Supplementary Material

Dynamic compression schemes for graph coloring

Harun Mustafa,[1,2,3,‡] Ingo Schilken,[1,‡] Mikhail Karasikov,[1,2,3] Carsten Eickhoff,[4,*] Gunnar Rätsch,[1,2,3,*] and André Kahles[1,2,3,*]

[1]Department of Computer Science, ETH Zurich, Zurich, Switzerland
[2]University Hospital Zurich, Biomedical Informatics Research, Zurich, Switzerland
[3]SIB Swiss Institute of Bioinformatics, Zurich, Switzerland
[4]Brown Center for Biomedical Informatics, Brown University, Providence, RI, USA
[‡]Equal contribution. [*]To whom correspondence should be addressed.

A Parallel wavelet trie construction

For this method, we define the *descendants* function $\mathbf{D} : \{1, \ldots, |V|\} \to 2^{\{1,\ldots,|V|\}}$ for the wavelet trie $T = (V, E)$ with nodes $V = \{(\alpha_j, \beta_j)\}_{j=1}^{n}$ by the recurrence

$$j \in \mathbf{D}(j),\ \forall j \in \{1, \ldots, n\},$$
$$k \in \mathbf{D}(j) \text{ and } \beta_k \neq \varepsilon \Rightarrow \{2k, 2k+1\} \subset \mathbf{D}(j).$$

The three steps in the merging operations are as follows:

A.1 Align

Given nodes (α'_j, β'_j) and (α''_j, β''_j), we compute their longest common prefix

$$\hat{\alpha} \leftarrow \mathbf{LCP}(\{\alpha'_j, \alpha''_j\}).$$

If $\alpha_j \neq \alpha'_j$, we let

$$\hat{\beta}'_j \leftarrow \underbrace{\alpha'_j[|\alpha_j| + 1] \cdots \alpha'_j[|\alpha_j| + 1]}_{|\beta'_j|}$$

and update the indices in T' by applying the transformation $j \leftarrow 2j + \alpha'_j[|\alpha_j| + 1]$ and updating all nodes $k \in \mathbf{D}(j)$ accordingly. We then let $\alpha'_j \leftarrow \alpha_j$ and $\beta'_j \leftarrow \hat{\beta}'_j$ and truncate the prefix in the newly created child nodes,

$$\alpha'_{2j} \leftarrow \alpha'_{2j}[|\alpha_j| + 2 :] \quad \text{if } \hat{\beta}'_j[1] = 0,$$
$$\alpha'_{2j+1} \leftarrow \alpha'_{2j+1}[|\alpha_j| + 2 :] \quad \text{if } \hat{\beta}'_j[1] = 1.$$

If $\alpha_j \neq \alpha''_j$, the second trie is processed accordingly.

A.2 Merge

If $\mathrm{rank}_1(\beta'_j, |\beta'_j|) = 0$ and $\mathrm{rank}_1(\beta''_j, |\beta''_j|) = 0$, then terminate. Otherwise, merge the two assignment vectors

$$\beta_j \leftarrow \beta'_j \beta''_j$$

A.3 Repeat

The merging algorithm is then performed on nodes n_{2j} and n_{2j+1} depth-first to continue the recursion.

If two wavelet tries constructed from bit vectors of different lengths are merged, this merging algorithm leads to the decoding of bit vectors with trailing zeros. Since we intend to use these vectors for representing the membership of the k-mers to various metadata, the presence of extra trailing zeros in the decoded bit vector does not represent false information.

B Wavelet trie updating

Internal insertion of rows is done similarly to parallel construction. The alignment step is identical, while merging proceeds by internal insertion of β''_j into β'_j instead of concatenation. When traversing down the tree, the index i_j at which β''_j is inserted is updated using $i_j \leftarrow \mathrm{rank}_1(\beta'_j, i_j)$ if traversing to the right and $i_j \leftarrow \mathrm{rank}_0(\beta'_j, i_j)$ if traversing to the left.

As an optimization for updating bits in a computed wavelet trie operates in three steps. Suppose we are given a set of entries $\{\mathcal{A}^{i_1}_{j_1}, \ldots, \mathcal{A}^{i_\ell}_{j_\ell}\}$ for the wavelet trie compressing $\mathcal{A}$ of size $n \times m$. Then

1. **Insert** Query the rows $\{\mathcal{A}^{i_1}, \ldots, \mathcal{A}^{i_\ell}\}$ and flip the appropriate bits in their decompressed forms. Concatenate these rows to produce a new matrix $\mathcal{A}'$ of size $(n + \ell) \times m$.
2. **Swap** For each i_k, swap the rows $(\mathcal{A}'^{i_k}, \mathcal{A}'^{n+k})$ by swapping bits in the βs.
3. **Remove** Remove the rows $\{\mathcal{A}'^{n}, \ldots, \mathcal{A}'^{n+\ell}\}$ to form the final matrix $\mathcal{A}''$ of size $n \times m$.

C Data used for evaluation

The datasets used to evaluate the performance of our compression schemes originate either from viruses (*Virus100–Virus50000*), bacteria (*Lactobacillus*) or human (*chr22+gnomAD* and *hg19+gnomAD*) and are chosen to test the methods on different color distributions, annotation matrix sizes and densities. They further reflect varying graph topologies and allow us to study the effect of topology-informed compression in a robust testing bed. We construct de Bruijn graphs of order $k = 63$ for each dataset and compare the compression performance of all methods by measuring the *compression ratio* for each dataset, defined as the ratio of the number of bits in an annotation matrix and the number of bits in its respective compressed representation.

C.1 Virus Datasets

The virus datasets are generated from publicly available GenBank (Clark *et al.*, 2016) complete genome data. The *Virus50000* dataset consists of the set of 53,412 virus strains present in GenBank on September 30th, 2016. The *Virus1000*, *Virus3000* and *Virus20000* datasets are randomly selected subsets of the *Virus50000* set. The *Virus100* dataset is a subset consisting of the first 100 genomes in *Virus1000*. On these datasets, the colors are defined to represent the membership to each genome ID, while the class indicator bits are defined by each genome's taxonomic genera. The rows of these annotation matrices are very sparse and present degenerate cases with respect to compressability by wavelet tries.

C.1.1 Virus50000

This dataset consists of 53,412 viral genome sequences downloaded from GenBank via the eUtils API on 09/30/2016. The search term applied was `txid10239 [orgn] AND "complete genome"`

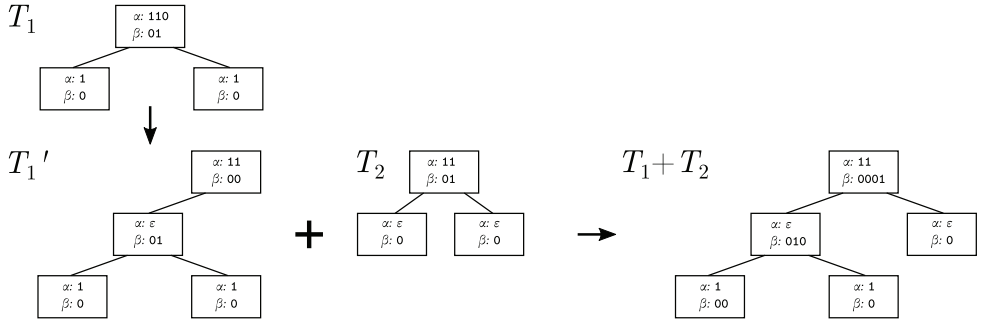


Fig. S-1: **Merging of wavelet tries T_1 and T_2 to form the wavelet trie $T_1 + T_2$.** Starting from the root node, the common prefix of the two α vectors is found and new β vectors are computed from their respective next significant bit after the common prefix. These become new parent nodes and the initial nodes' α vectors are updated to their respective remainders after removing the common prefix (e.g., the conversion from from T_1 to T_1'). When the two αs are equal, their respective βs are concatenated and the merging function is applied to their children. When a leaf is reached in one tree, but the equivalent node in the other tree is internal (e.g., the left child of the root in $T_1 + T_2$), the leaf is merged by appending or prepending additional zeros to the β vectors of all left ancestors. Note that extra leaf nodes producing trailing zeros in the decoded bit vectors are added during the merging process. See Section 2.3.1 for more details.

`[Title] NOT txid131567 [orgn]`. The data represents a set of sequences with both high variability to each other but also good pairwise conservation for sub-species. It reflects a wide range of variability and provides a good testing bed for the application within a colored de Bruijn graph setting. The graph corresponding to this dataset contains 622,587,315 nodes, 625,110,390 edges, and 1,359,843 unique edge colorings.

C.1.2 Virus1000, Virus3000, Virus20000

The *Virus1000* dataset is composed of 1000 virus genomes randomly selected from the *Virus50000* set, meant to study a graph whose topology is a series of almost mutually-exclusive loops with slight variation. The columns in this graph's annotation matrix indicate presence of edges in each of the virus genomes. Similar to the *Lactobacillus* dataset, the viruses were grouped by the first word of their names and the first species in each group was assigned as a backbone path. The resulting annotation bit matrix is very sparse and adjacent rows are either almost identical or almost mutually exclusive. This graph contains 30,310,634 unique k-mers, 30,347,373 edges, and 11,612 unique edge colorings.

The *Virus3000* and *Virus20000* datasets are supersets of *Virus1000*. The *Virus3000* graph contains 82,418,835 unique k-mers, 82,579,519 edges, and 52,187 unique edge colorings. The *Virus20000* graph contains 357,552,076 unique k-mers, 358,683,520 edges, and 537,344 unique edge colorings.

C.1.3 Virus100

This is a subset of the Virus1000 set containing only 100 virus strains used to facilitate the permutation tests in Section 3.1.1. This graph contains 2,954,719 unique k-mers, 2,956,113 edges, and 463 unique edge colorings.

C.1.4 Compressor update virus datasets

These datasets consist in subsets of the *Virus50000* dataset of varying sizes, but fixed number of columns. Given a number of columns C, 50 different contiguous subsets of C viruses are selected uniformly (i.e., a sliding window). Graphs and compressed annotations are then constructed for these sets and a randomly-selected virus genome not in any of the windows (`gi|1000033404|gb|KP975430.1|`) was selected to update the graph. Values of $C \in \{200, 500, 2000\}$ were used.

C.2 *Details for Lactobacillus*

This dataset is composed of 135 strains of bacteria in GenBank Clark *et al.* (2016) from the *Lactobacillus* genus, leading to a linear topology in the graphs with many shorter paths (*bubbles*) diverging from and reconnecting to the main reference genome paths. The colors are defined to represent the membership to genome IDs, while the class indicator bits are defined by each genome's species. The columns in this graph's annotation matrix indicate presence of an edge in each of the strains. Because of the low variability in the input sequences, they are represented as a graph with a predominantly linear topology and short variant paths (called *bubbles*). One genome from each of the species was chosen as a backbone path. The resulting graph had 134,951,429 unique k-mers, 135,369,397 edges, and 6,630 unique edge colorings (i.e., color combinations). See Supplementary Section D for a list of the bacterial strains used.

C.3 Human Data

For the human datasets, the *hg19* reference assembly is used as the main reference backbone, together with exome variants from the gnomAD dataset (Consortium *et al.*, 2016). The *hg19+gnomAD* dataset consists of the human autosomal portion of this data and *chr22+gnomAD* dataset consists of chromosome 22 only. On these datasets, the colors are defined to represent the membership to the reference chromosomes and the ethnic groups present in the gnomAD data. The colors corresponding to reference chromosomes are designated as the class indicators without adding additional columns (i.e., the sequence variant edges are additionally colored by their corresponding reference chromosome colors).

C.3.1 chr22+gnomAD

This graph consists of chromosome 22 from the `hg19` assembly of the human reference genome as the main reference backbone. To provide genetic variability, the set of exome variants from the gnomAD dataset were incorporated into the graph Lek *et al.* (2016). This larger dataset is meant to analyze the properties of the trie when the underlying graph is large, but with little variability. The columns in this graph's annotation matrix are defined to represent the membership of its edges to 9 ethnic groups defined in the dataset. The first column in the matrix is used to indicate edges which are present in the reference genome and serves as the backbone bit. The graph contains 178,196,890 unique k-mers, 180,023,641 edges, and 510 unique edge colorings.

C.3.2 hg19+gnomAD

This graph was constructed from the same datasets as the one described above, using data from the full human autosome. The same definition is used for the annotation matrix columns, with 9 columns being used to indicate edges observed in the defined ethnic groups and 22 prefix columns being used to indicate presence in the first 22 reference chromosomes as the backbone bits. This graph's topology was designed to be analogous to the Virus1000 dataset, but with $1000\times$ the number of rows and one-tenth of the number of annotation columns. It contains 5,714,136,751 unique k-mers, 5,728,489,633 edges, and 380,051 unique edge colorings.

D List of bacterial strains used

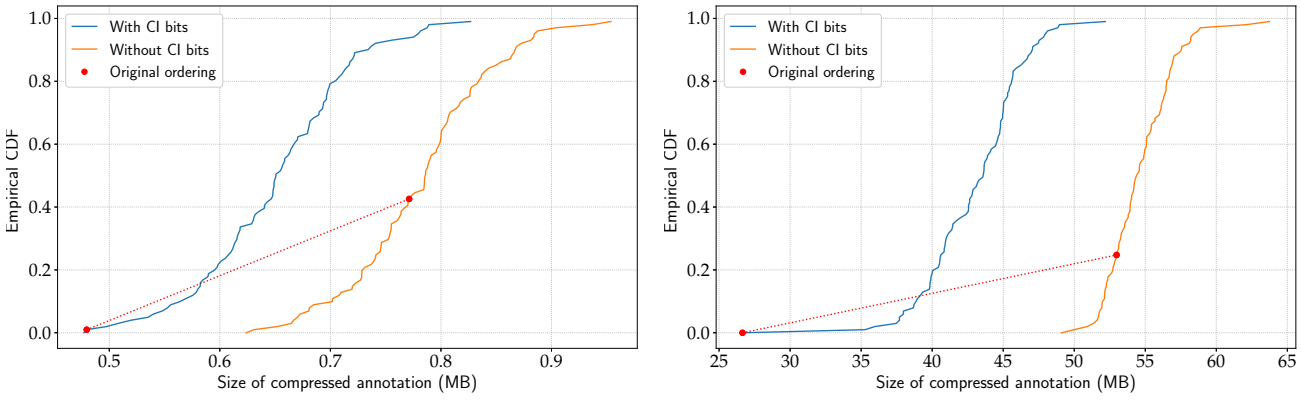
- *Lactobacillus acidophilus*
  - 30SC chromosome, complete genome
  - 30SC plasmid pRKC30SC2, complete sequence
  - La-14, complete genome
  - NCFM chromosome, complete genome
- *Lactobacillus amylovorus*
  - GRL 1112 chromosome, complete genome
  - GRL 1112 plasmid1, complete sequence
  - GRL 1112 plasmid2, complete sequence
  - GRL1118 chromosome, complete genome
  - GRL1118 plasmid1, complete sequence
  - GRL1118 plasmid2, complete sequence
- *Lactobacillus brevis*
  - ATCC 367, complete genome
  - ATCC 367 plasmid 1, complete sequence
  - ATCC 367 plasmid 2, complete sequence
  - KB290 DNA, complete genome
  - KB290 plasmid pKB290-1 DNA, complete genome
  - KB290 plasmid pKB290-2 DNA, complete genome
  - KB290 plasmid pKB290-4 DNA, complete genome
  - KB290 plasmid pKB290-5 DNA, complete genome
  - KB290 plasmid pKB290-7 DNA, complete genome
  - KB290 plasmid pKB290-9 DNA, complete genome
  - KB290 plasmid pKB290-3 DNA, complete genome
  - KB290 plasmid pKB290-6 DNA, complete genome
  - KB290 plasmid pKB290-8 DNA, complete genome
- *Lactobacillus buchneri*
  - NRRL B-30929 plasmid pLBUC01, complete sequence
  - NRRL B-30929 plasmid pLBUC03, complete sequence
  - NRRL B-30929 chromosome, complete genome
  - NRRL B-30929 plasmid pLBUC02, complete sequence
  - CD034 plasmid pCD034-2, complete sequence
  - CD034 plasmid pCD034-1, complete sequence
  - CD034 chromosome, complete genome
  - CD034 plasmid pCD034-3, complete sequence
- *Lactobacillus casei*
  - ATCC 334 plasmid 1, complete sequence
  - ATCC 334 chromosome, complete genome
  - BD-II chromosome, complete genome
  - BD-II plasmid pBD-II, complete sequence
  - BL23 chromosome, complete genome
  - LC2W chromosome, complete genome
  - LC2W plasmid pLC2W, complete sequence
  - LOCK919, complete genome
  - LOCK919 plasmid pLOCK919, complete sequence
  - W56, complete genome
  - W56 plasmid pW56, complete sequence
  - str. Zhang plasmid plca36, complete sequence
  - str. Zhang chromosome, complete genome
- *Lactobacillus crispatus* ST1, complete genome
- *Lactobacillus delbrueckii subsp. bulgaricus*
  - 2038 chromosome, complete genome
  - ATCC 11842 chromosome, complete genome
  - ATCC BAA-365 chromosome, complete genome
  - ND02 chromosome, complete genome
  - ND02 plasmid unnamed, complete sequence
- *Lactobacillus fermentum*
  - CECT 5716 chromosome, complete genome
  - F-6, complete genome
  - IFO 3956, complete genome
- *Lactobacillus gasseri* ATCC 33323 chromosome, complete genome
- *Lactobacillus helveticus*
  - CNRZ32, complete genome
  - DPC 4571, complete genome
  - H10 chromosome, complete genome
  - H10 plasmid pH10, complete sequence
  - R0052 chromosome, complete genome
- *Lactobacillus johnsonii*
  - DPC 6026 chromosome, complete genome
  - FI9785 plasmid p9785S, complete sequence
  - FI9785 chromosome, complete genome
  - FI9785 plasmid p9785L, complete sequence
  - N6.2, complete genome
  - NCC 533, complete genome
- *Lactobacillus kefiranofaciens*
  - ZW3 plasmid pWW1, complete sequence
  - ZW3 chromosome, complete genome
  - ZW3 plasmid pWW2, complete sequence
- *Lactobacillus paracasei subsp. paracasei*
  - 8700:2, complete genome
  - 8700:2 plasmid 1, complete sequence
  - 8700:2 plasmid 2, complete sequence
- *Lactobacillus plantarum*
  - 16, complete genome
  - 16 plasmid Lp16A, complete sequence
  - 16 plasmid Lp16C, complete sequence
  - 16 plasmid Lp16E, complete sequence
  - 16 plasmid Lp16F, complete sequence
  - 16 plasmid Lp16H, complete sequence
  - 16 plasmid Lp16L, complete sequence
  - 16 plasmid Lp16B, complete sequence
  - 16 plasmid Lp16D, complete sequence
  - 16 plasmid Lp16G, complete sequence
  - 16 plasmid Lp16I, complete sequence
  - JDM1, complete genome
  - subsp. plantarum P-8, complete genome
  - subsp. plantarum P-8 plasmid LBPp2, complete sequence
  - subsp. plantarum P-8 plasmid LBPp3, complete sequence
  - subsp. plantarum P-8 plasmid LBPp5, complete sequence
  - subsp. plantarum P-8 plasmid LBPp6, complete sequence
  - subsp. plantarum P-8 plasmid LBPp1, complete sequence
  - subsp. plantarum P-8 plasmid LBPp4, complete sequence
  - subsp. plantarum ST-III chromosome, complete genome
  - subsp. plantarum ST-III plasmid pST-III, complete sequence
  - WCFS1, complete genome
  - WCFS1 plasmid pWCFS101, complete sequence
  - WCFS1 plasmid pWCFS102, complete sequence
  - WCFS1 plasmid pWCFS103, complete sequence
  - ZJ316, complete genome
  - ZJ316 plasmid pLP-ZJ101, complete sequence

ZJ316 plasmid pLP-ZJ102, complete sequence
ZJ316 plasmid pLP-ZJ103, complete sequence
- *Lactobacillus reuteri*
  DSM 20016 chromosome, complete genome
  I5007, complete genome
  I5007 plasmid pLRI03, complete sequence
  I5007 plasmid pLRI02, complete sequence
  I5007 plasmid pLRI05, complete sequence
  I5007 plasmid pLRI06, complete sequence
  I5007 plasmid pLRI01, complete sequence
  I5007 plasmid pLRI04, complete sequence
  JCM 1112, complete genome
  SD2112 chromosome, complete genome
  SD2112 plasmid pLR585, complete sequence
  SD2112 plasmid pLR580, complete sequence
  SD2112 plasmid pLR581, complete sequence
  SD2112 plasmid pLR584, complete sequence
  TD1, complete genome
- *Lactobacillus rhamnosus*
  ATCC 8530 chromosome, complete genome
  GG, complete genome
  GG chromosome, complete genome
  Lc 705 chromosome, complete genome
  Lc 705 plasmid pLC1, complete sequence
  LOCK900, complete genome
  LOCK908, complete genome
- *Lactobacillus ruminis* ATCC 27782 chromosome, complete genome
- *Lactobacillus sakei* subsp. sakei 23K chromosome, complete genome
- *Lactobacillus salivarius*
  CECT 5713 plasmid pHN1, complete sequence
  CECT 5713 plasmid pHN2, complete sequence
  CECT 5713 chromosome, complete genome
  CECT 5713 plasmid pHN3, complete sequence
  UCC118 plasmid pSF118-20, complete sequence
  UCC118 plasmid pSF118-44, complete sequence
  UCC118 chromosome, complete genome
  UCC118 plasmid pMP118, complete sequence
- *Lactobacillus sanfranciscensis*
  TMW 1.1304 chromosome, complete genome
  TMW 1.1304 plasmid pLS1, complete sequence
  TMW 1.1304 plasmid pLS2, complete sequence

E List of viral strains used

See the GitHub repository for a list of viral strains used.

F Supplementary Figures and Tables



(a) *Virus100* **dataset** (b) *Lactobacillus* **dataset**

Fig. S-2: **Column-order effect on the Virus100 and Lactobacillus datasets** Shown are distributions of the file sizes of wavelet tries over 100 random permutations of the *Virus100* and *Lactobacillus* annotation matrix column orders. The red dashed lines link the points corresponding to the original column orderings in the annotation matrices. Adding class indicator (CI) bits as the matrix prefices in both datasets leads to decreases in the sizes of the compressed files, with the original ordering being optimal when CI bits are set as the initial columns. The original orderings are not optimal when CI bits are not set. In both datasets, the CDFs are not disjoint, so there exist permutations of the columns with CI bits that exhibit worse compression ratios than some permutations without CI bits set.

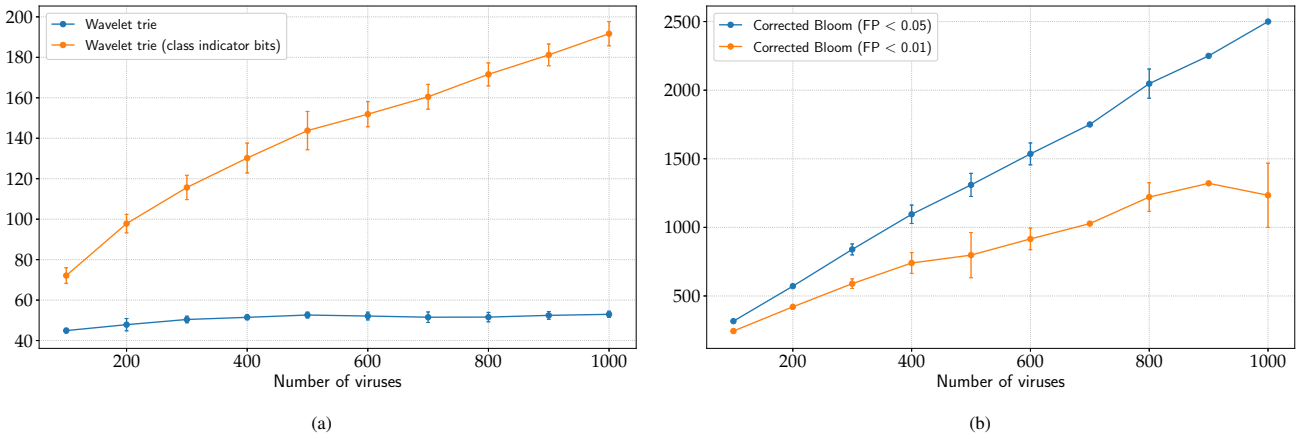


Fig. S-3: **Growth of compression ratios on the *Virus100* to *Virus1000* datasets.** See Figure 3 for the full plot on log-scale axes.

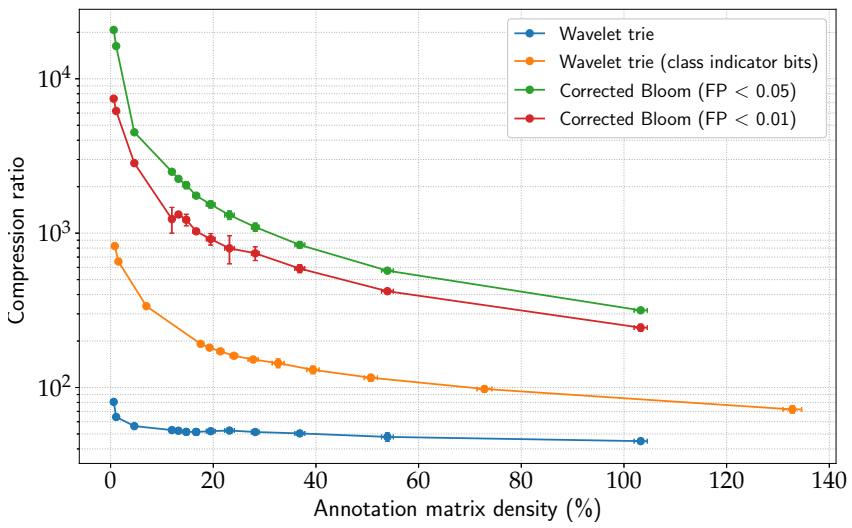


Fig. S-4: **Decrease in compression ratio with increasing annotation matrix density on the virus data sets.** Each data point represents a mean value for across the genomes of a virus collection of a given size. Error bars in both axis represent standard deviation.

	Bloom filter ($FPP < 0.05$)			**Bloom filter** ($FPP < 0.01$)			**Wavelet trie**	**Wavelet trie CI**	**VARI**	**Rainbowfish**
Data set	**Bits/edge**	**Time**	**Nbhd**	**Bits/edge**	**Time**	**Nbhd**	**Time**	**Time**	**Time**	**Time**
Virus100	0.36	1.575	207.333	0.44	1.338	156.261	0.0027	0.0025	0.0060	0.00447
Virus1000	0.49	5.571	175.155	0.82	4.795	122.594	0.0038	0.0047	0.0547	0.0275
Virus50000	2.58	430.59	117.655	7.41	591.609	96.271	0.010	0.0240	3819.2	N/A
Lactobacillus	0.95	2.051	124.688	1.40	1.834	105.560	0.0052	0.0055	0.0077	0.00665
chr22+gnomAD	0.45	1.085	99.121	2.41	1.406	82.272	N/A	0.0013	0.00054	0.00159
hg19+gnomAD	0.68	1.972	124.618	1.82	2.168	98.259	N/A	0.014	0.0018	0.00244

Table S-1. **Annotator query times (ms) for our proposed methods and competing methods Nbhd** refers to the number of neighbors traversed during graph-based Bloom filter error correction.

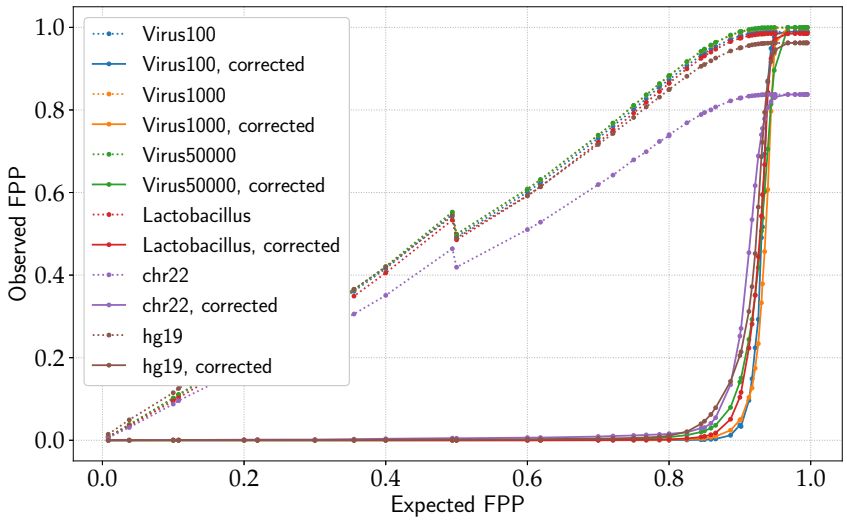
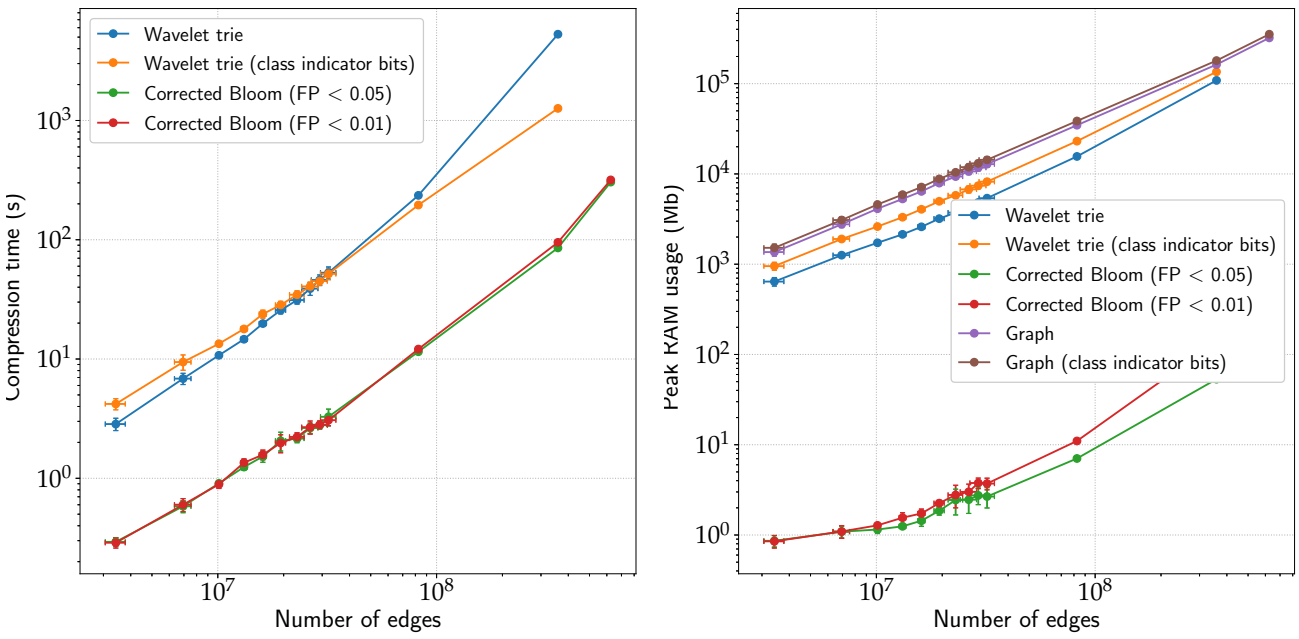


Fig. S-5: **Observed vs. expected per-bit false positive probabilities of the Bloom filters before and after correction.** The drop in the observed false positive probability at $E[FPP] = 0.5$ can be explained by the change in optimal number of hash functions at that point.



	Bloom filter ($FPP < 0.05$)		**Bloom filter** ($FPP < 0.01$)		**Wavelet trie**		**Wavelet trie CI**		**Graph**
Data set	**Time (s)**	**RAM (Mb)**	**Time (s)**	**RAM (Mb)**	**Time (s)**	**RAM (Mb)**	**Time (s)**	**RAM (Mb)**	**RAM (Mb)**
Virus100	0.260	0.929	0.242	0.934	2.31	558.51	4.14	829.09	1,191.49
Virus1000	3.029	2.387	2.935	3.805	48.13	5,012.53	47.93	7,659.39	12,150.47
Virus50000	304.352	192.2	317.475	912.20	17,721.3†	113,218.20†	16,825.4†	52,662.46†	321,500.80
Lactobacillus	18.557	1,024.0	19.121	35.17	208.14	31,726.50	274.62	45,171.55	59,052.50
chr22+gnomAD	37.408	93.60	64.773	115.06	233.99*	32,115.62	233.99	32,115.62	76,111.37
*hg19+gnomAD**	1,009.93*	1,841.17	1,431.11*	2,640.23	9,769.00*	146,001.03*	9,769.00*	146,001.03*	14,342.87*

Fig. S-6: **Compression time and peak RAM usage on the virus data sets.** RAM usage values for the graph and annotation compressor objects are reported separately. In the plots, each data point represents a mean value across the genomes of each virus collection of a given size (e.g., six draws of the *Virus100* dataset are averaged in both axes and represented by one point). Error bars in both axes represent standard deviation. To reduce RAM usage to fit with our computing system's 400GB job RAM usage limit, the wavelet trie for the *Virus50000* dataset with class indicator bits was computing in two steps, first computing the graph, then loading it in chunks to compress the annotation. Bloom filter runtimes are for a single thread, while wavelet trie runtimes are for ten threads. *: performed with a succinct graph representation. †: Constructed with an alternate method in which the graph was unloaded to disk, then read in chunks for wavelet trie construction.

References

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2016). Genbank. *Nucleic Acids Research*, **44**(Database issue), D67–D72.

Consortium, E. A., Lek, M., Karczewski, K. J., and Minikel, E. V. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**(7616), 285–291.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., OâŁ™Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**(7616), 285.