Tiling the Genome into Consistently Named Subsequences Enables Precision Medicine and Machine Learning with Millions of Complex Individual Data-Sets

Sarah Guthrie1,*, Abram Connelly1,*,2, Peter Amstutz1, Adam Berrey1, Nicolas Cesar1, Jiahua Chen1,3, Radhika Chippada1, Tom Clegg1,2, Bryan Cosca1, Jiayong Li1, Nancy Ouyang1,2, Jonathan Sheffi1, Brett Smith1, Ward Vandewege1,2, and Alexander Wait Zaranek1,2,**

1Curoverse Inc., Somerville, MA, USA
2Harvard Personal Genome Project, Boston, MA, USA
3University of Massachusetts Boston, Boston, MA, USA
*We consider these authors to have contributed equally to this work.
**Senior author and Corresponding author: sasha@curoverse.com

ABSTRACT

The scientific and medical community is reaching an era of inexpensive whole genome sequencing, opening the possibility of precision medicine for millions of individuals. Here we present tiling: a flexible representation of whole genome sequences that supports simple and consistent names, annotation, queries, machine learning, and clinical screening. We partitioned the genome into 10,655,006 tiles: overlapping, variable-length sequences that begin and end with unique 24-base tags. We tiled and annotated 680 public whole genome sequences from the 1000 Genomes Project Consortium (1KG) and Harvard Personal Genome Project (PGP) using ClinVar database information. These genomes cover 14.13 billion tile sequences (4.087 trillion high quality bases and 0.4321 trillion low quality bases) and 251 phenotypes spanning ICD-9 code ranges 140-289, 320-629, and 680-759. We used these data to build a Global Alliance for Genomics and Health Beacon and graph database. We performed principal component analysis (PCA) on the 680 public whole genomes, and by projecting the tiled genomes onto their first two principal components, we replicated the 1KG principle component separation by population ethnicity codes. Interestingly, we found the PGP self reported ethnicities cluster consistently with 1KG ethnicity codes. We built a set of support-vector ABO blood-type classifiers using 75 PGP participants who had both a whole genome sequence and a self-reported blood type. Our classifier predicts A antigen presence to within 1% of the current state-of-the art for in silico A antigen prediction. Finally, we found six PGP participants with previously undiscovered pathogenic BRCA variants, and using our tiling, gave them simple, consistent names, which can be easily and independently re-derived. Given the near-future requirements of genomics research and precision medicine, we propose the adoption of tiling and invite all interested individuals and groups to view, rerun, copy, and modify these analyses at https://curover.se/su921-j7d0g-swtofxa2rct8495.

Keywords: precision medicine, clinical screening, human genetics, machine learning

INTRODUCTION

Precision medicine requires a collaboration between researchers, clinicians, and patients [Brenner, 2007; Kohane, 2015]. Clinicians must incorporate genetic findings made by a community of scientists and other clinicians to inform diagnoses and treatments [Foley et al., 2015; Lee et al., 2015; Ellingford et al., 2015; Soden et al., 2014]. Patients must be able to understand their genetic diagnoses and treatment recommendations. Researchers should be able to generate, agree with, or dispute genetic findings guiding these recommendations [Rehm et al., 2015; Landrum et al., 2014]. For millions of genomes, they should be able to drill down into specific locations, examine whole genomes, and separate the population based...
on phenotypic data and intrinsic data discovered by machine learning algorithms. Extensive progress has been made to foster these connections, leading to the creation of resources such as ClinGen (Rehm et al., 2015), COSMIC (Forbes et al., 2015), LOVD (Fokkema et al., 2011), the UCSC Genome Browser (Karolchik et al., 2014), and EMBL-EBI (Brooksbank et al., 2014). However, as we are poised to receive genomic datasets on an unprecedented scale, there is an opportunity to revisit the representation of the genome and ensure it enables all the aforementioned use cases.

We feel this representation must be open and public, since public solutions encourage reproducibility and collaboration (Buck, 2015; McNutt, 2014; Angrist and Cook-Deegan, 2014; GA4GH, 2015a; Collins and Varmus, 2015). In addition, we feel this solution must be built using individual human genomic data. It should be able to represent all types of genomes well, including somatic and germline (Levy et al., 2007; Chaisson et al., 2015; Zhang et al., 2011). It must have a good machine representation that can support standardized queries and machine learning techniques. Standardized queries, such as those implemented by the Global Alliance for Genomics and Health (GA4GH) Beacon Project (GA4GH, 2015b) and GA4GH Matchmaker Exchange (MatchmakerExchange, 2015), are gaining importance as data volumes increase and more organizations are using these data in clinical settings. Machine learning techniques on genome sequences are used to predict phenotype (Gonzalez-Recio and Forni, 2011; Ornella et al., 2014; Yoon et al., 2012), generate disease prognoses (Abraham et al., 2012; Kourou et al., 2014), strengthen genome-wide association studies (Botta et al., 2014; Mittag et al., 2012; Pirooznia et al., 2012; Roshan et al., 2011), and elucidate ancestry (Hajiloo et al., 2013). Finally, we feel this whole genome representation must be consistent: two researchers or clinicians, given the same called genome, should be able to generate the same representation of that genome. Consistency vastly reduces the likelihood of error, and thus simplifies collaboration, annotation aggregation, comparison between genomes, and analyses that span multiple studies (Sboner and Elemento, 2015).

The VCF-like representation for called genomes, which includes variant call format files (Danecek et al., 2011) (VCF) and genome variation format files (Reese et al., 2010) (GVF), stores the places where the called genome differs from a mostly linear reference. This type of representation is excellent at storing annotations both on specific variants (such as allele frequency or predicted effects on protein sequence) and on genomic regions (such as introns and exons). It also represents quality and read depth for called genomes well. With additional software and pre-processing, these files can be used for queries and machine learning techniques. Unfortunately, this representation has unstandardized variant representations, so one variant can be represented many different ways (Danecek et al., 2011; github, 2015). This lack of standardization complicates machine learning, requiring a greater number of samples to find an effect. Finally, the use of a linear reference genome does not effectively represent highly variable regions, such as the major histocompatibility complex and cancer genomes (Church et al., 2015; Difithey et al., 2015).

Graphical representations are being developed to address these issues, including representing highly variant regions (Difithey et al., 2015). Graph-based designs are effective at conceptualizing and implementing pan-genomes, the genomes of entire species (Church et al., 2015; Marcus et al., 2014). Graphical representations will, over time, support queries and machine learning; however, the additional complexity of using graphs will slow adoption.

We propose tiling, a new representation of whole genome sequences, as a genomic abstraction for precision medicine. Tiling is designed to be a compromise between the VCF-like and graph representations that includes basic quality and phasing information. It provides identical subsequences with unique identical names, supports clinical analysis, and simplifies machine learning. Additionally, tiling allows us to define an individual’s genome more rigorously (Text S3), providing a probabilistic estimate of the compatibility of two genomes generated from the same sample. Here, as a demonstration of the tiling abstraction’s use in precision medicine, we tiled 680 public genomes, and expressed them in a preliminary GA4GH graphical representation and a GA4GH beacon query engine. We analyzed these 680 genomes using supervised and unsupervised machine learning, found 14 pathogenic BRCA variations, and give each variation a simple, consistent name. The open nature of our approach and data allows researchers to use tiling to assign variants consistent names in any gene or region of interest.

RESULTS

All results described here may be found, replicated, and rerun on different data using Arvados at http://curover.se/su92l-j7d0g-swtofxa2rct8495.
Tiling the Genome
Tiling abstracts a called genome by partitioning it into overlapping shorter segments, termed tiles. Here we chose each tile to be at least 250 reference bases long, with the beginning and ending 24 bases overlapping with adjacent tiles. These overlapping 24-mers are termed “tags” and are chosen to be unique: they are at least a 2 base distance from anywhere in the genome. Our choice of tags partitioned the GRCh37 human reference genome into 10,655,006 tiles, composed of 3.1 billion bases (an average of 314.5 bases per tile). Each tile is labeled with an MD5 hash digest of the sequence it contains, which we term its “variant value” and which serves as an unique, reproducible identifier for that tile. Since tags are unique, identical sub-sequences from different tile positions in the genome are guaranteed different unique names. Additionally, each tile is labeled with the number of tiles before it, which we term its “tile position”. One tile position can have multiple tile variants - one for each sequence observed at that position. We term the pan-genome of all tile variants for all tile positions a “tile library”. Note that choosing a different set of unique 24-mer tags would result in a different tiling. Therefore, each tile is also labeled with the set of tags used to create it, which we term a “tag set”. Detailed methods for tiling genomes are described in Text S1.

Tiling 680 public genomes from the 1000 Genomes Project and the Harvard Personal Genome Project (PGP) covered 14.13 billion tiles (4.519 trillion bases), resulting in a tile library containing 1.246 billion tile variants (0.656 trillion bases, 116.9 tile variants per tile position). On average, each new genome added 0.1719 new tile variants per position. 68.66% of these tile variants had no genomic variants but contained at least one uncalled base, which we consider to be a poorly sequenced region. By representing poorly sequenced regions outside the tile library, the tile library will take, at the most, 0.3904 billion tile variants (0.299 trillion bases, 36.64 tile variants per tile position, 0.0539 new tile variants per position per genome). The predicted growth of the tile library for one million tiled genomes is discussed in Text S2.

An individual called genome can be easily represented as an array referencing the tile library. Each position in the array corresponds to a tile position and points to the tile variant observed at that position for that individual. Since the variant value is the hash of the sequence observed at the tile position, the variant value and the tag set used to generate the tile may be used to unambiguously refer to subsections of the genome. This property provides consistency and reproducibility when representing genomes and important variants.

Annotations may be added to the tile library for both tile positions and tile variants. Each tile position may be annotated with its chromosome location to support translation between reference genomes and queries defined by chromosomal loci. Additionally, tile positions may be annotated with gene names, introns, and exons. Tile variants may be annotated similarly to variants in VCF-like representations. Annotations are added to and stored in the tile library, which is population-wide. Since individual sequences use pointers to tile variants in the tile library, all individuals with a pointer to a tile variant in an annotated library will be annotated identically with all tile variant and tile position annotations.

After tiling the 680 public genomes, we annotated all tiles with ClinVar database information, exome aggregation consortium (ExAC) database information, and the annotations produced from the clinical annotation of variants (CAVA) tool [Muenz et al., 2015]. We also expressed the BRCA1 and BRCA2 regions for the PGP whole genome sequences in the GA4GH graphical representation (Fig. [1], and exposed them in a query engine.

Organizing the PGP Phenotype data
We collected the PGP reported and predicted phenotypes, organized them into an SQLite3 database, and exposed them using Arvados at [http://curover.se/su921-j7d0g-k6xj9k9g015pmqh](http://curover.se/su921-j7d0g-k6xj9k9g015pmqh). This information is recollected periodically to produce snapshots of the PGP data. Each snapshot is about 200 MB, and included in this database snapshot are pointers, often in the form of a URL, to the whole genome files, user uploaded files, and reports for each. A snapshot is also included in the Arvados project associated with this paper ([http://curover.se/su921-j7d0g-swtofxa2rct8495](http://curover.se/su921-j7d0g-swtofxa2rct8495)).

Unsupervised Learning - Principal Component Analysis
We performed PCA on well-sequenced positions in the autosomal chromosomes for our 680 whole genomes. Our tiling contains 10,080,498 tile positions per phase for the autosomal chromosomes (20,160,996 total). We defined a tile position to be well sequenced if all tile variants at that position were free of no-calls for the entire population. This definition is very stringent, since 680 genomes, each with randomly distributed single-base no-calls in only 1% of tile positions, would be expected to have 11,867
Figure 1. Tile library built from the Harvard Personal Genome Project participants’ whole genomes for exon 11 of BRCA2, with the sequence of participant hu92C40A highlighted. The tile library is shown in the context of the UCSC Genome Browser (Karolchik et al., 2014). Participant hu92C40A has self-reported breast cancer and has a BRCA2 frameshift mutation. The two phases of participant hu92C40A are highlighted (one with a red dashed line and one with an orange solid line) and their sequences are shown – this participant is homozygous for all tile positions except 00.247.0b2e. Tile variants are ordered into columns by tile position. Tags are the smaller rectangles and tile variants are the larger ones. The zoomed-in tile variant is labeled with its tile position, tile variant value (which is the MD5 digest of the start tag, the variant sequence, and the end tag), and the HGVS labels for the genomic variation present in the tile. The number of haplotypes containing each tile variant are shown in parenthesis.

Well sequenced tile positions. At 2.5% no tile positions would be expected to be well sequenced. In our data, only 29,366 (0.146%) of the autosomal tile positions were well sequenced for all 680 whole genome sequences.

We then projected the 680 tiled sequences onto the first two principal components obtained from PCA (Fig. 2). The projection of the 502 1000 Genomes sequences results in clusters corresponding to ethnic background of the participants (Fig. S1). This replicates the ethnic clustering from the 1000 Genomes Consortium in 2012, who projected SNP data onto the first two principal components (1000GenomesProjectConsortium, 2012). The projection of the 178 PGP whole genome sequences also clusters the population based on ethnic background (Fig. S2). Colors shown in Fig. 2 for the PGP data were obtained by examining both the PGP reported ethnicities and their reported grandparent country of origin. Any participant reporting to be white and having at least one grandparent with a European country of origin is colored magenta. Three participants (hu38168C, hu92FD55, and huCA017E, colored olive) reported themselves to be Asian and have grandparents exclusively from China. One participant (huEA4EE5, colored dark orange) reported being Hispanic or Latino and has grandparents exclusively from Columbia. Finally, one participant (hu49F623, colored black) reported being Black or African American and has grandparents exclusively from Ethiopia.

Supervised Learning - ABO Blood Type Classifiers
75 of the 178 PGP called genomes had self-reported ABO blood types as follows: Type O (32), Type A (30), Type B (13), and Type AB (0). We trained two support vector machines (SVMs), one to predict A
Figure 2. Projection of well sequenced positions in autosomal chromosomes along the first two principal components for 680 whole genome sequences. In total, 29,366 tile positions are used out of 20,160,996 tiles per genome (10,080,498 tiles per phase) (0.146% of the genome). As expected from the 1000 Genome Project results (1000GenomesProjectConsortium [2012]), the first principal component separates the participants of African descent from the other participants. The second principal component separates the participants of European and Asian descent. PGP participants are marked with a circle, 1000 Genome participants are marked with a triangle. The African super population is colored with brown shades, the Ad Mixed American super population is colored with red shades, the East Asian super population is colored with green shades, the South Asian super population is colored in light blue shades, and the European super population is colored in purple shades. The legend labels each specific color with its 1000 Genomes population description, if applicable, and provides the number of whole genome sequences with this ethnicity in parenthesis. Note the PGP European population cluster is less precise than the 1000 Genomes cluster, and the PGP Ethiopian participant, though separate from the other clusters, does not fall into the African super population cluster. Interestingly, the PGP self reported ethnicities cluster consistently with the 1000 Genomes population ethnicity codes.

antigen presence and another to predict B antigen presence, using these 75 called genomes as training data. Each tile position that was well sequenced for the 178 PGP whole genomes (3,609,599 tile positions, 16.9% of total tile positions) was treated as a feature. To reduce the number of features without losing information, we further restricted them to tile positions that were well sequenced and had more than one tile variant at each position for the 178 PGP whole genomes (2,097,222 tile positions, 9.84% of total tile positions). We used both classifiers to predict the ABO type of the remaining 103 unlabelled PGP called genomes; the classifiers with the highest accuracies generated predictions as follows: Type O (103), Type A (23), Type B (28), Type AB (18).
We found 6 PGP participants with confidently called pathogenic tile variants, which we associated with no self-reports relating to cancer. Finally, we found participant huCD380F with familial breast-ovarian cancer, complementation group D1 fanconi anemia, and pancreatic cancer. A antigen positive called genomes in the training set as A antigen negative. It predicted 43 of the 103 unlabeled called genomes to have the A antigen phenotype.

The classifier that predicted B antigen presence with the highest accuracy (84.0% ± 36.7%) was an SVM with a linear kernel. This classifier had 1,786,803 non-zero coefficients (85.2% of the available tile positions). The largest coefficient magnitude was 2.11 ± 10^-5, 4 orders of magnitude less than the coefficients of the A antigen classifiers. This classifier did not misclassify any of the training called genomes, and it predicted 44 out of 103 called genomes have the B antigen phenotype. An in-depth summary of the classifier parameterization results are in Text S5.

**Clinical Analysis of the BRCA regions**

We found 6 PGP participants with confidently called pathogenic tile variants, which we associated with their simple, consistent tile variant values (Tables 1 and 2). All 6 variations were predicted by CAVA to have a severe impact. We considered a tile variant pathogenic if it contains a variation generating a premature termination codon in BRCA1 or BRCA2, excluding variations that generate a premature termination codon in exon 27 after codon 3010 of BRCA2 (Rebeck et al., 2015). Tile variants containing genomic variants that result in large in-frame deletions spanning 1 or more exons or that delete transcription regulatory regions would also be considered pathogenic tile variants; however, we did not observe any such tile variants in our population. 4 of the 6 pathogenic tile variants contained variations in ClinVar that were classified as pathogenic by multiple submitters. We submitted the two new variants to ClinVar, along with the Harvard PGP identifiers associated with the genomic sequences containing these variants.

We added an evaluation of these variants to GET-Evidence, a community-edited database of variants where researchers, the general public, and PGP participants can learn about and discuss the significance of variants in a public, collaborative forum (Ball et al., 2012). We informed the Harvard PGP staff of our findings. The PGP tracks the recommendations of the regulatory and ethical standards of the GA4GH and, thus, simplifies the process of adhering with these recommendations. Additionally, by using GET-Evidence, researchers can be assured that PGP participants will be notified of new findings in their genomes once they are well supported by evidence. The open nature of these methods and data allows researchers to evaluate their own genes and regions of interest.

First, we found participant huF6B09D is heterozygous for a BRCA2 Asn1198Lys frameshift variant in exon 11, which is not in ClinVar. huF6B09D is a female in her 40's with no self-reports relating to cancer, though she reports Hashimoto's thyroiditis. GET-Evidence reports she is also heterozygous for variant CASP10-V410I, which is documented by GET-Evidence to have a dominant, likely protective effect on breast cancer. Second, we found participant hu92C40A is heterozygous for a BRCA2 Phe1866Leu frameshift variant in exon 11, which is also not in ClinVar. hu92C40A is a 71 year-old female who reports having had breast cancer and breast biopsies in 2010. Third, we found participant huD3A569 is heterozygous for a BRCA2 Ser1982Arg frameshift variant in exon 11, which ClinVar associates with familial breast-ovarian cancer, complementation group D1 fanconi anemia, and pancreatic cancer. huD3A569 is a male in his 40's. He reports having 2 skin cancer incidents that occurred in his 30's and a mother who has had 20+ occurrences of skin cancer. Fourth, we found participant hu72C17A is heterozygous for a BRCA2 Lys3084Asn frameshift variant and/or splice region variant in exon 24. ClinVar associates this variant with familial breast-ovarian cancer. hu72C17A is a male in his 50's with no self-reports relating to cancer. Participant hu72C17A has a cell line, GM24077, available from Coriell. Fifth, we found participant hu82436A is heterozygous for a BRCA1 Asp1162Val frameshift variant in exon 10, which ClinVar associates with familial breast-ovarian cancer. hu82436A is a 33 year-old female with no self-reports relating to cancer, though she reports Hashimoto’s thyroiditis. Further details about all participants are in Text S4.

No PGP participants in our population reported having fanconi anemia, ovarian cancer, or pancreatic cancer, so we did not perform Fisher's exact tests for these phenotypes. Unsurprisingly, we did not find a significant correlation between BRCA frameshift presence and breast cancer in our relatively young cohort.
Table 1. Details about the pathogenic tile variants. Tile variants are named by their tile position, which is separated into tag set, “paths”, and “steps”, and the MD5 digest of their sequence (see Text S1 for the description of path and step). Each tile variant is shown with variations from the GRCh37 reference genome. All genomic variations contained within a pathogenic tile are shown regardless of the genomic variant’s likely pathogenicity. The predicted protein change was produced by CA VA (Muenz et al., 2015) and GET-Evidence [Ball et al., 2012]. Harvard PGP ids of the Harvard PGP participants who have these variants are shown, as well as their genotype for the tile (heterozygous is abbreviated to het; homozygous is abbreviated to hom). All loci use 0-based GRCh37 coordinates.

<table>
<thead>
<tr>
<th>Tile position, Variant Value</th>
<th>Loci Covered by Tile Variant</th>
<th>Harvard PGP IDs</th>
<th>Genomic Variants in Tile</th>
<th>Predicted Protein Change</th>
<th>Predicted Stop-Gain Variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>00.247.0b24, 91e94cf32f709225f17f37c159930451</td>
<td>chr13: 32,911,657 - 32,912,133</td>
<td>huFFB09D (het)</td>
<td>32,911,887 SNP A→G</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>00.247.0b2e, 3552b96c68c47f5e7783addc4c2877a2</td>
<td>chr13: 32,913,918 - 32,914,167</td>
<td>hu92C40A (het)</td>
<td>32,912,080 INDEL A→T</td>
<td>BRCA2 N119Kfs</td>
<td>BRCA2 1199</td>
</tr>
<tr>
<td>00.247.0b30, b0ee262cb5737ac19d609912fa2f0d40</td>
<td>chr13: 32,914,368 - 32,914,617</td>
<td>huB3A569 (het)</td>
<td>32,914,437 INDEL T→C</td>
<td>BRCA2 F1866Lfs</td>
<td>BRCA2 1873</td>
</tr>
<tr>
<td>00.247.0bc8, d16d51ee6d027ae066ea827dc181d038</td>
<td>chr13: 32,954,218 - 32,954,467</td>
<td>hu72C17A (het)</td>
<td>32,954,277 INDEL AACA→TT</td>
<td>BRCA2 K3084Nfsa</td>
<td>BRCA2 3086</td>
</tr>
<tr>
<td>00.2c5.04c0, 84fa1b2665461533586d77aa8c21393e</td>
<td>chr17: 41,243,883 - 41,244,132</td>
<td>huB2436A (het)</td>
<td>41,243,999 SNP T→C</td>
<td>BRCA1 K1183R</td>
<td>-</td>
</tr>
<tr>
<td>00.2c5.04cb, e318e2ddbb54bb3d6cc3a5ca6749482c5</td>
<td>chr17: 41,246,365 - 41,246,614</td>
<td>huCD380F (het)</td>
<td>41,246,531 INDEL T→C</td>
<td>BRCA1 V340Gfs</td>
<td>BRCA1 345</td>
</tr>
</tbody>
</table>

*aThis frameshift results in a stop-gain variant at BRCA2 3086 and removes the exon-splice site in Exon 24.

bThis frameshift results in the next variant changing from BRCA1 Lys1183Arg to Lys1183Glu and inserts a stop-gain variant at BRCA1 1209. Next variant will change if the phasing called in the GFF file is accurate.
Table 2. Details provided by ClinVar for the pathogenic tile variants. Tile variants are named by their tile position, which is separated into tag set, “paths”, and “steps”, and the MD5 digest of their sequence. If variants were not present in ClinVar as of 23 July, 2015, N/A is shown.

<table>
<thead>
<tr>
<th>Tile position, Variant value</th>
<th>HGVS identifier</th>
<th>Associated ClinVar Conditions</th>
<th>Clinical significance</th>
</tr>
</thead>
</table>
| 00.247.0b24, 91e94cf32f709225f17f37c159930451 | NC_000013.10: g.32911888A>G | • Familial cancer of the breast  
• Breast-ovarian cancer, familial 2  
• Hereditary cancer-predisposing syndrome | Benign |
| 00.247.0b2e, 3552b9fc68c47f5e7783addc4c2877a2 | N/A | N/A | N/A |
| 00.247.0b30, b0ee262cb5737ac19d609912fa2f0d40 | NC_000013.10: g.32914438delTinsC | • Familial cancer of breast  
• Breast-ovarian cancer, familial 2  
• Fanconi anemia, complementation group D1  
• Pancreatic cancer 2  
• BRCA1 and BRCA2 hereditary breast and ovarian cancer  
• Hereditary cancer-predisposing syndrome | Pathogenic |
| 00.247.0bc8, d16d51ee6d027ae066ea827dc181d038 | NC_000013.10: g.32954278_32954281delAA-CainsTT | • Breast-ovarian cancer, familial 2 | Pathogenic |
| 00.2c5.04c0, 84fa1b2665461533586d77aa8c21393e | NC_000017.10: g.41244000T>C | • Familial cancer of breast  
• Breast-ovarian cancer, familial 1  
• BRCA1 and BRCA2 hereditary breast and ovarian cancer  
• Hereditary cancer-predisposing syndrome | Benign |
| 00.2c5.04cb, e318e2ddb54bb3d6cc3a5ca6749482c5 | NC_000017.10: g.41246532dupT | • Familial cancer of breast  
• Breast-ovarian cancer, familial 1  
• BRCA1 and BRCA2 hereditary breast and ovarian cancer | Pathogenic |
We can already encode 680 ethnically diverse genomes in less than 84 MB per genome, and we estimate without requiring conversion to VCF-like representations. We demonstrated tiling supports common tools VCF-like genome representations, including GVF, VCF, and gVCF, by precomputing an inexpensive with 256 GiB of RAM, we can store 1 million genomes in RAM if we use less than 84 MB per genome. (Collins and Varmus, 2015). A genome tiling can naturally support these use cases, since it can be simply (RAM), extremely fast queries can be achieved. Given ten racks of standard hardware with 32 nodes, each with 256 GiB of RAM, we can store 1 million genomes in RAM if we use less than 84 MB per genome. We can already encode 680 ethnically diverse genomes in less than 84 MB per genome, and we estimate tiled genomes to be compressible to a few megabytes. A typical node costs about $8000, has a three year lifetime, and requires power, cooling, space and maintenance. Consequently, this encoding enables a fully amortized operating cost, supporting a variety of analyses and queries, of just $2.4 per genome per year for a million genome data warehouse. This cost is 10 times better than reported previously (Haussler et al., 2012) and does not take into account another order of magnitude in potential improvement with better tiling encodings (Text S2).

Public domain data is ideally suited to demonstrate tiling because it facilitates openness and collaboration (Buck, 2015; McNutt, 2014; Angrist and Cook-Deegan, 2014; GA4GH, 2015). A genome tiling can naturally support these use cases, since it can be simply converted to a graphical genome representation. To demonstrate that tiling is compatible with graphical encoding, we encoded the BRCA1 and BRCA2 region tilings for the PGP genomes into the preliminary GA4GH graphical database specification required for the HGVR pilot project.

Most clinicians and researchers rely on linear genome abstractions to hold annotation information. Unfortunately, reasoning about graphical coordinate systems, annotations on these coordinates, and annotations on specific sequences will require considerable deliberation by the clinical community. The simplest solution, and one that will be useful for the adoption of a new genome representation, is the ability to convert tiling into VCF-like genome representations, which we implemented for tiling. Converting a tile to a VCF-like genome representation can be done by aligning a tile variant sequence against the tile variant sequence of the desired reference. Both of these sequences are about 250 bases long, have the same 24 starting bases, and have the same 24 ending bases. Additionally, we can annotate specific tile variants to indicate they are a reference sequence. Therefore, we can obtain multiple standardized VCF-like genome representations, including GVF, VCF, and gVCF, by precomputing an inexpensive global alignment on a tile variant sequence against the tile variant associated with GRCh38 or GRCh37 at the same tile position.

Tiling’s use of tile positions enables many common use cases, such as annotation and querying, without requiring conversion to VCF-like representations. We demonstrated tiling supports common tools by annotating the BRCA1 and BRCA2 region tilings for 680 genomes with CAVA, an open-source variant...
annotated, and with information aggregated from ExAC and ClinVar. Since tile positions and tile variants are annotated at a population level via the tile library, these annotations apply to all 680 called genomes identically, removing the possibility of conflicting annotations within the same population as a source of ambiguity and error.

We demonstrate tiling enables machine learning by replicating the 1000 Genomes principal component results (1000GenomesProjectConsortium2012), and extending those results to the PGP data. Interestingly, we found the PGP self-reported ethnicities cluster consistently with the 1000 Genomes population ethnicity codes. This clustering demonstrates that self-reported data can be of sufficient quality for at least some machine learning applications in genomics. Since VCF-like and graph abstractions of whole genome sequences can represent one sequence in multiple ways (Danecek et al., 2011; Dilthey et al., 2015), data gathered from multiple studies using multiple tools can be hard to combine without ambiguity. This ambiguity further reduces the power of machine learning, requiring more samples to discover an effect. Tiling’s simplicity, naming consistency, and native matrix representation is extremely advantageous for machine learning applications.

As an example of tiling’s use in supervised machine learning, we built an ABO classifier on the PGP tiled genomes. Currently, in silico ABO blood type classifiers use known variants associated with ABO blood type phenotype to build their predictions. BOOGIE, a predictor using SNV databases, predicts the ABO blood type group with 94.2% accuracy for well sequenced whole genome sequences (Giollo et al., 2015). Our A antigen classifier was given 75 labeled whole genomes with no prior knowledge, and it selected, out of over 2 million tile positions across the genome, a position in the ABO gene. This classifier provides an accuracy of 93%, about one percent less than the BOOGIE ABO classifier, which relies on SNP-blood type databases. Our B antigen classifier did not have enough participants with the B antigen phenotype to accurately train the classifier.

We believe that adding feature selection, along with a larger and more varied training set, will increase the accuracies of our ABO blood type classifiers, since our current classifiers have a very large discrepancy between the number of features and the number of training sets. The Harvard PGP is currently in the process of releasing a blood type survey to participants and making their responses publicly available. We plan to incorporate these responses and grow our training set. Additionally, developing a less exclusive mechanism that incorporates poorly sequenced regions might allow the classifiers access to the underlying variants producing the ABO phenotype, which might also increase the accuracies of our classifiers. Finally, including known phenotypes, such as ethnicity, could strengthen the predictive accuracies of our classifiers.

We identified previously undetected, pathogenic BRCA mutations in 6 persons in the PGP. Using tiling, we provided simple, consistent identifiers of these variations, which we converted into human-friendly VCF-like representations and HGVS identifiers. We submitted the previously new variations to ClinVar, and added relevant clinical evidence to GET-Evidence, enabling the PGP to inform these 6 participants. We note that 6 participants with pathogenic BRCA mutations out of a total of 178 is a higher number of pathogenic mutations than the expected 2.67 participants in randomly selected, non-Jewish, healthy population (Frank et al., 2002). We hypothesize this high incidence rate results from the self-selection process of joining the PGP. One might think the PGP is enriched for the worried well; however, the high number of BRCA frameshifts found here might indicate participants are indeed ill.

Of the six PGP participants with BRCA frameshift mutations, three are male and three are female. The oldest female participant, hu92C40, reports having breast cancer at the age of 66. The second oldest female participant, huffB09D, does not report breast cancer. Participant huffB09D is also heterozygous for CASP10-V410I, which is reported to have a dominant, likely protective effect on breast cancer (Frank et al., 2006), but is reported to not protect against BRCA frameshift mutations (Engel et al., 2010). The youngest female participant, huCD380F, who is 33 years old, also does not report breast cancer. This 2:1 ratio of healthy to cancerous patients can be expected, given the ages of the two healthy participants (Levy-Lahad and Friedman, 2007), and the fact that one also has a possibly protective variant. Of the three males, two report no cancers, which also can be expected, given that BRCA variants have a greater increase of cancer risk in females than males (Levy-Lahad and Friedman, 2007). The other participant, huD3A569, reports multiple occurrences of skin cancer before the age of 40 and reports his mother, at age 72, has had over 20 basal cell carcinomas and squamous cell carcinomas on her head and face. He is a carrier of a BRCA2 variant, which is reported to strongly increase the likelihood of skin cancers (Levy-Lahad and Friedman, 2007; Ginsburg et al., 2010).
The two Fisher’s exact tests we performed examined the correlation between breast cancer and having a BRCA frameshift variant (Table S2) and the correlation between Hashimoto’s thyroiditis and having a BRCA frameshift variant (Table S1). We did not observe a significant correlation between breast cancer and having a BRCA frameshift variant. Given the small size of our cohort (Table S1), the young age of the participants with the BRCA frameshift variants, and the number of factors contributing to breast cancer, we are not surprised by these results. We estimate 131 females need to join the PGP to achieve a power of 95% with Fisher’s exact test (current power is 48%). We observed a significantly increased likelihood of presenting with Hashimoto’s thyroiditis if the patient has a BRCA frameshift variant. This correlation has both supporting (Chen et al., 2013) and contradicting (Sarlis et al., 2002) evidence, and we suggest further investigation into the effect BRCA frameshift variants have on Hashimoto’s thyroiditis and possibly other diseases.

Tiling, by using sequence hashes, provides a consistent and unambiguous representation of genomic sequences. Because tiling uses the hash of the sequence to refer to the sequence, tiling reduces variant naming ambiguities. A tile variant’s sequence allows anyone to generate a simple and identical identifier of that sequence, which can then be associated with more human-friendly names provided by HGVS or other standards bodies. While a VCF-like representation (Danecek et al., 2011) can give the same underlying sequence a variety of names, tiling vastly simplifies information aggregation and collaboration by using a tag set to give each sequence exactly one name.

Finally, tiling is, to our knowledge, the first decomposition of a genome into smaller pieces that can be easily verified by molecular biology techniques. Exome capture (Hodges et al., 2007) and tiling arrays (Mockler and Ecker, 2005) have already demonstrated the promise of using tag-like sequences in genomics. Our 24-base tags were inspired by PCR primers. Therefore, we postulate that in the process of designing an optimal tag set, we may consider the utility of these tags as PCR primers for high-throughput amplification and sequencing. Thus, one tag set may represent a called genome and be used to biologically query a sequence.

We have demonstrated tiling can support clinical screening, supervised machine learning, and unsupervised machine learning on the whole genome. Furthermore, this approach is scalable to millions of individual complex data-sets and can, thus, enable population-wide precision medicine and genomic screening. We hope our open approach to tiling, in addition to collaborations with the GA4GH and the wider genomics community, will ensure tiling addresses the problems faced by the scientific and medical communities. We invite all interested individuals and groups to contribute to the open-source project at http://github.com/curoverse/lightning

**METHODS**

**Tiling**

We tiled 680 called, whole genome sequences, aligned to GRCh37. 178 of these called genomes were tiled using the Harvard Personal Genome Project (PGP) CGI-var files (obtained at http://curover.se/su92l-j7d0g-nf54gdds53j03tc via Arvados). 69 of these genomes were tiled using the 1000 Genomes Project CGI-var files (obtained from Complete Genomics at ftp://ftp2.completegenomics.com/vcf_files/Build37\_2.0.0/), and 433 were tiled using the 1000 Genomes Project CGI-var files (obtained from NCBI at ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/data/). In total, 502 called genomes are from the 1000 Genomes and 178 called genomes are from the Harvard Personal Genome Project (PGP). All CGI-var files were converted to GFF (general feature format) files, which we tiled as described in Text S1. We preserved the phase called by Complete Genomics, found in each GFF file. Further details, source code for tiling, and source code for other custom functions can be found at https://github.com/curoverse/lightning and http://curover.se/su92l-j7d0g-swtofxa2rct8495.

**Pythonic Tiling Representation for Machine Learning**

For use in supervised and unsupervised machine learning, we converted the 680 tiled called genomes into python numpy arrays: with one row per called genome in the population and one column per tile position. To preserve space, we hashed each tile variant value with a non-negative integer for each tile position. All phase A variant calls were placed on the left half of the array, and phase B variant calls were placed on the right half. Tile variants that spanned multiple tile positions (described in Text S1) were assigned a non-negative integer greater than their position to ensure tiles spanning into a different position would not
We used the definition of a clearly pathogenic variant used by Rebbeck et al. (2015). Each tile variant falling in these regions was aligned back to GRCh37 and converted to gVCF, then annotated using CAVA (Clinical Annotation of Variants) (Muenz et al., 2015), an open-source variant annotation tool, downloaded from the Oxford Genomics Centre on March 30, 2015. The default impact definitions, ensembl file, dbsnp, and GRCh37 reference provided by CAVA were used. Additional annotations from ExAC and ClinVar were imported using custom-built open-source software that can be accessed at https://github.com/curoverse/lightning. When called genomes from the PGP were found to contain a pathogenic variant of high quality, the variant existence was first confirmed by converting the called genome to VCF using cgatools and comparing them with ExAC and ClinVar using bcftools isec (Details in Text S6). Then, if this variant was not already reported by GET-Evidence (Genome-Environment- Trait Evidence) (Ball et al., 2012), the variant was added and annotated, the relevant PGP reports were updated, and PGP participants were informed. Additionally, we performed Fisher’s exact tests for any phenotype reported by two or more participants with pathogenic BRCA1/2

Unsupervised Learning - Principal Component Analysis
To avoid capturing the noise generated by sequencing artifacts, we ran PCA only on tile positions where all 680 called genomes were well sequenced. Spanning tiles (described in Text S1) that contained a no call removed all tile positions it spans from being well-sequenced. The numpy arrays were then encoded using a one-hot encoding scheme (Snoek et al., 2012). The resulting matrix was used as input for principal component analysis, implemented by scikit-learn (Pedregosa et al., 2011), and projected onto the first two principal components. The 502 callsets from the 1000 Genomes Project were colored based on their assigned population ethnicities (obtained from the 1000 Genomes Project at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130502_1000genomes_phase1_release/technical_documentation/1000G_phase1_germline_sample_info.xlsx). The 178 PGP callsets were colored after visually examining their PCA projection, while crossValidating using their reported ethnicity and ancestry.

Supervised Learning - ABO Blood Type Classifiers
To avoid capturing the noise generated by sequencing artifacts, we removed tile positions with any poorly sequenced tile variants in any of the 178 genomes. Since each tile position is considered a feature, we removed all tile positions where the variant value was the same in all 178 genomes, since the genomic sequence is identical for those 178 genomes in that small region. To avoid training on random noise, we used the scikit-learn StandardScaler functionality to transform the data by centering all tile variant values around zero and dividing non-constant features by their standard deviation. We fit on (found the mean and standard deviation of) the training data and transformed both the training and unlabeled data. Blood types for the PGP participants were scraped from the Harvard PGP site on April 29, 2015 using scrapy. 75 PGP participants have self-reported blood types associated with their account; their whole genome sequence, including mitochondrial DNA and sex chromosomes, were used as training inputs to the classifiers.

Due to the low number of training callsets, we used leave-one-out cross-validation to determine the optimal classifier type and parameter values for the A-antigen and B-antigen classifiers. We tested different parameters for three types of classifiers implemented by scikit-learn (Pedregosa et al., 2011): SVC (linear and radial basis function kernels, varying C), NuSVC (linear and radial basis function kernels, varying nu), and LinearSVC (l1 or l2 penalty, varying C). SVC implements C-Support Vector Classification, where C is the penalty parameter of the error term. NuSVC implements Nu-Support Vector Classification, where nu is an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors. LinearSVC implements C-Support Vector Classification with a linear kernel, where C is the penalty parameter of the error, and allows the user to specify the norm used in penalization. Using the classifier and the parameter which provided the highest accuracy (as measured by leave-one-out cross-validation), we predicted the blood type for the remaining unlabeled 103 PGP callsets.

Personalized Medicine and Clinical Analysis of the BRCA region
We chose to examine the BRCA regions for possible pathogenic variants in our 680 called genomes. We used the definition of a clearly pathogenic BRCA1/2 mutations used by Rebbeck et al. (2015). Each tile variant falling in these regions was aligned back to GRCh37 and converted to gVCF, then annotated using CAVA (Clinical Annotation of Variants) (Muenz et al., 2015), an open-source variant annotation tool, downloaded from the Oxford Genomics Centre on March 30, 2015. The default impact definitions, ensembl file, dbsnp, and GRCh37 reference provided by CAVA were used. Additional annotations from ExAC and ClinVar were imported using custom-built open-source software that can be accessed at https://github.com/curoverse/lightning. When called genomes from the PGP were found to contain a pathogenic variant of high quality, the variant existence was first confirmed by converting the called genome to VCF using cgatools and comparing them with ExAC and ClinVar using bcftools isec (Details in Text S6). Then, if this variant was not already reported by GET-Evidence (Genome-Environment-Trait Evidence) (Ball et al., 2012), the variant was added and annotated, the relevant PGP reports were updated, and PGP participants were informed. Additionally, we performed Fisher’s exact tests for any phenotype reported by two or more participants with pathogenic
mutations in BRCA1/2 and for any phenotype associated with a ClinVar condition for a pathogenic variant, with the exception of hereditary cancer-predisposing syndrome. We were unable to test for hereditary cancer-predisposing syndrome given the little familial information in the PGP dataset.

ACKNOWLEDGMENTS

We thank all members of the Personal Genome Project, Berkeley AMPLab, Global Alliance for Genomics and Health (GA4GH), and Curoverse communities for their help and advice. In particular Madeleine Price Ball, Pete Estep, Frank Nothaft, Dave Patterson, Deanna Church and Heidi Rehm provided invaluable comments and discussion that improved the manuscript.

REFERENCES


support vector machines for disease risk prediction in genome-wide association studies: Concerns and opportunities. *Hum Mutat*, 33:1708–1718.


