Selecting representative subsets of genomic loci

Galip Gürkan Yardımcı[1], Jacob Schreiber[2], Jeff Bilmes[3], and William Stafford Noble[1,2]

[1]Department of Genome Sciences, University of Washington
[2]Paul G. Allen School of Science and Engineering, University of Washington
[3]Department of Electrical Engineering, University of Washington

1 Introduction

The human genome is long, and many genomic assays are too expensive to perform in a genome wide manner. Consequently, we aim to develop computational methods for selecting a representative set of genomic loci, i.e., a compact set of loci that represent the full diversity of the properties of the genome. This necessarily means that the selected are not redundant, for if they were redundant, they would not be an efficient representation. Such a representative set of genomic loci will be useful, for example, in benchmarking or systematically applying an assay for measuring enhancer activity, such as STARR-seq [1], CREST-seq [3], or crisprQTL [4]. Rather than testing the assay on a subset of loci selected at random, we propose that testing on only a representative set provides a better estimate of the assay's performance on the genome as a whole.

The scale at which we perform locus selection depends upon properties of the assay we are working with. Some methods, such as STARR-seq, can be targeted to small (∼1 kb-sized) loci selected anywhere in the genome, whereas other methods, such as CREST-seq, must be applied in a "tiling" fashion to ∼1 kb-sized loci within a specified region or small set of regions. Accordingly, we consider two variants of the set selection task: choosing loci of size ∼1 kb for a targeted assay or choosing loci of size ∼1 Mb for a tiling assay.

To address this selection problem, we use a class of methods based on submodular function maximization. A set function is submodular if it has a diminishing returns property; i.e., the incremental function value's gain of adding an element s to a set A becomes smaller as the size of the set A becomes larger. More specifically, given a finite set $S = \{s_1, s_2, ..., s_n\}$, a discrete set function $f : 2^S \to \mathbb{R}$ is submodular whenever

$$f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B), \;\; \forall A \subseteq B \subset S, s \notin B.$$

Submodularity is sometimes said to be a discrete counterpart to convexity. Although maximization of a submodular function, given a cardinality constraint on the selected subset, is in general NP-hard, it has been shown that the greedy procedure will yield a subset whose objective value is within $1 - 1/e$ of the optimal subset, and that this is the best approximation one can make unless P=NP [10].

In this work, we optimize a *facility location* submodular objective function (defined in Methods). We begin with an initial set V of loci (the *ground set*) and search for a subset A that maximizes the facility location objective, subject to a cardinality constraint $|A| = c$. The facility location, function operates on pairwise similarities between pairs of loci and is a well-known submodular function. Due to the submodularity, there is a guarantee when optimizing it using the greedy procedure.

The similarities that we use for our facility location function are based on the epigenomic state of the genome at each locus. This state is defined using two different sets of features. The first set of features captures genome state in a cell type-agnostic manner and comes from a latent representation of the genome learned by an imputation approach. The second set of features are obtained from histone ChIP-seq experiments performed on K562 cells and captures the state of the epigenome in a cell type-specific manner.

We show that by selecting representative subsets from the full human genome at low resolution and from individual gene loci at high resolution, we can obtain non-redundant representative subsets that cover various types of elements across the genome. We offer qualitative evidence that our representative subsets achieve uniform coverage across the epigenomic landscape, and we measure the abundance of rare elements in our representative subsets against randomly selected subsets, showing that rare elements are better represented in our representative subset.

2 Methods

2.1 Representing the genome

The input to our selection task is a set V of n genomic loci. These loci cover the part of the genome we wish to summarize. We use two complementary sets of features to represent each locus. First, we use latent factors derived from Avocado [13], which is a deep tensor factorization model that embeds 1014 tracks of histone ChIP-seq and DNase-seq data into a 110-dimensional latent space at 25 bp resolution. The Avocado latent factors are defined at three different length scales; hence, each 25-bp locus is represented via 25 factors at 25-bp resolution, 40 factors at 250-bp resolution and 45 factors at 5-kbp resolution, for a total of 110 factors. Accordingly, a 1-kb locus corresponds to $25 \times (1000/25) = 1000$ 25-bp factors, $40 \times (1000/250) = 160$ 250-bp factors, and 45 5-kbp factors, for a total of 1205 factors. Similarly, a 1-Mb locus corresponds to $(1000000/25) \times 25 = 1000000$ 25-bp factors, $(1000000/250) \times 40 = 160000$ 250-bp factors, and $(1000000/5000) \times 45 = 9000$ 5-kbp factors, for a total of 1,169,000 factors. Note that, unlike in the original description of Avocado, we use latent factors trained with a non-negativity constraint similar to [6].

The Avocado latent factors are trained on many different cell types. In order to capture properties of the genome that are cell type-specific, we use a panel of ten histone modification ChIP-seq measurements that have been performed in K562 cells. These histone marks capture active, bivalent and repressive regions of the genome. The full set of marks we use are H2A.Z, H3K27ac, H3K27me3,H3K36me3,H3K4me1,H3K4me2, H3K4me3,H3K79me2, H3K9ac,H3K9me3. These tracks are obtained from the ENCODE consortium data portal [2]. There are multiple processed version of Histone ChIP-seq assays, and we opt to use signal p-value tracks.

2.2 Subset selection

From a given collection V of genomic loci, we aim to select a representative subset $A \subset V$ of a given size $|A| = k$. The facility location function is defined as

$$f(A) = \sum_{v \in V} \max_{a \in A} \phi(v, a), \tag{1}$$

where $\phi(\cdot, \cdot)$ is a non-negative similarity function. In this work, we use the squared Pearson correlation between v and a. We optimize the objective functions using the accelerated greedy algorithm [9] as implemented in apricot v0.3.0 [12].

2.3 Evaluation

In addition to the Avocado latent factors, we also characterize each genomic locus using the unsupervisedly-learnt dynamic Bayesian network, Segway [5]. In particular, we use the Segway annotations from a recently described analysis of 164 human cell types based on data from the ENCODE Consortium [7]. In this analysis, each cell type is annotated independently, using all available histone modification and transcription factor ChIP-seq, chromatin accessibility and replication timing data sets. For a cell type with M available data sets, the annotation contains $10 + 2\sqrt{M}$ state labels. Note that we use these “raw” labels, rather than the smaller set of automatically aggregated and interpreted labels described by Libbrecht *et al.*.

Additionally, we characterize representative and random subsets according to genome subcompartments as identified by the Hi-C assay [11]. Genomic subcompartments capture the relative positioning of large scale loci within the nucleus. However, Rao et al. have shown that individual subcompartments are enriched for different histone marks and other features of epigenomic state. Thus, the subcompartments offer an orthogonal measurement of epigenomic state, different from the features we use for subset selection.

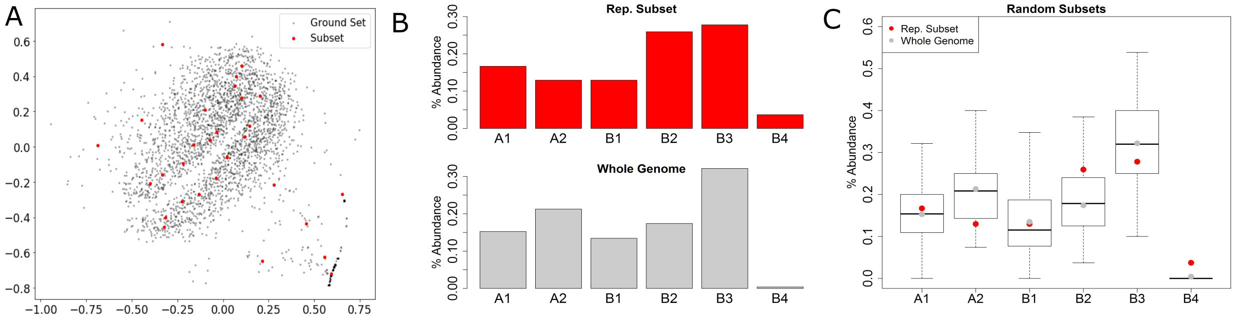

Figure 1: **A cell type-agnostic summary of the full genome.** A) Two-dimensional embedding of the ground set obtained from breaking the human genome into 1MB windows. The elements of the representative subset are shown as red dots. B) Abundance of each subcompartment across the human genome and the representative subset. C) Abundances of each subcompartment across 100 randomly selected hundred subsets, shown as a boxplot. Red dots indicate the corresponding abundance of that subcompartment in the representative subset. Gray dots indicate the abundance across the whole genome.

Results

A cell type-agnostic summary of the full genome

We performed cell type-agnostic selection of the human genome assembly hg19 at 1 megabase (Mb) resolution. The human genome was broken into 1 Mb wide non-overlapping windows resulting in approximately 3000 windows that make up the ground set V. Avocado latent factors were averaged across each window to summarize the cell type-agnostic epigenomic properties within each window.

Using the facility location function objective function with the Pearson correlation coefficients as a similarity measure, we selected a representative subset of 30 elements that covers 1% of the human genome. We visualized elements of the ground set and the representative subset in two dimensions via the UMAP dimensionality reduction method [8] (Fig. 1A). We observed that the elements of the representative subset are distributed evenly across elements of the ground set in this low-dimensional projection. We note that the representative subset covers rare elements that make up the lower right part of the scatter plot in Fig 1A.

To quantify the ability of the representative subset to cover rare elements, we used chromatin subcompartment annotations, which comprise an orthogonal characterization of epigenomic state from chromosome conformation capture assays. We found that the representative subset is covered more evenly by different subcompartments compared to the prevalence of each subcompartment across the full genome. More strikingly, we found that the rarest subcompartment B4 is more abundantly represented in our representative subset, validating the ability of the submodular selection approach to producing a better representation of the full genome (Fig 1B). Lastly, we computed the coverage of each subcompartment in 100 randomly selected subsets, and we found that the randomly selected subsets do not exhibit the properties of the representative subset outlined above (Fig 1C).

A cell specific summary of gene loci

Next, we performed a high resolution summary of the genome centered around gene loci in a cell type-specific fashion. We broke five different 2Mb genomic loci centered around five genes (BCL11A, HBE1, LMO2, MYC and RBM38) into 8000 250 bp non-overlapping windows, yielding five different ground sets. For each gene locus, we selected representative subsets that cover 1% of each ground set. We used signal from 10 different histone ChIP-seq experiments from the K562 cell line to capture the epigenomic state of 250 bp elements. Facility location was the objective function, using the Pearson correlation coefficient as a similarity measure.

We visualized the ground set embedded into two dimensions and colored each element of the set based on Segway annotations. This embedding reveals that elements that shared annotations largely cluster together, validating the ability of our features to capture the epigenomic state (Fig 2A). Next, we visualized the

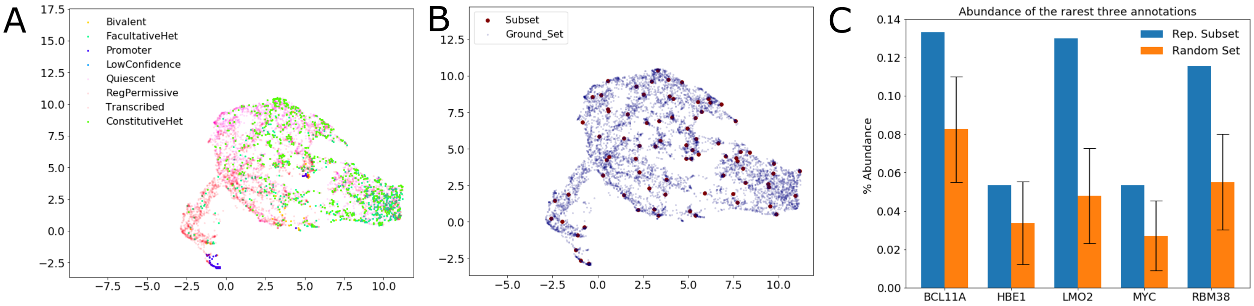

Figure 2: **A cell specific summary of gene loci.** A) Two-dimensional embedding of 250 bp windows tiled across the 2Mb locus surrounding the BCL11A gene. Each 250 bp element is colored according to Segway annotations. B) Representative subset selected from set of 250 bp windows across the same BCL11A gene locus C) Abundance of the three rare annotations types in the representative subset, shown with a blue bar, and the abundances of the same three annotations in 100 randomly selected subsets, shown with orange bars with error bars indicating the variation in the randomly selected subsets.

elements of the representative subset on the same embedding and again qualitatively confirmed our method's ability to achieve a better representation of different elements across the gene locus.

Next, we tested our method's ability to select Segway annotations that are underrepresented in each gene locus. For each gene locus, we computed the prevalence of the rarest three Segway annotations in the representative subset. The rarest three Segway annotations make up 3–8% of each gene locus. We repeated this calculation for 100 randomly selected subsets. We found that rare elements are better represented in the representative subset compared to randomly selected subsets, confirming our method's ability to select elements that capture rare but diverse genomic loci from a wide selection of different types of elements.

References

[1] Cosmas D Arnold, Daniel Gerlach, Christoph Stelzer, Łukasz M Boryń, Martina Rath, and Alexander Stark. Genome-wide quantitative enhancer activity maps identified by starr-seq. *Science*, 339(6123):1074–1077, 2013.

[2] Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Kriti Jain, Ulugbek K Baymuradov, Aditi K Narayanan, et al. The encyclopedia of dna elements (encode): data portal update. *Nucleic acids research*, 46(D1):D794–D801, 2017.

[3] Yarui Diao, Rongxin Fang, Bin Li, Zhipeng Meng, Juntao Yu, Yunjiang Qiu, Kimberly C Lin, Hui Huang, Tristin Liu, Ryan J Marina, et al. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nature methods*, 14(6):629, 2017.

[4] M. Gasperini, A. J. Hill, J. L. McFaline-Figueroa, B. Martin, S. Kim, D. Jackson, A. Leith, J. Schreiber, W. S. Noble, C. Trapnell, N. Ahituv, and J. Shendure. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell*, 176:377–390, 2019.

[5] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–476, 2012.

[6] Chandrashekhar Lavania and Jeff Bilmes. Auto-summarization: A step towards unsupervised learning of a submodular mixture. In *SIAM International Conference on Data Mining (SDM-2019)*, May 2019.

[7] M. W. Libbrecht, O. Rodriguez, Z. Weng, M. Hoffman, J. A. Bilmes, and W. S. Noble. A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types (preprint in advance of publication). *bioRxiv*, 2016.

[8] L. McInnes and J. Healy. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 2018.

[9] M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. *Optimization Techniques*, pages 234–243, 1978.

[10] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

[11] S. S. P. Rao, M. H. Huntley, N. Durand, C. Neva, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 59(7):1665–1680, 2014.

[12] J. M. Schreiber, J. Bilmes, and W. S. Noble. apricot: Submodular selection for data summarization in python. *arXiv*, 2019. https://arxiv.org/abs/1906.03543.

[13] J. M. Schreiber, T. J. Durham, J. Bilmes, and W. S. Noble. Multi-scale deep tensor factorization learns a latent representation of the human epigenome. *bioRxiv*, 2018. https://www.biorxiv.org/content/early/2018/07/08/364976.