Genomic Data Standards Resources and Initiatives Cited in the Supplemental Information to the Genomic Data Sharing Policy

IMPORTANT NOTE: The National Institutes of Health makes no endorsement of the non-NIH-funded genomic data standards resources and/or initiatives included in this document.

Categories	Resources and Initiatives
NIH-Funded	Common Data Element Resource Portal: http://www.nlm.nih.gov/cde/ NIH encourages the use of common data elements (CDEs) in clinical research, patient registries, and other human subject research in order to improve data quality and opportunities for comparison and combination of data from multiple studies and with electronic health records. This portal provides access to NIH-supported CDE initiatives and other tools and resources that can assist investigators developing protocols for data collection.
	Clinical Genome Resource (ClinGen): http://www.nih.gov/news/health/sep2013/nhgri-25.htm In 2013, the NIH National Human Genome Research Institute (NHGRI) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) awarded three grants totalling over $25 million to support a consortium of research groups to design and implement a framework for evaluating variants that are relevant to patient care (e.g., play a role in disease). International Collaboration for Clinical Genomics (ICCG): http://www.iccg.org/about-the-iccg/clingen/ ICCG was awarded as a grant under ClinGen. ICCG was charged with developing standard formats for gathering and depositing data in ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/). It will work with a variety of different stakeholder groups, including clinical laboratories and existing locus-specific databases, to obtain robust data sets on genomic variants and disease associations. It will also develop standards to analyze variants and determine whether they are potentially disease-causing and medically informative.

	National Center for Biotechnology Information Archives a. Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/info/submission.html): an NIH data repository that archives and distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data. b. Database of Genotypes and Phenotypes (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document_name=HowToSubmit.pdf): a database originally designed to archive and distribute coded genotype, phenotype, exposure, and pedigree data from genome-wide association studies. dbGaP now accepts additional types of data such as copy number variants and large-scale sequencing data. c. Database of Short Genetic Variations (http://www.ncbi.nlm.nih.gov/projects/SNP/how_to_submit.html): a database that includes single nucleotide variations, microsatellites, and small-scale insertions and deletions. dbSNP provides population-specific frequency and genotype data, experimental conditions, molecular context, and mapping information for both neutral variations and clinical mutations. d. GenBank Genome Assemblies (http://www.ncbi.nlm.nih.gov/genbank/wgs.submit): a genetic sequence database that provides an annotated collection of publicly available DNA sequences. e. Sequence Read Archive: http://www.ncbi.nlm.nih.gov/books/NB47529/) SRA stores raw sequencing data as well as alignment information in the form of read placements on a reference sequence.
	National Database for Autism Research (NDAR): http://ndar.nih.gov/standards.html/ The National Database for Autism Research (NDAR) is an NIH-funded research data repository that aims to accelerate progress in autism spectrum disorders (ASD) research through data sharing, data harmonization, and the reporting of research results. NDAR also serves as a scientific community platform and portal to multiple other research repositories, allowing for aggregation and secondary analysis of data.
	NIH Big Data to Knowledge (BD2K) Initiative: http://bd2k.nih.gov a. Group 1: Enabling Data Utilization https://www.youtube.com/watch?v=za0xQbI79vo&feature=youtu.be: tasked with developing new policies that better encourage data and software sharing, developing a catalog of research datasets that will enable researchers to find and cite datasets, and establishing frameworks to support community-based data and metadata standards. b. Frameworks for Community-Based Standards Efforts Workshop (http://bd2k.nih.gov/pdf/frameworks_for_comm_based_standards_efforts_report.pdf): the overall goal of this workshop was to learn what has and has not worked in community-based standards efforts, and to address and discuss specific issues that pertain to formulating, conducting, and maintaining standards.
	The National Cancer Informatics Program (NCIP): http://cbiit.nci.nih.gov/ncip/about-ncip The semantic infrastructure and interoperability framework enable data integration across different specialties and institutions by providing standard vocabularies, common data elements, clinical case-report forms, data models, and definitions. The interoperability framework follows a widely used approach of employing services that are independent of any particular information-technology platform.

	The NIH Standards Information Resource (NSIR) (Under development) The purpose of this initiative is to collect, organize, and make publicly available trusted, systematically organized, and curated information about data-related standards that are widely-used in biomedical research and related activities. This information resource is intended to help investigators identify and choose data-related standards that are well-suited to their needs.
Biomedical Research	1000 Genomes: http://www.ncbi.nlm.nih.gov/pubmed/21653522; http://www.ncbi.nlm.nih.gov/pubmed/19505943 a. The variant call format (VCF) is a generic format for storing DNA polymorphism data such as SNPs, insertions, deletions and structural variants, together with rich annotations. VCF data are usually stored in a compressed manner and can be indexed for fast data retrieval of variants from a range of positions on the reference genome. The format was developed for the 1000 Genomes Project and has also been adopted by other projects such as UK10K, dbSNP and the NHLBI Exome Project. VCFtools is a software suite that implements various utilities for processing VCF files, such as validation and merging and comparing files as well as providing a general Perl API. b. The Sequence Alignment/Map (SAM) format is a generic alignment format for storing read alignments against reference sequences and supporting short and long reads (up to 128 Mbp) produced by different sequencing platforms. It is flexible in style, compact in size, efficient in random access and is the format in which alignments from the 1000 Genomes Project are released. SAMtools implements various utilities for post-processing alignments in the SAM format, such as indexing, variant caller, and alignment viewer, and thus provides universal tools for processing read alignments.
	Genomic Standards Consortium (GSC): http://gensc.org/ GSC is an open-membership working body formed in September 2005. The goal of this International community is to promote mechanisms that standardize the description of genomes and the exchange and integration of genomic data.
	The Global Alliance for Genomics and Health (GA4GH): http://oicr.on.ca/oicr-programs-and-platforms/global-alliance-genomics-and-health-ga4gh Data Working Group: concentrates on data representation, storage, and analysis, including working with platform development partners and industry leaders to develop standards that will facilitate interoperability.
	Human Genome Variation Society (HGVS): http://www.hgvs.org/rec.html Members of the Society have formulated guidelines and recommendations on a number of topics, particularly for the nomenclature of gene variations and guidelines for variation databases.
Clinical	American College of Medical Genetics (ACMG): http://pathology.ucla.edu/workfiles/News/ACMG-NGS-Guidelines-2013.pdf ACMG has developed professional standards and guidelines to assist clinical laboratories with the validation of next-generation sequencing methods and platforms, the ongoing monitoring of next-generation sequencing testing to ensure quality results, and the interpretation and reporting of variants found using these technologies.

	Health Level 7 International (HL7): http://www.hl7.org/index.cfm?ref=nav HL7 is dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services.
	Healthcare Information Technology Standards Panel (HITSP): http://hitsp.org/ HITSP is a cooperative partnership between the public and private sectors. The Panel was developed for the purpose of harmonizing and integrating standards that will meet clinical and business needs for sharing information among organizations and systems.
Phenotype Ontology	PhenX: https://www.phenx.org/ PhenX provides the scientific community with recommended, standard high-priority measures of phenotypes and exposures for use in genome-wide association studies and more generally, epidemiological and biomedical research.