In-Depth Tiling Methods

Tiling Terminology

Tiling abstracts a called genome, or a callset, by cutting it into overlapping shorter segments, termed "tiles". We require each tile to cover at least 250 reference bases and share 24 bases with the tile sequentially before it and 24 bases with the tile after it. These 24-mer sections are termed "tags" and are chosen to be unique: they are at least a 2 base distance from anywhere in the genome. Each tile is labeled with an MD5 digest of the sequence it contains, which we term its "variant value". Additionally, each tile is labeled with the number of tiles before it, which we term its "position". One position can have multiple tile variants - one for each sequence observed at that position. We term the set of all positions and all tile variants to be a tile library. We represent individual callsets as arrays of variant values, one variant value for each tile position. Each callset in a population is represented using one tile library. Since choosing a different set of unique 24-mer tags results in a different tiling, each tile is also labeled with the set of tags used to create it.

For indexing purposes, we split the concept of a position into two values: a path and a step. We chose paths to be cytogenetic bands, defined by UCSC (`http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/cytoBand.txt.gz`). The step integer is a counter of the number of tile positions already passed in that path. All integers used here are represented in hex and are 0-indexed. Through the rest of this description, we will represent a tile position by a period-separated integer: 00.000.0000 ([tag set integer].[path integer].[step integer]). A tile variant will be represented by a tile position integer followed by the tile variant sequence MD5 digest: 00.000.0000.<MD5 digest>([tag set integer].[path integer].[step integer].[variant value MD5 digest]).

Tiling a Reference Sequence

Using the ENCODE Mapability data (DCC accession number: wgEncodeEH000608) generated in April 2010 [1, 2], we chose a tag set such that each 24-mer tag is only observed once in hg19 and each tag is at least 2 mismatches away from every other tag. For parallelization purposes, we separated hg19 into paths: one path per cytogenetic band (defined by UCSC). All further methods (shown in Fig. 1) were performed once for each path.

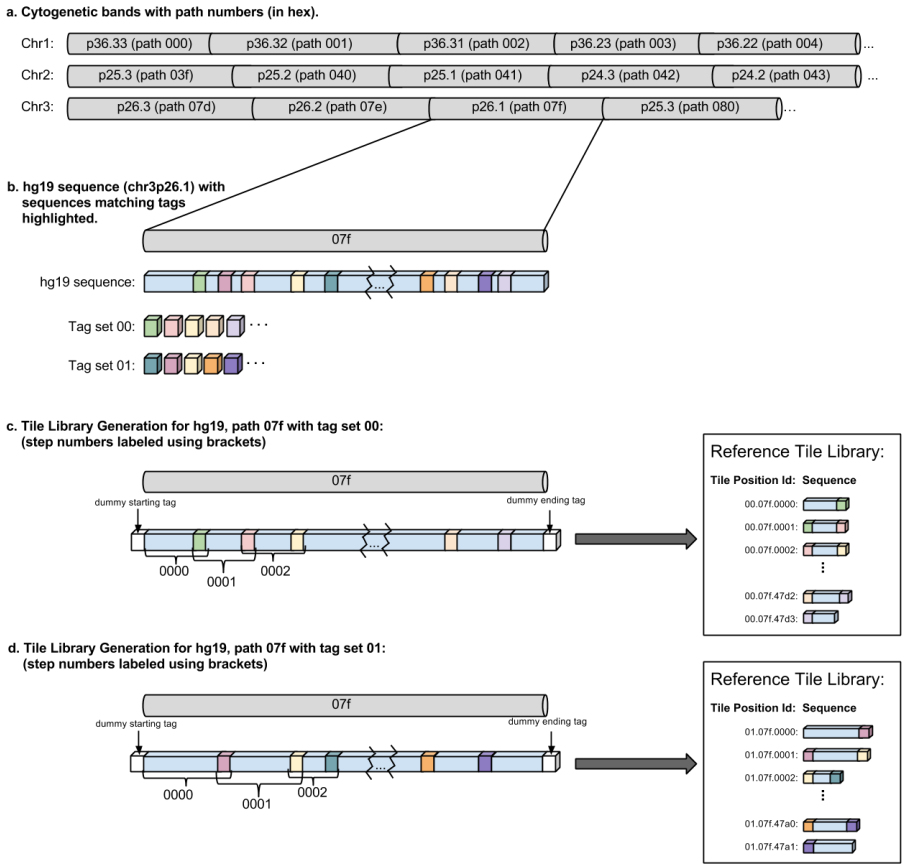

Figure 1. Tiling a reference genome. a. Cytogenetic bands labeled with their path number. Each band was treated independently. **b.** hg19 sequence highlighted where the sequence matches tags. Two tag sets (00 and 01) are shown. **c.** Tiling of hg19 on path 07f with tag set 00. White boxes indicate the 24-base dummy tags on either side of the sequence. Step numbers are labeled using brackets. The final tile library is labeled with positions ([path].[tag set].[step]). **d.** Tiling of hg19 on path 07f with tag set 01. White boxes indicate the 24-base dummy tags on either side of the sequence. Step numbers are labeled using brackets. Final tile library is labeled with positions ([path].[tag set].[step]). Note that this reference tile library can be combined with the reference tile library in **c**, since the tile position ids are different.

To ensure the entire sequence was captured, the reference sequence was capped on both sides with 24 dummy bases, serving as automatic start and end tags. We chose the dummy base tag as our first tag, skipped 202 bases, then scanned the sequence for the next 24 bases that appeared in our tag set. This 24-mer becomes the end tag of the first tile and the start tag of the next tile. We repeat until the end 24 dummy bases are reached. This greedy algorithm enforces the minimum base length requirement for each tile and ensures each non-edge tile overlaps with the tile preceding it and the tile following it.

A different set of unique 24-mer tags, which generates a different reference tile library, is also shown in Fig. 1. Note that the alternate tile library is given a different tag set integer (01).

Tiling a Variant Call File

The principal behind extending the tile library is fairly simple: first we create a temporary tile variant at tag set y, path x, and step 0. Then we scan through the sequence until a tag is reached. If the sequence we just scanned does not exist in the library at position $y.x.0$, we save the temporary tile variant under its sequence hash at position $y.x.0$. We start a new temporary tile variant at tag set y, path x, and step 1, then continue until the end of the path has been reached.

However, most variant call files do not store the entire human sequence, just the variations from a human reference. This difference may be sidestepped by initializing the temporary tile variant with the appropriate reference sequence at the tile variant's position. When a variation is read from the variant call file, we check if the variation position is in the current temporary tile variant. If it is, we alter the tile variant sequence accordingly and continue. Otherwise, we know we passed at least one tag. We save our temporary tile variant, if necessary, and iterate through tile variants, incrementing the step by one each time until we reach the tile position containing the variation.

When a variation occurs on a tag, a new issue arises. The tag set was chosen to be unique, so accepting an altered tag invalidates the definition of the tile library. We chose to solve this problem by allowing tile variants to span multiple steps when the tags that would normally end it contain a variation. By default, we assume that each tile variant spans exactly one step. If we find a variation that affects the ending tag, we find out how many more steps the variation effects. We add this to the number of steps spanned by the current tile variant. We then extend the current tile variant sequence to include the reference sequence of the steps the tile variant now spans into. We alter the tile variant sequence according to our current variation, and then continue. When we reach the end of the current tile variant, we go to the next step (our current step + the number of steps the last tile variant spanned). We perform this procedure, summarized in Fig. 2, for each path in each human.

No calls can be treated in one of two ways: assume all no-calls are reference, or consider no-calls to be variation. The first method conserves space, but is

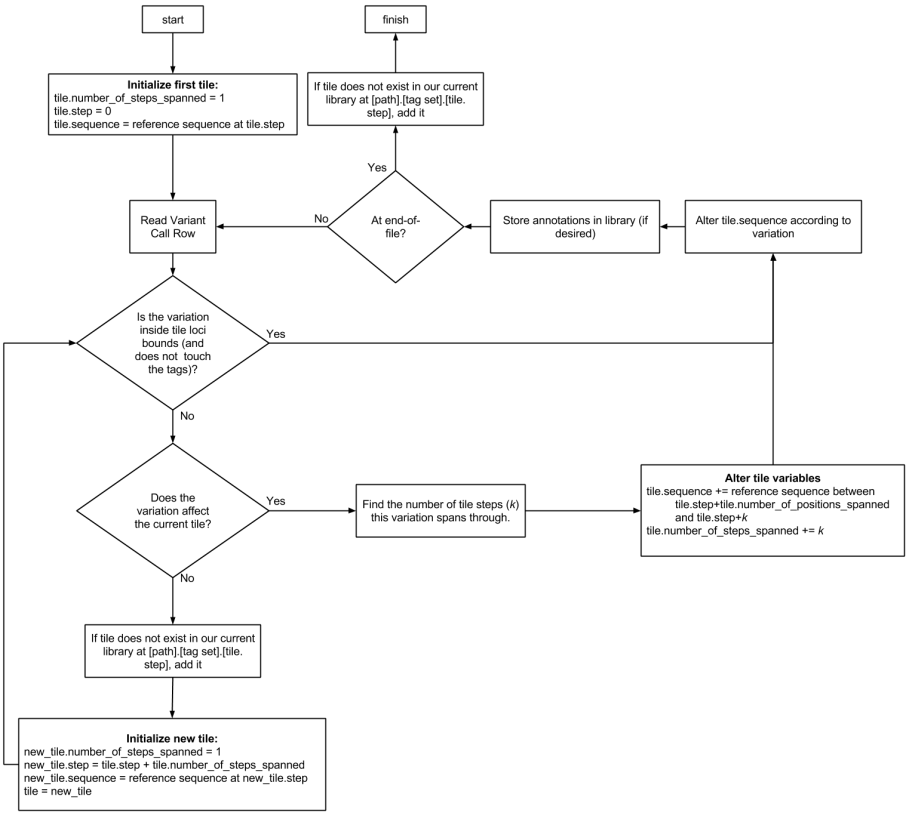

Figure 2. Flowchart for tiling a variant call file. Assumes the file only has one callset, ignores no-calls, is monophasic, and covers only one path.

lossy. The second method is lossless, but can lead to an exponential growth in the number of tiles, since variation that affects the tags automatically creates tiles that span multiple positions. We compromised by altering the tile sequence for no calls, but treating tags that were composed of reference and no-calls as normal tags. Only a definite variation (SNP, SUB, or INDEL) resulted in a spanning tile.

The procedure described above assumes that the phases are well known for the entire sequence; however, most current genomic data is partially or completely unphased. If the experimenters do not want to rely on the variant caller to correctly call phases, and instead wish to treat the entire sequence or loci ranges as unphased, they may add tile variants exactly as before, with regular expressions in their sequences instead of raw sequences. If any part of a spanning tile is unphased, the entire spanning tile is treated as unphased. We plan to incorporate more specific the phasing information at a higher level as a phase group annotation.

Once a tile library for a population is generated, genomic sequences can be represented compactly as arrays of tile variant values, one for each phase. Fig. 3 describes the tiling of two phased, called sequences. A visualization of the tile library for the *BRCA1* region, and a callset expressed using the tile library can be seen in Fig. 4 and an interactive version for *BRCA1* and *BRCA2* for 174 PGP callsets, GRCh38, GI388428999, GI528476586, and GI262359905_rc is at `http://science.curoverse.com/tiling/brca/pgp-graph`.

References

1. Derrien T, Estelle J, Marco Sola S, Knowles DG, Raineri E, Guigo R, et al. Fast computation and applications of genome mappability. PLoS One. 2012;7(1): e30377. doi: 10.1371/journal.pone.0030377.

2. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414): 57-74.

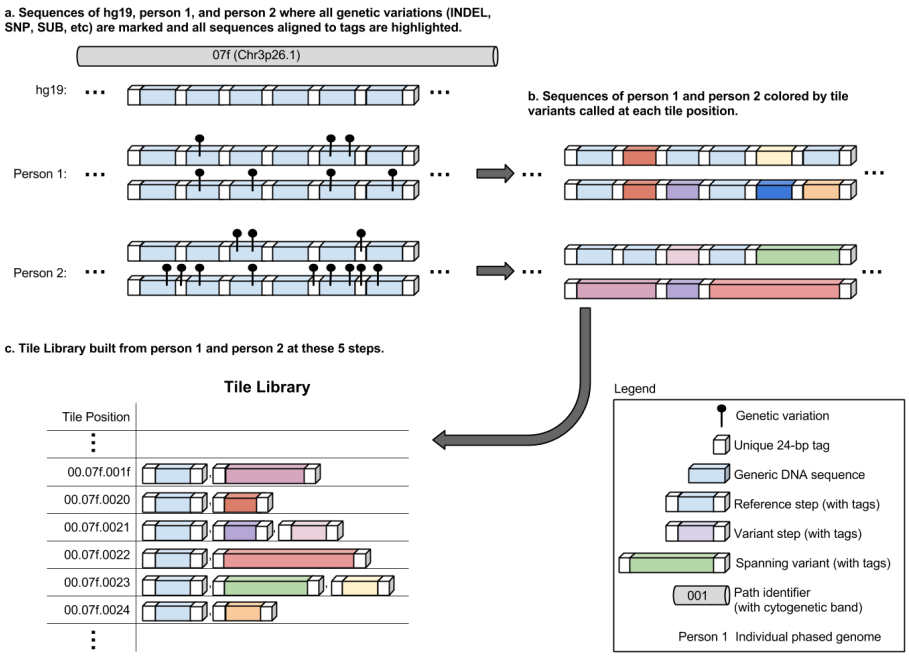

Figure 3. Tiling of two human sequences. All genetic variations, including insertions and deletions (INDELs), single nucleotide polymorphisms (SNPs), and substitutions (SUBs), are represented by the black lollipop symbol. In this simple example, if two symbols are in the same horizontal position, they represent the same genetic variation. Note that tiles are considered the same only if they have exactly the same set of genetic variations. Sharing a subset of genetic variations is not sufficient. If a variation occurs on a tag, the tile before it becomes a spanning tile: spanning multiple steps.

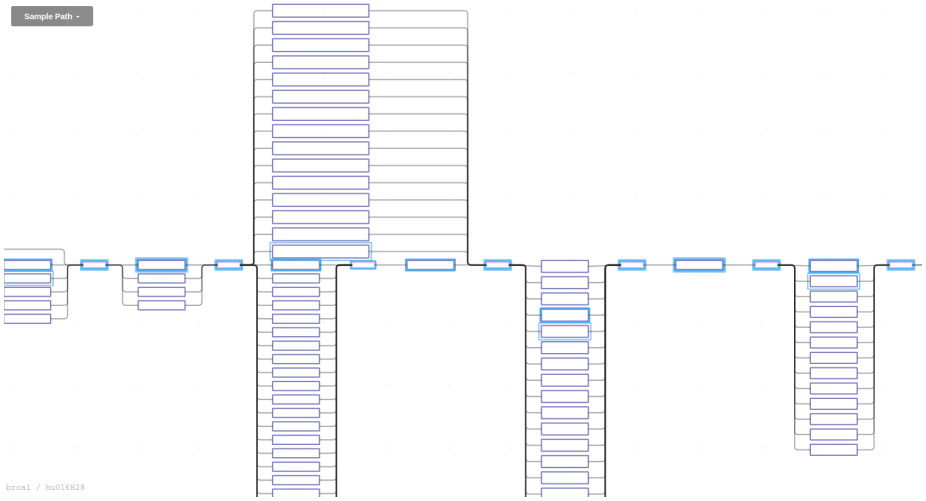

Figure 4. Visualization of a small portion of the *BRCA1* tile library built from the 174 PGP participants, with hu016B28 highlighted. Phase 1 is highlighted with the inner blue overlay; phase 2 is highlighted with a lighter outer blue overlay. Tags are colored in light pink and tile variants are colored in light purple. Tile variants are ordered into columns by tile position. Spanning tiles are shown above the center row. The center row is the GRCh37 reference sequence. Note this library section is one of high variance.