution must be exercised to retain variants from a geographical sub-population where the disorder may occur at significantly higher frequency. For example, the carrier frequency for the recessive disorder glutaric aciduria is estimated to be 1 in 150 worldwide but in the subpopulation, Old Order Amish of Lancaster County, Pennsylvania, 1 in 10 are carriers and the disease prevalence is significantly higher[87]. Thus, this variant could be mistakenly filtered if a laboratory used a local population as a source of variant frequency.

5.1.3 Pathogenicity prediction tools - additional details

As mentioned in the preceding sections, the majority of informatics pipelines developed by clinical laboratories utilize pathogenicity prediction programs to evaluate the potential impact of a variant on protein structure or function. Pathogenicity prediction programs are helpful for identifying variants more likely to disrupt gene structure or the resulting protein product; however, their sensitivity and specificity are low[88-91]. It is important to understand the algorithm employed by each pathogenicity prediction tool. Using a combination of tools can be informative; however, results from tools that use the same algorithm should not be taken as independent evidence that a variant impact prediction is accurate. For example, many laboratories will use both SIFT and PolyPhen2 to analyze data sets, even though these programs use similar algorithms, thereby giving similar impact predictions. **Where possible the laboratory should consider using more than one prediction program with each taking a different approach to predicting pathogenicity**. Predictions are based on multiple criteria including the biochemical nature of the variant, phylogenetic information, and perhaps, structural information. These programs have limitations such as a lack of capability to integrate the influence of other neighboring variants that may promote or mitigate a structural or functional change. For example, in one study, it was estimated that *in silico* prediction programs, such as PolyPhen2 and SIFT, accurately predict whether a variant is damaging or benign in approximately 71% of the cases examined[92]. Therefore, it is recommended that results from prediction programs not be used as the sole source to filter or classify variants in the absence of other supportive data. Additionally, if a protein X-ray crystallographic structure is available, molecular structural modeling may help to elucidate the pathogenic effect of a missense variant[39].

Nature Biotechnology: doi:10.1038/nbt.3237

5.1.4 Knowledge curation

When performing genome and exome analysis, it is customary to limit the analysis of the patient-derived sequence to "clinically relevant" genes. The "medical exome" refers to the portion of the exome that contains genes known to be clinically relevant[83, 93, 94]. Historically, these genes were derived from the OMIM database or the HGMD. These and similar databases are essential resources but many are not curated with sufficient clinical rigor. For example, it is now well established that many variants listed as pathogenic are actually benign and some of these variants may have been the basis for assignment of gene-disease association. For many genes, the published literature for disease-associated genes does not contain sufficient evidence for a definitive association or even likely association in some cases. Large community efforts (e.g. ClinVar[83], ClinGen: Clinical Genome Resource Program (http://www.iccg.org/about-the-iccg/clingen/, accessed August 18, 2014), in collaboration with NCBI, are under way to address these issues and develop clinical grade databases. Some laboratories have developed custom algorithms that limit the analysis to smaller subsets of genes based on the patient phenotype[84, 95]. **The workgroup recommended that filtration algorithms should utilize databases containing reported pathogenic variants (e.g., HGMD) to minimize the possibility that disease-associated variants are inappropriately filtered.** Some programs will tell the user when a gene or a variant has been reported in the literature (e.g. HGMD and OMIM), and then the user can evaluate the data and decide if it matches the patient phenotype.

5.1.5 Validation of computational tools

The workgroup recommended that methods selected for annotating a sequence must be evaluated to demonstrate that variant attributes are properly assigned. This is done

Nature Biotechnology: doi:10.1038/nbt.3237

using the data output from control samples previously sequenced with known, experimentally confirmed variants. New tools can then be evaluated to determine whether they assign expected annotations to these known variants. Software tools and databases used for deriving annotations are regularly updated. **A revision to a database or analysis algorithm used by the laboratory may affect the annotation process; consequently, the data analysis pipeline must be re-validated before the adoption of any updated data sources or software**. These changes are not always announced or obvious, which presents a challenge to the laboratory in maintaining a validated test. Therefore, **if web-based tools are unable to provide version control, the workgroup recommends that clinical laboratories bring the software or datasets in-house to document version changes.**

The types and number of annotations used by clinical laboratories differ. A standardized minimum set of annotations has not been established within the laboratory community. For example within the work group, one laboratory used 80 annotation fields, while another used 127 fields, with some of these fields overlapping and others unique. Even within overlapping fields, the use of unrestricted strings often results in incompatible coding of field data, or differences in ontology and syntax.

Common examples of annotation fields that can be derived computationally include:

- genomic sequence coordinates
- transcript nucleotide and amino acid position
- types of variant (e.g., synonymous, non-synonymous/missense, and predicted loss-of-function such as non-sense, frameshift or splicing variants that are likely to prematurely truncate a protein through reading frame modification or disrupted

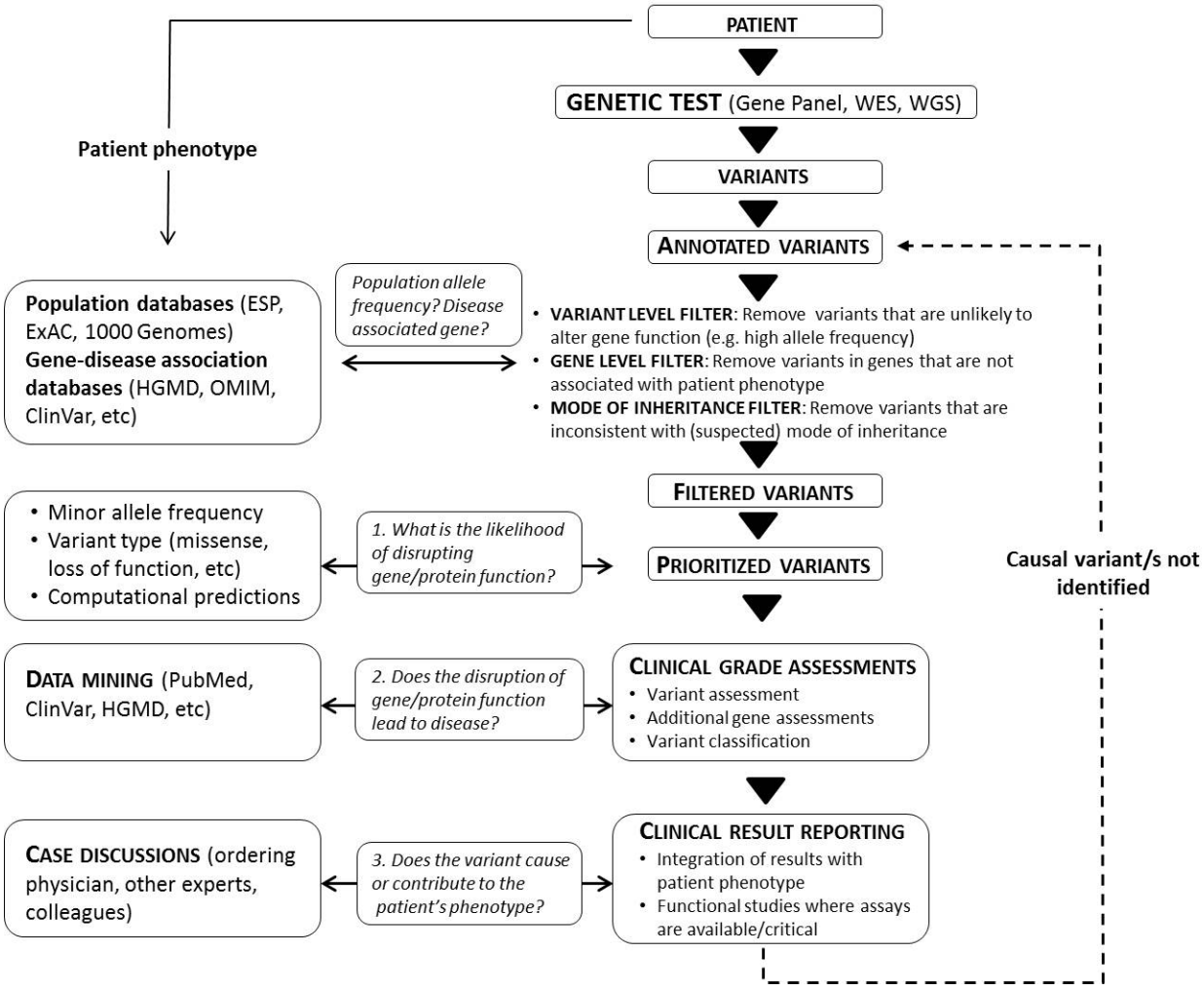
Nature Biotechnology: doi:10.1038/nbt.3237

splicing). These should be described relative to a particular transcript of the gene, often the single canonical transcript chosen by the laboratory.

- protein function annotation/ predicated impact on the protein based on pathogenicity prediction programs (damaging, neutral, etc.)
- predicted effect on splicing
- nucleotide conservation
- error detection annotations
    - e.g. quality scores, mapping quality, depth of coverage, and other evidence to support a variant call
- allele frequencies from population datasets
- observed segregation in the patient's family

The gene containing the sequence variants is also annotated for exome and genome sequencing. Annotations are used for filtration or removal of irrelevant genes or variants from analysis and to create a prioritized (or ranked) list according to the likelihood of the gene and/or variant being associated with the indication for testing (Supplementary Figure 3).

Nature Biotechnology: doi:10.1038/nbt.3237



Supplementary Figure 3: Model for tertiary analysis of sequence variants.

The purpose of tertiary analysis is to identify those variants to be reported back to the physician for medical decision making. The set of variants identified during secondary analysis are filtered and prioritized, taking into consideration knowledge of their gene associations.

5.2 Clinical Assessment and Result Reporting

Next, variants are analyzed based on the annotated information to identify those that are relevant to the patient and the reason the test was ordered. In contrast to the automated, high throughput filtration and prioritization above, this final assessment process is still fairly manual

Nature Biotechnology: doi:10.1038/nbt.3237

and requires persons with expertise with respect to the genes and diseases in question. The end product of this assessment is typically a report of all relevant information available for a given variant, including the annotations discussed above. Assessment of published literature is critical, and provides clinically significant information such as co-segregation of a variant with disease, and functional assays. Computational predictions contribute to manual annotation by experts in clinical and laboratory genetics but are typically only used as supporting evidence, because the accuracy of most tools is suboptimal. Additional details about the type of information collected during clinical variant assessment had been described elsewhere[5, 76, 77]. The number of variants remaining after annotation, filtration and prioritization ranges from a few to several hundred, depending on the test. Minimal additional evaluation is required for well-characterized variants known to be disease associated. For others, a manual review, as described above, is typically needed to determine which ones are clinically relevant to the patient (noted as "clinical grade assessment" in Supplementary Figure 3). **The workgroup recommended that clinical assessment for disease association be performed by personnel with relevant clinical expertise. In some instances, this can be a collaborative activity among the laboratory professional and others that may include a physician(s), genetic counselor(s), and/or informatician(s).** The primary purpose of the clinical assessment is to determine what will be included in the laboratory report.

5.2.1 Variant classification

The current guideline for variant classification of Mendelian diseases is a five- tiered system (benign, likely benign, a variant of uncertain significance, likely pathogenic or pathogenic)[14, 15]. The CAP, ACMG, and AMP developed an updated guidance for variant

classification (available at https://www.acmg.net/docs/Standards_Guidelines_for_the_Interpretation_of_Sequence_Variants.pdf). This new guideline was initiated, in part, by the recognition that the existing guidelines do not provide sufficient recommendations to guide the evaluation of evidence used to classify variants.

It is important to differentiate between variant classification (e.g. assessing whether the variant is deleterious) and clinical result interpretation (assessment of one or more variants in the context of the clinical presentation and other test results). For example, benign variants would probably not be reported. Those sequence variant(s) most likely determined to be related to the patient's phenotype would be reported and interpreted in the report, linking their relevance to the indication for testing and other information known about the patient and the family. The type and level of evidence used to specify that a variant is associated with the indication for testing will vary for many reasons. For example, the level of evidence needed for a variant in a gene known to be associated with the disorder in question may be lower than for a variant in a novel or noncoding gene or when reporting carrier status or disease risk in an otherwise healthy individual. **Laboratories should consider three essential questions when assessing variants that are identified during the annotation and prioritization process:**

1. **Does the variant disrupt or alter the normal function of the gene in a manner consistent with the understanding of the disease mechanism?**
2. **Does this disruption lead to, or predispose a patient to, a disease or other outcome relevant to human health?**
3. **Does this health outcome have relevance to the patient's clinical presentation and indication for NGS testing?**

Nature Biotechnology: doi:10.1038/nbt.3237

Clinical result reporting was not a focus of the workgroup meeting but there was some discussion about current result reporting challenges. Standards for clinical reporting are only beginning to emerge and additional work is necessary[5]. The challenge is the distillation of complex information to a format that can be readily understood by a clinician and useful for informing medical decisions. The complexity of some NGS test results and their limitations may require that the ordering physician consult with laboratory professionals with the relevant expertise. The workgroup recommended that a collaborative relationship be established prior to ordering of the test. This provides the opportunity for the ordering physician to be kept informed about the uses and limitations of the test. The work group developed a description of the general steps that take place during clinical result reporting (Supplementary Figure 3), including the integration of the patient's clinical presentation data, gene assessment in the context of the patient (e.g. integration of a patient's family history, when relevant), and results from functional studies when assays are available (e.g. enzyme testing, biochemistry).

The workgroup recommended that pathogenic variants and variants of uncertain significance should be reported for heritable conditions. The workgroup discouraged the reporting of benign variants. Laboratories should consider confirming all reportable variants using Sanger sequencing or another method[3]. Likely benign variants may be reported at the discretion of the laboratory, but if reported, they must be clearly distinguished from other variants and when applicable, note that the presence of a disease-associated variant may not have been detected. **The workgroup also recommended that the laboratory have strategies to reclassify or to monitor the reclassification of variants as new data become available to inform the analysis of findings.**

Nature Biotechnology: doi:10.1038/nbt.3237

5.2.2 Other findings: Implications for test result reporting and incidental findings

Some genes and variants may be associated with more than one disease, for example the *apoE* gene (hypercholesterolemia and Alzheimer disease), requiring the laboratory to consider the disclosure of information not related to the indication for testing[96, 97]. Other criteria are used in the reporting of pharmacogenetic results because the associated variants are not related to a disease state. Additionally, a combination of variants (haplotype), and not individual variants, determines metabolizer status, thus the combinations and phase of variants must be considered. In 2014, pharmacogenetic testing is primarily performed using other methods. One of the challenges for NGS is its current weakness in defining phase. Phasing variants based on the relatively short read sizes of current instrumentation is challenging although some methods do exist[98].

NGS, particularly when it is applied to exome and genome sequencing, may identify secondary or incidental findings that reveal carrier status, non-paternity, or a significant risk for a disease that is not related to the reason the test was ordered. In these instances, the workgroup recommended that the laboratory develop a policy describing how these data will be handled in terms of what is to be reviewed by the laboratory and what would be reported to the clinician and ultimately the patient. If a laboratory will report secondary findings, optimization and validation of the clinical test should include those regions in which incidental findings may be found. The ACMG published a policy that certain incidental findings obtained from exome and genome clinical testing should be reported. The policy provided a list of those diseases, genes and variant types thought to be clinically actionable[99] and recommended reporting incidental findings associated with variants known or expected to be pathogenic.

5.2.3 Clinical Validation

Once the informatics pipeline for a clinical NGS test has been established and optimized, the next step is test validation. This topic was previously addressed[3,100]. NGS platforms, software, and supporting data are continuously evolving. **The informatics pipeline must be revalidated before the adoption of any new, updated, or re-optimized software or databases.** In some instances, only downstream processes need to be revalidated (e.g., a change in the annotation software should not influence the quality of the alignment protocols)[3]. As a consequence, the laboratory must integrate the decision to adopt changes, re-optimize, and re-validate into the overall workflow and projected costs in providing clinical NGS services.

6. Discussion

The informatics pipeline is an integral component of NGS. NGS can be a powerful tool for identifying sequence variations associated with a medical condition that may not be found using other available clinical testing methods. False positive and negative results can occur. When false positive results are likely, confirmatory testing using a different technology can be integrated into the testing algorithm. On the other hand, a false negative result can occur when clinically relevant findings are filtered out during the course of the analysis. These can be more difficult to detect but a rigorous optimization and validation of the test can minimize this occurrence and provide some sense of the likelihood for these to occur.

Clinical laboratory professionals typically understand the parameters associated with achieving a reliable analytic test result, but many are less experienced in the field of bioinformatics or the curation of a set of sequence variations that occur in genes that are not initially targeted for analysis, as is often the case for exome and genome analysis. The

Nature Biotechnology: doi:10.1038/nbt.3237

workgroup recommended that laboratory professionals work closely with informaticians to assure the quality and reliability of the informatics pipeline as a NGS test is being developed and optimized.

There are two general steps in the analysis of NGS data. The first is the determination of the genotype including sequence variations that differ from a reference sequence, and the second is the analysis of the genotype to determine the loci and variation(s) that are relevant to the patient in question. This latter step, in part, requires consideration of data obtained from external databases. Currently, there is no comprehensive, publically available, curated variant database to support variant interpretation. This type of "clinical grade" database is needed in order to ensure useful and reliable diagnostic testing in general, and NGS in particular, as the returned amount of data quickly exceeds a single laboratory's ability to properly assess all variants within a reasonable turnaround time. The databases that are currently consulted were originally developed for research applications. Nonetheless, these databases have been employed for clinical NGS testing with users typically cross-checking the data obtained against primary peer-reviewed literature to assess its relevance and validity. Manual retrieval and evaluation of data are time consuming, thus a useful feature of a "clinical-grade" database is the capacity to extract data and use it in an automated process for analysis and interpretation[101]. Efforts are underway to create databases designed for clinical applications to address these needs. For example, ClinGen, a joint effort between NCBI and several grantees funded through NGHRI (http://www.iccg.org/about-the-iccg/clingen, accessed August 18, 2014), are working to enhance the new ClinVar database (http://www.ncbi.nlm.nih.gov/clinvar/, accessed August 18, 2014), into a comprehensive and clinical grade resource. ClinVar currently archives reports of the relationships among human variations and phenotypes, along with supporting evidence,

Nature Biotechnology: doi:10.1038/nbt.3237

submitted by laboratories or curation projects. Interpretation of clinical NGS data is also available from commercial entities; however these organizations need to meet appropriate regulatory and professional standards.

While efforts like ClinVar can address some of the issues associated with data sharing, there is also a need for a gene-centric database that would allow clinical laboratories to annotate genes and curate gene-disease relationships. Clinical laboratories should be able to share information about the genes they are analyzing and aggregated assertions about variants, but may have some challenges sharing patient-level variant observations, due to patient confidentiality requirements[76]. However, the ClinGen Resource is developing additional approaches to support the sharing of patient-level data that ensure patient privacy is protected.

Widespread use of genomics in the clinical setting will also require appropriate decision support systems to help clinicians interpret possibly pathogenic genomic variants, integrate genomic information into diagnosis, and guide selection of preventative and personalized/stratified therapeutic options. Most clinical decision support systems consist of three parts: a dynamic knowledge base; an inference engine based on consensus evidence rules and requirements to determine the pathogenicity for each type of variant; and an appropriate mechanism for communication with the health-care professional (or patient)[76, 102]. In genomic terms, this might equate to: a database (or databases) of genotype–phenotype associations, an analysis pipeline to prioritize a list of candidate variants of interest to a particular patient, and a user-friendly portal for inputting, accessing, and visualizing patient data both at the diagnostic laboratory and the clinic. Standardized representation of genomic and non-genomic patient data is essential to ensure reliable computer-based interpretation and processing[101, 103].

Nature Biotechnology: doi:10.1038/nbt.3237

Another shortcoming identified by the workgroup, but not addressed in the primary discussion, is the practical difficulty of sharing variant-level data among laboratories during test development and patient testing. This sharing is essential for inter-laboratory comparison of data to determine the concordance among laboratories to identify variants. Current file specifications (e.g., VCF, GVF) do not provide a strict enough definition of parameters to allow data comparison[12, 13]. For example, some laboratories deposit all variant calls, including some outside the intended reportable range, into their VCF file with minimum filtering. Other laboratories deposit only those variant calls within their intended reportable range after filtering to remove those that do not meet certain quality criteria. To address this issue, **the workgroup recommended that a new effort be initiated to establish a "clinical-grade" VCF or equivalent file format specification to facilitate interoperability of clinical laboratory and health IT systems. This will facilitate data sharing among laboratories and with proficiency testing programs for quality assurance, with databases that are used to support variant interpretation, and for other purposes.** These other purposes may include outsourcing of variant data for downstream informatics analysis and interpretation, deposition of genomic data to a medical database, and messaging to a patient's electronic medical record or to cloud storage for future analysis as warranted by new data or indications for testing. While it is not likely that the variant file alone generated during NGS sequencing will be the primary means for messaging genomic data sets from the clinical laboratory to other entities, its content needs to be standardized to facilitate interoperability. In developing standards for genomic data representation, there should be compliance with established practices for the description and exchange of electronic health information. As a consequence of this recommendation, the CDC,

Nature Biotechnology: doi:10.1038/nbt.3237

in collaboration with other federal partners, organized and is actively facilitating a national workgroup tasked with meeting these objectives.

The principles and recommendations described in this document are relevant to the design and optimization of the informatics pipeline based on current platforms and software tools. It is expected that at some point in time there will be robust end-to-end solutions able to handle the informatics demands of NGS available through integrated software packages developed for clinical applications. This would help to reduce the burden associated with in-house test development. It is likely that alignment will become a more accurate and simplified process as read lengths increase with advances in the sequencing chemistry and instrumentation. This may also reduce the laboratory's cost to assemble and optimize an informatics pipeline including the significant need for services provided by an informatician. Data sharing to build up a reliable set of genotype/phenotype correlations will always be important.

The recommendations in this guideline can be implemented by clinical genetic testing laboratories to improve the development and optimization of their informatics processes. Endorsement of these recommendations as part of professional or regulatory guidelines could assure widespread standardization of the laboratory informatics processes.

Acknowledgements

The research was supported in part by an appointment to A.S.G. to the Research Participation Program at the CDC administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the US Department of Energy and the CDC. H.L.R. was supported in part by National Institutes of Health grants U01HG006500 and U41HG006834.

7. References

1. Stitziel, N.O., Kiezun, A. & Sunyaev, S. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol.* **12** (2011).

2. Collins, F.S. & Hamburg, M.A. First FDA authorization for next-generation sequencer. *N. Engl. J. Med.* **369**, 2369-2371 (2013).

3. Gargis, A.S. *et al*. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat. Biotechnol.* **30**, 1033-1036 (2012).

4. Centers for Medicare and Medicaid Services. US Department of Health and Human Services. Part 493—Laboratory Requirements: Clinical Laboratory Improvement Amendments of 1988. 42 CFR §493.1443-1495.

5. Rehm, H.L. *et al*. ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med.* **15**, 733-747 (2013).

6. Jennings, L., Van Deerlin, V.M., Gulley, M.L. & College of American Pathologists Molecular Pathology Resource, C. Recommended principles and practices for validating clinical molecular pathology tests. *Arch. Pathol. Lab. Med.* **133**, 743-755 (2009).

7. Mattocks, C.J. *et al*. A standardized framework for the validation and verification of clinical molecular genetic tests. *Eur. J. Hum. Genet.* **18**, 1276-1288 (2010).

8. Coonrod, E.M., Durtschi, J.D., Margraf, R.L. & Voelkerding, K.V. Developing genome and exome sequencing for candidate gene identification in inherited disorders: an integrated technical and bioinformatics approach. *Arch. Pathol. Lab. Med.* **137**, 415-433 (2013).

9. Cock, P.J., Fields, C.J., Goto, N., Heuer, M.L. & Rice, P.M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767-1771 (2010).

10. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

11. Narzisi, G. *et al*. Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly. doi: http://dx.doi.org/10.1101/001370.

12. Danecek, P. *et al*. The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

13. Reese, M.G. *et al*. A standard variation file format for human genome sequences. *Genome Biol.* **11**, R88 (2010).

14. Richards, C.S. *et al*. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet. Med.* **10**, 294-300 (2008).

15. Kearney, H.M. *et al*. American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genet. Med.* **13**, 680-685 (2011).

16. Clinical and Laboratory Standards Institute. Nucleic Acid Sequencing Methods in Diagnostic Laboratory Medicine; Approved Guideline, MM09-A2 (2014).

17. Lubin, I.M. *et al*. Clinician Perspectives about Molecular Genetic Testing for Heritable Conditions and Development of a Clinician-Friendly Laboratory Report. *J. Mol. Diagn.* **11**, 162-171 (2009).

18. Chen, B. *et al*. Good laboratory practices for molecular genetic testing for heritable diseases and conditions. *MMWR* **58**, 1-37 (2009).

19. Wang, J. *et al*. Clinical application of massively parallel sequencing in the molecular diagnosis of glycogen storage diseases of genetically heterogeneous origin. *Genet. Med.* **15**, 106-114 (2013).

20. Cui, H. *et al*. Comprehensive next-generation sequence analyses of the entire mitochondrial genome reveal new insights into the molecular diagnosis of mitochondrial DNA disorders. *Genet. Med.* **15**, 388-394 (2013).

21. Zhang, W., Cui, H. & Wong, L.J. Comprehensive one-step molecular analyses of mitochondrial genome by massively parallel sequencing. *Clin. Chem.* **58**, 1322-1331 (2012).

22. Sule, G. *et al*. Next-generation sequencing for disorders of low and high bone mineral density. *Osteoporosis Int.* **24**, 2253-2259 (2013).

23. Jones, M.A. *et al*. Molecular diagnostic testing for congenital disorders of glycosylation (CDG): Detection rate for single gene testing and next generation sequencing panel testing. *Mol. Genet. Metab*. **110**, 78-85 (2013).

24. Jones, M.A. *et al*. Targeted polymerase chain reaction-based enrichment and next generation sequencing for diagnostic testing of congenital disorders of glycosylation. *Genet. Med.* **13**, 921-932 (2011).

25. Chin, E.L.H., da Silva, C. & Hegde, M. Assessment of clinical analytical sensitivity and specificity of next-generation sequencing for detection of simple and complex mutations. *BMC Genet.* **14** (2013).

26. Valencia, C.A. *et al*. Comprehensive Mutation Analysis for Congenital Muscular Dystrophy: A Clinical PCR-Based Enrichment and Next-Generation Sequencing Panel. *PLoS One* **8,** e53083 (2013).

27. Meyer, M., Stenzel, U., Myles, S., Prufer, K. & Hofreiter, M. Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res.* **35**, e97 (2007).

28. Craig, D.W. *et al*. Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* **5**, 887-893 (2008).

29. Cronn, R. *et al*. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* **36** (2008).

30. Harismendy, O. & Frazer, K.A. Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. *Biotechniques* **46**, 229-231 (2009).

31. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939-946 (2012).

32. Mamanova, L. *et al*. Target-enrichment strategies for next-generation sequencing (vol 7, pg 111, 2010). *Nat. Methods* **7**, 479-479 (2010).

33. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40** (2012).

34. Binladen, J. *et al*. The Use of Coded PCR Primers Enables High-Throughput Sequencing of Multiple Homolog Amplification Products by 454 Parallel Sequencing. *PLoS One* **2**, e197 (2007).

35. Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J. & Knight, R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods* **5**, 235-237 (2008).

36. Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S. & Hoekstra, H.E. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS One* **7**, e37135 (2012).

37. Bystrykh, L.V. Generalized DNA Barcode Design Based on Hamming Codes. *Plos One* **7**, e36852 (2012).

38. Jun, G. *et al*. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839-848 (2012).

39. Zhang, V. in Next Generation Sequencing. (ed. L.-J.C. Wong) 79-96 (Springer New York, 2013).

Nature Biotechnology: doi:10.1038/nbt.3237

40. Sudmant, P.H. *et al*. Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-646 (2010).

41. Church, D.M. *et al*. Modernizing Reference Genome Assemblies. *PLoS Biol.* **9**, e1001091 (2011).

42. Pabinger, S. et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* (2013).

43. Yu, X.Q. *et al*. How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? *BioData. Min.* **5** (2012). doi: 10.1186/1756-0381-5-6.

44. Ruffalo, M., LaFramboise, T. & Koyuturk, M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* **27**, 2790-2796 (2011).

45. Li, H. & Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.* **11**, 473-483 (2010).

46. Flicek, P. & Birney, E. Sense from sequence reads: methods for alignment and assembly (vol 6, pg S6, 2009). *Nat. Methods* **7**, 479-479 (2010).

47. Merriman, B., Rothberg, J.M. & Team, I.T.R.D. Progress in Ion Torrent semiconductor chip based sequencing. *Electrophoresis* **33**, 3397-3417 (2012).

48. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851-1858 (2008).

49. Homer, N., Merriman, B. & Nelson, S.F. BFAST: An Alignment Tool for Large Scale Genome Resequencing. *PLoS One* **4**, A95-A106 (2009).

50. Burrows M, W.D. A block-sorting lossless data compression algorithm. *Technical Report 124*, Digital Equipment Corporation (1994).

51. Ferragina, P. & Manzini, G. Opportunistic data structures with applications, FOCS '00 Proceedings of the 41st Annual Symposium on Foundations of Computer Science (2000), http://people.unipmn.it/manzini/papers/focs00draft.pdf, accessed June 30, 2014.

52. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359 (2012).

53. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

54. Li, R.Q. *et al*. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967 (2009).

55. Chaisson, M.J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13** (2012).

56. Oliver GR. Considerations for clinical read alignment and mutational profiling using next-generation sequencing [v2; ref status: indexed, http://f1000r.es/NMpsFc] F1000Research 2012, 1:2 (doi: 10.12688/f1000research.1-2.v2).

57. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv:1303.3997v2 [q-bio.GN] (2013).

58. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

59. Francey, L.J. *et al*. Genome-wide SNP genotyping identifies the Stereocilin (STRC) gene as a major contributor to pediatric bilateral sensorineural hearing impairment. *Am. J. Med. Genet. A.* **158A**, 298-308 (2012).

60. Nielsen, R., Paul, J.S., Albrechtsen, A. & Song, Y.S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443-451 (2011).

61. Neuman, J.A., Isakov, O. & Shomron, N. Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. *Brief. Bioinform.* **14**, 46-55 (2013).

62. Homer, N. & Nelson, S.F. Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. *Genome Biol.* **11** (2010).

63. DePristo, M.A. *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491-498 (2011).

64. McKenna, A. *et al*. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-1303 (2010).

65. Hamada, M., Wijaya, E., Frith, M.C. & Asai, K. Probabilistic alignments with quality scores: an application to short-read mapping toward accurate SNP/indel detection. *Bioinformatics* **27**, 3085-3092 (2011).

66. Liu, X.T., Han, S.Z., Wang, Z.H., Gelernter, J. & Yang, B.Z. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS One* **8**, e75619 (2013).

67. Zook, J.M. *et al*. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* (2014).

68. Zook, J.M., Samarov, D., McDaniel, J., Sen, S.K. & Salit, M. Synthetic spike-in standards improve run-specific systematic error analysis for DNA and RNA sequencing. *Plos One* **7**, e41356 (2012).

69. Lysholm, F., Andersson, B. & Persson, B. FAAST: Flow-space Assisted Alignment Search Tool. *BMC Bioinformatics* **12** (2011).

70. Gilissen, C., Hoischen, A., Brunner, H.G. & Veltman, J.A. Disease gene identification strategies for exome sequencing. *Eur. J. Hum. Genet.* **20**, 490-497 (2012).

71. O'Rawe, J. *et al*. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* **5**, 28 (2013).

72. Altmann, A. *et al*. A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Hum. Genet.* **131**, 1541-1554 (2012).

73. Lyon, G.J. & Wang, K. Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress. *Genome Med.* **4** (2012).

74. Frampton, M. & Houlston, R. Generation of artificial FASTQ files to evaluate the performance of next-generation sequencing pipelines. *PLoS One* **7**, e49110 (2012).

75. Fajardo, K.V.F. *et al*. Detecting false-positive signals in exome sequencing. *Hum. Mutat.* **33**, 609-613 (2012).

76. Bean, L.J., Tinker, S.W., da Silva, C. & Hegde, M.R. Free the data: one laboratory's approach to knowledge-based genomic variant classification and preparation for EMR integration of genomic data. *Hum. Mutat.* **34**, 1183-1188 (2013).

77. Duzkale, H. *et al*. A systematic approach to assessing the clinical significance of genetic variants. *Clin. Genet.* **84**, 453-463 (2013).

78. Siepel, A. *et al*. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034-1050 (2005).

79. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073-1082 (2009).

80. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

81. Yandell, M. *et al*. A probabilistic disease-gene finder for personal genomes. *Genome Res.* **21**, 1529-1542 (2011).

82. McLaren, W. *et al*. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-2070 (2010).

83. Landrum, M.J. *et al*. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* (2013).

84. Bell, C.J. et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci. Transl. Med.* **3**, 65ra64 (2011).

85. Woolfe, A., Mullikin, J.C. & Elnitski, L. Genomic features defining exonic variants that modulate splicing. *Genome Biol.* **11** (2010).

86. Hu, H. *et al*. VAAST 2.0: Improved Variant Classification and Disease-Gene Identification Using a Conservation-Controlled Amino Acid Substitution Matrix. *Genet. Epidemiol.* **37**, 622-634 (2013).

87. Kolker, S. *et al*. Diagnosis and management of glutaric aciduria type I--revised recommendations. *J.Inherit. Metab. Dis.* **34**, 677-694 (2011).

88. Flanagan, S.E., Patch, A.M. & Ellard, S. Using SIFT and PolyPhen to Predict Loss-of-Function and Gain-of-Function Mutations. *Genet. Test. Mol. Bioma.* **14**, 533-537 (2010).

89. Ohanian, M., Otway, R. & Fatkin, D. Heuristic methods for finding pathogenic variants in gene coding sequences. *J. Am. Heart Assoc.* **1**, e002642 (2012).

90. Castellana, S. & Mazza, T. Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. *Brief. Bioinform.* **14**, 448-459 (2013).

91. Vihinen, M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics* **13** (2012).

92. Thusberg, J., Olatubosun, A. & Vihinen, M. Performance of Mutation Pathogenicity Prediction Methods on Missense Variants. *Hum. Mutat.* **32**, 358-368 (2011).

93. Santani, A., Gowrishankar, S. , da Silva, C. , Mandelkar, D., Sasson, A., Sarmady, M., Shakhbatyan, R., Tinker, S., Church, D., Funke, B., Hegde, M. The Medical Exome Project: From concept to implementation. *American Society of Human Genetics 2013 Meeting Abstract* (2013).

94. Berg, J.S. *et al*. Processes and preliminary outputs for identification of actionable genes as incidental findings in genomic sequence data in the Clinical Sequencing Exploratory Research Consortium. *Genet. Med.* **15**, 860-867 (2013).

95. Saunders, C.J. *et al*. Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units. *Sci. Transl. Med.* **4** (2012).

Nature Biotechnology: doi:10.1038/nbt.3237

96. Cash, J.G. *et al*. Apolipoprotein E4 Impairs Macrophage Efferocytosis and Potentiates Apoptosis by Accelerating Endoplasmic Reticulum Stress. *J. Biol. Chem.* **287**, 27876-27884 (2012).

97. Nalls, M.A. *et al*. A Multicenter Study of Glucocerebrosidase Mutations in Dementia With Lewy Bodies. *JAMA Neurol.* **70**, 727-735 (2013).

98. Browning, S.R. & Browning, B.L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703-714 (2011).

99. Green, R.C. *et al*. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15**, 565-574 (2013).

100. Aziz, N. *et al.* College of American Pathologists' Laboratory Standards for Next-Generation Sequencing Clinical Tests. *Arch. Pathol. Lab. Med.* [Epub ahead of print], (2014).

101. Moorthie, S., Hall, A. & Wright, C.F. Informatics and clinical genome sequencing: opening the black box. *Genet. Med.* **15**, 165-171 (2013).

102. Sintchenko, V. & Coiera, E. Developing decision support systems in clinical bioinformatics. *Methods Mol. Med.* **141**, 331-351 (2008).

103. Kawamoto, K., Lobach, D.F., Willard, H.F. & Ginsburg, G.S. A national clinical decision support infrastructure to enable the widespread and consistent practice of genomic and personalized medicine. *BMC Med. Inform. Decis. Mak.* **9**, 17 (2009).

Nature Biotechnology: doi:10.1038/nbt.3237