UNIVERSITY OF CALIFORNIA,
IRVINE

Scaffolding of a *Peromyscus Leucopus* genome using Hi-C

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Mathematical, Computational, and Systems Biology

by

Yuan Tao

Thesis Committee:
Professor Anthony Long, Chair
Associate Professor Ali Mortazavi
Assistant Professor J.J. Emerson

2018

Dedication

To

My parents, Jun Tao and Aiju Xu,

Who offer unconditional love and support to me,

To

All my dear friends,

Who always cheer me up when I'm down,

To all the researchers,

For we are standing on the shoulders of them,

And to the charming nature,

For the incredible "design" of life and everything

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my committee chair, Professor Anthony Long, who has the intelligence and persistent patience to guide me during this journey to my degree. I feel very fortunate to have had an opportunity to work with him. He has always been open-hearted and helpful. He unreservedly shared his experience with me in scientific thinking and keeping balance between research and life.

I would like to thank my committee members, Professor Ali Motarzavi and Professor J.J. Emerson, as well as my labmate Robert Linder and the entire J.J. lab for stimulating discussion and valuable assistance to implement my experiments.

Thanks to Professor Alan Barbour, who has been studying the pathogenesis of *Borrelia* infections and Lyme disease for years and started the *Peromyscus Leucopus* project with my mentor, Professor Anthony Long. I would like to thank James Baldwin Brown for unpublished data where I get continued his work.

In addition, a thank you to the program Mathematical, Computational, and Systems Biology (MCSB) and all its faculty and supporting staff, which reintroduced me to Biology. Without this program, I would not be able to focus my research interest in Biology and find myself where I am today.

Research reported in this thesis was supported by National Institute of Allergy and Infectious Diseases of the National Institutes of Health (NIH) under grant numbers R21 AI126037 (A.D.L. and A.G.B.) and U54 AI065359 (A.G.B.), by the Bay Area Lyme Foundation, and with institutional funds of the University of California Irvine and the University of South Carolina. This work was also made possible, in part, through access to the Genomics High Throughput Facility Shared Resource of the Cancer Center Support Grant (P30CA-062203) and NIH shared instrumentation grants 1S10RR025496-01, 1S10OD010794-01, and 1S10OD021718-01. We thank Parth Sitlani for technical assistance. We thank Russ Corbett for constructing the Hi-C libraries.

ABSTRACT OF THE THESIS

Scaffolding of a *Peromyscus Leucopus* Genome using Hi-C

By

Yuan Tao

Master of Science in Mathematical, Computational, Systems Biology

University of California, Irvine, 2018

Professor Anthony Long, Chair

White-footed mice (*Peromyscus leucopus*) are the critical host for black-legged ticks, which carry and spread the bacterium that causes Lyme disease. A high-quality genome assembly of *P. leucopus* is a prerequisite for the genetic experiments which will identify host factors that determine the suitability of *P. leucopus* to serve as a host. We created a high-quality assembly of the white-footed mouse genome by combining Hi-C data with an existing draft assembly to generate chromosome-length scaffolds for the *Peromyscus* genome. We then validated this new assembly by comparing it to a previously published genetic linkage map. A synteny analysis between our assembly and the published genome for the common brown rat (*Rattus norvegicus*) supports previous results that the two genomes are very similar. Furthermore, the assembly reveals that chromosomes 16 and 21 reside on a single scaffold, suggesting a possible fusion between the two chromosomes in the past. In summary, this new assembly represents a novel high-quality genomic resource for mouse research as well as a starting point for investigating the genotypic basis of the ability of *Peromyscus* to serve as a reservoir host for Lyme disease.

CHAPTER 1 - INTRODUCTION

White-footed mouse (*Peromyscus leucopus*), also known as the wood mouse, is a rodent that widely distributed among eastern and central North America. It belongs to the genus *Peromyscus*, which is distributed over almost every type of terrestrial habitats, and provides several examples of ecological adaptation [1]. Biological features of *Peromyscus* have resulted in this species being intensively studied, including its ecology, evolution, physiology, reproductive biology, cancer biology and behavioral neuroscience, giving it the label 'the *Drosophila* of North American mammology' [2].

Due to its large geographic range and long lifespan, *Peromyscus* species carry several pathogens important to public health. The deer mice (*Peromyscus maniculatus*) is a primary reservoir of Yersinia pestis [3] and an enzootic reservoir of plague [4] in California. There have been studies on white-footed mice as a key reservoir of *Borrelia burgdorferi*, which causes lyme disease in nature since 1980s [5, 6, 7]. This species is good to study for ecology of this pathogen, predictions about its geographic spread, and field-based prevention strategies. Recent field experimental study reveals that the occurrence of antibodies to outer surface protein A (OspA) is rare in field animals [8], which suggest that it could be an appropriate target for a field-administered transmission-blocking vaccine directed at *P. leucopus* and other reservoirs [9].

Despite increased attention to *P. leucopus* as the carrier of Lyme disease, a lack of genetic and genomic resources has limited studies in this field. In the case of the whole *Peromyscus* genus, the NCBI genome database shows a total of 1 draft genome and 8 organelle genomes. Previously, a *Peromyscus maniculatus bairdii* genome was reported to be 2,473.54 Mb with over 30 thousand scaffolds [10]. A linkage map of *Peromyscus* has

been created using hybrids between *P. maniculatus* females and *P. polionotus* males consisting of 196 markers [11]. However, the detailed gene structures and comparative genomics of *Peromyscus* species are poorly explored and reported. And there is not even a draft assembly for *P. leucopus*, the reservoir for Lyme disease.

Without whole-genome sequencing, the genomic basis of key adaptations remains difficult to identify. In the last few years, the developments in genome sequencing technologies and the improvements of assembly algorithms make it possible to study non-model organisms. A hybrid approach to sequencing and assembly using both short and long reads [12] allowed for a hybrid assembly of the *P. leucopus* genome (unpublished). This assembly is 2,474.05 Mb in total length and consists of 4216 contigs (thus it is already a considerable improvement over the Baylor assembly). The goal of my Master's thesis was to see if we could use the long-range contact information from Hi-C to obtain chromosome sized scaffolds in a cost-efficient way [13, 14, 15].

In this study, de novo assembly of the white-footed mouse genome is presented, and its genome is described, including a comparison with the linkage map of its close relative in *Peromyscus* genus, the deer mouse (*Peromyscus Leucopus*) and a synteny map with the genome of its distant but better sequenced relative in rodent family, rat (*Rattus Norvegicus*) [16]. I was able to use Hi-C libraries to obtain chromosome length scaffolds for *P. leucopus*, and thus greatly enhance the value of this genomic resource. Based on this improved highly contiguous annotated assembly future RNA-seq experiments can be better contextualize. Furthermore, it is now possible to carry out population genetics and look at things like linkage disequilibrium in wild and colony animals.

CHAPTER 2 - METHODS AND MATERIALS

3d-DNA

We used 3d-dna to scaffold the genome using our Hi-C libraries [14]. We set aside the contigs shorter than 10kb (as per recommendations) and then used Hi-C data to split, anchor, order, and orient the remaining contigs. We set the iterative number of misjoin corrections to be 7, compared to 2 for the human genome and 9 for the *Aaedes aegypti* genome [14]. We suspect *P. leucopus* has 24 chromosomes [11], the same number of chromosomes as *Peromyscus maniculatus*. We ran the 3d-dna software with chromosome varying from 22-27, and chose the best final scaffolding based on aligning scaffolds to the *P. maniculatus* linkage map (below). To examine the impact of sequence coverage on scaffolding we both down-sampled the p1 and p2 Hi-C libraries by half, and conversely pooled the two libraries (called pp) to increase coverage.

Salsa

Since we noticed that 3d-dna made a lot of local modifications (i.e., rearranged blocks within finished high confidence contigs), we wanted to know if other Hi-C scaffolding programs also have this problem. We tried SALSA [15] to assemble the genome using the same set of Hi-C libraries (p1, p2, pp). To compare the effect of misjoin corrections on finished contigs, we ran the SALSA software with and without misjoin correction.

Validating scaffolds using a *Peromyscus maniculatus* linkage map

To validate our scaffolds and to assign scaffolds to traditional chromosome names we employed *P. maniculatus* linkage map [11], a close relative to *P. leucopus*. We placed 196 of

the genetic markers from the linkage map that mapped consistently onto our scaffolds using bwa [17] and bowtie2 [18]. We visually compared the marker orders in the linkage map and scaffolds.

Comparing different scaffolding approaches

We measured the robustness of the scaffolding via dotplots as implemented in Mummer v4 [19] and visualized Hi-C contact matrix using juicebox [20]. For each both 3d-dna and Salsa, we scaffolded using three different set of libraries (p1, p2, pp). We scaffolded both with and without misjoin correction switch in SALSA. Since we don't know the true genome for this species, we examined robustness by comparing the assemblies to one another. We further summarized the performance of each scaffolding approach using BUSCO v3 [21].

Syntenic analysis

Synteny maps were created using SynChro [22]. We mapped all predicted ORFs to our scaffolds using blat [23]. The *Rattus Norvegicus* data were downloaded from Genbank Assembly Database (GCA_000001895.4). We modified the synteny plotting code to simplify the resulting figures.

CHAPTER 3 - RESULTS

Hi-C scaffolded genome

Chromosome number

After we ran 3d-dna with different target number of chromosomes, a chromosome number of 24 (matching out expectation) results in the most consistent splitting of chromosomes. With N=24, chromosome 8 is split into two scaffolds 8a and 8b. These two scaffolds are consistent with the linkage map of *P. maniculatus* that splits chromosome 8 into two linkage groups. In that mapping paper, the authors showed via *in situ* hybridizations to metaphase chromosome spreads that linkage groups 8a and 8b map to the same physical chromosome. It highlights the need to study why neither linkage mapping nor scaffolding is able to join those two regions. Furthermore, with N=24, chromosomes 16 and 21 are fused into a single scaffold. Increasing the number of chromosome does not split these two chromosomes, and a visual examination of the Hi-C contact data for the presumed regions where the join occurs suggests these two regions are physically attached (Figure 3.4A). We conclude that in *P. leucopus* there is a chromosome fusion relative to *P. maniculatus* or that the *P. maniculatus* linkage map is incorrect.

Varying Hi-C coverage and source libraries produces different scaffolds

As we expected, Hi-C libraries play an important role in the scaffolding results. We obtained 10X and 4X of raw sequence coverage (and 341,768X and 140,112X of span coverage) from two independently constructed Hi-C libraries. Our total coverage was about double that of the mosquito Hi-C scaffolding experiment in the original 3D-DNA study [14], seemingly in the recommended range of coverage. Furthermore, the N50 of our contigs is comparable to

those employed in the literature. In contrast to the published "state of the art" we employed two independently constructed libraries - there is currently no community consensus on the value of biological replicates of Hi-C libraries. We compared the assemblies of chromosome 3 using dot plots to looks at the effect of coverage, replication, assembler, and assembler switches on the resulting assemblies (Figure 3.1). Three patterns are apparent. First, regardless of approach and data Hi-C scaffolding is robust in a longer-range ordering, but prone to local inversions and re-arrangement. These events are visibly obscured in dot-plots of entire genomes versus a known reference (since the errors plot very close to a diagonal), it would be valuable in future work to have some numerical "quality control" summary of the extent of this problem. Second, relative to Salsa, 3d-dna seems to produce assemblies that are more consistent with one another when comparing different libraries or down-sampled libraries. Finally, our scaffolds based on combining two independent libraries is the only example of an assembly that is robust to down-sampling. Although we cannot say with certainty, it seems like biological replicates of the Hi-C library, the 3D-DNA software, and higher total sequencing coverage than is generally employed may lead to a better set of final scaffolds.

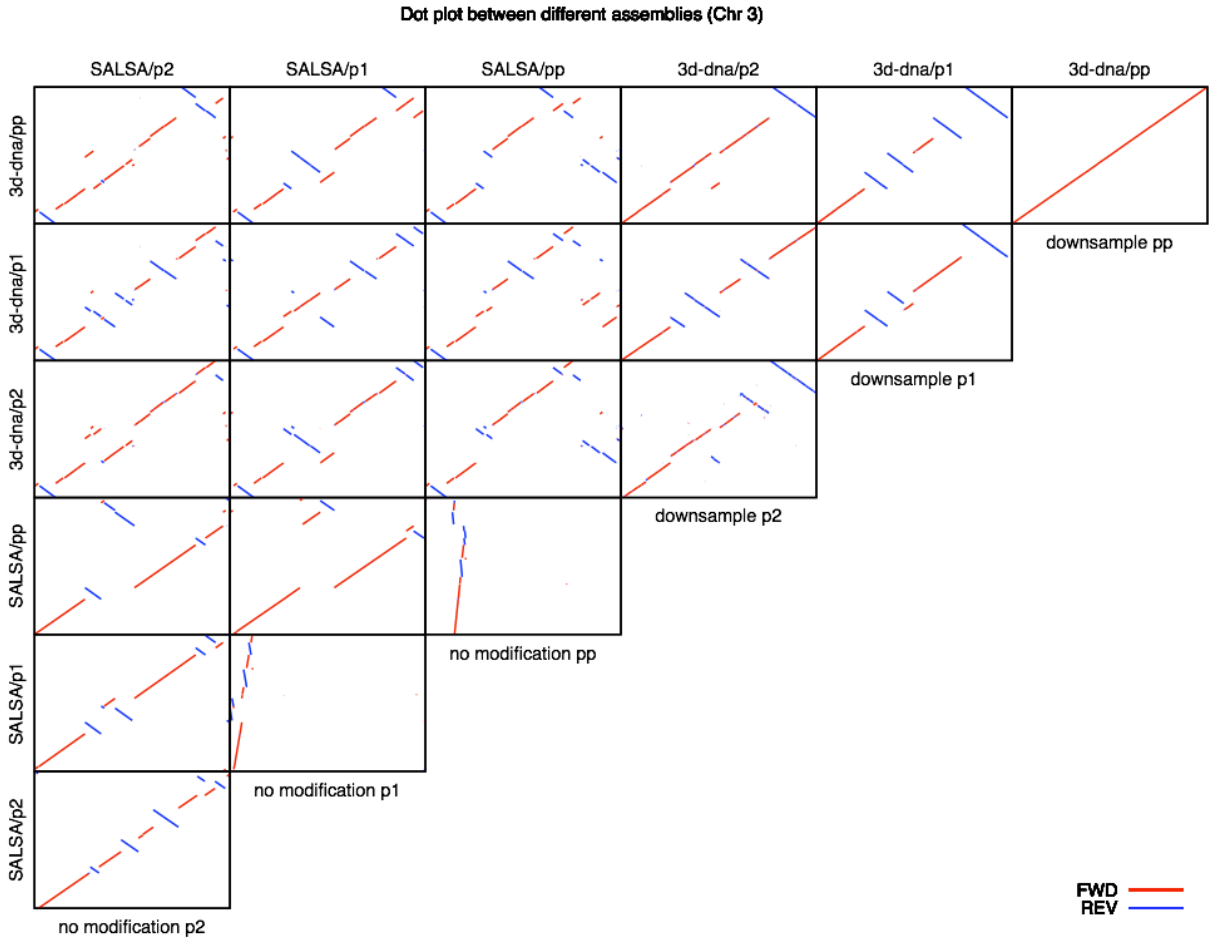

Figure 3.1 Dot plot between different assemblies (Chromosome 3)
SALSA and 3d-dna are the two different programs. Forward matches are plotted in red, reverse matches in blue. We see that comparisons between 3d-dna assemblies show that they are more consistent with one another than SALSA. The 3d-dna down-sampling on the diagonal shows that Hi-C library coverage can impact scaffolding, perhaps converging when both libraries are combined. The SALSA no modification experiment shows that allowing a scaffolder to split contigs has a large impact on the final assembly.

BUSCO analysis and cumulative scaffold length plot

We carried out a BUSCO analysis of our scaffolded genomes in order to quantitatively measure the annotation completeness of our assemblies (Table 3.1). We observe that scaffolding the genome has little impact on BUSCO scores, presumably since the large

majority of genes are located in contigs that are not impacted by scaffolding. Based on our examination of the data and BUSCO scores the assembly we present for the remainder of the manuscript is the one produced using Hi-C with N=24 chromosomes, 3d-dna, and the two libraries combined.

Table 3.1: Summary of BUSCO scores for different assemblies.
The 3d-dna/pp assembly produces the most complete single copy mammalian BUSCO scores. BUSCO determines the percentage of mis-assembled transcripts by trying to align transcripts from highly conserved proteins to each assembly. C: complete BUSCOs, S: complete single copy BUSCOs, D: complete duplicate copy BUCOs, F: fragment copy BUSCOs, M: missing copy BUSCOs. The table is sorted in descending order of the 'C' column. Scaffolding the genome only has a small impact on BUSCO scores.

Assembly	C	S	D	F	M
contigs	94.60%	92.10%	2.50%	2.10%	3.30%
3d-dna/pp	94.50%	92.50%	2.00%	2.00%	3.50%
SALSA/p2	94.40%	92.50%	1.90%	1.80%	3.80%
3d-dna/p1	94.30%	92.30%	2.00%	2.10%	3.60%
3d-dna/pp*	94.20%	92.20%	2.00%	2.10%	3.70%
SALSA/pp	94.10%	92.10%	2.00%	2.20%	3.70%
3d-dna/p2	93.80%	91.60%	2.20%	2.70%	3.50%

We also plotted the cumulative scaffold length of different scaffolded genomes (as well as the original contigs).

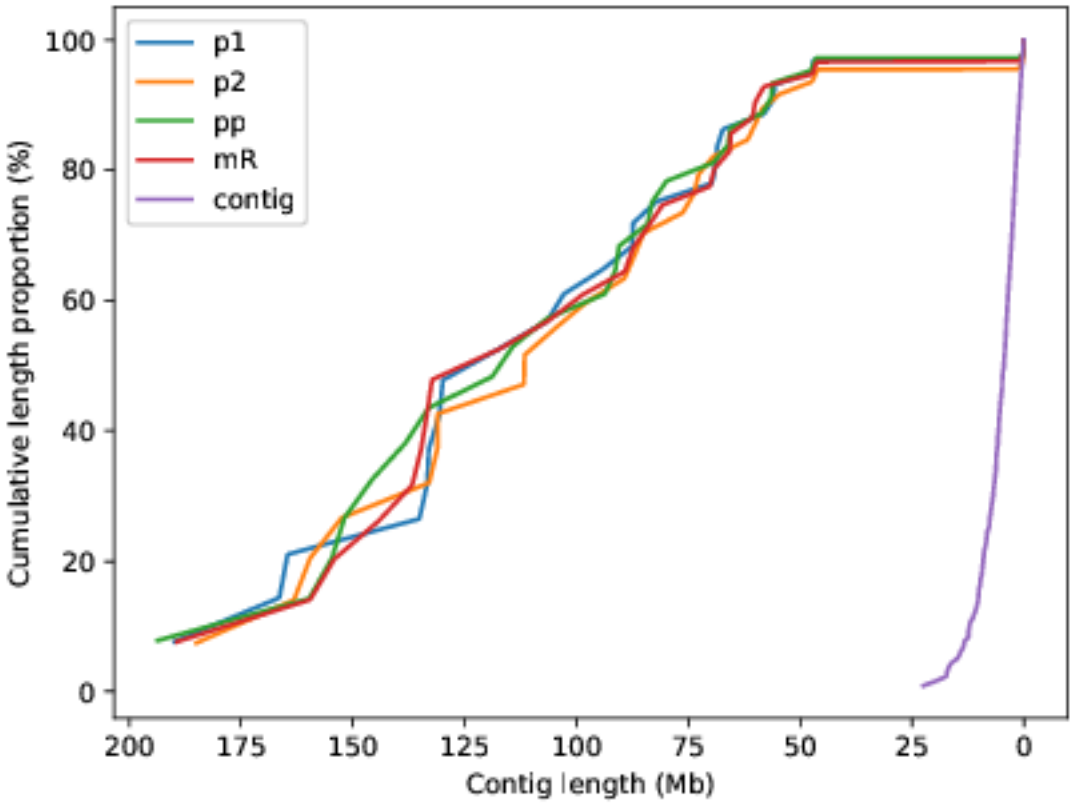

Figure 3.2: Cumulative scaffolds length plot of different 3d-dna assemblies.
The pp assembly has the largest N25 and N75.

Summary of scaffolding

Comparing different scaffolded result of 3d-dna and SALSA, the 3d-dna pp scaffolded genome assembly was chosen for highest coverage and better software. This assembly contained 24 chromosome length scaffolds. The total length of all chromosome length scaffolds was 2,405,334,684 bp, represents 97.18% of the assembly. The scaffold N50 lengths reached 114,273,790 (Table 3.2).

Table 3.2: Statistics for the *Peromyscus Leucopus* assemblies.
We set aside the tiny scaffolds, and assembled other scaffolds using Hi-C to create huge, chromosome-length scaffolds and additional small scaffolds.

	Draft contigs	Final assembly	Chromosome length scaffolds	Small scaffolds	Tiny scaffolds
Total base pairs	2,474,055,010	2,475,164,510	2,405,334,684	66,909,757	2,920,069
% total	100%	100%	97.18%	2.70%	0.12%
Number of contigs/scaffolds	4216	1856	24	1170	662
Min length	1,000	1,000	46,490,564	10,022	1,000
Mean length	586,825	1,333,602	100,222,279	57,188	4,411
N50 length	3,921,871	114,273,790	NA	85,932	8,455
Max length	22,229,712	193,658,164	193,658,164	861,000	10,000

Comparison to linkage map and synteny with Rat suggests a possible chromosome fusion relative to *P. maniculatus*

We compared our assembly to a genetic map of *P. maniculatus* [11]. Of the 196 markers in the genetic map, 129 markers could be mapped in the raw assembly contigs, while 137 markers mapped to the new assembly, 134 markers could be unambiguously mapped in chromosome length scaffolds (Figure 3.3). Our assembly agreed with the genetic map for 128 of these 134 markers mapped to the correct chromosome. Exceptions may due to the difference between the two species. Notably, there is a chromosome fusion in our scaffolding relative to *P. maniculatus* (Figure 3.4A). The next step was to check if the fusion is biological or due to the scaffolding methods.

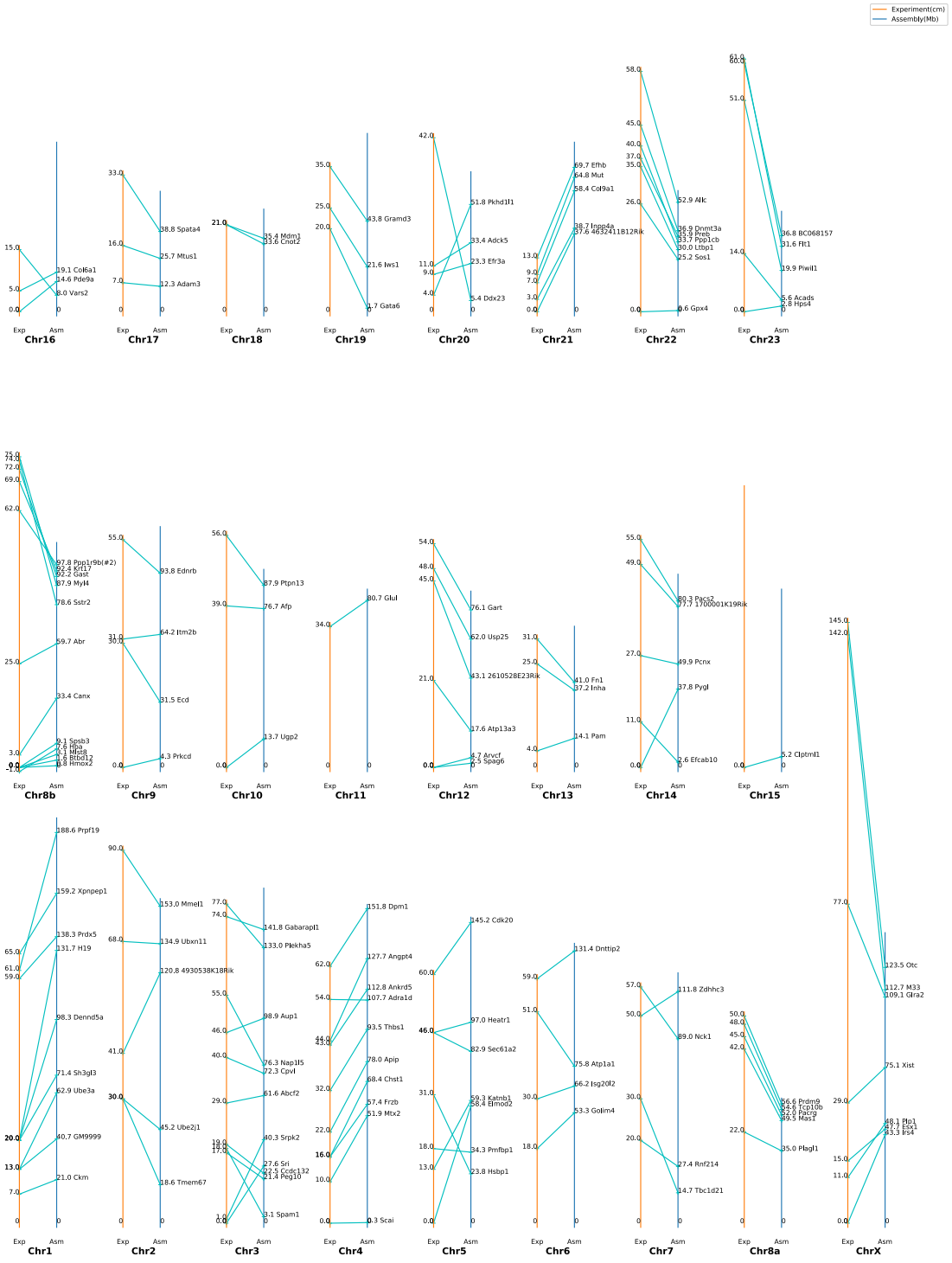

Figure 3.3: Comparative relationships of the linkage map with the molecular map of the *P. leucopus*

The linkage map is indicated in orange, while molecular map is in blue.

In order to check whether the fusion was real or due to the scaffolding methods, we visualized the Hi-C contact map of the proposed fusion scaffold. We see that the contacts between the two chromosomes are at the average level as other with-in chromosome contacts (Figure 3.4C). Thus, the Hi-C information supports the fusion event.

Then, we read about the comparative relationships of the genetic map of *Peromyscus* with the molecular map of *Rattus* and *Mus* [8]. We noticed that *P. maniculatus* chromosome 16 is orthologous to *Rattus* chromosome 20 while *P. maniculatus* chromosome 21 is orthologous to *Rattus* chromosome 9 (Figure 3.4). We did the syntenic analysis between *Rattus Norvegicus* and our assembly (*Peromyscus Leucopus*). Maps of our scaffolded *P. leucopus* assembly and of the 21 *R. norvegicus* chromosomes based on the positions of 9491 Reciprocal Best-Hits (RBHs) and 49,023 orthologous pairs demonstrate highly conserved synteny for the 2 species.

We found that our putative fusion scaffold is orthologous to both *Rattus* chromosome 20 and chromosome 9 (Figure 3.4C and Figure 3.5A). Furthermore, the BGBG ("brown-grey-brown-grey") area in the synteny plot is very suspicious (Figure 3.4B). We suspected it is a mis-orientation of one contig that caused the fusion.

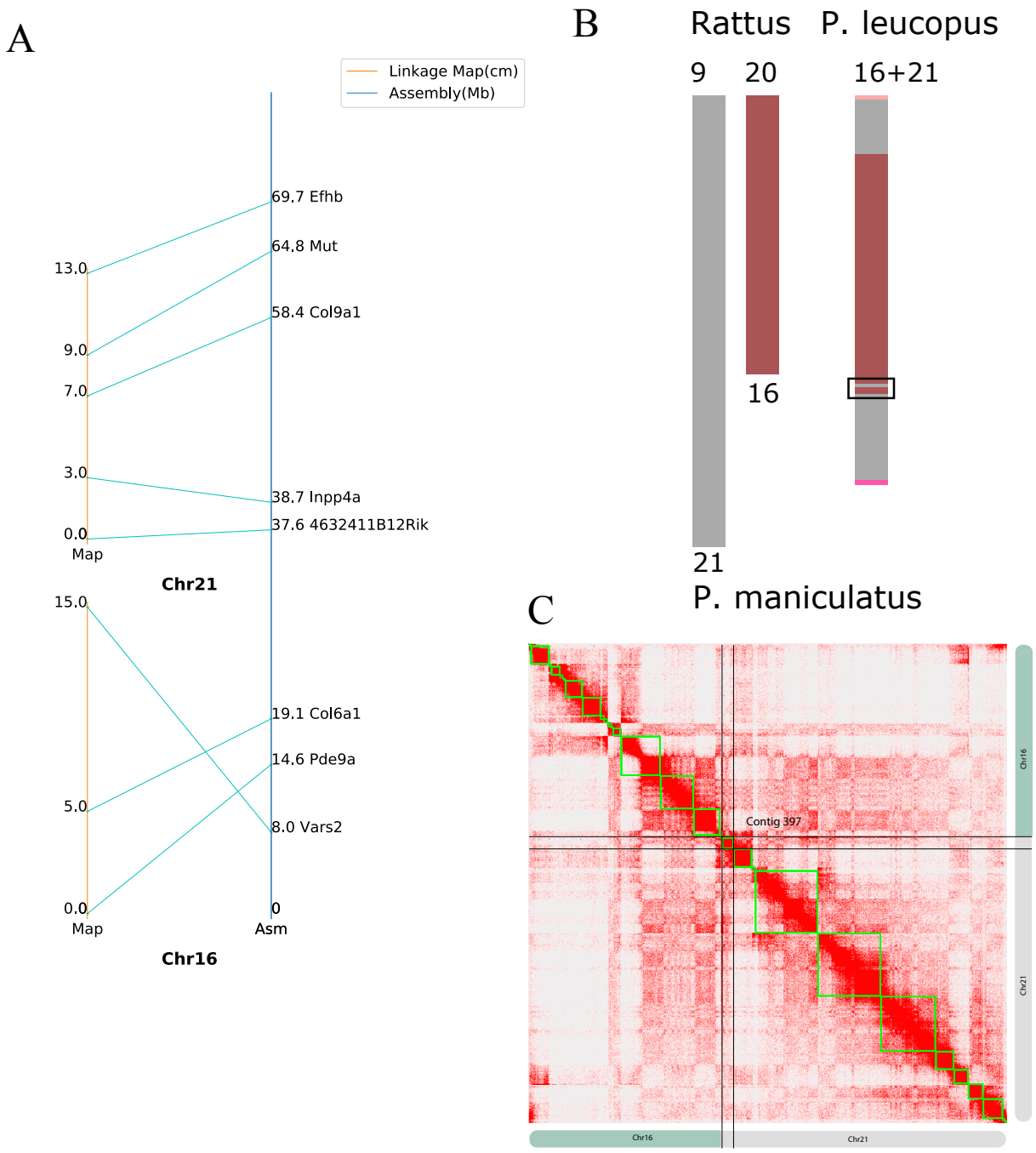

Figure 3.4: Various evidences about the proposed fusion
A) The proposed fusion scaffold mapping to linkage map. **B)** The synteny blocks of the proposed fusion scaffold between *R. norvegicus* and *P. leucopus.* **C)** The Hi-C contact map of the proposed scaffold.

We then located the switch area of the synteny map in the assembly. It's exactly one contig C397 with orthologs from the two chromosomes that causes the inverted GB ("gray-brown") pattern in the middle. We validated the guess by removing this contig from the raw assembly and repeating the scaffolding. The result showed that the fusion of two chromosomes still exists, which suggests that the fusion is not caused by this specific contig but the enormous contacts within the scaffolds, while the BGBG pattern has gone, leaving a clear BG ("brown-grey") pattern. We concluded that we detected the clues for a chromosome fusion and C397 is kind of "inverted" in the middle of it. More experiments are needed to validate them.

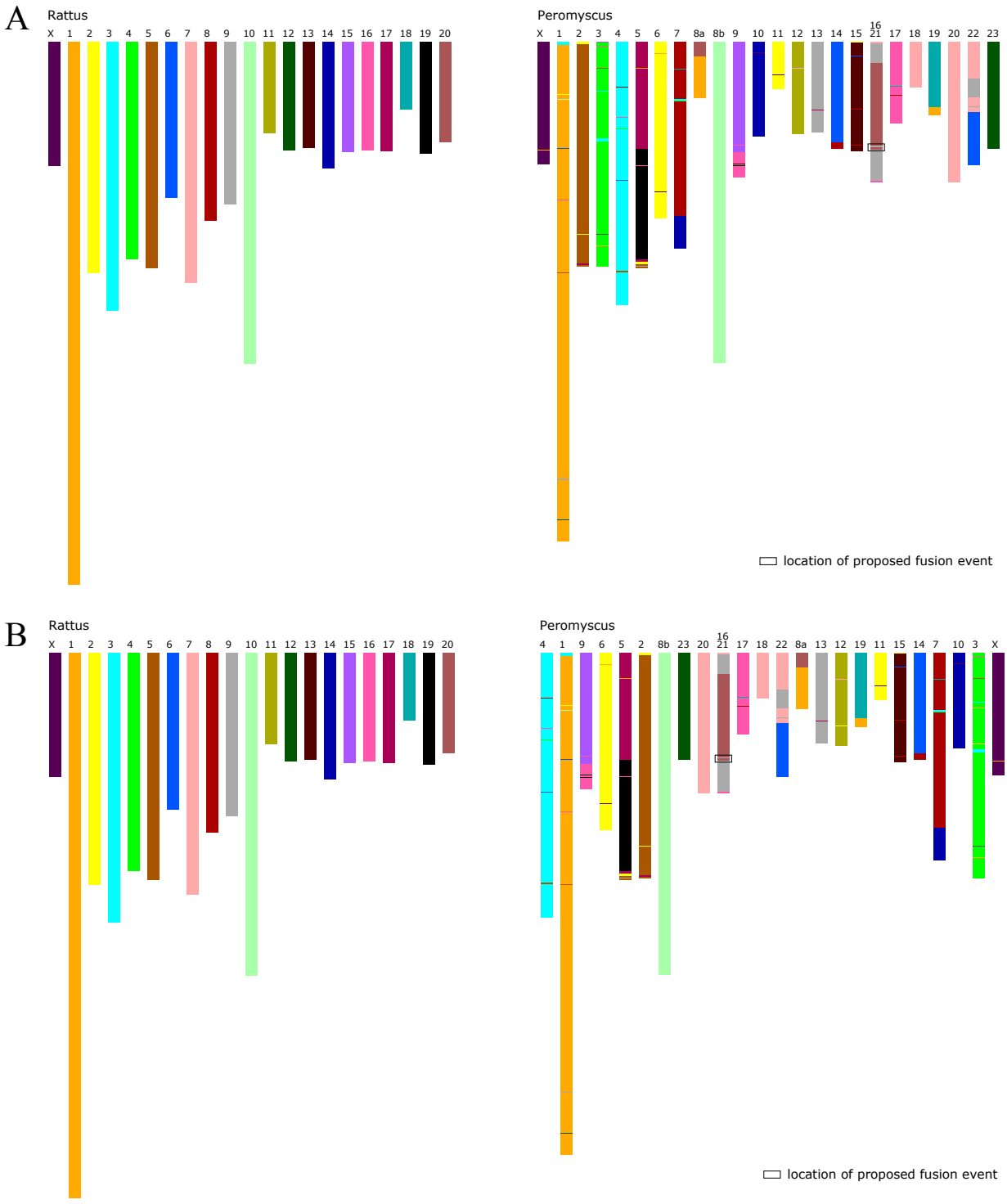

Figure 3.5 Synteny plots between *R. norvegicus* and *P. leucopus*.
A) *P. leucopus* genome in chromosome order. **B)** *P. leucopus* genome in Hi-C contact order. **Scaffolds unclearly separated in synteny plot**

The syntenic analysis between *R. norvegicus* and *P. leucopus* reveals an interesting pattern of the assembly (Figure 3.5B). As we rearranged the P. leucopus genome in scaffold order which is the order in Hi-C contacts map, we observed there are some "connections" at the ends of chromosomes. For example, the end of Chr4 and the start of Chr1 are in blue; the end of Chr5 and the start of Chr2 are in brown; the end of Chr20 and the start of Chr16+21 (the fusion chromosome) are in pink; the end of Chr16+21 and the start of 17 are in rose, etc. We concluded that it might be real or it is caused by unclear cutting at the edges.

CONCLUSIONS

We have successfully generated a high-quality genome scaffolding for *P. leucopus*. The scaffolded genome is of 2405.33 Mb in length, with 24 chromosome length scaffolds accounting for 97.18% of the total assembly. This scaffolded genome is a valuable resource for further research in *Peromyscus*. This assembly generally agrees with the published *Peromyscus maniculatus* linkage map, although our assembly places linkage groups 8a and 8b onto separate chromosomes, and fuses linkage groups 16 and 21 into a single chromosome. A syntenic analysis between *P. leucopus* and *Rattus Norvegicus* provides insight into genome structures and genome evolution. While further experiments will be necessary to validate the structural differences between our *P. leucopus* Hi-C scaffolded assembly and a *P. maniculatus* linkage map, we provide evidence that our conclusions are valid.

On the basis of our results, Hi-C libraries allow for assemblies consisting of chromosome-sized scaffolds. Obtaining such scaffolds is both rapid and inexpensive. Despite its good performance over long distance, it is important to bear in mind that the best Hi-C scaffolder (3D-DNA) may be necessarily introducing false positive re-arrangements and other edits into pre-existing high quality contigs. Furthermore, it seems very likely that our final assembly contains local re-arrangements and inversions. It is conceivable that deeper coverage and replicate Hi-C libraries can reduce the rate at which these errors occur but these are not standard practices in the field (nor do we prove this).

And we observed a clear pattern of scaffolds unclearly separated in their boundaries in the Hi-C contact map order (Figure 3.5B). We doubt if it is a problem caused by Hi-C scaffolding unable to cut in the boundary or if it is real 3d connection revealed by Hi-C.

Clearly fixing these problems is a question that future Hi-C assemblers should address based on scaffolding genomes.

Our scaffolded genome will help RNA-seq analyses in this system, and provide a framework for linkage mapping host / pathogen interacting genes. Ultimately this scaffolded genome will allow the genetic identification of genes that mediate host response to pathogens, identify genomic regions suitable for Crispr-based mutagenic chain reaction constructs, and allow us to characterize the the population genetics of this fascinating species.

Bibliography

1. Bedford NL, Hoekstra HE (2015) Peromyscus mice as a model for studying natural variation. Elife. https://doi.org/10.7554/eLife.06813
2. Dewey MJ, Dawson WD. Deer mice: "the *Drosophila* of North American Mammalogy" Genesis. 2001; 29:105–109. doi: 10.1002/gene.1011.
3. Pollitzer R, Meyer KF (1961) The ecology of plague. In: Studies in Disease Ecology, May JM (editor) New York: Hafner Publishing Company, pp 433-590
4. Danforth, M., Tucker, J., & Novak, M. (2018). The Deer Mouse (Peromyscus maniculatus) as an Enzootic Reservoir of Plague in California. EcoHealth.
5. Levine JF, Wilson ML, Spielman A. Mice as reservoirs of the Lyme disease spirochete, Am J Trop Med Hyg, 1985, vol. 34 (pg. 355-60)
6. Donahue JG, Piesman J, Spielman A. Reservoir competence of whitefooted mice for Lyme disease spirochetes, Am J Trop Med Hyg, 1987, vol. 36 (pg. 92-6)
7. Bunikis, J. *et al. Borrelia burgdorferi* infection in a natural population of *Peromyscus leucopus* mice: a longitudinal study in an area where Lyme borreliosis is highly endemic. *J. Infect. Dis.***189**, 1515–1523 (2004)
8. Kurtenbach K, Dizij A, Voet P, Hauser P, Simon MM. Vaccination of natural reservoir hosts with recombinant lipidated OspA induces a transmission-blocking immunity against Lyme disease spirochaetes associated with high levels of LA-2 equivalent antibodies, Vaccine, 1997, vol. 15 (pg. 1670-4)
9. Baum, E., Hue, F., & Barbour, A. G. (2012). Experimental Infections of the Reservoir Species *Peromyscus leucopus* with Diverse Strains of *Borrelia burgdorferi*, a Lyme Disease Agent. mBio, 3(6), e00434–12. http://doi.org/10.1128/mBio.00434-12
10. Baylor College of Medicine, 2013 https://www.ncbi.nlm.nih.gov/assembly/GCF_000500345.1/
11. Kenney-Hunt, J., Lewandowski, A., Glenn, T. C., Glenn, J. L., Tsyusko, O. V., O'Neill, R. J., et al. (2014). A genetic map of Peromyscus with chromosomal assignment of linkage groups (a Peromyscus genetic map). Mammalian Genome, 25(3-4), 160–179. http://doi.org/10.1007/s00335-014-9500-8
12. Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Research, 44(19), e147. http://doi.org/10.1093/nar/gkw654
13. Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nature Biotechnology, 31(12), 1119–1125. http://doi.org/10.1038/nbt.2727
14. Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science, 356(6333), 92–95. http://doi.org/10.1126/science.aal3327
15. Ghurye, J., Pop, M., Koren, S., Bickhart, D., & Chin, C.-S. (2017). Scaffolding of long read assemblies using long range contact information. BMC Genomics, 18, 527. http://doi.org/10.1186/s12864-017-3879-z
16. Ramsdell, C., Lewandowski, A., Glenn, J., Vrana, P., O'Neill, R., and Dewey, M. (2008). Comparative genome mapping of the deer mouse (Peromyscus maniculatus) reveals

greater similarity to rat (Rattus norvegicus) than to the lab mouse (Mus musculus). BMC Evol. Biol. 8:65. doi: 10.1186/1471-2148-8-65
17. Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics, Epub.
18. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012, 9:357-359.
19. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, et al. (2018) MUMmer4: A fast and versatile genome alignment system. PLOS Computational Biology 14(1): e1005944. https://doi.org/10.1371/journal.pcbi.1005944
20. Neva C. Durand, James T. Robinson, Muhammad S. Shamim, Ido Machol, Jill P. Mesirov, Eric S. Lander, and Erez Lieberman Aiden. 2016. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Systems 3(1).
21. Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., et al. (2017). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Molecular Biology and Evolution, 35(3), 543–548. http://doi.org/10.1093/molbev/msx319
22. Drillon, G., Carbone, A., & Fischer, G. (2014). SynChro: A Fast and Easy Tool to Reconstruct and Visualize Synteny Blocks along Eukaryotic Chromosomes. PLoS ONE, 9(3), e92621–8. http://doi.org/10.1371/journal.pone.0092621
23. Kent WJ. BLAT - the BLAST-like alignment tool. *Genome Res.* 2002 Apr;12(4):656-64.