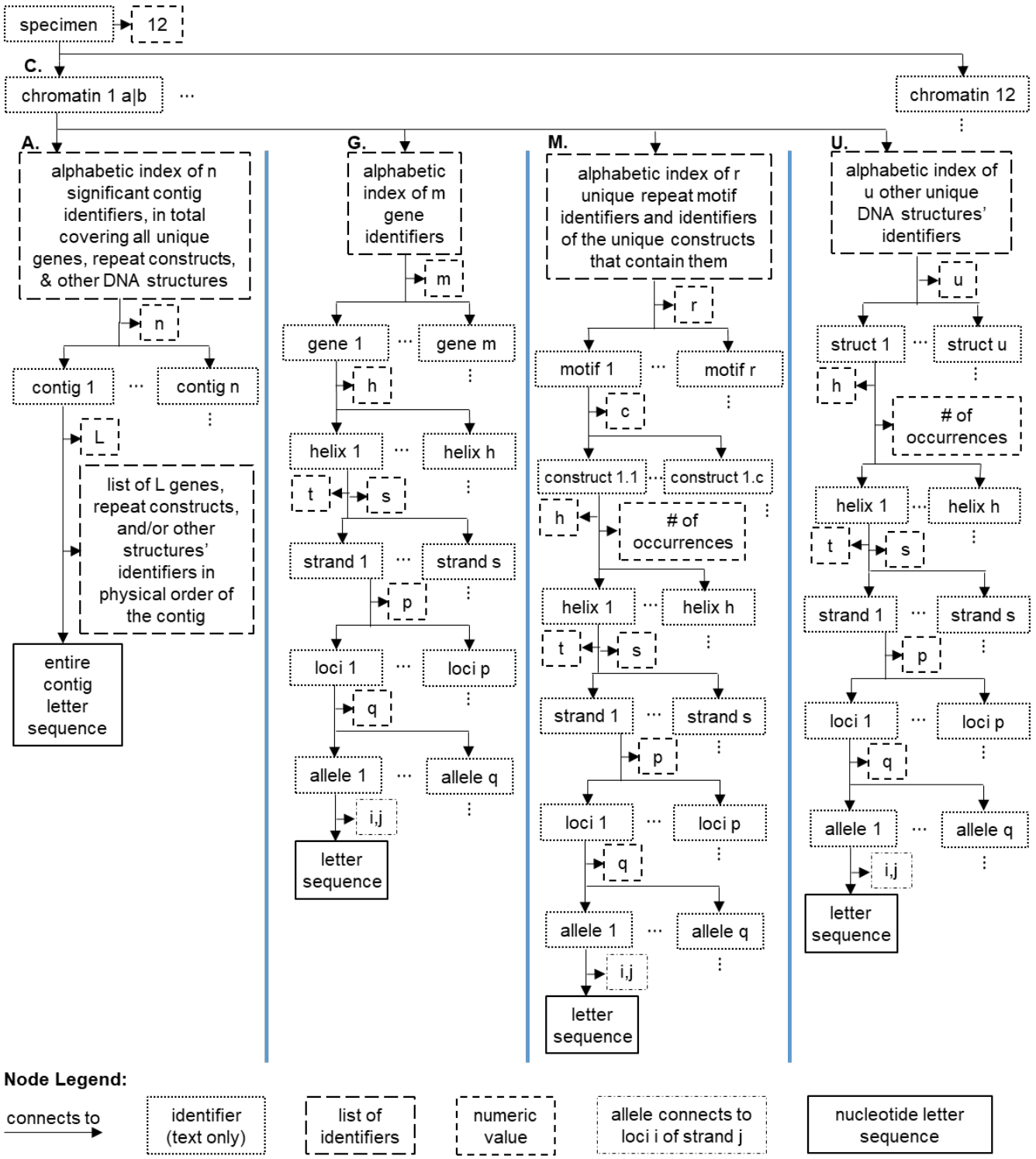
Data structure for annotated DNA, draft #9 10/18/2021
Richard Frost, Vista CA. tangent.vectors@gmail.com
 https://geneticdistance.org

Proposed hierarchical data structure for 2 specific and 1 catch-all category of sequences found within scanned DNA of fruiting perennial plants – and the contigs in which they are found. *Ficus carica* is used as an example. To implement it I suggest zipped hierarchical FASTA format utilizing CSV files. See Rationale below.



Rationale

The current whole genome data structures utilized by sources on GenBank are unsuitable for annotated storage, retrieval, and analysis. The above diagram is one possible solution. It is a post-assembly construction containing only annotated structures and contigs that are significant to the structures – omitting superfluous sequences with completely redundant information.

At the highest level is the specimen name and the number of chromatins. The work of Mori et al [1] demonstrates for *Ficus carica* that C = 12, one of which will be configured in either male or female form. Following Mori's convention, the chromosomes (level C) would be enumerated 1a or 1b, 2, 3, …, 12.

Each chromatin contains a hierarchy of structures including DNA genes, repeat motifs, and other nucleotide structures of unknown function. These are to be placed in sections G, M, and U – with the latter functioning as a catch-all category. The contigs in which these elements were identified reside in section A. The identifiers of contigs can follow local conventions such as enumerated names, e.g. ctg100420066. The same is true of identifiers for sections G, M, and U except in possibly rare cases that the sequencing lab is aware of a name in common use.

Regarding linear continuity of a whole chromosome, I have yet to see a construction for a linear physical ordering of all chromatin DNA – and in fact some of the genes could be separated within the chromatin in different hierarchies (e.g. fibers) as they are in mammals [2]. Thus there is no expectation for a contig to span the complete DNA of a single chromatin.

The m genes of a particular chromatin placed in section G are the simplest to annotate. Each is composed of up to h helixes (can vary by m), which in turn can be notated by type t (e.g. A, B, Z, …) and the number of strands s in the helix. For example in *F. carica* h=1 and s=2, but for *Prunus domestica* h=3 and s=2 [3]. In some species the number of strands s can vary with h or higher order indices.

Each strand is composed of p physically ordered loci (can vary by strand), which have q alleles (can vary by locus) with one end anchored at the strand locus – and one or more of which has the other end anchored at a corresponding locus of another strand in the helix, thus forming one or more base pairs. The ordered pair annotation i, j denotes a base pair connection from a particular location to locus i of corresponding strand j.

Repeat structures have an additional level of complication in that they are categorized by repeat motif, and within that by unique structures exhibiting the motif. Although the repeat motifs are numerous there are relatively few unique motifs. Whether or not the ordering of motifs among the repeats is biologically significant is unknown (to me) as of this writing. Also unknown is whether specific repeats preceding or following genes have biologic function. This physical ordering could be important to capture.

1. Mori K, Shirasawa K, Nogata H, Hirata C, Tashiro K, Habu T, et al. Identification of RAN1 orthologue associated with sex determination through whole genome sequencing analysis in fig (Ficus carica L.). Scientific reports. 2017;7(1):1-15.

2. Zheng H, Xie W. The role of 3D genome organization in development and cell differentiation. Nature Reviews Molecular Cell Biology. 2019;20(9):535-50.

3. Zhebentyayeva T, Shankar V, Scorza R, Callahan A, Ravelonandro M, Castro S, et al. Genetic characterization of worldwide Prunus domestica (plum) germplasm using sequence-based genotyping. Horticulture Research. 2019;6(1):12.