What's in a Genotype?: An Ontological Characterization for Integration of Genetic Variation Data

Matthew H. Brush[1], Chris Mungall[2], Nicole Washington[2], Melissa A. Haendel[1]
[1]Oregon Health & Science University, 3181 S.W. Sam Jackson Park Rd., Portland, OR 97239, USA
[2]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

ABSTRACT

Exploration of the mechanistic basis of biology and disease has long leveraged the concept of a genotype, which represents the genetic composition associated with a physical trait. Translational research efforts rely increasingly on the ability to integrate genotype-phenotype data across systems and organism communities, but are hindered by the lack of a shared, computable model of the information coded into genotype representations. Here, we present the efforts of the Monarch Initiative to build GENO, an ontological model of genotype information. The Monarch Initiative is a collaborative effort to integrate data from diverse resources to leverage model systems for disease research based on their phenotypes. The genotype model we have developed is based on decomposing the different types of information represented in a genotype, is interoperable with existing OBO Foundry ontologies, and utilizes modeling from orthogonal ontologies to describe a broad range of attributes of these sequences. We describe the features and utility of such an approach toward the integration of diverse genotype data with a broad spectrum of related biomedical data.

1 INTRODUCTION

Historically and today, biologists have explored the basis of biology and disease by correlating genotype with phenotype, wherein a genotype represents the genetic composition of a phenotype - a physical trait as realized in a certain environment. This paradigm has supported research in human and model organism systems, and translational approaches are emerging to apply knowledge across these communities toward an understanding of human biology. The Monarch Initiative[1] represents one such translational effort, aimed at integrating data from diverse resources to drive the identification of models for disease research, and discovery of novel connections between genes, environments, and disease.

A number of groups have made progress in standardizing model organism database (MOD) representations, such as the Generic Model Organism Database consortium (GMOD, www.gmod.org) and Intermine (Smith et al. 2012), yet the MODs continue to vary both in their descriptions of genotypes, and how they are linked to related data such as phenotypes. Furthermore, these models lack the ontological underpinnings to support reasoning and inference. Therefore, there is a need for a consensus genotype model that can support inference across disparately recorded aspects of genotypes as they relate to gene expression and phenotypes. The Monarch Initiative is situated at the intersection of these diverse information systems and has developed an ontological model of genotype information called GENO[2], to support data integration needs and provide a common framework for the community to structure genotype-related data.

In this paper, we will first disentangle a clear conceptual understanding of genotypes as information entities that specify variation in a genome (henceforth the term 'genotype' will refer to such information entities and not the genetic material they are about). We then describe our ontological representation of this information in the context of existing ontological frameworks and related data types. Finally, we describe the features and utility of this model toward the integration and analysis of diverse genotype data with a broad spectrum of related biomedical data. Notation conventions used in this document include `courier font` for ontological classes, and ***bold italicized text*** for emphasis of important terms and concepts.

2 RESULTS

2.1 What's In a Genotype?

There is a general consensus amongst biologists that genotypes represent heritable genomic sequence variation linked to one or more physical traits. In searching for a more precise characterization beyond this, we find many and varied interpretations (Mahner et al. 1997). We take the view that a genotype is an abstract information entity that represents an entire genome sequence in terms of its variation from some reference genome sequence. This view reflects a careful analysis of a diverse set of human and model organism genotype data, and discussions within the ontology community at large.

2.1.1 Decomposing the 'Sequence Content' of a Genotype

The information content of a genotype is complex, and encoded in a precise syntax that represents information through a defined nomenclature and structure (see Table 1). We characterize this information as being of two types: the primary ***sequence content***, and secondary ***sequence attribute content*** that describes features of these sequences. The sequence content of a genotype can be understood as a mapping between syntactic elements of a genotype and extents of genomic sequence they represent. We believe this conceptualization will facilitate a clear understanding by biologists across disciplines, who share a basic view of genomic architecture that can anchor their understanding of

[*] To whom correspondence should be addressed: brushm@ohsu.edu

[1] www.monarchinitiative.org

[2] GENO is accessible at http://purl.obolibrary.org/obo/geno.owl

Organism	Genotype
zebrafish	fgf8a$^{ti282a/+}$; shha$^{tq252/tq252}$ (AB)
mouse	B6.Cg-Shh$^{tm1(EGFP/cre)Cjt}$/J
worm	daf-2(e1370) III; fog-2(oz40) V
human	ATP1A3(NM_152296.3) [c.946G>A, p.Gly316Ser]+[=]

Table 1: Example genotypes showing syntaxes used by four organismal communities.

the sequences represented in a genotype. In addition, such mappings will guide our ontological modeling by clarifying the often ambiguous sequence referents in genotypes. Finally, this approach will facilitate computational operations on genotype data by linking them to referent sequence information, which will be critical to novel approaches to analyze genotype data by efforts such as the Monarch Initiative.

At the highest level, a genotype can be conceptually decomposed into a ***variant component*** and ***reference component*** - each of which is itself a collection of sequences (Figure 1A). The ***variant component*** of a genotype represents all known variant elements that are linked to some phenotypic outcome, typically organized into variant locus complements (e.g. the gene locus complement fgf8a$^{ti282a/+}$). In GENO, we call this the `genomic variation complement` of the genotype. The reference component, which we term the `reference genome`, offers a genomic context in which these variations are associated with observed trait(s). Importantly, these variant and reference components allow us to resolve a genotype into a final variant genome sequence – the ***resolved sequence content*** of a genotype. Conceptually, this resolution is achieved through a 'find and replace' operation in which the sequences of the variant component are substituted for the corresponding sequences in the reference genome. Following this top-level breakdown into reference and variant components, the `genomic variation complement` can be further decomposed into one or more `variant single locus complements`, representing the set of all complementary loci where at least one variant exists (Figure 1B). This complement is typically a pair of sequences for diploid organisms (i.e. the two variants of a locus on maternal and paternal chromosomes). The `variant single locus complement` has member parts that are the individual complementary loci - at least one of which is a `variant locus`. This `variant locus` results from it having as part some `sequence alteration`, whose extent is only those bases that vary between the variant locus and the reference sequence specified in the reference genome. This 'parts list' of a genotype outlined above represents one of many ways that a genome sequence can be partitioned into simpler elements. But it is important because it decomposes the genome precisely into those units that are of interest to geneticists seeking to understand the link between genetic variation and phenotypic traits. For this reason, this ***genotype partonomy*** will form the core of our ontological model, as described in 2.2.

2.1.2 The 'Sequence Attribute Content' of a Genotype

In addition to its sequence content, a genotype also encodes secondary ***sequence attribute content*** describing additional information about its sequences. These can include information about zygosity of a variant locus, or its relative chromosomal location. For example, the zebrafish genotype in Table 1 describes a heterozygous complement at the fgf8a gene locus (ti282a/+), and a homozygous complement at the shha locus (tq252/tq252). And the worm genotype in this table is comprised of two variant loci, daf-2 and fog-2, which are indicated to reside on chromosomes 3 and 5 respectively. Such information is also incorporated into our genotype modeling efforts, as described below.

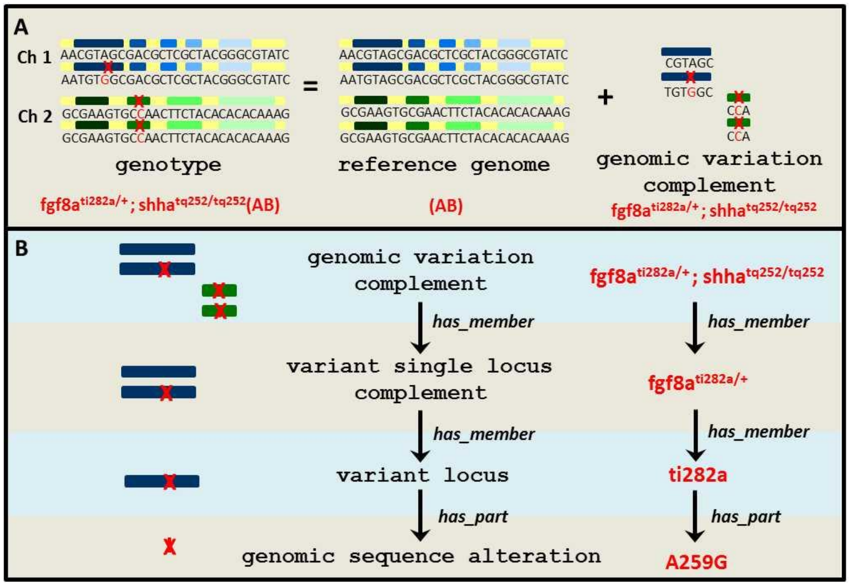

A
Ch 1
AACGTAGCGACGCTCGCTACGGGCGTATC
AATGTGGCGACGCTCGCTACGGGCGTATC
Ch 2
GCGAAGTGCCAACTTCTACACACACAAAG
GCGAAGTGCCAACTTCTACACACACAAAG
=
AACGTAGCGACGCTCGCTACGGGCGTATC
AATGTAGCGACGCTCGCTACGGGCGTATC
GCGAAGTGCGAACTTCTACACACACAAAG
GCGAAGTGCGAACTTCTACACACACAAAG
+
CGTAGC
TGTGGC
CCA
CCA
genotype
fgf8a$^{ti282a/+}$; shha$^{tq252/tq252}$(AB)
reference genome
(AB)
genomic variation complement
fgf8a$^{ti282a/+}$; shha$^{tq252/tq252}$
B
genomic variation complement
fgf8a$^{ti282a/+}$; shha$^{tq252/tq252}$
has_member
has_member
variant single locus complement
fgf8a$^{ti282a/+}$
has_member
has_member
variant locus
ti282a
has_part
has_part
genomic sequence alteration
A259G

Figure 1: Conceptual Decomposition of Genotype Sequence Content. **(A)** Top level breakdown into reference and variant components. **(B)** Further decomposition of the `genomic variation complement` into its more fundamental parts. Examples of a zebrafish genotype and its compositional parts according to our model are shown in red text.

2.2 Ontological Foundations of a Genotype Model

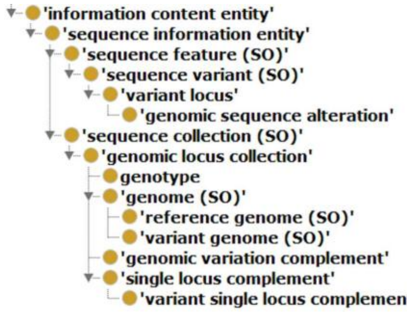
2.2.1 The SO Framework for Biological Sequences

The varied conceptions of a genotype mentioned above are due in part to the fact that the biological sequences that genotypes specify are themselves inconsistently understood (Hoehndorf et al. 2009). A precise characterization of biological sequences will be critical to building our ontology, as we will model genotypes in terms of their compositional sequence elements. Our basis for understanding these sequences is the Sequence Ontology (SO), an OBO Foundry ontology that models structural and functional genomic sequence features and their attributes. The SO views sequences as abstract entities representing a specific linear ordering of monomers, which are encoded in information artifacts such as text or database records that are about sequence macromolecules (Bada et al 2012). In Basic Formal Ontology terms (BFO, Smith et al., 2002), SO sequences are `information content entities` (ICEs) - generically dependent continuants that exist independent of space and time, and that stand in relation of ***aboutness*** to some entity.

GENO is implemented in OWL within this SO framework and guided by OBO Foundry principles (Smith et al. 2007). Notably, the SO is currently undergoing major refactoring to accommodate a parallel representation of physical sequences, and also expand its modeling of genetic variation (Bada et al. 2011). Current representation of variation in the SO is limited. While it includes classes such as ‘genotype’, ‘sequence alteration’, and ‘allele’, the precision and logical encoding of these representations is not sufficient for reasoning over rich variation data, or integration with phenotype data in a way that will supporting novel types of analyses. Towards these goals, GENO will align with and extend representation of variation found in the SO. We are collaborating with SO developers to work toward interoperable representations of genetic variation.

2.2.2 Modeling a Genotype and Its Parts

A first step in building our model is to decide where to place a `genotype` class within the BFO and SO framework (Figure 2). A genotype is an ICE that specifies some `genome`, however we do not see it as on par with the `genome` itself. As discussed above, genotypes indirectly resolve to a sequence through some conceptual operation on its reference and variant components. For this reason, we consider a genotype to be an SO `sequence collection`, but not a direct subtype of SO `genome`. Placement of the remaining core GENO sequence classes is relatively straightforward, following from the compositional breakdown outlined above and in Figure 1. Briefly, a `reference genome` is a subtype of `genome` that bears a `reference role`. A `genomic variation complement` is a `sequence collection` comprised of `variant single locus complements`, which are also `sequence collections` that contain as members at least one `variant locus`. A `variant locus` is a type of SO `sequence variant` whose extent is delimited by specific coordinates in the genome of an organism. Finally, a `genomic sequence alteration` is classified as a type of `variant locus` that varies along its entire extent. Definitions and axioms for each of these core GENO classes can be further explored in ontology at http://purl.obolibrary.org/obo/geno.owl.



- 'information content entity'
  - 'sequence information entity'
    - 'sequence feature (SO)'
      - 'sequence variant (SO)'
        - 'variant locus'
          - 'genomic sequence alteration'
    - 'sequence collection (SO)'
      - 'genomic locus collection'
        - genotype
        - 'genome (SO)'
          - 'reference genome (SO)'
          - 'variant genome (SO)'
        - 'genomic variation complement'
        - 'single locus complement'
          - 'variant single locus complement'

Figure 2: Integration of GENO modeling under SO class structure. Current SO classes indicated by ‘(SO)

2.2.3 Modeling to Support Data Ingest, Query, and Analysis

On top of its core genotype partonomy, GENO implements additional modeling to allow linking of genotype sequences to their attributes and related biomedical concepts. This will provide a framework for integration and analysis of genotype data captured in MODs and human variation databases with other valuable types of biomedical information. A key area of related modeling is the linking of genotype sequence elements to phenotypes. We have built design patterns for linking variant sequences at all levels of the GENO partonomy to phenotypic outcomes, as represented by a number of established and high-quality phenotype ontologies (e.g. Human Phenotype Ontology, HPO; Mammalian Phenotype Ontology, MP). Here, the logic encoded in GENO will support additional inferencing of relationships between genotype components and phenotypes, to enhance capabilities of systems such as Monarch. For example, it is desirable for purposes of query and analysis to establish links between a phenotype annotation asserted on a full genotype, and the more fundamental components of that genotype such as an individual variant locus. This process, which we call ‘phenotype propagation’ (Washington et al. 2009), is enabled by the definition of composed relations using OWL property chains, allowing GENO to support the inference of relationships between a phenotype annotation and individual variant loci or genes. These inferred links can be used to support a range of search, display, and analysis functionality in the Monarch system.

Another goal of GENO is to support linking genotype variant sequences to their genomic coordinates. The ability to precisely locate a variant locus on a genome build would add great value to genotype data, particularly if the genome is well annotated with additional information. In this way, we might enable applications that feature visual navigation of genotype data in the context of a genome browser, and inferencing about structural and functional features of variants based on overlapping genome annotations.

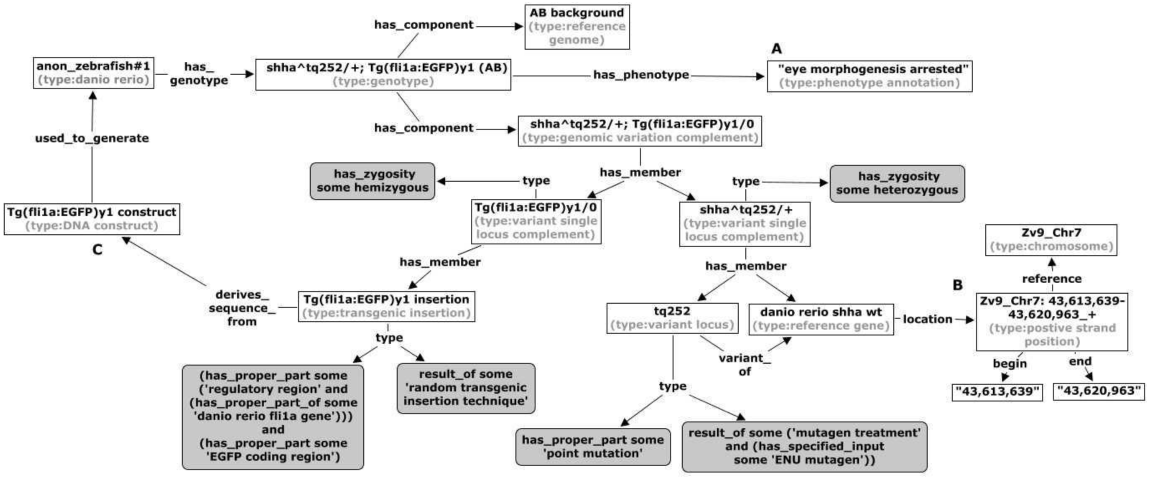

has_component
AB background
(type:reference genome)
A
anon_zebrafish#1
(type:danio rerio)
has_ genotype
shha^tq252/+; Tg(fli1a:EGFP)y1 (AB)
(type:genotype)
has_phenotype
"eye morphogenesis arrested"
(type:phenotype annotation)
used_to_generate
has_component
shha^tq252/+; Tg(fli1a:EGFP)y1/0
(type:genomic variation complement)
has_zygosity some hemizygous
type
has_member
type
has_zygosity some heterozygous
Tg(fli1a:EGFP)y1 construct
(type:DNA construct)
C
Tg(fli1a:EGFP)y1/0
(type:variant single locus complement)
shha^tq252/+
(type:variant single locus complement)
Zv9_Chr7
(type:chromosome)
has_member
has_member
reference
B
derives_ sequence_ from
Tg(fli1a:EGFP)y1 insertion
(type:transgenic insertion)
tq252
(type:variant locus)
danio rerio shha wt
(type:reference gene)
location
Zv9_Chr7: 43,613,639-43,620,963_+
(type:postive strand position)
type
variant_ of
begin
end
(has_proper_part some ('regulatory region' and (has_proper_part_of some 'danio rerio fli1a gene'))) and (has_proper_part some 'EGFP coding region')
result_of some 'random transgenic insertion technique'
type
"43,613,639"
"43,620,963"
has_proper_part some 'point mutation'
result_of some ('mutagen treatment' and (has_specified_input some 'ENU mutagen'))

Figure 3: Application of GENO to describe an instance of a complex zebrafish genotype. White squares are instances, with ontological type indicated in gray text below. Rounded gray boxes are class expressions that compose unnamed types. The "shha$^{tq252/+}$; Tg(fli1a:EGFP)y1 (AB)" genotype exhibits two variant loci on an AB background: a heterozygous point mutation at the shha gene locus (tq252), and a random transgenic insertion of a fli1a:EGFP reporter construct. The genotype is decomposed into its fundamental elements according to the GENO core partonomy, and these are linked to various attributes including a phenotypic outcome (A), chromosomal coordinates using FALDO design patterns (B), and experimental reagents (C).

2.3 Representation of Genotype Data using GENO

The utility and quality of the ontological framework outlined above will ultimately be measured by its ability to structure complex instance data in a manner that will support diverse data ingest, integration, and analysis applications. We have tested our model by using it to describing genotypes and related data from humans and various MODs, including ZFIN zebrafish and MGI mice. Figure 3 illustrates one example of how a complex ZFIN genotype is decomposed into its fundamental units of variation using GENO, and how these can be linked to related entities such as experimental provenance, genomic location, and phenotypic outcomes. D2R mapping to the GENO model is also underway to support publication of genotype data as RDF.

3 DISCUSSION AND CONCLUSIONS

A key contribution of this work relates to its disentangling conceptual, terminological, and syntactic aspects of genotype representation, which show remarkable heterogeneity across the biomedical research community. ***Conceptually***, we believe our approach can accommodate a number of similar and competing variant representations as subsets or extensions of GENO, including those used to capture human and model organism variation data. ***Terminologically***, we provide precise meaning for many variation-related terms that are ambiguous or inconsistently used (e.g. 'allele', 'locus'). ***Syntactically***, in order to integrate and operate across diverse genotype data, we are developing a generic genotype syntax that can support precise, granular mappings across diverse sources, support cross-species computational analysis, and identify areas where MOD syntaxes can be improved. Together, our work will offer a precise characterization of genotypes through ontological and syntactic models that will support understanding, integration, and analysis of genetic variation data across the biomedical domain.

ACKNOWLEDGEMENTS

This work was funded by an NCATs contract via Booz Allen Hamilton Award # 90177520 and NIH # 1R24OD011883-01. We are grateful to OBI and SO developers, and Jim Balhoff for their intellectual contributions.

REFERENCES

Smith RN., Aleksic J., Butano D.,…& Micklem G. (2012). Intermine: A Flexible Data Warehouse System for the Integration and Analysis of Heterogeneous Biological Data. *Bioinformatics*, **28**(23), 3163-3165

Mahner M., Kary M. (1997). What exactly are genomes, genotypes and phenotypes? And what about phenomes? *Theor Biol.* **186(1):**55-63.

Eilbeck K., Lewis S., Mungall CJ., Yandell M., Stein L., Durbin R., Ashburner M. (2005) The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology.* **6**:R44

Hoehndorf, R., Kelso, J., and Herre, H. (2009) The ontology of biological sequences.*BMC Bioinform* **10**:377

Bada M., Eilbeck K. (2012). Efforts Toward a More Consistent and Interoperable Sequence Ontology. Proceedings of the 3rd International Conference on Biomedical Ontology (ICBO 2012), KR-MED Series Graz, Austria, July 21-25, 2012.

Smith B., Grenon P. (2002). Basic formal ontology. Draft. Downloadable at http://ontology. buffalo. edu/bfo.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., . . . & Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, **25(11),** 1251-1255.

Washington, NL., Haendel, MA., Mungall, C.J., Ashburner, M., Westerfield, M., & Lewis, SE. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS biology*, ***7*(11),** e1000247.