Paper DS11

Evolution of Genomic Data Mappings from PF to GF in SDTM v3.4: Concepts, Challenges, and Controlled Terminology with Genetic Variation Examples

Rama Kudaravalli, Merck & Co., Inc., Rahway, NJ, USA

ABSTRACT

The Genomics Findings (GF) domain in SDTM v3.4 was introduced to address the need for standardized representation of genomic data in clinical research. It is designed to handle various types of genomic data, including genomic summary measures, and replaces the PF domain from the provisional SDTM Implementation Guide for Pharmacogenomics and Pharmacogenetics (SDTMIG-PGx). This evolution from PF to GF reflects the need for more comprehensive and standardized representation of genomic data in clinical trial data standards. One of the key challenges is to ensure that the domain can accommodate a wide range of genomic findings while maintaining consistency and standardization. Controlled terminology plays a crucial role in the GF domain, providing a standardized vocabulary for describing genomic findings.

This paper aims to illustrate the concepts, challenges, and controlled terminology associated with representing the genetic variations and other genomic data in a standardized and comprehensive manner in GF domain in clinical trials.

INTRODUCTION

Genomic medicine is an emerging field that involves using genomic information about an individual as part of their clinical care. It focuses on the collective impact of a person's genes on their body. Genetic testing is useful in clinical trials by identifying and following treatment effectiveness, increasing the possibility of early detection, diagnosis, and intervention, and guiding drug therapy to improve efficacy and avoid adverse effects.

SDTMIG-PGx version 1.0, a provisional guidance was introduced in May 2015 by CDISC to map the genomic data collected from clinical trials to various SDTM domains. The CDISC has provided the supporting controlled terminology for PGx guide in December 2015. To map the data related to biospecimen collection and handling BE, BS, RELSPEC domains were designed. PB, SB domains were used to map biomarker data and PF domain for genetic information in PGx guide. Later, in 2021 a single GF domain was introduced in SDTMIG v3.4 which replaced the PF domain. The biomarker domains, PB and SB were deprecated as these biomarkers were not specific to genomics. PG domain was deprecated as the methods mapped under this domain were beyond genomic testing.

GF is a findings domain that contains data related to the structure, function, evolution, mapping, and editing of subject and non-host organism genomic material of interest. This domain includes but is not limited to assessments and results for genetic variation and transcription, and summary measures derived from these assessments. It is used for findings from characteristics assessed from nucleic acids and may include subsequent inferences and/or predictions about related proteins/amino acids. To implement the GF domain, we need to understand the genomic concepts and controlled terms in mapping GF domain.

As the regulatory environment for genomic-based therapies is complex and constantly evolving, we need the personnel with appropriate knowledge to understand, integrate and interpret the genomic data for analysis. Standardizing genomic data in a clinical trial is essential for ensuring data quality, reproducibility, and interoperability. To provide appropriate focus, scope, and relevance for this paper, we would like to show how the data for genomic findings mapping is evolved from PF to GF domain and discuss how to map genetic variations and genotyping.

GENETIC VARIATIONS

Genetic variation is the difference in DNA among individuals. The vast majority of the DNA letters in peoples' genomes is identical, but a small fraction of those letters varies. This genomic variation accounts for some of the differences among people, including important aspects of their health and susceptibility to diseases. To understand the genetic variations, we need to know several key concepts.

CONCEPTS

The central dogma of molecular biology describes the flow of genetic information within a biological system. The principal foci of this dogma are genetics, transcriptomics, and proteomics. The genetic information flows unidirectionally from DNA to RNA to Protein. The DNA is transcribed to mRNA by transcription process. The RNA is translated to protein by translation process.

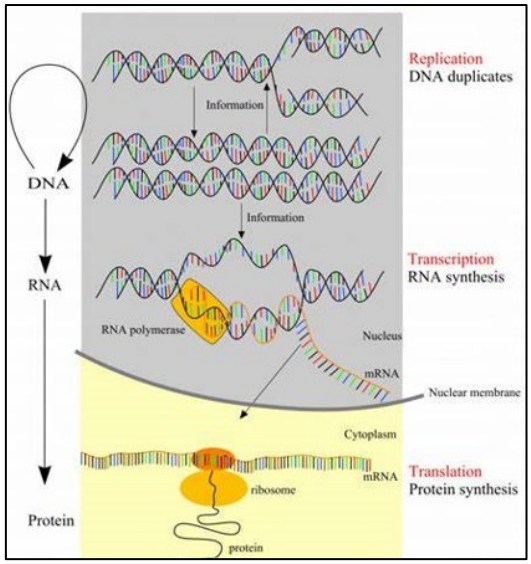

Figure 1. Central dogma of biology

Genetics is the study of genes, genetic variation, and heredity in organisms. Transcriptomics is the study of the transcriptome, the complete set of RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell, using high-throughput methods, such as microarray analysis. Comparison of transcriptomes allows the identification of genes that are differentially expressed in distinct cell populations, or in response to different treatments. Proteomics is the large-scale study of proteomes, which are sets of proteins produced in an organism. The scope for genomic findings, GF is genetics and transcriptomics.

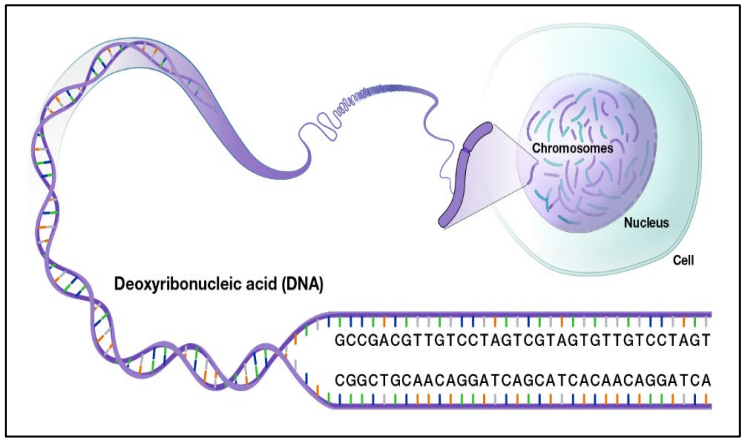

Figure 2. Genome, DNA and Chromosomes

A genome is the complete set of DNA instructions found in every cell. DNA is made of four different chemicals (called nucleotides or bases), each represented by a different letter: adenine (A), thymine (T), cytosine (C) and guanine (G). The order of these letters (i.e., the DNA sequence) encodes the information that instructs each cell what to do and when to do it. The genome's DNA is packed into structures called chromosomes. The gene is considered as the basic unit of inheritance. Genes are passed from parents to offspring and contain the information needed to specify physical

and biological traits. A person's genome consists of 23 pairs of chromosomes of which 22 pairs are called as autosomes and one pair is called sex chromosomes. One copy of each chromosome is inherited from each parent to offspring.

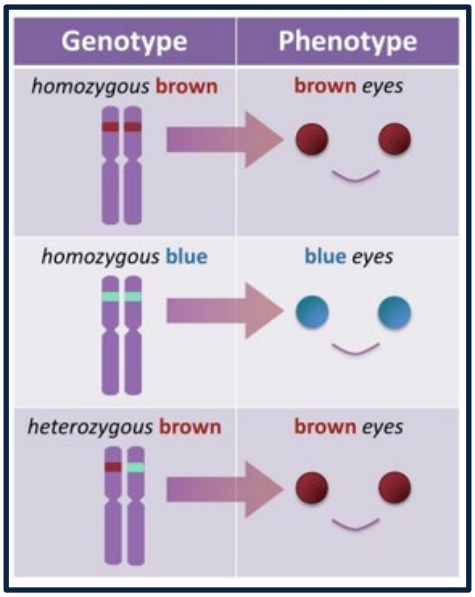

Figure 3. Allele

An allele is one of two or more versions of DNA sequence (a single base or a segment of bases) at a given genomic location. An individual inherits two alleles, one from each parent, for any given genomic location where such variation exists. If the two alleles are the same, the individual is homozygous for that allele. If the alleles are different, the individual is heterozygous. Genotype is the combination of alleles that they possess for a specific gene. Phenotype is the combination of their observable characteristics or traits. While an organism's genotype is directly inherited from its parents, phenotype is merely influenced by genotype. Environmental factors can also affect phenotype.

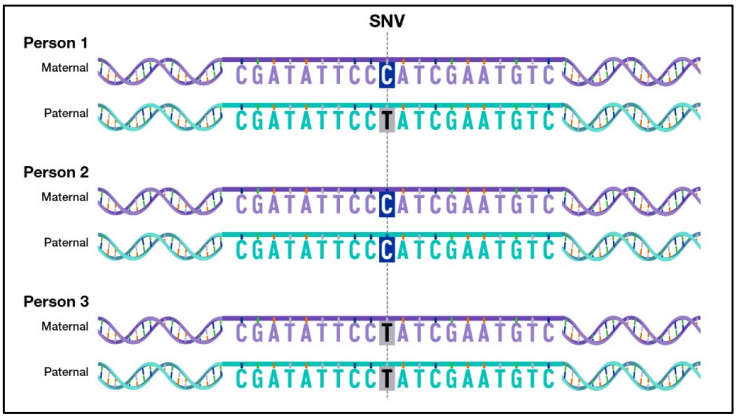

Figure 3. Single Nucleotide Variations

There are multiple types of genomic variants. The smallest genomic variants are single-nucleotide variants (SNVs). Each SNV reflects a difference in a single nucleotide (or letter). For a given SNV, the DNA letter at that genomic position might be a C in one person but a T in another person. SNVs are the most common type of genomic variation. A subtype of SNVs is called a single-nucleotide polymorphism (SNP; pronounced "snip"). To be considered a SNP, a SNV must be present in at least 1% of the human population. As such, SNV is a more general term that includes both relatively common (such as SNPs) and rare single-nucleotide differences. Irrespective of their frequency they are generally referred as SNVs.

A SNV can be either common or rare and can occur in either germline or somatic cells. In non-coding regions, SNVs can impact mRNA sequence and the level of gene expression, meaning they result in disease susceptibility and can signal a higher risk of cancer. SNVs in coding regions can be synonymous substitutions or nonsynonymous substitutions. Synonymous substitutions do not alter the amino acid sequence, while nonsynonymous substitutions can be one of these types: missense substitutions result in proteins that are malfunctioning while nonsense substitutions result in prematurely stopped codons. In nonstop mutations the stop codon is erased to produce a long nonfunctional protein.

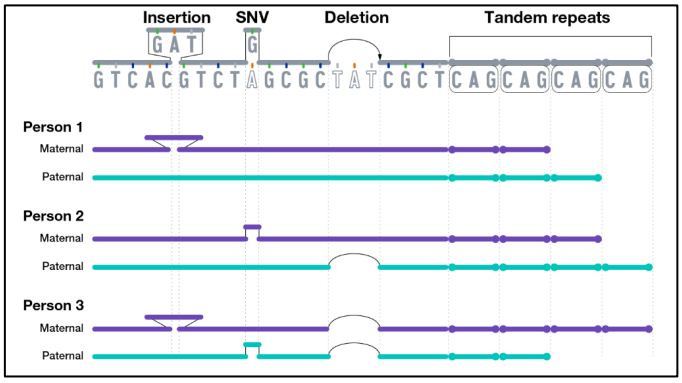

Figure 4. Genomic Variations - Indels

Insertion/deletion variants are also known as “Indels” have an extra or missing DNA nucleotides in the genome, respectively, and typically involve fewer than 50 nucleotides. Indels are less frequent than the SNVs but have greater impact on health and disease. One of the most common types of insertion/deletion variants are tandem repeats also known as microsatellites. Tandem repeats are short stretches of nucleotides that are repeated multiple times and are highly variable among people. Short copy number variation is the trinucleotide repeat of the CAG base pairs in the huntingtin gene responsible for the neurological disorder Huntington's disease.

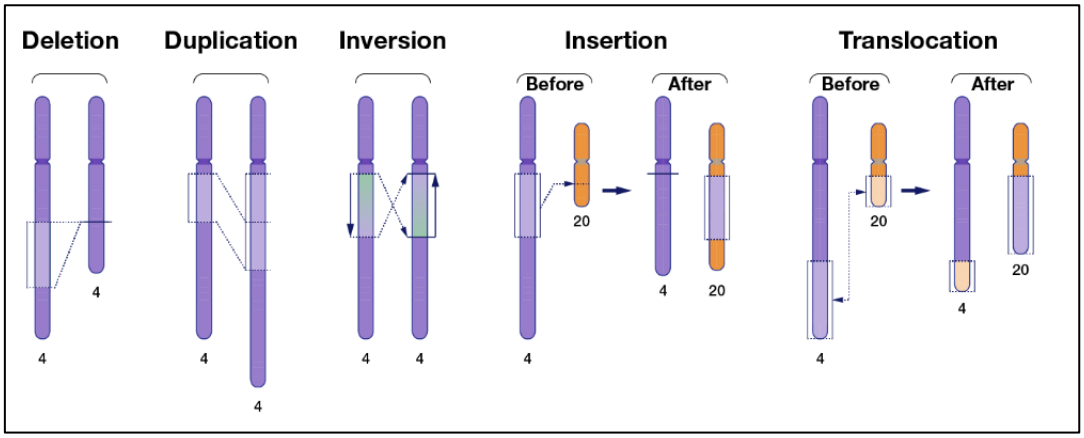

Figure 4. Structural Variations

Genomic variation also extends beyond small stretches of nucleotides to larger chromosomal regions. These large-scale genomic differences are called structural variants and involve at least 50 nucleotides and as many as thousands of nucleotides that have been inserted, deleted, inverted, or moved from one part of the genome to another. Tandem repeats that contain more than 50 nucleotides are considered structural variants; in fact, such large tandem repeats account for nearly half of the structural variants present in human genomes. When a structural variant reflects differences in the total number of nucleotides involved, it is called a copy-number variant (CNV). CNVs are distinguished from other structural variants, such as inversions and translocations, because the latter types often do not involve a difference in the total number of nucleotides.

A reference genome also known as a reference assembly, is a digital nucleic acid sequence database, assembled by

scientists as a representative example of the set of genes in one idealized individual organism of a species. It represents the sequence of one copy of each chromosome, useful for detecting variants in a person's genome sequence and serves as an important data-analysis tool.

CONTROLLED TERMINOLOGY

The controlled terminology used for mapping genomic data to GF domain can be obtained from CDISC terminology, HGNC database and other external databases. The CDIC has provided the controlled terms for GFTEST, GFTESTCD, GFTSTDTL, GFINHERT, GFSYMTYP, GFANMETH etc., and these controlled terms can be extensible so one can have the flexibility to map the data as per the requirement. In addition to these variables, CDISC provides controlled terminology to GF variables for genomic specimen material type, method used, units etc., to be compliant with regulatory agencies. The gene symbol has to be obtained from HGNC (HUGO Gene Nomenclature Committee) database to map to GFSYM variable. Gene Symbol, GFSYM cannot be mapped to GFTEST or GFTESTCD. Each gene with an approved HGNC symbol has its own Symbol Report that contains manually curated data and links to many other external biomedical resources. External databases like Ensembl, UCSC genome browsers will provide the reference values to map GFGENREF, GFSEQID, GFPRVID variables. The controlled terminology is helpful in submitting the standardized genomic data to the agency. It is evolving as the need for new genomic tests is growing in clinical research.

MAPPING DIFFERENCES

The main difference between PF and GF domain is change in name. New GF domain comes with addition of new variables, concepts, and the associated controlled terminology. It provides a standardized way to represent genomic data, ensuring consistency and accuracy across studies. GFLNKGRP is a new variable to link related records across the domains. The NHOID (Non-Host Organism Identifier) is a variable, specifically designed to identify non-host organisms, such as bacteria, viruses, or parasites, in genomic data. When genetic tests provide results related to non-host organisms found in subject samples, the NHOID variable is populated to distinguish these findings from the subject's genetic data and helps in accurately representing and analyzing genomic data from non-host organisms. The NHOID variable replaces the PFNPSCES and PFNSTRN variables from PF domain. PFGENRI and PFGENTYP are replaced by GFSYM and GFSYMTYP variables. Genomic symbol, in GFSYM variable must be obtained from HGNC (HUGO Gene Nomenclature Committee) database. It cannot be mapped to GFTEST or GFTESTCD. The PFMUTYP variable is replaced by GFINHERT which collects the inheritability of the organism. GFCHROM variable is added to collect the name or number of the chromosome. The GFSTREFC, GFSTREFN are added to GF domain to collect the reference results in standard format. The values populated in GFGENREF, GFSEQID, and GFPVRID should reflect the level of granularity collected (e.g., version, build, patch, release) to support interpretation of the reported result. PFREFSEQ, PFRSNUM are replaced by GFSEQID, and GFPVRID variables respectively whereas GFGENREF is a newly added variable to GF domain. PFALLELC variable is replaced by GFCOPYID, is an identifier used to differentiate between copies of a genetic target of interest present on homologous chromosomes. PFGENLI and PFGENTRG variables were retired.

STUDYID	DOMAIN	USUBJID	GFSEQ	GFREFID	GFTESTCD	GFTEST	GFTSTDTL	GFORRES	GFORREF	GFSTRESC	GFSTREFC
ABC-006	GF	ABC-006-00003	1	NA18537	SNV	Single Nucleotide Variation	GENOTYPE	T/T	G/G	T/T	G/G
ABC-006	GF	ABC-006-00004	2	NA07000	SNV	Single Nucleotide Variation	GENOTYPE	G/T	G/G	G/T	G/G
ABC-006	GF	ABC-006-00005	3	NA00131	SNV	Single Nucleotide Variation	GENOTYPE	G/G	G/G	G/G	G/G

GFGENREF	GFCHROM	GFSYM	GFGENLOC	GFSEQID	GFPVRID	GFNAM	GFSPEC	GFMETHOD	VISITNUM	VISIT	GFDTC	GFDY
GRCh38.p13	4	ABCG2	4:88131171	ENSG00000118777	rs2231142	ACME LABS	DNA	REAL-TIME POLYMERASE CHAIN REACTION	1	SCREENING	2020-06-25	-3
GRCh38.p13	4	ABCG2	4:88131171	ENSG00000118777	rs2231142	ACME LABS	DNA	REAL-TIME POLYMERASE CHAIN REACTION	1	SCREENING	2020-06-25	-3
GRCh38.p13	4	ABCG2	4:88131171	ENSG00000118777	rs2231142	ACME LABS	DNA	REAL-TIME POLYMERASE CHAIN REACTION	1	SCREENING	2020-06-25	-3

TABLE 1. Genetic Variation

In the above table 1, the genomic DNA sample obtained from blood is used to detect the SNVs in ABCG2 gene by doing real-time PCR method for three subjects. Each subject has a different genotype at the genetic locus of the interest. The inheritability of the variation is considered to be germline as the sample is DNA. The nucleotide variant is located on chromosome number 4, gene location is provided in GFGENLOC. The GFSEQID is the identifier for the sequence used as the reference to identify the genetic variation in the result. The GFPVRID identifier for the single

nucleotide variant that has been characterized in an external database.

STUDYID	DOMAIN	USUBJID	PFSPID	PFTESTCD	PFTEST	PFTSTDTL	PFGENRI	PFGENTYP	PFCAT
ABC-006	PF	ABC-006-00003	1	GENVAR	Genetic Variation	INSERTION	CALR	GENE	GENOMIC ABERRATION TESTING
ABC-006	PF	ABC-006-00003	1	GENVAR	Genetic Variation		JAK2 V617F	GENE	GENOMIC ABERRATION TESTING
ABC-006	PF	ABC-006-00003	1	GENVAR	Genetic Variation	DELETION	KDM6A	GENE	GENOMIC ABERRATION TESTING
ABC-006	PF	ABC-006-00003	1	GENVAR	Genetic Variation	SINGLE NUCLEOTIDE VARIANT	KIT	GENE	GENOMIC ABERRATION TESTING

STUDYID	DOMAIN	USUBJID	GFSPID	GFTESTCD	GFTEST	GFTSTDTL	GFSYM	GFSYMTYP	GFCAT
ABC-006	GF	ABC-006-00003	1	GENVAR	Genetic Variation	INSERTION	CALR	GENE WITH PROTEIN PRODUCT	GENOMIC ABERRATION TESTING
ABC-006	GF	ABC-006-00003	1	GENVAR	Genetic Variation		JAK2 V617F	GENE WITH PROTEIN PRODUCT	GENOMIC ABERRATION TESTING
ABC-006	GF	ABC-006-00003	1	GENVAR	Genetic Variation	DELETION	KDM6A	GENE WITH PROTEIN PRODUCT	GENOMIC ABERRATION TESTING
ABC-006	GF	ABC-006-00003	1	GENVAR	Genetic Variation	SINGLE NUCLEOTIDE VARIANT	KIT	GENE WITH PROTEIN PRODUCT	GENOMIC ABERRATION TESTING

TABLE 2. Genetic Variation – PF to GF changes

STUDYID	RDOMAIN	USUBJID	QNAM	QLABEL	QVAL	QORIG
ABC-006	PF	ABC-006-00003	PFGENRIO	Genetic Region of Interest Other	**ChrX:44732820**	CRF
ABC-006	PF	ABC-006-00003	PFGENRIO	Genetic Region of Interest Other	**Chr4:55593395**	CRF
ABC-006	PF	ABC-006-00003	PFGENRIO	Genetic Region of Interest Other	**Chr19:13054627**	CRF

GFORRES	GFCHROM	GFGENLOC	GFSTRESC	GFSPEC	GFMETHOD	GFDTC
DETECTED	**19**	**Chr19:13054627**	DETECTED	DNA	NEXT GENERATION SEQUENCING	2021-08-05
NOT DETECTED			NOT DETECTED	DNA	NEXT GENERATION SEQUENCING	2021-08-05
DETECTED	**X**	**ChrX:44732820**	DETECTED	DNA	NEXT GENERATION SEQUENCING	2021-08-05
DETECTED	**4**	**Chr4:55593395**	DETECTED	DNA	NEXT GENERATION SEQUENCING	2021-08-05

TABLE 3. Genetic Variation – SUPPPF to GF changes

The genetic variation example in table 2 shows the mapping change for values from PFGENRI and PFGENTYP to GFSYM And GFSYMTYP. In table 3, the other genetic region of interest information was mapped to SUPPPF.QVAL is moved to the core GF domain variables, GFCHROM and GFGENLOC. The addition of these two new GF variables gives opportunity to map the additional information collected related to the genetic variation into the core domain.

CONCLUSION

The GF domain provides a standardized way to represent genomic data, ensuring consistency and accuracy across studies. It facilitates the integration of genomic data with other clinical data, enhancing the ability to conduct comprehensive analyses. Addition of new variables in GF domain gives the flexibility to map additionally collected information to the core domain and makes it convenient for analysis. By using controlled terminology and predefined structures, the GF domain helps improve the quality and reliability of genomic data. It aligns with regulatory requirements, making it easier for researchers to submit genomic data to regulatory agencies. Overall, the future of the GF domain in SDTM is geared towards enhancing the quality and usability of genomic data in clinical research, ultimately contributing to better patient outcomes and advancements in personalized medicine.

RECOMMENDED READING

CDISC SDTM PGxIG V1.0
CDISC SDTMIG v3.4

CONTACT INFORMATION
Rama Kudaravalli
Merck & Co., Inc., Rahway, NJ, USA
Rama.kudaravalli@merck.com