CCGG Supplementary Material

Hang Su[1,2,4], Ziwei Chen[1,4], Jaytheert Rao[1,4], Maya Najarian[1], John Shorter[3], Fernando Pardo Manuel de Villena[3] and Leonard McMillan[1]

[1]Department of Computer Science, University of North Carolina, Chapel Hill, NC 27514, USA; [2]Curriculum of Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, NC 27514, USA; [3]Department of Genetics, University of North Carolina, Chapel Hill, NC 27514, USA

Contact: mcmillan@cs.unc.edu
[4]These authors contributed equally to this work

Content

1. Core Genome Size Estimation in Collaborative Cross
2. Graphical Genome Construction
    - Anchor Selection
    - Edge Creation
    - Adding CC Path
    - Graph Annotation
    - Graph Manipulation
    - Graph Genome Compression
3. Supplementary Figures
4. Supplementary Tables

Core Genome Size Estimation in Collaborative Cross

As previously reported we have constructed multi-string Burrows-Wheeler Transforms (msBWTs) for all lanes and pair ends of the Illumina reads sets from (Srivastava et al. 2017) and (Keane et al. 2011). From these we dynamically-construct (and subsequently cache) a sampled FM-index that supports efficient queries into the raw-sequence data. We elected to make sequences of a uniform size of 45 base pairs, hereafter referred to as 45-mers. Firstly, we built an

occurrence count matrix of every non-overlapping 45-mers (summing both forward and reverse complements) from the Mus musculus reference genome, using the msBWT for each of the 69 sequenced CC samples and 20 replicates (Srivastava et al. 2017). We next constructed a corresponding occurrence matrix for these same 45-mers and their reverse complement in all of the eight founder genomes. The kmer count occurrence matrices were used first to select candidate anchors according to the rule that they were unique in reference assembly and appear in more than 3 reads in *every* sequenced sample.

The number of anchor candidates depends on how many samples are taken into account. We simulated the procedure by sequentially adding the sample columns by randomizing the combination orders for n times (n = 15). We measure the number of anchor candidates as a function of the sample numbers that taken into account and extrapolate the size of the core genome of Collaborative Cross population. For n = 1 to 89, a panel of n-1 samples is taken into consideration and the n-th sample is chosen from samples that haven't been considered; then the number of anchor candidates presented in samples of the panel are counted. Since the number of anchor candidates in sample A shared with B and C can be different from the corresponding number of anchor candidates in B shared with A and C. The procedure is repeated by randomizing the sequential adding order of the samples. However, the final number of anchor candidates in all CC samples depends only on the known data instead of their sequential adding orders. The trend of anchor candidates number decay was only used to estimate the size of core CC genome.

The number of candidate anchors were divided by the total number of 45mers in the reference genome to estimate the fraction of core genome in CC population. The simulated sequential data were extrapolated by fitting the exponential decay function:

$$F = a \times e^{\frac{-n}{t}} + d$$

where a is the amplitude of the exponential decay, t is the decay speed constant and d is the size of the core genome for n approximate to infinity. n is the number of samples taken into account which are continuously extrapolated. The results suggest that the size of CC core genome converge to a plateau value of $27.93\% \pm 0.24\%$.

Graphical Genome Construction

Anchor Selection and Mapping

Anchor nodes represent topologically sorted sequence fragments that are conserved within the population of assemblies represented in our CCGG pangenome. The kmer count occurrence matrices were used first to select candidate anchors according to the rule that 1) they were unique in every reference assembly and 2) appear in more than 3 reads in *every* sequenced sample. We allow for the potential high counts of anchors, the possible private gene duplications fragments, within a single CC line. The number of conserved unique 45-mers in each chromosome are shown in the second column of Table 1. Many conserved unique 45-mers were adjacent. We next reduced the set of anchor candidates by only keeping the first and last 45-mers of these contiguous runs. As shown in the third column of Table 1, this reduced the anchor candidate set by 35%.

Next, we applied a recursive subdivision approach to establish the monotonicity of anchor sequences in the other 7 draft genomes. We first sorted the candidate anchors on a common contig according to their positions in the standard reference assembly. We partitioned these sorted candidates into 10 equal-size intervals and randomly sampled a 45-mer near the boundary of each. We mapped these 45-mers to the other 7 founder genome assemblies. We dropped and replaced any of the ten 45-mers whose mapped coordinates were not increasing in every genome. Once a coarse level set of anchors were established, the process was repeated to establish ten new anchors between each of the established ten from the previous level by partitioning the remaining sorted anchor list between the 2 bounding mapped anchors into ten intervals. This process was repeated until 8 new anchors could not be found at the next finer level. Finally, we removed any anchor that overlapped an adjacent anchor in any genome. The final anchor count for each chromosome is shown in the fourth column of Table 1.

Overall, we generated an ordered list of Anchor nodes, which are unique in each founder genome and conserved across all CC samples and 8 founders with monotonically increasing order. Each anchor node was then annotated with its position in all eight genomes.

Edge Creation

Edges contain sequences that lie between a source and a destination node. The initial edges were created by first extracting the sequences between the anchor nodes. Identical sequences were merged into a common edge and a list of founder strains sharing the edge was maintained as the

strain edge attribute. Each initial edge was also annotated by its source and destination anchor nodes with *src* and *dst* attributes respectively.

In many cases all eight initial founder edges merged into a single common edge, which we call a *collapsed* edge. These collapsed edges are called *core* sequences in other pangenome representations. Likewise all of the CCGG anchor nodes would also be considered core sequences. However, the converse is not true. Unlike anchors, collapsed edge sequences may not appear in every CC sample, may not be unique, or may not be ordered relative to other nodes and edges in the graph. We annotate these collapsed edges into 3 categories. If the read counts from any 45-mers along a fully collapsed edge are less than 3, it's *type* attribute is set to "missing". Otherwise, if any 45-mers are not unique in the genome, their type is set to "with repeats". And, if all the 45-mers in the collapsed edge are conserved and unique, their type is set to "conserved". We found 2,781,442 collapsed edges across 21 genomes. 72.3% of these collapsed edges are annotated as "Conserved", 26.1% of these collapsed edges are annotated as "Missing", and 1.6% of them are annotated as "With repeats". These results suggest many conserved edges appear in the graphical genome. These collapsed edges can be promoted to anchor nodes if needed when more genomes are added to the graphical genome.

Finally, edges were added before the first anchor node of each contig to contain all sequence data before it, in each of the founder assemblies. The source of these edges was set to the implicit *SOURCE* node and its strain attribute is set according to its founder.

Adding CC paths

CC genomes can be regarded as a mosaic of their 8 founder genomes. We annotated CC path in the graphical genome based on the hapfile reconstructed by imputation from the genome sequence of the corresponding founder inbred strain. SNP intensities were used to calculate the probability of genome region descent from each of the 8 founder strains []. We merged the genotype probability files from genome sequencing data, and annotated the CC path in the graphical genome according to the imputed haplotype regions. For overlapping boundary in the hapfiles, we decide the boundary by taking the mean value of this region.

CCGG integrates sequence commonality and haplotype transitions to understand CC recombination. We traversed the graph and regenerate the hapfiles from the CCGG. The results suggest it provides novel information, such as in some recombination region, the two haplotypes share the same sequences.

Graph Annotation

Gene and Exon Annotation

Next, we annotated anchor nodes and edges with genes, exons, and interspersed genomic repeat types (Smit et al. 1996). The Gene and Exon intervals of TSL1 transcripts were obtained from Ensembl Biomart according to their start and end coordinates in the mouse reference genome. Both anchor nodes and edges on the B path can also be represented as intervals in the reference genome. Given a gene region, we first found out all the anchor nodes and edges whose ending position was larger than the start of the target gene/exon, and the starting position is less than the

end of the target. We enumerated the anchor node and edge list and recorded the relative position between each anchoredge and gene by using a SAMtools compatible *cigar* string (Li et al. 2009). We compacted the gene name, gene orientation and the cigar string by using the separator '|' and annotated the corresponding anchor nodes or edges by the 'gene' attribute. Similarly, we annotated the overlap of each anchor node and edge with annotated exons from TSL1 transcripts. We recorded the exon name, orientation and the cigar string of their relative position in each node and edge. Both gene and exon attributes of anchor nodes and edges are represented as lists since they can be annotated with multiple genes or exons. Overall, 46036 genes and 726152 exons were annotated in the CCGG.

As shown in Figure S3, the first 89 bases of exon ENSMUSE00000758727, which is encoded on the positive strand, overlap with the B path edge E01.36917861. The cigar string representing this overlap is "1S89M" and the 'exon' attribute value of this edge is "ENSMUSG00000003974|-|40M5S".

Repeat-masker Annotation

We also annotated anchor nodes and edges according to their overlap with interspersed genomic repeat types as annotated by repeat-masker regions. Overall, 5700130 Repeat-masker regions were annotated in the CCGG. The Repeat-masker intervals for Mus reference genome were obtained from http://www.repeatmasker.org/. As described above, we overlaid the overlapping repeat regions to anchor nodes and edges. We recorded the class of the repeat, the name of the matching interspersed repeat, the sequence orientation and the cigar string describing the relative position between the repeat interval and the anchor/edge. We concatenated the repeat class and

repeat name by using symbol '@' and connected orientation and cigar string by using separator '|'. For example, anchor node A01.00071188 lies completely within a repeat region MIRb, the repeat class of SINE/MIR on the negative string. It is annotated with "SINE/MIR@MIRb|-|45M".

Variants Annotation

Variants along non-collapsed edges whose parallel reference path was fewer than 1000 base pairs were annotated as follows. An alignment with minimum Levenshtein edit distance was found for all paths between a given pair of anchors and the GRCm38 reference path between those same anchors. A SAMtools compatible cigar string (Li et al. 2009) was constructed for the alignment and saved as a "variant" attribute on each edge. As a result, every edge along the 'B' path should contain a 'variant' string of the form 'N=' where N represents the length of the edge's sequence. Other paths will contain mismatches ('X'), insertions ('I'), and deletions ('D') sufficient to transform the sequence on the alternate path to the reference sequence.

Alignments for longer paths were generally found to be less informative of the actual sequence variants, and, thus, we chose to insert floating nodes within those long gaps for further compression (See Graph Compression section). Floating nodes create new parallel segments between alternative and reference edges, where we further perform pairwise alignment in those regions. Overall, 27,285,809 of the 30,344,738 edges (89.9%) in the CCGG include at least one variant from the reference.

Graph Manipulation

The CCGG pan-genome is represented as a graph data structure, which is serialized as one or more standard FASTA files. Genomic attributes and annotations for both nodes and edges are stored in a dictionary encoded as a JSON string in each sequence's FASTA header. Edges require the 'src', 'dst', and 'strain' attributes, all other annotations are optional. Typical annotations for edges include type, gene, exon, repeat class, and variant. Nodes are typically annotated with their known build position in a linear genome where available.

We provide a Python-based API for accessing, traversing, and editing the CCGG, as well as a command-line interface (CLI) for extracting various paths, subpaths, and other symbolic genomic representations from the CCGG. The FASTA files are parsed and a graph representation is built from them. Components of the graph are four dictionaries: nodes, edges, incoming, and outgoing. The nodes and edges dictionaries are indexed by anchor node names and edge names respectively and return a dictionary of attributes. These attributes are represented as a JSON string in the sequence's FASTA header. An attribute of each edge is the names of its source and destination nodes. These required key values are used to dynamically construct the incoming and outgoing dictionaries during parsing. A second required key is 'strain' which returns a list of one or more paths/genomes that include the edge. The incoming and outgoing dictionaries represent the graph's topology. Both incoming and outgoing are indexed by node name (either anchor, floating, source, or sink) and they return a list of adjacent edge names. The graph is traversed in the forward direction by iterating through the outgoing edges of a node while testing whether each edge lies along the desired path. When a matching edge is found, its destination node is used to access the outgoing dictionary and the process repeats until either the

sink node or some other condition is reached. The graph can be similarly traversed in reverse order using the incoming.

Thus, by traversing the graph, we can abstract genome contigs for certains strain between “SOURCE” and “SINK” nodes. We can recover the 8 founder genomes from the graph with significantly storage saving; we can also abstract the heterozygosity genomes of CC samples that can be represented by 2 different paths in the CCGG. For example, CC001 have two path CC001a, and CC001b in chromosome 1, and thus we can constructed 2 heterozygous genomic contig in chromosome 1 for CC001. To extract a sequence contig from the graphical genome, one can perform simple commands from the CLI by specifying the node and file edges, the chromosome and strain path information. For example, to extract the sequence for strain CC001 from chromosome 1 following the 0 path (path 0 for homozygous genomes, path 0 or 1 for heterozygous genomes), one can use the command “python CCGG_CLI.py NodeFile_Chr1.fa EdgeFile_Chr1.fa 1 --sequence 001 0 sequenceCC001Chr1.txt” and dump it to the file named "sequenceCC001Chr1.txt".

We can also traverse all possible paths in certain genomic regions, such as a gene region that interested, and construct all haplotypes in that region for further comparative analysis (Figure S6). Taking Gene Dusp18 as an example (3.8Mb - 3.9Mb, chromosome 11, 6255 bp long in the reference genome), first we find the bounding anchor, A11.00086559 and A11.00086698 (3895156 bp - 3901411 bp) for this gene, we then traverse the graph and constructed all the haplotypes given the strain attributes. Further, by traversing the graph, we can identify highly variable regions, such as the interval between anchor A11.00086572 and A11.00086576 (3895741 - 3895921 bp) with 8 haplotypes, as well as the conserved coding regions such as the

interval between A11.00086609 A11.00086614 (3897406 - 3897631bp) in exon ENSMUSE00000682312.

Graphical Genome Compression

We compressed the graphical genome by adding floating nodes in the long gaps where any path between anchor pairs are longer than 1000bp. These long paths usually contain sequences that are shared by more than one genome. We partitioned the path sequences in a single anchor gap into non-overlapping 45-mers. Noted that if the sequence lengths are not multiples of 45 bp, we allow 45-mers overlap in the middle of the sequence. Each distinct 45-mer are assigned with an index based on their sequence contents. By sequentially scanning through each path, we can find shared 45mers between multiple paths and merge the continuous shared 45mers into a large interval. Here we excluded all the 45mers that contain ambiguous bases.

The compression itself involves adding a new floating node to the graph and connecting the shared sequence to one end and the divergent sequences to the other end. Floating nodes are connector nodes without annotations or sequence content, which only exist in the ‘src’ or ‘dst’ attributes of edges. We inserted floating nodes in the start and the end of each shared sequences. One such example can be seen in Figure 1C where the cluster of 2 outgoing edges of A01.01979848 were found to have an identical sequence at both their starts and ends. The two edges go from spanning the whole anchor to anchor distance to converging on a pair of new floating nodes somewhere between and thus lead to a set of collapsed edge that represents the identical sequence on the two edges. We topologically sorted the floating nodes based on their position in the graphical genome, and then inserted the floating nodes sequentially. By inserting

floating nodes, we partitioned the long paths into shorter segments, and collapsed the identical sequences into a new set of edges. Thus the graphical genome are further compressed and the edge length are further reduced. As shown in Figure 2B and D, the length of edges are greatly reduced by inserting floating nodes, especially for chromosome X.

Supplementary Figures

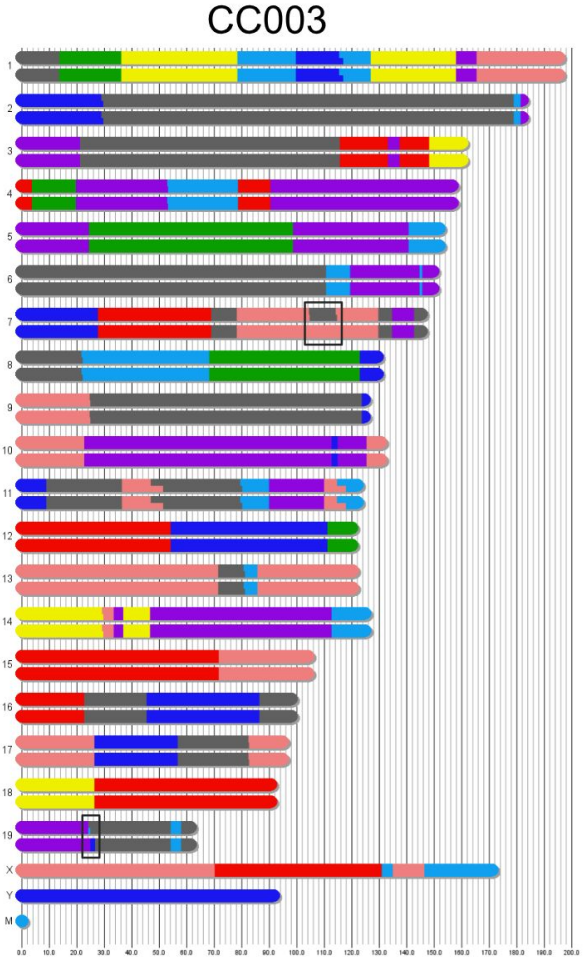

Figure S1. An example Genome of a Collaborative Cross (CC) mouse strain. The CC is a genetic reference population derived from a common set of 8 genetically diverse inbred founders. The genome of each CC strain is, in general, a mosaic of these founder genomes. Each founder is represented by a standard color and a single letter label as follows: yellow and 'A' represents the A/J, gray and 'B' is used for C57BL/6J, pink and 'C' for 129s1/SvImJ, dark blue and 'D' for NOD/ShiLtJ, light blue and 'E' for NZO/HlLtJ, green and 'F' for CAST/EiJ, red and 'G' for PWK/PhJ, and purple and 'H' for WSB/EiJ. The ideogram above represents the founder mosaic of a single CC003 male sample sequenced in (Srivastava et al. 2017). Most CC strains have residual heterozygous regions as indicated by the overlaid boxes shown above. This suggests two possible genomic sequences, one for each haplotype. The sticky-end overhangs within a chromosome (ex. the pink-to-black and pink-to-light-blue transitions on chromosome 11) represent ambiguities in estimating the exact recombination boundaries.

Figure S5 Collaborative Cross Graphical Genome Construction Workflow **A) Anchor Selection.** The unique and conserved candidate anchors were selected by using the kmer count occurrence matrices. The adjacent anchor candidates were collapsed by only keeping the first and last of these contiguous runs. The selected candidates were mapped to other 7 draft genome by using a recursive subdivision approach. The candidate anchors with monotonically increased order in the other 7 draft genomes were selected as anchor nodes, which are defined as conserved, unique and monotonically ordered 45-mers. **B) Edge Creation.** The initial edges were created by extracting the sequences between anchor pairs. Identical sequences were merged into a common edge. A list of founder strains sharing the edge was maintained as the *strain* attribute. Each initial edge was also annotated by its source and destination anchor nodes with *src* and *dst* attributes respectively.

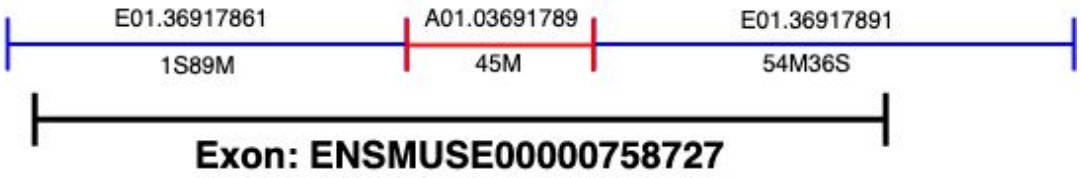

Figure S3 Biological Feature Annotation. Genes, exons and repeat maskers can be viewed as the intervals in the linear reference genome (the B path in CCGG). We annotated these biological features with the overlapped anchors and edges on B path and recorded their relative positions by using cigar strings. As shown in this picture, the first 89 bases of exon ENSMUSE00000758727, which is encoded on the positive strand, overlap with the B path edge E01.36917861. The cigar string representing this overlap is "1S89M", and the 'exon' attribute value of this anchor is "ENSMUSE00000758727|+|1S89M". Similarly, we annotated the overlap of each anchor node and edge with genes from TSL1 transcripts and the repeat masker intervals by using “gene” and “repeatclass” attributes. All of the attributes of these features are represented as lists since they can be annotated with multiple features.

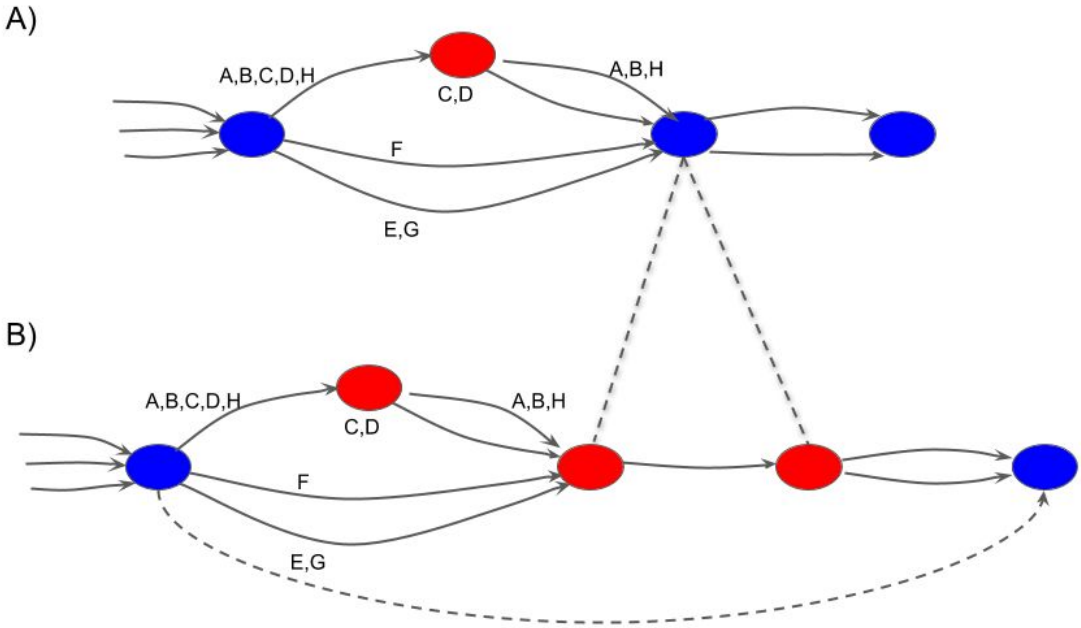

Figure S2. Demoting an Anchor. In the figure, blue nodes reference anchor nodes and red nodes represent floating nodes. In A) it is determined that the anchor node in the middle is no longer conserved and therefore needs to be demoted. This would occur, for example, if you found the dotted path in B) to exist in your Graphical Genome. The existence of this path would mean that the middle anchor in A) is no longer conserved among all sequences in the genome. To complete the demotion, the anchor node is replaced with an edge which is connected to two new floating nodes. The incoming edges of the old anchor are now given to the new floating node on the left and the outgoing edges from the old anchor are given to the floating node on the right.

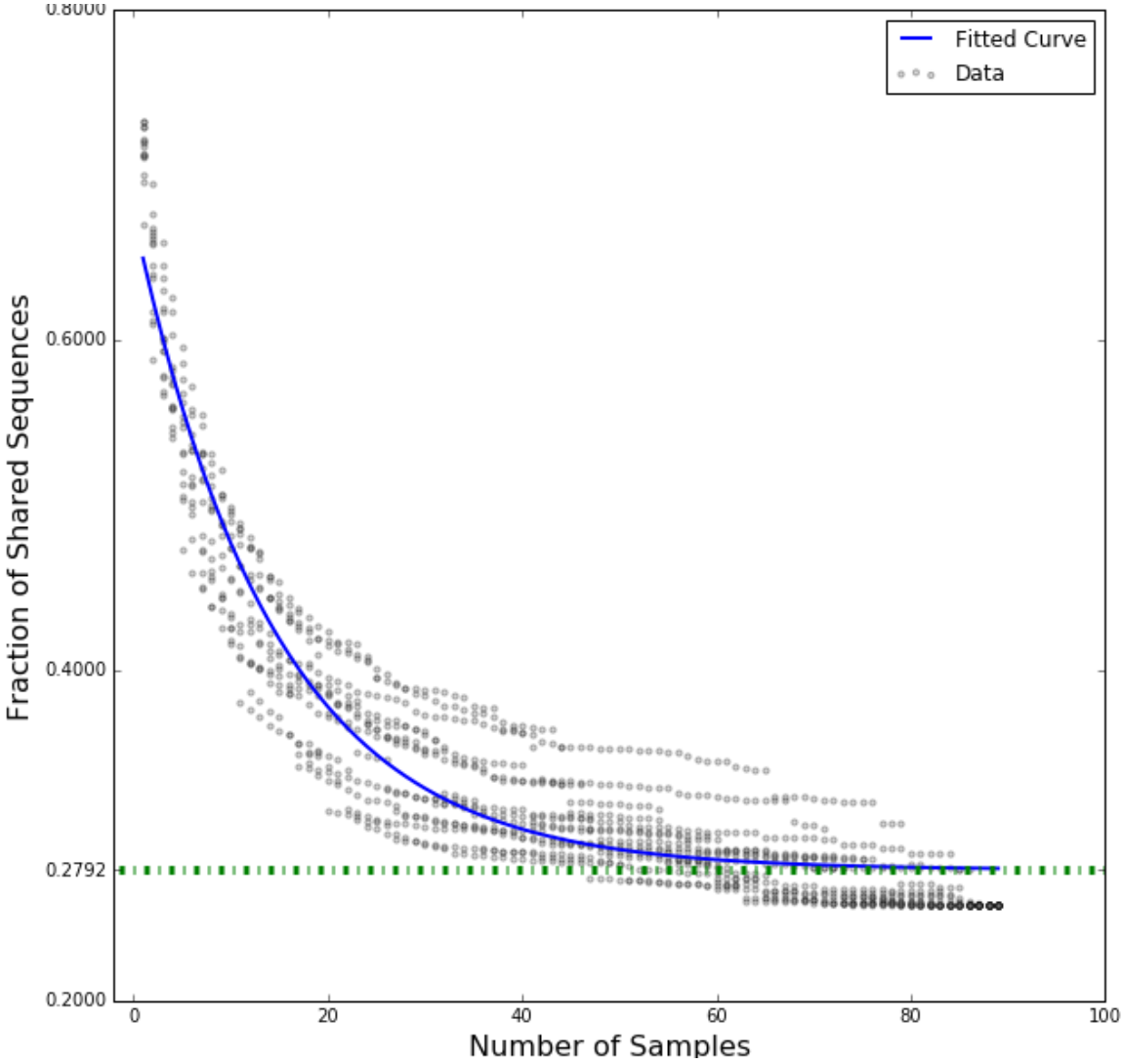

Figure S4 Collaborative Cross Conserved Genome. The fraction of anchor candidates, which are conserved in all 89 CC samples and unique in reference genome is plotted as a function of the number n of CC samples sequentially added (details in Material and Methods). For each n, circles are the value obtained for different sample combinations. The continuous curve represents the non-linear least-squre fit of the function $F = a \times e^{\frac{-n}{t}} + d$, where the best fit was obtained with a = 0.39678587, t = 14.38789651, c = 0.27916513. The estimated GBS core genome size is shown as a dashed line (95% confidence interval = 27.67% - 28.16%).

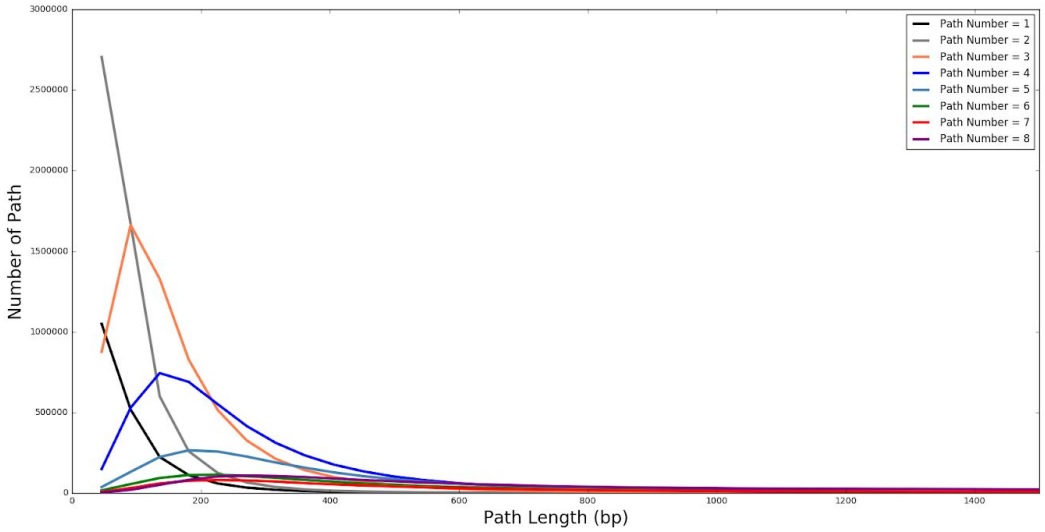

Figure S7 Path Length Distribution. The distribution of path lengths based on the number of distinct paths between anchors is shown above. As the number of paths increases, the peaks of the length distribution shifts to the right. This suggests genomic regions having large gaps separating anchors exhibit greater genetic diversity.

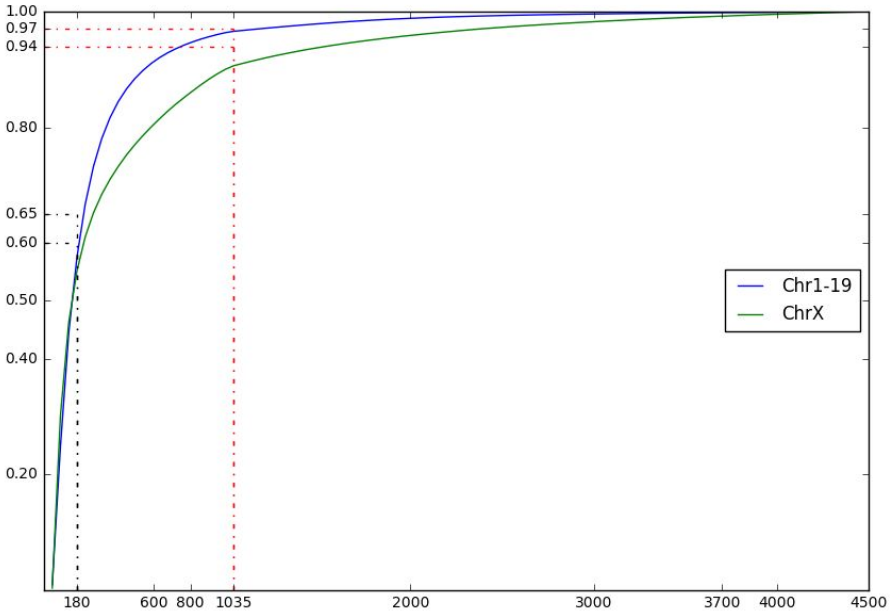

Figure S8 Cumulative Distribution of Edge Length. The blue curve denoted the edge length in autosomes (chromosome 1-19), while the green curve denoted the edge length in chromosome X. Comparing with B), the long gaps in chromosome X are further compressed after adding floating nodes.

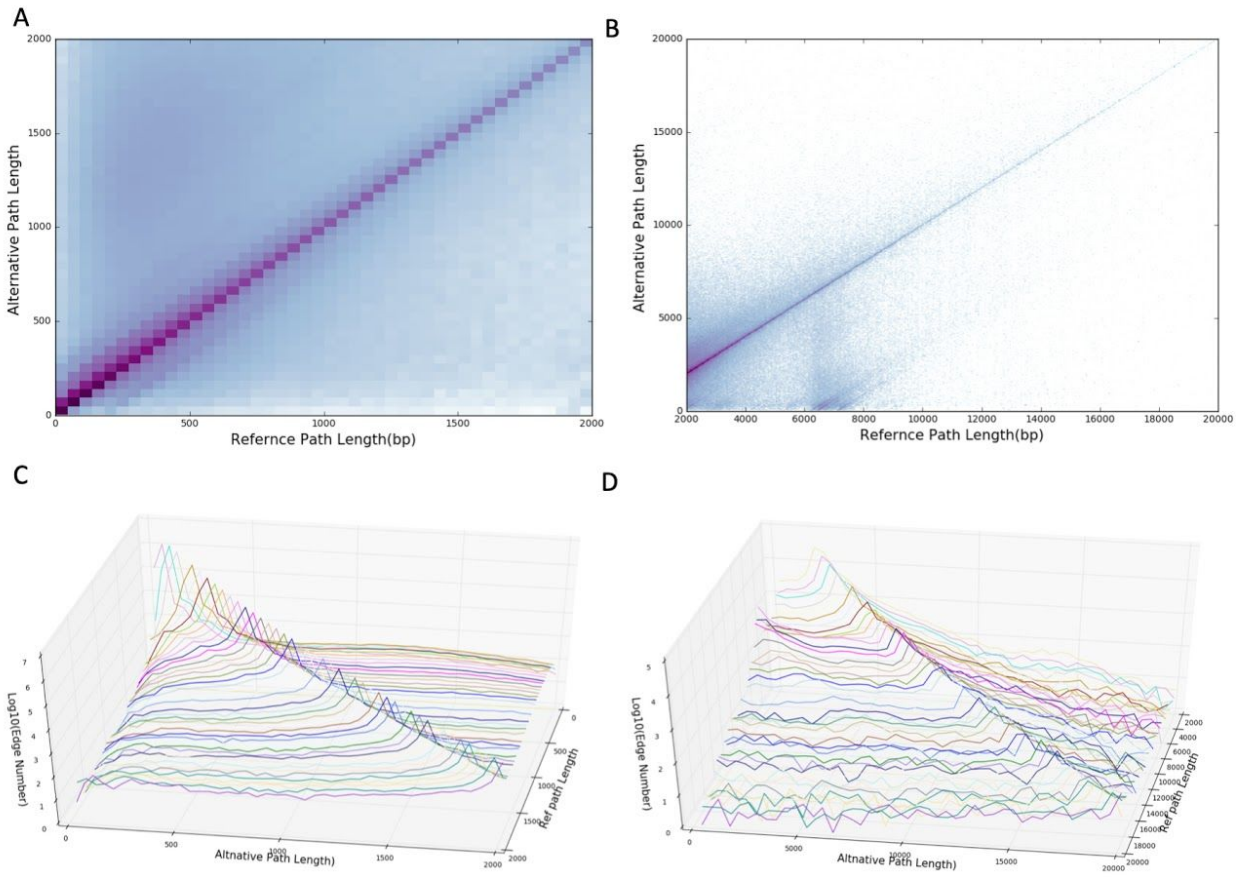

Figure S5 **Path lengths in short and long anchor intervals.** We plot the distribution of reference path lengths versus alternative path lengths for short intervals in **A)** and **C)** , where reference path length is less than 1000bp. The same distribution is shown for long Intervals in **B)** and **D).** For short intervals, the density is concentrated along the diagonal showing that reference and alternative paths are usually with similar length. For long intervals, the density is more dispersed suggesting insertions and deletions are more common in long intervals. The density is also more concentrated below the diagonal indicating alternative paths tend to be shorter than reference paths over 6000 bp. This suggests an excess of deletion events in those alternative paths.

Supplementary Table

Table 1: Graph Statistics. This table provides graph statistics for each step of the Graphical Genome Construction. The number of candidate 45-mers that are unique and conserved in each chromosome is recorded in the 'Candidate' column. The number of collapsed 45mers, which are not adjacent to each other is recorded in 'Collapsed' column. After filtering the unmapped and non-monotonically increased 45-mers, the number of anchor nodes is recorded in 'Anchors' column. The edges between anchors are extracted and collapsed and the number of edges is recorded in 'Edges' column. After compressing and adding the floating nodes, the simplified edge numbers are recorded in the 'Simplified' column.

Contig	Candidates	Collapsed	Anchors	Paths	Edges
Chr1	1218407	793670	719505	1995376	2339195
Chr2	1255880	791769	718250	1944457	2205332
Chr3	925215	616790	553441	1575928	1877551
Chr4	957604	623644	566460	1571966	1810652
Chr5	962893	625721	561112	1590482	1835997
Chr6	903590	590565	538322	1494561	1738998
Chr7	901319	574732	517691	1416486	1605912
Chr8	813525	530775	482921	1368578	1575589
Chr9	865715	555932	514167	1406830	1565759
Chr10	799008	526287	482150	1325553	1533677
Chr11	895244	564678	526441	1443748	1601468
Chr12	695657	459804	419750	1192424	1384932
Chr13	749132	491003	442715	1228702	1398549
Chr14	733561	471157	427084	1162109	1365860
Chr15	657846	424380	388202	1068242	1248647
Chr16	592542	386902	355440	974697	1148319
Chr17	606135	388629	343623	968378	1119346

Chr18	556844	369483	341015	969280	1112618
Chr19	403417	259138	239151	649043	723629
ChrX	109483	96773	74697	317209	1152708
ChrY	1927	1717	1717	1718	1718

Table S1: Ambiguous Bases and Repeats. This table recorded the statistics for intervals between two anchor nodes with repeats or ambiguous bases. The number of intervals for the four categories, whether the interval has repeat annotation and whether the ambiguous bases existed in the sequences are recorded.

Repeat and Ns	Intervals with Repeats	Intervals without Repeats
Intervals with Ns	754797	71259
Intervals without Ns	4483017	3915016

Table S2. **Gap Statistics for CC042.** This table summarizes the gap statistics for CC042. It includes the total number of uncollapsed gaps, which may be solved by our probes, total number of resolved gaps and total number of gaps that are consistent with the previous genotype data

Contig	Un-collapsed	Resolved	Consistent
Chr1	566585	395701	395134
Chr2	558578	390446	389665
Chr3	442101	296089	295621
Chr4	449676	337466	336849
Chr5	450188	311604	311068
Chr6	426996	300754	300263
Chr7	406981	297683	297194
Chr8	387177	269656	269163
Chr9	404313	286080	285770
Chr10	384370	267984	267562
Chr11	412185	292815	292531
Chr12	335355	230291	229878
Chr13	350902	243728	243353
Chr14	332482	233401	233091
Chr15	306603	211545	211313
Chr16	281578	208153	207852
Chr17	271117	189757	189412
Chr18	273189	189640	189347
Chr19	188721	133917	133741
ChrX	76876	10620	10597

Table S3: Sequence Length Comparison. Comparison of the total number of base pairs in Original and Graphical Genome compressed representation of the eight founder strains on each genome. Note that Chromosome Y was excluded because we did not perform any compression on it. The percentage denotes the relative size of the compressed sequence against the original.

Chromosome	Original	Compressed	Percent of Original
Chr1	788098812	1608113409	49%
Chr2	693360916	1491443672	46.49%
Chr3	660823486	1311452157	50.38%
Chr4	635228078	1268859285	50.06%
Chr5	627117854	1250870743	50.13%
Chr6	613858672	1226867494	50.03%
Chr7	629239517	1184777264	53.11%
Chr8	517986603	1056476299	49.03%
Chr9	462622807	1013255708	45.66%
Chr10	505965264	1069946153	47.29%
Chr11	451581528	1001994858	45.07%
Chr12	496386226	969223061	51.21%
Chr13	480735626	973326191	49.39%
Chr14	487772824	980838501	49.73%
Chr15	400778551	846693825	47.33%
Chr16	372883386	794842312	46.91%
Chr17	405326004	783839985	51.71%
Chr18	345291484	730697427	47.26%
Chr19	226368679	490201464	46.18%
ChrX	86743911	1318081905	6.58%

Table S4: Frequency of Recombination Events in Gene Body

	Intragenic region	Extragenic region	Total
Sequence Length /bp (Reference Genome, chr1-chr19, chrX)	1122110280 (42.6%)	1511666392 (57.4%)	2633776672
Recombination Num	3442 (46.8%)	3911 (53.2%)	7353