Review

Navigating Web-Based Resources for Genetic Testing of Chromosome Abnormalities, CNVs and Gene Mutations

Bixia Xiang, PhD;[1]* Fang Xu, PhD;[2] Wenqi Zeng, MD, PhD;[3] Dan Zi, MD;[4,5] Deqiong Ma, MD, PhD[6]*

[1] *BayCare Laboratories, LLC, Tampa, FL*
[2] *Department of Genetics, Yale University School of Medicine, New Haven, CT*
[3] *Department of Molecular Genetics at Quest Diagnostics, San Juan Capistrano, CA*
[4] *Department of Obstetrics and Gynecology, Guiyang Medical University, Guizhou, China*
[5] *College of Pharmacy, University of South Florida, Tampa, FL*
[6] *Department of Pathology, Montefiore Medical Center, Albert Einstein College of Medicine, Yeshive University, New York, NY*

Current clinical genetic and genomic testing involves genome-wide evaluation of chromosomal abnormalities, copy number variants (CNVs) and gene mutations. The major challenge facing genetic laboratory directors, physicians and counselors is to distinguish pathogenic variants from variants of unknown clinical significance (VOUS) and benign polymorphic variants. Various genetic and genomic databases were generated and maintained to facilitate the interpretation process. Those databases typically present collections of specific types of genetic abnormalities with cross references all relevant clinical findings and biological knowledge. This paper outlines the prevailing web-based resources used for genetic and genomic testing results interpretation in three categories: chromosomal abnormalities, CNVs, and gene mutations. Routine routes on utilizing these web resources in clinical setting are provided and some limitations are discussed.
[N A J Med Sci. 2014;7(4):163-170. **DOI:** *10.7156/najms.2014.0704163]*

Key Words: *genetic and genetic testing, CNV, gene mutations, genetic variants, clinical interpretation, web-based databases*

OMIM AND UCSC GENOME BROWSER

Rapid technology advances have brought clinical genetic testing to the genomic era. Nowadays cytogenetic analysis employs not only the traditional chromosome karyotyping and Fluorescence In Situ Hybridization (FISH) but also array Comparative Genomic Hybridization microarrays (aCGH) and single-nucleotide polymorphism (SNP) microarrays, which enable the identification of copy number variants (CNVs) with a much higher resolution and becomes the first tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies.[1] Molecular methods such as Southern Blot, polymerase chain reaction (PCR), Sanger Sequencing, and Multiplex Ligation-dependent Probe Amplification (MLPA) are still in use particularly for gene-specific analysis. Massive parallel sequencing or next- or third-generation sequencing (NGS) has made a tremendous progress in the characterization of molecular basis of genetic related conditions. It has been recently applied to clinical diagnosis for children with rare diseases of unknown etiology and patients with refractory cancers.[2-4] Another immediate area with great potential will be pharmacogenomics for cancer treatment although much work still needs to be done to establish its clinical diagnostic and prognostic role.[5,6] Coming along naturally with each technology are those compatible databases developed with different format, from different resources, to record the clinical experiences, to generate new knowledge, to facilitate interpretation of genetic and genomic testing results.

Two most fundamental web-based resources are Online Mendelian Inheritance in Man (OMIM) and UCSC Genome browser. OMIM (http://www.omim.org/) is a continuously updated online catalog of human genes and genetic disorders with focus on genotype-phenotype correlation. OMIM contains information on all known Mendelian disorders and over 12,000 genes. It contains full-text summaries of information from the scientific literatures and provides links to the references as well as other genomic resource tools. In a clinical setting, it has been used as the first step to find the known information for any gene. OMIM data are commonly used as tracks or hyperlinked text inside many bioinformatics tools. The UCSC Genome Browser (http://genome.ucsc.edu) provides convenient access to human genome sequence, annotations, and bioinformatics tools all together that enable detailed analysis of genomic data. It serves as a data aggregator for annotating and visualizing regions of interest across publicly available or custom-built data sets. The heavily annotated human genome data can be displayed

Received: 09/17/2014; Revised: 10/22/2014; Accepted: 10/25/2014
***Corresponding Authors:** Bixia Xiang, PhD, FACMG, Director of Cytogenetics Laboratory, BayCare Laboratories LLC, 5455 W. Waters Ave, Tampa, FL 33634. Tel: 813-443-8094; Fax: 813-443-8095.
(Email: bixia.xiang@baycare.org)
Deqiong Ma, MD, PhD, Molecular Pathology and Cytogenetics lab, Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, NY 10461. Tel: 718-405-8074; Fax: 718-405-8075.
(Email: dema@montefiore.org)

graphically as 'tracks' align to the genomic sequence and grouped according to common features, such as Variation, Phenotype and Literature, Gene predictions, et al. It allows users to add and view any given piece of genome at any scale and any type of annotations. These functionalities are crucial to the daily application in the clinical genetic/genomic testing setting. In addition to these two prominent websites, other frequently used web-based resources (in **Table 1**) for chromosomal abnormalities, CNVs, and gene mutations interpretation are outlined and their limitations are discussed.

Table 1. Web Resources for Genetic and Genomic Testing.

Databases	Weblinks	Clinical application		
		Chromosome	CNV	Mutation
Online Mendelian Inheritance in Man (OMIM)	http://www.omim.org/	√	√	√
The UCSC Genome Browser	http://genome.ucsc.edu	√	√	√
The Database of Chromosomal Mosaicism	http://mosaicism.cfri.ca	√		
The Database of small Supernumery Marker Chromosomes	http://ssmc-tl.com/sSMC.html	√		
The Atlas of Genetics and Cytogenetics in Oncology and Hematology	http://www.infobiogen.fr/services/chromcancer/	√		
Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer	http://cgap.nic.nih.gov/Chromosomes/Mitelman	√		
Database of Genomic Variants (DGV)	http://dgv.tcag.ca/dgv/app/home		√	
The International Standards for Cytogenomic Arrays (ISCA) Consortium	https://www.iscaconsortium.org/		√	
The Database of Genomic Structural Variation (dbVar)	http://www.ncbi.nlm.nih.gov/dbvar/studies/nstd37/		√	√
DECIPHER	https://decipher.sanger.ac.uk/syndromes#overview		√	√
ANNOVAR	http://wannovar2.usc.edu/			√
SeattleSeq	http://wannovar2.usc.edu/			√
Exomiser	http://www.sanger.ac.uk/resources/databases/exomiser/			√
SNP-Nexus	http://www.snp-nexus.org/			√
Exome Variant Server (EVS)	http://evs.gs.washington.edu/EVS/			√
1000 genome	http://www.1000genomes.org/data			√
dbSNP	http://www.ncbi.nlm.nih.gov/SNP/			√
Human Gene Mutation Database (HGMD)	http://www.hgmd.cf.ac.uk/ac/index.php			√
Locus/Disease/Ethnic/Other-Specific Databases (LSDB)	http://www.hgvs.org/biblio.html			√
BioMuta	https://hive.biochemistry.gwu.edu/tools/biomuta/index.php			√
Catalogue of Somatic Mutations in Cancer (COSMIC)	http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/			√
Clinical Genome Resource (ClinGen)	http://www.clinicalgenome.org/			√
MedGen	http://www.ncbi.nlm.nih.gov/medgen/	√	√	√

ATLAS AND DATABASES FOR CHROMOSOMAL ABNORMALITIES

Chromosomal heteromorphisms have been recognized for over four decades.[7,8] They are defined as heritable variations at specific chromosomal regions with no proven impact on phenotype. Common heteromorphisms include hetero-chromatin variations of chromosomes 1, 9, 16 and Y and also prominent short arms, satellites and stalks on acrocentric chromosomes. A survey administered by the Cytogenetics Resource Committee of the College of American Pathologists and the American College of Medical Genetics and Genomics (ACMG) summarized the reporting practices of chromosome heteromorphisms.[9] A comprehensive view of human chromosome heteromorphisms has been provided in the Human Chromosome Variation: Heteromorphisms and Polymorphism which is a desk reference for clinical cytogenetic laboratories.[10] For constitutional syndromic chromosomal abnormalities including whole chromosome aneuploidy and structural chromosome aberrations, there have been plenty of publications and web sources to summarize their genotype-phenotype correlations. For chromosome mosaicism and small supernumerary marker chromosome (sSMC), specifically designed websites are available to help interpret the results. Additionally, for somatic chromosomal abnormalities in cancer, the reported clonal chromosomal abnormalities in various type of cancer have also been compiled into web databases.

Chromosome Mosaicism

Chromosome mosaicsim is defined as a condition when an individual is found to have two or more cell populations with divergent chromosome contents, such as monosomy or trisomy in a portion of cells and a normal karyotype in the remaining cells.[11] Conventional metaphase G-banding analysis and interphase FISH are the most reliable techniques to identify mosaicism on a cell-by-cell basis. With the advent of aCGH and SNP microarrays, the array-based testing has also been used to detect mosaicism of submicroscopic abnormalities.[12] It is always difficult and sometime impossible to predict the clinical presentation of individuals with chromosomal mosaicsim because the mosaic pattern was mostly detected in the submitted peripheral blood specimen and rarely further defined in other tissues. The Database of Chromosomal Mosaicism (http://mosaicism.

cfri.ca) provides systematic data about clinical outcome specific to each chromosome. It summarizes known effects for a specific chromosome mosaicism and can be considered a guide to patients and families. At each chromosome link, it reviews some of the relevant cases reports in the literature, discusses the occurrence of confined placental mosaicism and the potential implications of uniparental disomy, and provides links to useful information about relevant gene maps or associated genetic disorders. A list of references is provided at the bottom of each page with links to the abstracts in PubMed for more details on the case reports and studies cited.

Small Supernumerary Marker Chromosome (sSMC)

A sSMC is a marker chromosome detectable by conventional cytogenetic method but its chromosomal origin and gene content remain uncharacterized due to its small size.[13] Further molecular characterization of an sSMC using FISH and aCGH is required for interpreting its clinical significance.[14,15] The Database of small Supernumerary Marker Chromosomes (http://ssmc-tl.com/sSMC.html) is created by Institute of Human Genetics in German with more than 5,250 sSMC cases collected in the database. The aims of the database are to collect all available sSMC case reports, define critical regions for partial trisomy or tetrasomy due to the presence of sSMC, and provide information for patients and clinicians. The online sSMC database has chromosome-specific pages with cases classified by four categories: cases without clinical findings, cases with clinical findings, cases with unclear clinical correlation and cases with neocentromeres. At the beginning of each chromosome-specific page, there are schematic drawings describing the presently known dosage sensitive centromere-near regions. For all sSMC cases, detailed cytogenetic information, clinical symptoms, and related references are provided.

Cancer Cytogenetics

The correlation of somatic clonal chromosomal abnormalities with different types of human malignancies has been widely recognized. To catalog recurrent chromosomal abnormalities in cancer in a systematic and concise way, the Atlas of Genetics and Cytogenetics in Oncology and Hematology (http://atlasgeneticsoncology.org/) has been made by cytogeneticists, molecular biologists and clinicians in oncology, hematology and pathology. This web-based resource reviews and summarizes genes involved in cancer, cytogenetics and clinical entities in cancer, i.e. leukemia, solid tumor and cancer-prone disease.[16] It presents concise and updated information on recurrent chromosome abnormalities and involved gene rearrangements in various types of cancers. It also contains 'Deep Insights' with traditional review articles focusing on a specific aspect such as chromothripsis, centrosome, autophagy and so forth, 'Case Reports' dedicated on rare cytogenetic anomalies in hematological malignancies, links toward websites and databases devoted to cancer and genetics, and education materials in genetics. The information helps cytogeneticists to comment on chromosomal findings for cancer classification and guide treatment decision making for clinicians. In the database, entities can be easily accessed either by theme (cancer genes, leukemia, solid tumor, cancer-prone disease) or by chromosome number. In the latter case, chromosomes are displayed in numerical order and cancer genes involved in the specific chromosome may be displayed in alphabetical order or in physical order from pter to qter. Two search formats, a quick/simple search and an advanced search, are also available on the Atlas home page. The advanced search offers a combination of forms and menus of search terms for complex queries. In addition, at the bottom of the page it provides a useful link to Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (http://cgap.nci.nih.gov/Chromosomes/Mitelman) which provide tools to search for recurrent chromosome aberrations in cancer.

THE DGV-ISCA-DECIPHER ROUTE FOR CNVs

"Extensive CNVs exist in the human genome" was first reported and evidenced by microarray studies about ten years ago.[17,18] Cytogeneticists have long known and practiced the concept of big CNV through the recognition of extra or missing chromosome fragments with causative conditions. Using genome-wide array technologies to detect the small CNVs in the clinical setting were adopted by the clinical cytogenetics community immediately. A CNV is defined as the DNA segment >1kilobase (kb) in size with copy number differing from two copies. The presence of CNV likely reflect errors from DNA recombination and replication machineries.[19] CNVs could cause phenotypes by gene dosage, gene disruption, gene fusion, and position effects; they are bound to have vital role in Mendelian diseases, sporadic diseases, complex diseases, disease susceptibility, and drug response.[20,21]

aCGH and SNP array technologies have been validated and widely used in the clinical diagnoses of constitutional cytogenomic abnormalities.[22,23] The ACMG Practice Guidelines have recommended microarray as the first-tier test for patients with developmental delay and intellectual disability, congenital anomalies, and dysmorphic features.[24] Current array technologies allow reliable detection of CNVs larger than 50 kb and the reported diagnostic yield for pediatric patients are around 10-25% which outperform the 5–6% yield by karyotyping and subtelomeric FISH.[25] The American College of Obstetricians and Gynecologists Committee on Genetics recommended microarray analysis to replace the need for fetal karyotype for patients with a fetus with one or more major structural abnormalities identified on ultrasonographic examination and who are undergoing invasive prenatal diagnosis.[26] Besides, some studies also show that genomic characterization of the structural abnormalities aided in the prediction of clinical outcomes for prenatal genetic counseling.[27] Many online databases to catalog and search for CNVs in normal and/or disease populations have been developed to facilitate the CNVs interpretation in a clinical setting.

CNV Databases for General Population

Database of Genomic Variants (DGV) (http://dgv.tcag.ca/dgv/app/home) defines structural variations as genomic alterations involving segments of DNA

that are > 50 base pair (bp). DGV provides a comprehensive summary of structural variation with greater than 50 bp and less than 3 Megabase (Mb) (10 Mb for inversions) from the general population. The database is continuously updated with new data from peer reviewed studies. Since its start from 2004, DGV contains 109,863 CNVs and 238 inversions collected from 55 published studies where 107 array platforms were used.[28] While using DGV database, it is important to remember the following facts: the CNVs being identified are through different platforms, the CNV data sets are reported at a sample by sample level, and the CNV calls with similar boundaries are merged across the sample sets. Only variants of the same type are merged, therefore, inversions, gains and losses are merged separately; sample level calls that overlap by >= 70% are merged in this process. If several different platforms and approaches are used within the same study, these data sets are merged separately. Due to the fact that the probe coverage and resolution may differ significantly among more than one hundred platforms, the boundaries of CNVs reported in DGV are often inaccurate. It is also often difficult to know for sure if a variant found using different platforms is the exact same as annotated in DGV. Many CNVs have overestimated boundaries, which leads to an exaggeration of the number of features overlapping CNVs. Therefore, a CNV listed in the DGV does not mean that a similar CNV cannot be disease causing in a patient sample. Similarly, a lack of CNVs in a specific region of the database does not necessarily mean there are no common CNVs at that locus. Specifically, the BAC arrays based dataset tend to significantly overestimate the size of variants. They are less reliable and did not include an estimation of the false discovery rate. The DGV therefore does contain data that represent false positives. Generally speaking, CNVs detected in many studies or by independent platforms are most likely real. Large variants identified in a single sample by a single study represent either extremely rare variants or may be false positives. At the same time, there are still a lot of smaller CNVs (<30 kb) that remain to be identified.[28]

Besides DGV, clinical laboratories also frequently refer to other two benign CNV databases. One is "CHOP CNV dataset" which presents a CNV database detected in 2,026 disease-free individuals using a uniform high-density, SNP-based oligonucleotide microarrays and computational process.[29] CHOP CNV database catalogued and characterized 54,462 individual CNVs, 77.8% of which were identified in multiple unrelated individuals. Another one is the "Itsara 2009 CNV Data" which identified CNVs in ~2,500 individuals by using Illumina SNP data, with an emphasis on ‘‘hotspots’’ prone to recurrent mutations.[30] This study finds variants larger than 500 kb in 5%–10% of individuals and variants greater than 1 Mb in 1%–2%. This sample size permits a robust distinction between truly rare and polymorphic but low-frequency CNV. A significant fraction of individual CNVs larger than 100 kb are rare and both gene density and size are strongly uncorrelated with allele frequency. Although large CNVs are generally deleterious, the size of CNVs alone cannot be used as a predictor of pathogenicity because such variations commonly exist in normal individuals. Together, those benign CNV databases are available to be imported as online tracks and provide a useful resource in distinguishing CNVs with pathologic significance from normal variants.

CNV Databases for Individuals with Constitutional Conditions

The International Standards for Cytogenomic Arrays (ISCA) Consortium (https://www.iscaconsortium.org/) now as The International Collaboration for Clinical Genomics (www.iccg.org) provides large publicly available database and forum where clinicians and researchers can share knowledge to expedite the understanding of CNV in patients with intellectual disability, autism, and developmental delay. Diagnostic laboratories performing chromosome microarray testing are the major data contributors and users. So far the ISCA database contains over 13,000 CNVs identified from over 28,000 patients, as well as information on the clinical interpretation of each CNV as determined by the submitting laboratories. The submission of clinical information is encouraged but not required. Electronic forms with tools to facilitate genotype and phenotype data submission are designed and free available for use. CNV calls with their clinical interpretations are available through the Database of Genomic Structural Variation (dbVar) at the NCBI (http://www.ncbi.nlm.nih.gov/dbvar/studies/nstd37/) and the UCSC Genome Browser. Phenotype information is also available within the public database for a subset of cases. Data within the ISCA database are curated on several different levels. The "ISCA Curated Pathogenic CNVs" represent pathogenic CNVs that have been assessed by the ISCA evidence-based review committee. The ISCA pathogenic, likely pathogenic, uncertain, likely benign, and benign CNV tracks include imbalances that have been interpreted as such by the ISCA submitting clinical laboratories. At present, these tracks have not been reviewed by the evidence-based review committee. The "ISCA Curated Benign CNVs" track includes imbalances that are known to be variable in normal populations based on the DGV and/or other databases and have been reviewed by the ISCA Review Committee. The ISCA list of pathogenic and benign regions is available online (http://www.ncbi.nlm.nih.gov/dbvar/studies/nstd45/) or as a user defined track.

DECIPHER stands for DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (https://decipher.sanger.ac.uk/syndromes#overview). It is an interactive web-based database to facilitate the interpretation of data from genome-wide analyses. It utilizes the nearly completed human reference sequence via Ensembl and other genome browsers to define which genes are involved in a specific CNV (microdeletion / microduplication) and which sequence variants are positioned within a gene or regulatory element. DECIPHER currently contains 70 syndromes which are supported by more than 10,000 cases contributed by a global network of >200 academic centers. Each contributing center has a nominated rare disease clinician or clinical geneticist who is responsible for overseeing data entry and membership for their center. DECIPHER enables a flexible approach to data-sharing. With patient consent, positional genomic information together with a brief description of the

associated phenotype becomes viewable without password protection. DECIPHER is equipped with a powerful search engine and can be rapidly searched by phenotype, by syndrome, by overlapping position, by cytogenetics band, and by gene. Under each genotype, the Browser shows the genes involved, DECIPHER Syndromes, SNVs, InDels, DECIPHER CNVs, dbSNP, Population CNVs, ISCA, HGMD, ClinVar, LSDB Variants, Research data, etc., as tracks. The patient's CNV is displayed in the context of above both normal and pathogenic CNV reported at that locus thereby facilitating differentiation between the pathogenic and benign CNV, also provide the overview and citation for the final report of the pathogenic CNV.

The DGV-ISCA-DECIPHER route has been routinely used to interpret the CNV findings. All detected CNVs can be classified into three categories: pathogenic or likely pathogenic, VOUS, and likely benign CNV or benign CNV. In general, a CNV is defined as pathogenic or likely pathogenic if it (1) overlaps with a genomic region associated with a well-established syndrome listed in DECIPHER, OMIM morbid database, ISCA pathogenic database and/or internal pathogenic databases, (2) is large in size (> 3 Mb) with a rich gene content, (3) or contains a gene or part of a gene implicated in a known disorder. A CNV is defined likely benign CNV or benign CNV if it (1) overlaps with known polymorphic CNVs listed in the DGV, (2) is identical to the ones detected in healthy family members, (3) is gene-poor and contains no known disease-causing genes. All remaining CNVs can be classified as VOUS. The potential clinical significance of VOUSs will be evaluated by using PubMed literature search either by involved gene name or genomic location to capture any possible finding which is newly described and not included in any databases yet.

Cancer Cytogenomics Profiling

Cancer genomes typically contain somatic CNVs with two features: 1) the CNVs with chromosome arm length level occurs 30 times more frequent than CNVs with focal level, and 2) the average number of CNVs per tumor type is 40 CNVs per genome across all cancer type.[31]

Difficulty in interpreting complex somatic CNV data remains the largest obstacle for the widespread application of array technologies in hematology and oncology specimens. Currently, knowledge generated from cancer Atlas and WHO classification of tumors,[32] clinical reports of case series, and evidence-based reviews were used for diagnostic interpretation of somatic CNVs.[33,34] Since 2009, a committee of Cancer Genomics Consortium (CGC) (http://www.urmc.rochester.edu/ccmc/) was formed with a aim to set up a platform-neutral cancer cytogenomic database suitable for user to share cancer microarray data and to carry out multicenter cancer genome research.[35]

WEB RESOURCES FOR INTERPRETING GENE MUTATIONS FROM NGS

An efficient pipeline is essential for NGS analysis which involves base-calling, read alignment, variant calling and variant annotation. Recently, some clinical laboratory-based NGS analysis and reporting pipelines have been proposed.[36,37] A comprehensive survey has been conducted by several groups for an in-depth comparison of these existing pipelines.[38-40] In addition, an effort has been made among a total of more than 30 international groups through a "CLARITY challenge" competition in the past several years towards developing standard for best practices in analysis, interpretation and reporting of clinical genome sequencing. Based on their recent summary, a general convergence on most elements of the analysis and interpretation process has been reached. However, only two groups identified the consensus candidate variants in all disease cases. Obviously, the general accepted pipelines are still under ongoing fine-tuning in order to optimize and then standardize the interpretive and reporting process.[2]

To assist clinical laboratories to overcome some of the challenges and allow more uniform practice, ACMG has published clinical laboratory standards for NGS, which involves three components: sample preparation, sequencing, and data analysis.[41] In general, the evidence-based strategy could be simplified as four interpretation elements: 1) allele frequency information from the general population and disease-specific population; 2) pathogenic prediction by computational and predictive tools; 3) family and segregation data; and 4) the knowledge on disease-specific genotype-phenotype correlations, inheritance, penetrance, case-control and functional studies of mutations from potential candidate genes. Hereby, the following content focuses exclusively on the web-based resources that allow us to obtain ACMG proposed evidences for downstream variant interpretation.

Variant Annotation

The output from variant calling in a NGS pipeline is named as Variant Call Format (VCF) file, which needs to be decoded through annotation so that the types of variants along with their coverage and quality information could be present as a readable and understandable format. Currently, several annotation tools have been made available free online. The widely accepted annotation tool is ANNOVAR,[42] which has been integrated into the pipelines in the majority of sequencing facilities worldwide. The output provides the interpretation evidence for two major categories required by ACMG guideline including the allele frequency from control population databases and pathogenicity prediction scores for missense variants from a variety of tools including SIFT,[43] PolyPhen2 HDIV and HVAR,[44] LRT,[45] MutationTaster,[46] MutationAssessor,[47] FATHMM,[48] GERP++,[49] PhyloP[50] and Siphy,[51] which are retrieved from dbNSFP.[52] Other commonly used annotation tools include SnpEff [53] and VEP (http://useast.ensembl.org/info/docs/tools/vep/index.html/).

Some web servers have been developed to facilitate users analyze VCF files without using command-line tools, including ANNOVAR (http://wannovar2.usc.edu/), SeattleSeq (http://wannovar2.usc.edu/), Exomiser (http://www.sanger.ac.uk/resources/databases/exomiser/), SNP-Nexus (http://www.snp-nexus.org/), Additionally, web services developed specifically for clinical interpretation of genomic variants and clinical report generation are available for commercial purpose, such as Tute Genomics (http://

tutegenomics.com/) and Ingenuity Variant Analysis (http://www.ingenuity.com/products/variant-analysis/).

Variant Databases for General and Disease Populations

NGS pipeline annotate gene mutations by using both general and disease populations. Four well-known general population based databases are Exome Variant Server (EVS, http://evs.gs.washington.edu/EVS/), 1000 genome (http://www.1000genomes.org/data), dbSNP (< 50 bp, http://www.ncbi.nlm.nih.gov/SNP/) and dbVar (> 50 bp, http://www.ncbi.nlm.nih.gov/dbvar/). There are five commonly used disease databases: ClinVar, OMIM, Human Gene Mutation Database (HGMD), Locus/Disease/Ethnic/Other-Specific Databases (LSDB) and DECIPHER. HGMD is the most comprehensive resource for disease-causing variants identified in patients with cross reference to ClinVar, OMIM and LSDB. There are four sequence databases: NCBI Genome, RefSeq Gene, Locus Reference Genomic and MitoMap which are used as reference sequence to generate VCF file. In addition, for somatic changes, BioMuta (https://hive.biochemistry.gwu.edu/tools/biomuta/index.php) collects sequence features from the Catalogue of Somatic Mutations in Cancer (COSMIC), ClinVar, UniProtKB, and through biocuration of information available from publications.[54] It provides a framework for automated and manual curation and integration of cancer-related sequence features.

Cautions should be given while using the general population databases and disease-specific databases. The allele frequency cutoffs from the control databases (e.g. >5% stand-alone as benign; absence from controls as moderate evidence for pathogenicity) has been utilized as filters to help the clinical diagnostic laboratories determine the potential pathogenic or benign effect of variants. While the majority of rare variant (< 1%) are not detected by 1000 genome project simply because some regions could have no calls due to a low or incomplete coverage from the 4X sequencing read depth. Meanwhile, most EVS project participants are patients with lung and heart diseases. Therefore, the interpretation has to be cautious for rare variants either not being present in 1000 genome project or present in EVS. It is always recommended to generate the internal databases for both disease-specific populations and controls using the same sequencing platform via the same pipeline. It allows for a more accurate evaluation of variant allele frequency in populations with a comparable coverage and quality scores to rule out the potential spurious calls in certain genomic regions.

The current disease-specific databases appear to be way more complicated with a large variability in types of variants, accuracy, update frequency and curation process.[55] In addition, the majority of the clinical laboratories set up their own databases without open access to the publics. To optimize the use of these resources, a centralized resource of clinically annotated genes and variants, which is named as the Clinical Genome Resource (ClinGen) is created to improve our understanding of genomic variation. This ClinGen database (ClinGenDB) is managed under an expert curation system following a consistent evidence evaluation to make ClinVar a reliable resource for variants through standardized submission and classification.

***In Silicon* Prediction Tools**

The understanding of molecular basis for the vast majority of the genetic traits is either limited or incomplete. NGS does advance the discovery of the potential disease-causing variants. Given that the traditional experimental approaches are time consuming, many *In Silicon* tools have been developed during the past five years to identify functional effects and disease relationships for missense variants. In general, these tools differ in the prediction models but all are designed with a combination of various common features on sequence conservation, amino acid physiochemical properties, secondary structure, structure stability, B-factor, solvent accessibility, protein domain model, functional residues or splicing sites to predict the deleterious variants.[55]

Prediction tools for missense variants

Based on the report from CLARITY challenge, more than 80% teams applied both SIFT and PolyPhen for missense pathogenicity prediction.[2] The SIFT method utilized a machine learning attribute to assess the positional conservation sorely. PolyPhen is a similar method that addresses the problem of inter-dependence between sequences in an alignment, accounting for sequences in the alignment with high similarity. However, it is possible that the same variant could be predicted as deleterious by SIFT (sift<=0.05) but as benign by PolyPhen (pp2_hdiv<=0.452) or vice versa. Comparison studies have been carried out to investigate the sensitivity and specificity of SIFT and PolyPhen for loss-of-function and gain-of-function variant prediction. The findings suggested that SIFT and PolyPhen might be useful in prioritizing changes that are likely to cause a loss of protein function, their low specificity means that their predictions should be interpreted with caution and further evidence to support or refute pathogenicity should be sought before reporting novel missense changes.[56] More recently, dbNSFP is developed for functional prediction and annotation of all potential missense variants in the human genome. Until now, it includes 87,347,043 missense and 2,270,742 essential splice site variants.[52] This database is available from https://sites.google.com/site/jpopgen/dbNSFP. In addition, ACMG also gives some weight on the evidence from nucleotide conservation prediction tools such as Phylop[50] and GERP[57]. Although based on the findings from CLARITY challenge, there was no significant difference in the success of the teams using both SIFT and Polyphen and those who used one or the other or some other tools such as PhyloP, likelihood ratio test scores (LRT), MutationTaster, GERP and in-house developed tools.

Prediction tools for splicing sites

Another type of variant which could have impact on protein expression and function is located around splicing sites. In general, variants within 12 bp upstream of splice acceptor region or 6 bp downstream of splice donor region of exons could have the potential to impact the splicing. In addition, variants roughly in 20-50 bp upstream of the acceptor, where

the branch site is normally located, are highly likely to affect splicing but identification of the branch site is more challenging. So far about 14% to 15% of all hereditary disease alleles in HGMD were annotated as splicing variants,[58] which makes the application of splice prediction tools of particular importance. As mentioned, ACMG new guideline provides a list of seven tools for splice site prediction. CLARITY challenge contest found the groups that utilized a suite of splice prediction tools, such as the maximum entropy model MAXENT,[59] ExonScan,[60] or positional distribution analysis (Spliceman)[61] were more likely to have identified potential pathogenic mutations, particularly in the *TTN* gene in Family 1, which is the variant being missed by many teams due to various reasons though.

Genotype-Phenotype Databases

Another important resource is online genotype-phenotype databases. So far the most comprehensive one is MedGen (http://www.ncbi.nlm.nih.gov/medgen/) from NCBI, which facilitates the search for possible diagnosis and differentiation diagnosis or candidate genes for distinctive traits and genotype-phenotype correlation. The newly updated version of DECIPHER enhances the features to enable the sequence variants being databased with the query function for genotype-phenotype correlation as well. Another online trait-based candidate gene searching tool is "Phenomizer" (http://compbio.charite.de/phenomizer/), which could be helpful to develop or search for the trait- or syndrome-specific candidate genes.

CONCLUSIONS

In summary, web-based resources are available for the interpretation of constitutional and somatic chromosomal abnormalities and CNVs, despite a reliable database for somatic CNV is still under development. While there are many resources for interpreting gene mutations, a lot of more effort is required to standardize the annotation and interpretation of variants from NGS in a clinical setting.

CONFLICT OF INTEREST
None.

REFERENCES

1. Miller DT, Adam MP, Aradhya S, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet. 2010;86(5):749-764.
2. Brownstein CA, Beggs AH, Homer N, et al. An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. Genome Biol. 2014;15(3):R53.
3. Wertheim GB, Daber R, Bagg A. Molecular diagnostics of acute myeloid leukemia: it's a (next) generational thing. J Mol Diagn. 2013;15(1):27-30.
4. Johansen Taber KA, Dickinson BD, Wilson M. The promise and challenges of next-generation genome sequencing for clinical care. JAMA Intern Med. 2014;174(2):275-280.
5. Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. J Pathol Inform. 2012;3:40.
6. Gillis NK, Patel JN, Innocenti F. Clinical implementation of germ line cancer pharmacogenetic variants during the next-generation sequencing era. Clin Pharmacol Ther. 2014;95(3):269-280.
7. Craig-Holmes AP, Shaw MW. Polymorphism of human constitutive heterochromatin. Science. 1971;174(4010):702-704.
8. Lubs HA, Kinberling WJ, Hecht F, et al. Racial differences in the frequency of Q and C chromosomal heteromorphisms. Nature. 1977;268(5621):631-633.
9. Brothman AR, Schneider NR, Saikevych I, et al. Cytogenetic heteromorphisms: survey results and reporting practices of giemsa-band regions that we have pondered for years. Arch Pathol Lab Med. 2006;130(7):947-949.
10. Herman E. Wyandt VST. Human Chromosome Variation: Heteromorphism and Polymorphism. 2nd ed: Springer Netherlands; 2011.
11. Biesecker LG, Spinner NB. A genomic view of mosaicism and human disease. Nat Rev Genet. 2013;14(5):307-320.
12. Conlin LK, Thiel BD, Bonnemann CG, et al. Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. Hum Mol Genet. 2010;19(7):1263-1275.
13. Liehr T, Claussen U, Starke H. Small supernumerary marker chromosomes (sSMC) in humans. Cytogenet Genome Res. 2004;107(1-2):55-67.
14. Reddy KS, Aradhya S, Meck J, Tiller G, Abboy S, Bass H. A systematic analysis of small supernumerary marker chromosomes using array CGH exposes unexpected complexity. Genet Med. 2013;15(1):3-13.
15. Malvestiti F, De Toffol S, Grimi B, et al. De novo small supernumerary marker chromosomes detected on 143,000 consecutive prenatal diagnoses: chromosomal distribution, frequencies, and characterization combining molecular cytogenetics approaches. Prenat Diagn. 2014;34(5):460-468.
16. Huret JL, Ahmad M, Arsaban M, et al. Atlas of genetics and cytogenetics in oncology and haematology in 2013. Nucleic Acids Res. 2013;41(Database issue):D920-924.
17. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004;36(9):949-951.
18. Sebat J, Lakshmi B, Troge J, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525-528.
19. Du RQ, Jin L, Zhang F. [Copy number variations in the human genome: their mutational mechanisms and roles in diseases]. Yi Chuan. 2011;33(8):857-869.
20. Almal SH, Padh H. Implications of gene copy-number variation in health and diseases. J Hum Genet. 2012;57(1):6-13.
21. Mikhail FM. Copy number variations and human genetic disease. Curr Opin Pediatr. 2014.
22. Xiang B, Li A, Valentin D, Nowak NJ, Zhao H, Li P. Analytical and clinical validity of whole-genome oligonucleotide array comparative genomic hybridization for pediatric patients with mental retardation and developmental delay. Am J Med Genet A. 2008;146A(15):1942-1954.
23. Xiang B, Zhu H, Shen Y, et al. Genome-wide oligonucleotide array comparative genomic hybridization for etiological diagnosis of mental retardation: a multicenter experience of 1499 clinical cases. J Mol Diagn. 2010;12(2):204-212.
24. Manning M, Hudgins L. Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities. Genet Med. 2010;12(11):742-745.
25. Ahn JW, Bint S, Bergbaum A, Mann K, Hall RP, Ogilvie CM. Array CGH as a first line diagnostic test in place of karyotyping for postnatal referrals - results from four years' clinical application for over 8,700 patients. Mol Cytogenet. 2013;6(1):16.
26. Committee Opinion No. 581: the use of chromosomal microarray analysis in prenatal diagnosis. Obstet Gynecol. 2013;122(6):1374-1377.
27. Li P, Pomianowski P, DiMaio MS, et al. Genomic characterization of prenatally detected chromosomal structural abnormalities using oligonucleotide array comparative genomic hybridization. Am J Med Genet A. 2011;155A(7):1605-1615.
28. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014;42(Database issue):D986-992.
29. Shaikh TH, Gai X, Perin JC, et al. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. Genome Res. 2009;19(9):1682-1690.

30. Itsara A, Cooper GM, Baker C, et al. Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet. 2009;84(2):148-161.
31. Beroukhim R, Mermel CH, Porter D, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010;463(7283):899-905.
32. Sabattini E, Bacci F, Sagramoso C, Pileri SA. WHO classification of tumours of haematopoietic and lymphoid tissues in 2008: an overview. Pathologica. 2010;102(3):83-87.
33. Bajaj R, Xu F, Xiang B, et al. Evidence-based genomic diagnosis characterized chromosomal and cryptic imbalances in 30 elderly patients with myelodysplastic syndrome and acute myeloid leukemia. Mol Cytogenet. 2011;4:3.
34. Kolquist KA, Schultz RA, Slovak ML, et al. Evaluation of chronic lymphocytic leukemia by oligonucleotide-based microarray analysis uncovers novel aberrations not detected by FISH or cytogenetic analysis. Mol Cytogenet. 2011;4:25.
35. Xiang B, Leon A, Li Mea. Atlas of Cytogenomics in Oncology and Hematology: a Platform-Neutral Clinical Cancer Genomics Database. Cancer Genet. 2012;205(7-8):420.
36. Bean LJ, Tinker SW, da Silva C, Hegde MR. Free the data: one laboratory's approach to knowledge-based genomic variant classification and preparation for EMR integration of genomic data. Hum Mutat. 2013;34(9):1183-1188.
37. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med. 2013;369(16):1502-1511.
38. O'Rawe J, Jiang T, Sun G, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 2013;5(3):28.
39. Pabinger S, Dander A, Fischer M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 2014;15(2):256-278.
40. Daber R, Sukhadia S, Morrissette JJ. Understanding the limitations of next generation sequencing informatics, an approach to clinical pipeline validation using artificial data sets. Cancer Genet. 2013;206(12):441-448.
41. Rehm HL, Bale SJ, Bayrak-Toydemir P, et al. ACMG clinical laboratory standards for next-generation sequencing. Genet Med. 2013;15(9):733-747.
42. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
43. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-1081.
44. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-249.
45. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553-1561.
46. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7(8):575-576.
47. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118.
48. Shihab HA, Gough J, Cooper DN, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34(1):57-65.
49. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12):e1001025.
50. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20(1):110-121.
51. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25(12):i54-62.
52. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat. 2013;34(9):E2393-2402.
53. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80-92.
54. Wu TJ, Shamsaddini A, Pan Y, et al. A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE). Database (Oxford). 2014;2014:bau022.
55. Peterson TA, Doughty E, Kann MG. Towards precision medicine: advances in computational approaches for the analysis of human variants. J Mol Biol. 2013;425(21):4047-4063.
56. Flanagan SE, Patch AM, Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. Genet Test Mol Biomarkers. 2010;14(4):533-537.
57. Cooper GM, Stone EA, Asimenos G, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15(7):901-913.
58. Stenson PD, Mort M, Ball EV, et al. The Human Gene Mutation Database: 2008 update. Genome Med. 2009;1(1):13.
59. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol. 2004;11(2-3):377-394.
60. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. Systematic identification and analysis of exonic splicing silencers. Cell. 2004;119(6):831-845.
61. Lim KH, Fairbrother WG. Spliceman--a computational web server that predicts sequence variations in pre-mRNA splicing. Bioinformatics. 2012;28(7):1031-1032.