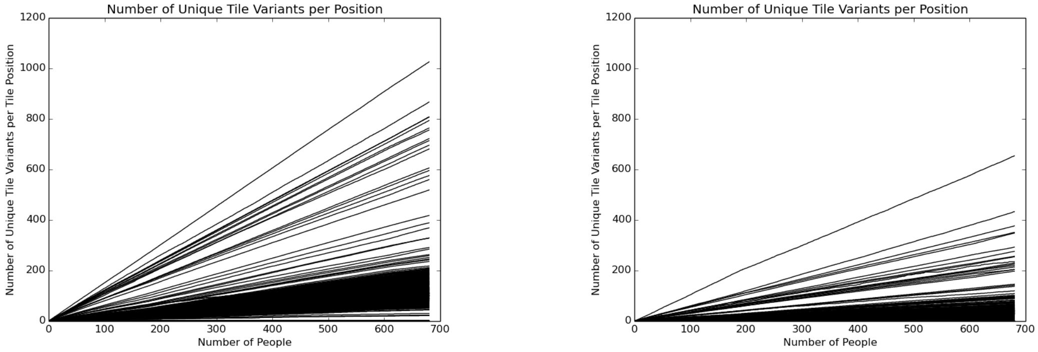
Upper-bounding Estimates for Storing 1 Million Tiled Genomes

Here we estimate the amount of memory needed to store 1 million tiled genomes and their tile library, with and without quality information. We consider three cases, one where quality information is stored elsewhere, one where the quality information is contained in the tile library and another where the quality information is maintained at the genome level. We assume the growth of the tile library (in the number of bases and the number of tile variants it contains) is upper bounded by a linear function. Our observations over the 680 tilings support this assumption (Fig. 1a, 1b, 2a, and 2b). We also assume that the advances in whole genome sequencing will preserve, if not increase, the number of well sequenced tile positions in each genome (80.4%).

The Haussler Million Genome Warehouse [1] estimates a storage cost of $25 per genome per year, which assumes 1 GB per called genome, or 1 PB for 1 million genomes. Our cost estimations assume an average node cost of $8000 per year. Each node has 256 GB, and is sold in racks of 32 nodes each. We note that using these numbers, storing 1 million genomes that use 1 GB each yields a storage cost of $31.49 per genome per year. A simplistic representation of tiled genomes is estimated to use 53 MB per genome and require 120 TB for the tile library containing tile variants for 1 million genomes ($5.63 per genome per year). If we assume that only 20% of poorly sequenced tile variants with a genomic variant are novel tile variants, the tile library size can be reduced to 31 TB and each genome can be represented in 39 MB, yielding an amortized cost of $2.31 per genome per year.

Quality Information Stored Outside the Tiling Representation

If we assume quality information is stored outside of the tiling representation, for example, with the collection of reads used to generate the called genome, the tile library grows slowly. We calculated an upper bound for the growth of tile variants by adding the number of well sequenced positions and the number of poorly sequenced tile variants that contained a genomic variant (Fig. 1b). Extrapolating from our observed linear tile variant rate of growth per genome (where each genome adds an average number of 0.054 new tile variants at each tile position, Fig. 1b), 1 million genomes will require 16 bits to represent each



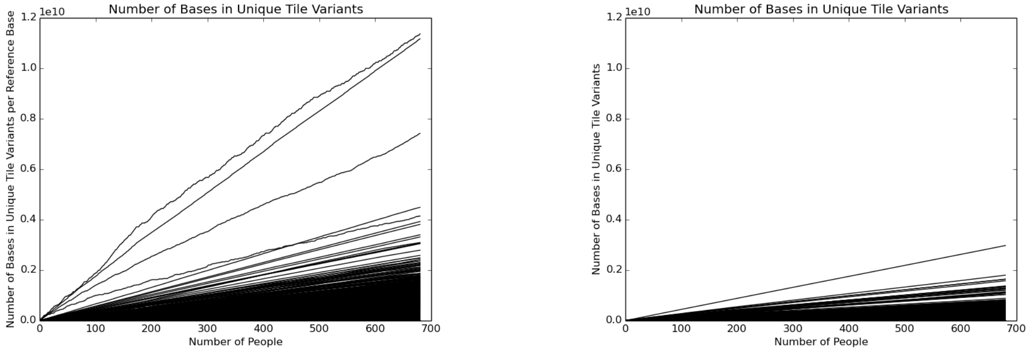
(a) Number of unique tile variants per tile position when no calls create new tile variants. Note that the majority of paths have less than 200 tile variants for 680 genomes, though there are outliers, which have less than 1100 tile variants for 680 genomes. This high growth is mostly a result of poorly sequenced regions, since a tile with no calls is treated as a tile with variants and tiles are only considered the same if all no calls and normal genomic variants are identical.

(b) Upper bound on the number of unique tile variants per tile position when no calls are represented in the called genome files. Note that the majority of paths have less than 100 tile variants for 680 genomes, though there are outliers, which have less than 700 tile variants for 680 genomes. This figure describes an upper bound on the number of unique tile variants, since we assume that a tile variant that includes poorly sequenced regions and a genomic variant is not already present in our tile library. We predict this assumption is likely to be untrue for the majority of tile variants.

Figure 1. Number of Unique Tile Variants per Position. Each sub-figure shows one line for each of the 863 paths. Note that all paths grow linearly.

tile position. Using this estimate, we can represent the called sequence of one phase of one genome in this cohort of one million genomes in 21.3 MB (42.6 MB for a normal biallelic genome).

The tile library for 680 which does not need to store quality information has at most 0.299 trillion bases. This library is composed of 390 million tile variants (332 million of which were poorly sequenced tile variants with a genomic variant in our tile library, which we assume are not already present in the library). If we represent this library with the number of bits defined by Shannon entropy [2], and assume that each A, T, C, and G occur with the same probability (which is the least optimal according to Shannon entropy), each base in the tile library requires 2 bits to represent. If we assume the number of bases in this optimized tile library grows linearly at the rate we observed in the 680 genomes (Fig. 2b), the tile library for the million genomes will include 440 trillion bases, which we can represent in 110 TB. These bases compose 0.574 trillion tile variants. Note that we did not use the MD5 digest in the tile library to conserve space (since the MD5 digest requires 128 bits to represent). We therefore need to associate each sequence with its MD5 digest and the integer used in each tiled genome representation, which add 9.19 TB and 1.15 TB respectively. This simplistic



(a) Total number of bases in the tile library when no calls create new tile variants. Note that most paths include less than 2 million bases.

(b) Upper bound on the number of bases in the tile library when no calls are represented in the called genome files. Note that most paths include less than 1 million bases. Similarly to Fig. 1b, this figure describes an upper bound on the number of bases required, since we assume that a tile variant that includes poorly sequenced regions and a genomic variant is not already present in our tile library. We predict this assumption is likely to be untrue for the majority of tile variants.

Figure 2. Number of Bases in each Unique Tile Variant. Each sub-figure shows one line for each of the 863 paths. Note that all paths grow linearly.

representation of the Tile Library uses 120 TB. We emphasize this is a gross overestimation, since the tile library contains a great deal of repetition and may be further compressed. Additionally, we expect most of the poorly sequenced tile variants that have a genomic variant to already be present in our tile library, eliminating the need to represent them.

1 million genomes and one copy of their tile library can be stored in random access memory (RAM) using 20 racks of standard hardware with 32 nodes per rack and 256 GiB of RAM per node, leaving 1 TB of RAM free for computation. We estimate this to have an amortized operating cost of $5.12 per genome per year. If we assume that only 25% of the poorly sequenced tile variants that had a genomic variant were not already present in the tile library, we can represent each biallelic genome in 39 MB and store the tile library in 36.5 TB. 1 million of these genomes could be stored using 10 racks of standard hardware and leave 5.4 TB of RAM free for computation, which has an amortized operating cost of $2.56 per genome per year.

Quality information kept in the tile library

Our preliminary tile library stored quality information in the tile library. Extrapolating from our observed linear tile variant rate of growth per genome (where each genome adds an average number of 0.172 new tile variants at each tile position, Fig. 1a), 1 million genomes will require 18 bits to represent each tile position. Using this estimate, we can represent one phase of one genome in this cohort of one million genomes in 24 MB (48 MB for a normal biallelic genome).

Storing the quality information in the tile library created a tile library composed of 0.656 trillion bases (0.498 trillion well sequenced bases and 0.158 trillion poorly sequenced bases). This library is composed of 1.25 billion tile variants (57.9 million well sequenced variants and 1.19 billion poorly sequenced variants). If we represent this library with the number of bits defined by Shannon entropy [2], and assume that each A, T, C, and G occur with the same probability (which is the least optimal according to Shannon entropy), each base in the tile library requires 2.32 bits to represent. If we assume the number of bases in the tile library grows linearly at the rate we observed in the 680 genomes (Fig. 2a), the tile library for the million genomes will include 965 trillion bases, which we can represent in 279.2 TB. These bases compose 1.83 trillion tile variants. Note that, again, we did not use the MD5 digest in the tile library to conserve space (since the MD5 digest requires 128 bits to represent). We therefore need to associate each sequence with its MD5 digest and the integer used in each tiled genome representation, which add 29.3 TB and 4.12 TB respectively. This simplistic representation of the Tile Library uses 313 TB. We emphasize this is a gross overestimation, since the tile library contains a great deal of repetition and thus is optimal for further compression.

1 million genomes and one copy of their tile library can be stored in random access memory (RAM) using 45 racks of standard hardware with 32 nodes per rack and 256 GiB of RAM per node, leaving 8 TB free for computation. We estimate this to have an amortized operating cost of $11.52 per genome per year.

Quality information kept in each tiled genome

Using our calculations from Quality Information Stored Outside the Tiling Representation, we can store one phase of one genome in our 1 million genome cohort in 21.3 MB (42.6 MB for a normal biallelic genome). However, we now need to hold quality information with each called genome. Since we assume the percentage of well sequenced tile positions will either be stable or increase, we can state that for 80.4% of tile positions, we can represent their quality in exactly 1 bit. The remaining 19.6% tile positions will require more bits, since we want to indicate which bases are poorly sequenced. The average number of bases in a poorly sequenced tile we measured in our 680 genomes is 518. If we represent each one of these bases with the number of bits defined by Shannon entropy [2], using the probability that any base is poorly sequenced to be the frequency of poorly sequenced bases in poorly sequenced tiles, each base can be represented in an average of 0.9 bits. The total memory used for storing quality information is 122 MB per genome ($1 * 0.804 * n + 0.196 * n * 518 * 0.9$bits, where n is the number of positions). Thus each genome will use 164 MB.

Our calculations from Quality Information Stored Outside the Tiling Representation also apply to the size of the tile library, which we calculate to use a maximum of 121 TB. 1 million genomes and one copy of their tile library can be stored in random access memory (RAM) using 35 racks of standard hardware with 32 nodes per rack and 256 GiB of RAM per node, leaving 2.3 TB free for computation. We estimate this to have an amortized operating cost of $8.96 per genome per year.

References

1. Haussler D, Patterson DA, Diekhans M, Fox A, Jordan M, Joseph AD, et al. A Million Cancer Genome Warehouse. Technical Report No. UCB/EECS-2012-2011.

2. Shannon, CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3): 379-423.