Collaborative Cross Graphical Genome

Hang Su[1], Ziwei Chen[2], Jay Rao[2], Fernando Pardo Manuel de Villena[3], Leonard McMillan[1,2]

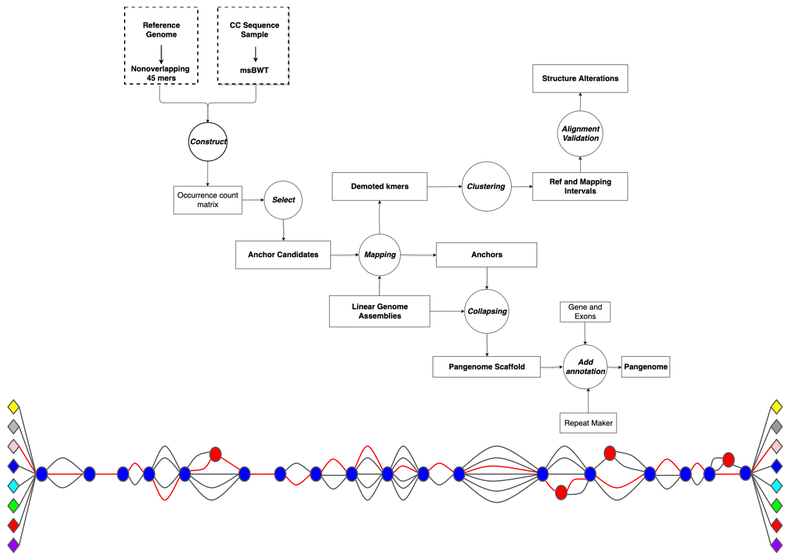
[1] Curriculum in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599

[2] Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599.

[3] Department of Genetics, University of North Carolina, Chapel Hill, NC 27599.

Abstract

Inexpensive and fast genome sequencing has enabled the assembly of multiple intra-specific genomes that capture more genetic diversity than any single reference assembly. It has been widely suggested that pangenome reference assemblies, which incorporate multiple genomes into a single representation, are the path forward, but there are few standards for, or instances of practical pangenome representations suitable for large eukaryotic genomes. We present a graph-based pangenome representation and an instance of it for a widely-used recombinant-inbred mouse genetic reference population known as the Collaborative Cross. Our pangenome representation leverages existing standards for genomic sequence representations with backward-compatible extensions to describe graph topology and genome-specific annotations along paths. It packs 82 genomes in a single graph representation which directly captures important notions relating genomes such as identity-by-descent between mouse strains, and highly variable genomic regions. The introduction of special anchor nodes with sequence data provide a valid position-reference framework that divided large eukaryotic genomes into homologous segments and addresses most of the graph-based position reference issues, such as *monotonicity, backward-compatibility, spatiality*. Paralleled edges between a pair of anchors contains all variants and provide effective way for orthogonal genome comparison and visualization. Furthermore, our graph structure allows annotations to be placed in multiple genomic contexts and simplifies their maintenance as the assembly improves. The CC reference pangenome provides an open framework for combining and incorporating new genome sequences and an effective representation for searching, annotating, comparing and visualizing biological features at different scales, which also facilitates tool chain development for downstream analysis.



CCGG: A single ***directed, k-partite*** graph representation for 82 mouse genomes, including 8 founder genomes and 74 CC genomes which are mosaic of the 8 founders

Nodes: Anchors: ***conserved, unique 45mers*** with ***monotonically*** increasing order
Floating nodes, Source nodes and Sink nodes

Edges: Sequences between nodes, sequence diversity.

Path: A series of nodes and edges sharing common source and destination anchors

Gap: Sequences between two anchors

Genomic Contig: a Path from a Source to a Sink node

Collaborative Cross Graphical Genome

CCGG Overview

A) Fragment of NodeFile_Chr1.fa

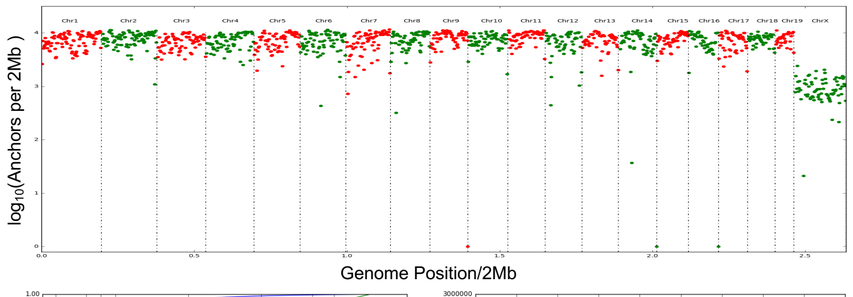
```
...
>A01.01979840:{"A":"87291523","C":"90544851","B":"89092801","E":"89980229","D":"100381609
","G":"85912473","F":"89402313","H":"89616328","gene":["ENSMUSG00000036206|+|45M"]}
AGCTGCCACATTTCCTGCCATTCAGGCAGTCGCTGGTTGTTTGTTT
>A01.01979844:{"A":"87291703","C":"90545031","B":"89092981","E":"89980409","D":"100381789
","G":"85912653","F":"89402493","H":"89616588","gene":["ENSMUSG00000036206|+|45M"]}
CACAATGATACTGGAATGGAAAAATGCTGTCAAACAACTCCTTTT
>A01.01979848:{"A":"87291883","C":"90545211","B":"89093161","E":"89980589","D":"100381969
","G":"85912833","F":"89402673","H":"89616488","gene":["ENSMUSG00000036206|+|45M"]}
CACCATCCCTAATGTACTTGTCAACCAGAATCCTCCAGGTTCAGC
>A01.01979853:{"A":"87292198","C":"90545436","B":"89093386","E":"89980814","D":"100382194
","G":"85913058","F":"89402898","H":"89616913","gene":["ENSMUSG00000036206|+|45M"]}
GTTTGTGCAAGAAGGAGAGGCAAGTGAGCCTGGGAGGGGGTTGGG
>A01.01979856:{"A":"87292243","C":"90545571","B":"89093521","E":"89980949","D":"100382329
","G":"85913193","F":"89403038","H":"89617048","repeatclass":["LTR/ERVL-MaLRRM_TTD|+|10M3
5S"],"gene":["ENSMUSG00000036206|+|45M"]}
GCAGATGGGAGGAGCCTGGGCTGTCAGGATTTCTTGGTTATTTTCA
>A01.01979858:{"A":"87292333","C":"90545661","B":"89093611","E":"89981039","D":"100382419
","G":"85913283","F":"89403128","H":"89617138","repeatclass":["LTR/ERVL-MaLRRORR1F|-|45M"
],"gene":["ENSMUSG00000036206|+|45M"]}
CATCATTGGGTGCTGAGATCCCTGGGAGACACAGAAACCAGAGA
>A01.01979860:{"A":"87292423","C":"90545751","B":"89093701","E":"89981129","D":"100382509
","G":"85913373","F":"89403218","H":"89617228","repeatclass":["LTR/ERVL-MaLRRORR1F|-|45M"
],"gene":["ENSMUSG00000036206|+|45M"]}
AGATGAATTCCTTTTAGGGGCGACCCTACCACCGATTACCTTCTG
...
```

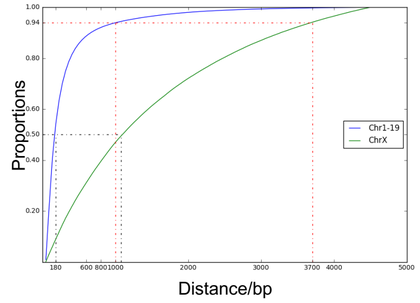
B) Fragment of EdgeFile_Chr1.fa

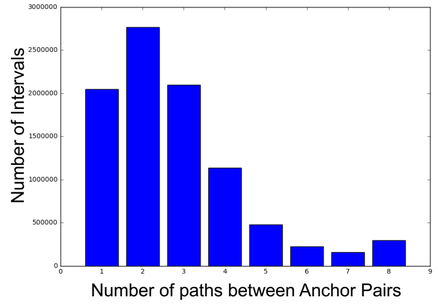
```
...
>L01.41725700:{"src":"A01.01979844","dst":"B01.00417257","strain":["074","075","072","073",
"071","022","C","078","010","011","038","013","058","059","016","017","018","019","056","05
0","037","052","053","029","036","051","020","068","035","E","A","060","063","062","065","D
","G","F","083","082","081","080","030","003","002","001","007","006","049","044","046","00
8","043","028","048","025","023","012","027","005","055","015"]}
GATAGTTCACAACAAAAAATTTAATTTTACTCGTTTCTTTCTAGAA
>L01.41725701:{"src":"B01.00417257","dst":"A01.01979848","strain":["074","072","049","004",
"003","078","010","011","038","013","058","059","016","017","018","056","050","037","052",
"053","043","036","051","035","A","060","C","062","E","D","G","083","068","081","080","003",
"002","001","007","006","005","022","046","008","029","028","040","025","023","012","027",
"020"]}
GGCACGGTACAAAGGGGAATGAGCACACGACTGCCCTTCTTTTTAAATGACCTTTAGCCTCCACTTTCTACCTACGCTGTGGTGGGGGG
>L01.41725800:{"src":"B01.00417257","dst":"B01.00417258","strain":["075","065","073","071",
"082","F","015","019","055","030"]}
GGCAGGTGCAAAGGGGAATGAGCACACGACTGCCCCTCTTTTT
>L01.41725801:{"src":"A01.01979844","dst":"B01.00417258","strain":["061","076","B","070","H
","024","039","021","047","033","009","031","045","034","057","044","042","026","041","032
"],"variants":"90=","gene":["ENSMUSG00000036206|+|90M"]}
GATAGTTCACAACAAAAAATTTAATTTTACTTGTTTCTTTCTAGAAGGCAGATGCAAAGGGGAATGAGCACAGCTGACTGCCCCTCTTTTT
>L01.41725802:{"src":"B01.00417258","dst":"A01.01979848","strain":["061","076","075","065",
"073","078","071","082","024","026","021","015","019","047","055","045","031","009","034",
"032","057","044","F","042","039","041","030","H","033","B"],"variants":"45=","gene":["ENSMU
SG00000036206|+|45M"]}
AAATGACCTTTAGCCTCCACTTTCTACCTACGCTGGGGTGGGGGG
...
```

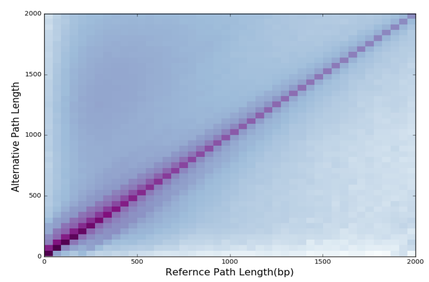
C)

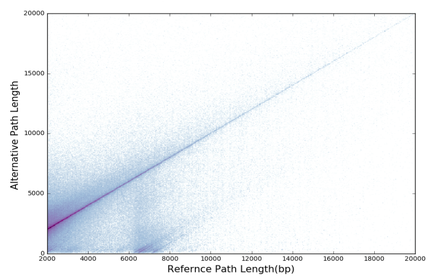
Graph Properties



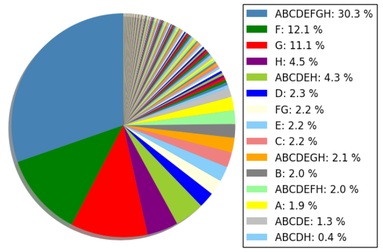




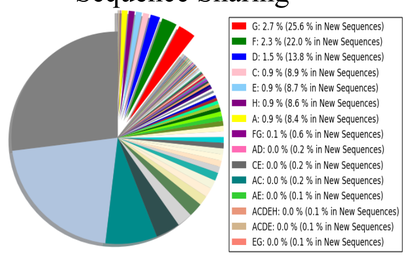




Edge Sharing

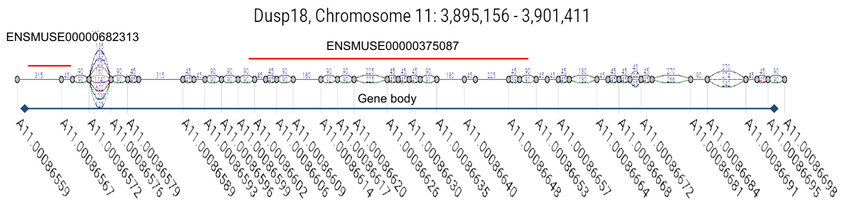


Sequence Sharing

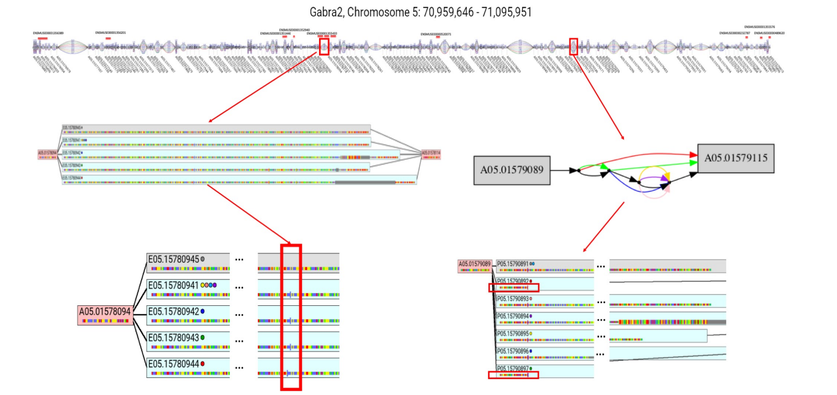


CCGG Graph Structure in Functional Regions

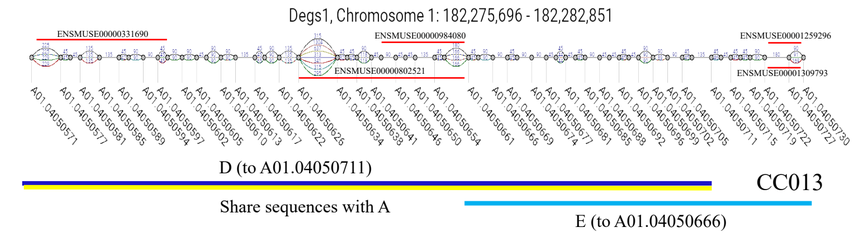
Genomic Structure: identify conserved and highly variable regions



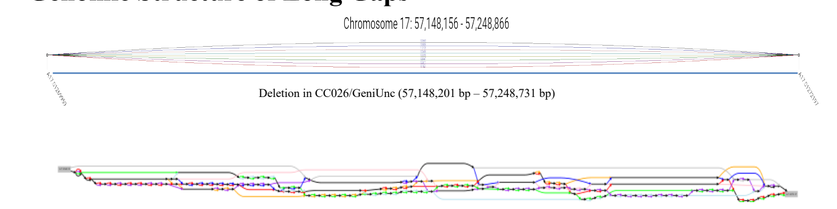
Genomic Structure and Variants representation



CC Recombination in more than 2000 genes



Genomic Structure of Long Gaps



- Reference Pangenome for the Collaborative Cross
- Graph reveals sequence commonality and diversity
- Functional variants and recombination analysis in CCGG
- CCGG establish a better standard for genome annotation and coordinates
- Downstream tool chains: aligners, variant calling…