Assembly of Complete KIR Haplotypes from a Diploid Individual by the Direct Sequencing of Full-Length Fosmids

Richard J. Hall[1], Kevin Eng[1], Chul-woo Pyo [2],; Dave Roe[3], Martin Maiers[3] Cynthia Vierra-Green[4], Daniel E. Geraghty[2], Swati Ranade[1]
[1]Pacific Biosciences, Menlo Park, USA, [2]Fred Hutchinson Cancer Research Center, Seattle, USA, [3]National Marrow Donor Program, USA, [4]Immunobiology Research, Center for International Blood and Marrow Transplant Research, Minneapolis, USA

Abstract

We show that linearizing and directly sequencing full-length fosmids simplifies the assembly problem such that it is possible to unambiguously assemble individual haplotypes for the highly repetitive 100-200 kb killer Ig-like receptor (KIR) gene loci of chromosome 19. A tiling of targeted fosmids can be used to clone extended lengths of genomic DNA, 100s of kb in length, but repeat complexity in regions of particular interest, such as the KIR locus, means that sequence assembly of pooled samples into complete haplotypes is difficult and in many cases impossible.

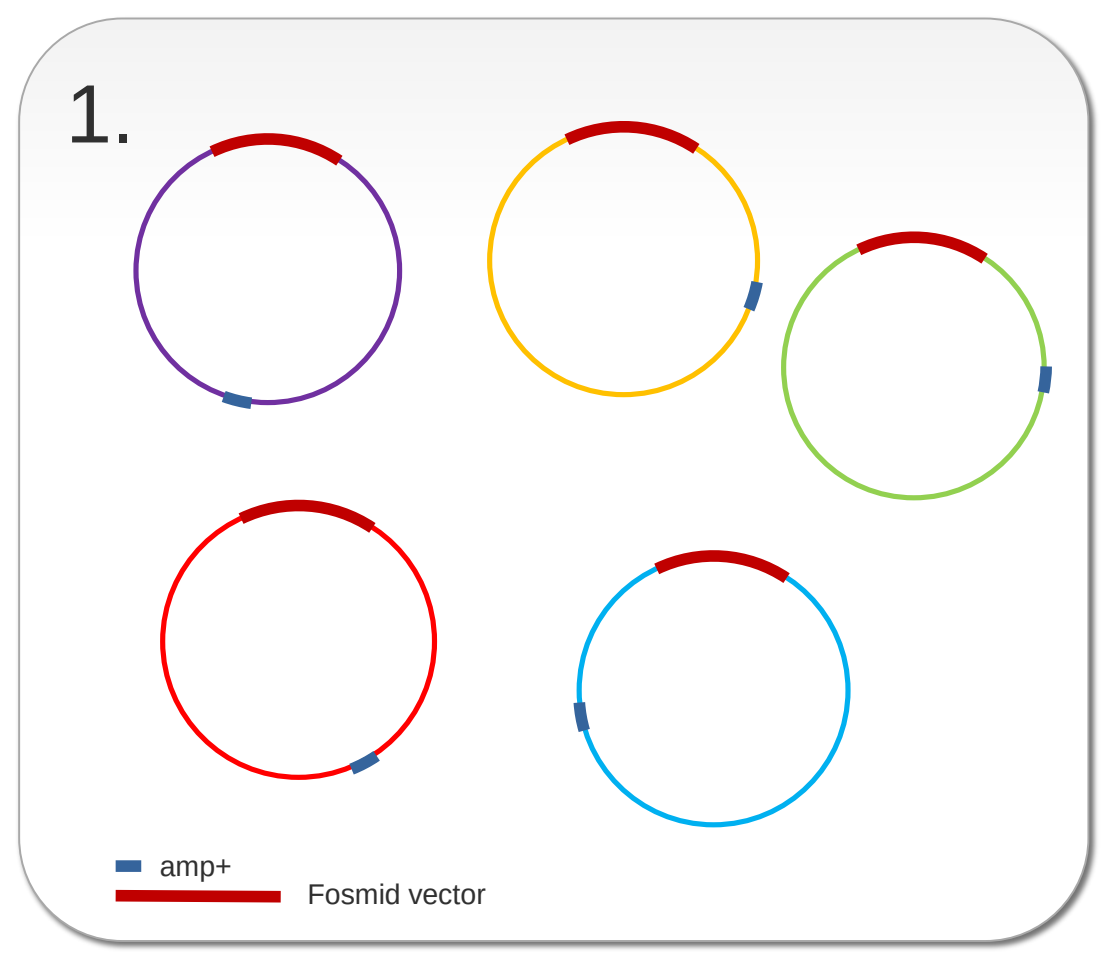
The current maximum read length generated by SMRT® Sequencing exceeds the length of a 40 kb fosmid; it is therefore possible to span an entire fosmid in one sequencing read. Shearing, sequencing and assembling fosmids in a shotgun approach is prone to errors when the underlying sequence is highly repetitive. We show that it is possible to directly sequence linearized fosmids and generate a high-quality consensus by simple alignment, removing the need for an error-prone assembly step. The high-quality sequence of complete fosmids can then be tiled into full haplotypes.

We demonstrate the method on DNA samples from a number of individuals and fully recover the sequence of both haplotypes from a pool of KIR fosmids. The ability to haplotype and sequence complex immunogenetic regions will bring exciting opportunities to explore the evolution of disease associations of the immune sub-genome. This simple and robust approach can be scaled-up allowing a complex genomic region to be sequenced at a population level. We expect such sequencing to be valuable in disease association research.

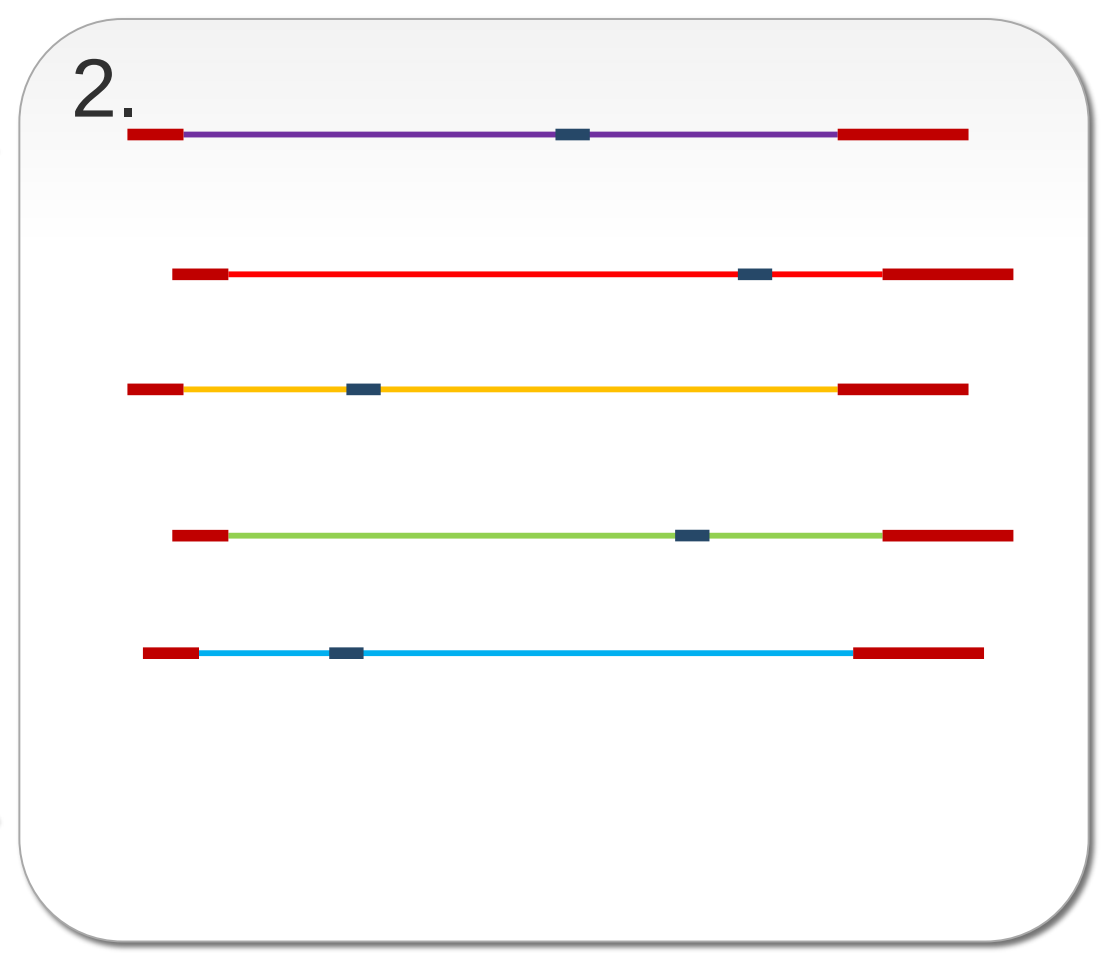
Library Generation

A set of fosmids covering both haplotypes of the KIR region were selected (http://sciscogenetics.com/technology/fosmid-based-resequencing/). While it is possible to shear and shotgun sequence each individual fosmid, the repetitive nature of the KIR region results in assembly problems when processing data from fosmids pooled before library preparation. With the number of fosmids for an individual in the ~11-23 range, preparing libraries and sequencing each individual fosmid is not practical when studying multiple individuals.

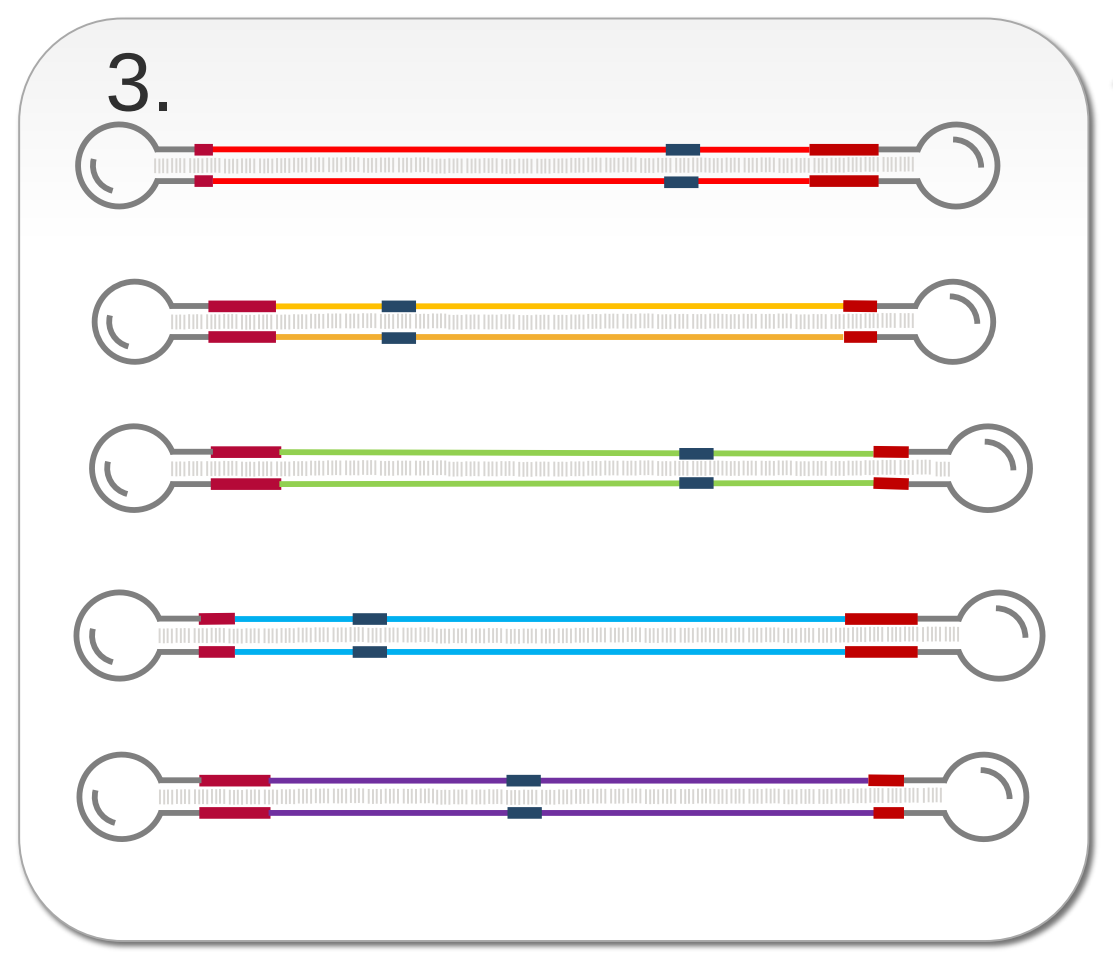
The assembly problem can be circumvented if the library preparation forgoes shearing for linearizing full-length fosmids, allowing multiple fosmids to be sequenced in a single library preparation.



1. Fosmids covering both haplotypes for a single individual are isolated and pooled. A fosmid vector joins the beginning and end of the target sequence. An amp+ selection vector is placed randomly within the target sequencing.

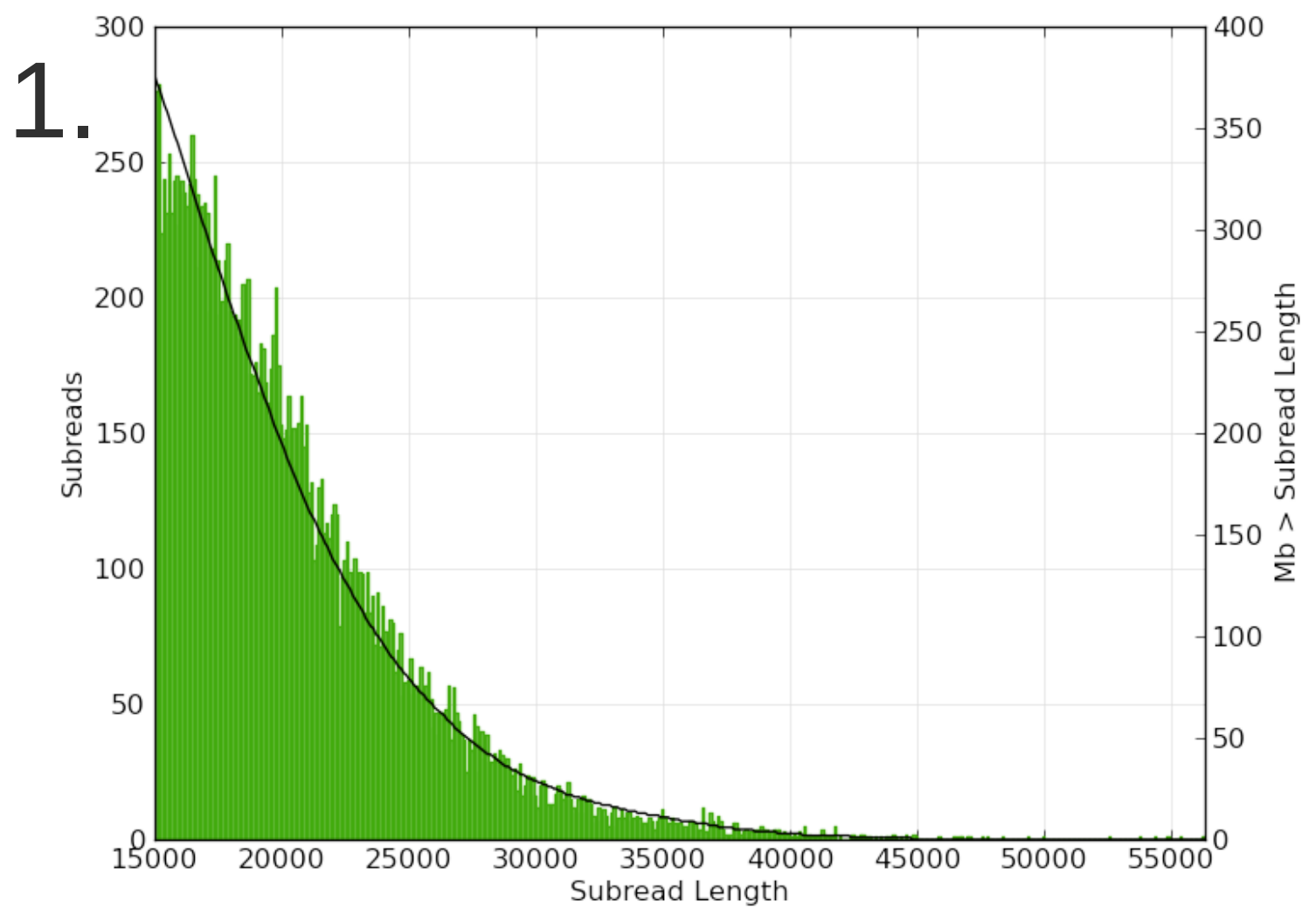


2. A restriction site within the fosmid vector is used to linearize the circular fosmids.



3. Linearized fosmids are blunt ended and used as input to standard SMRT® Sequencing template prep, resulting in full-length fosmid SMRTbell™ templates with vector sequence adjacent to the hairpin adapter.

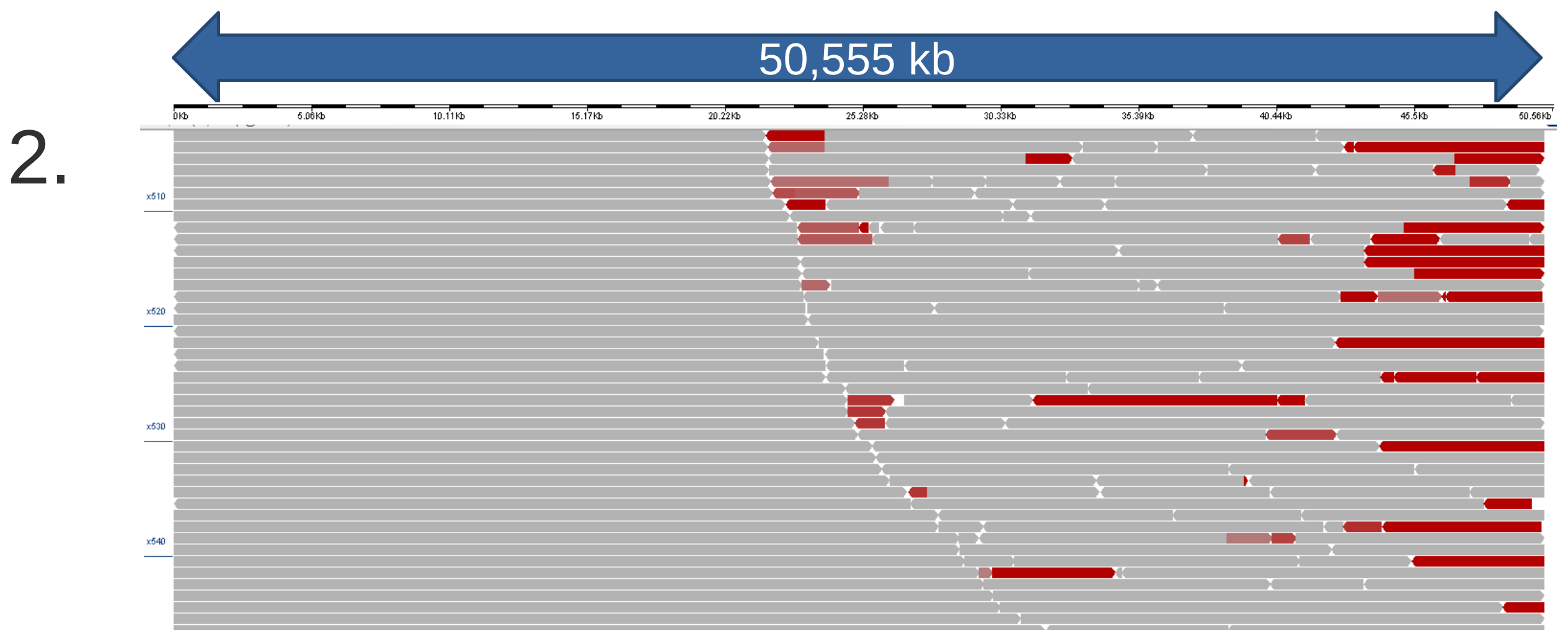
Sequencing



Mean Subread Length	20,462
Total Bases	381,846,530
Subread N50	20,247
Number of Reads	18,661

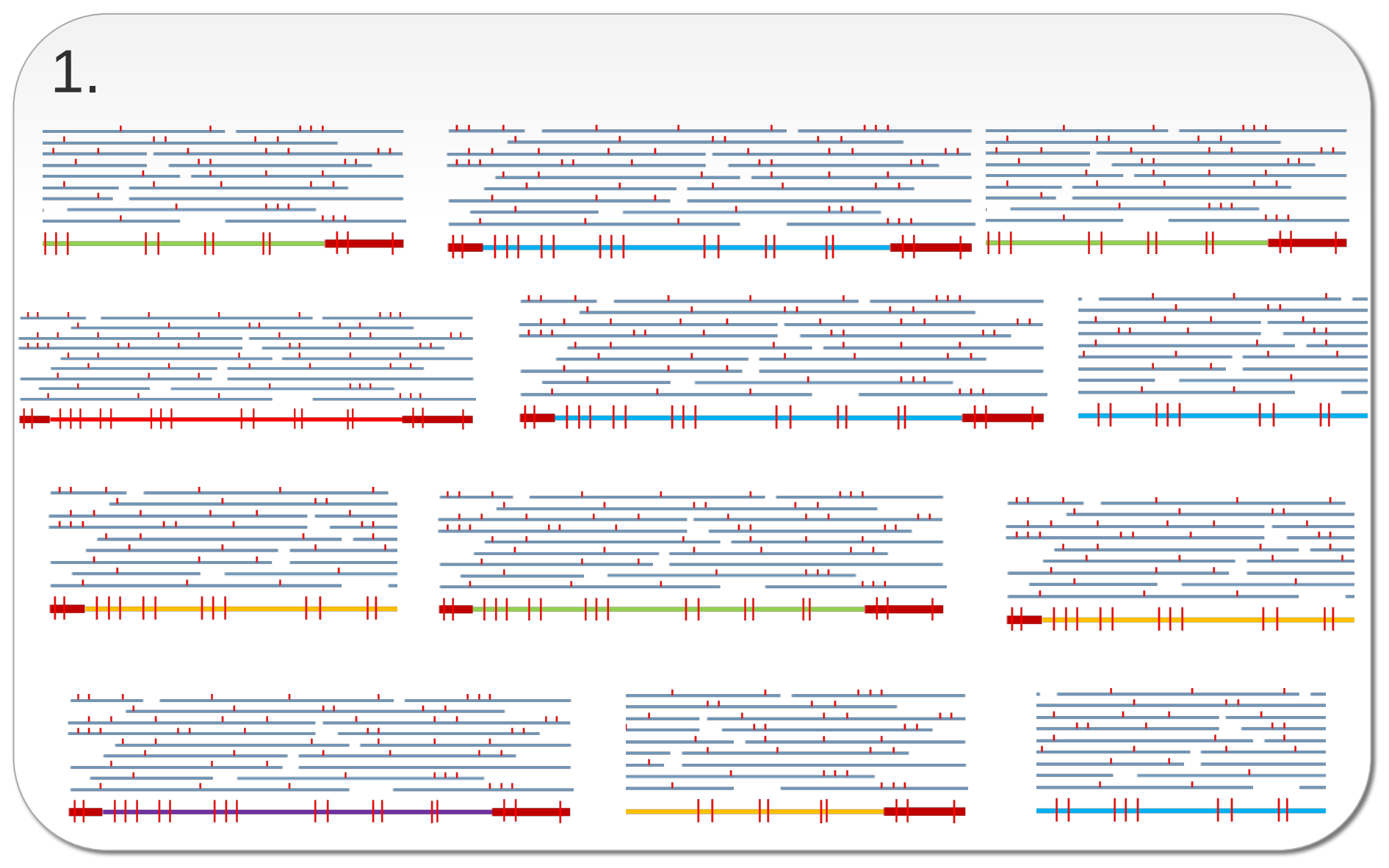
*Subreads filtered at 15,000

1. Example from the sequencing of two SMRT® Cells for 11 pooled fosmids for a single individual. Statistics are shown after an initial >15 kb filtering of subreads. The subreads in the extreme of the distribution are long enough to span the entire fosmid within the SMRTbell template structure.



2. Mapping of subreads to a ~50 kb read spanning a complete fosmid.

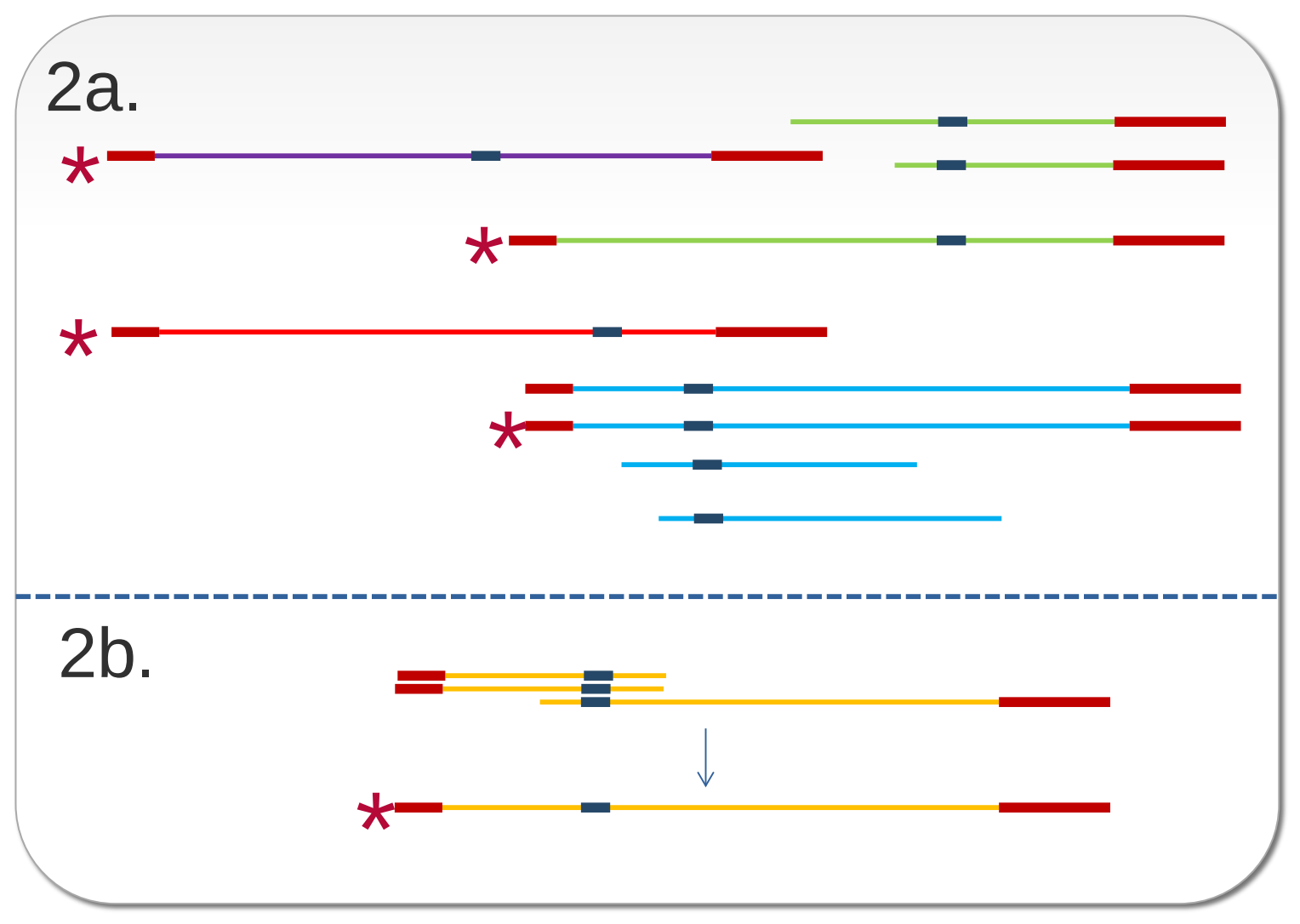
Data Processing

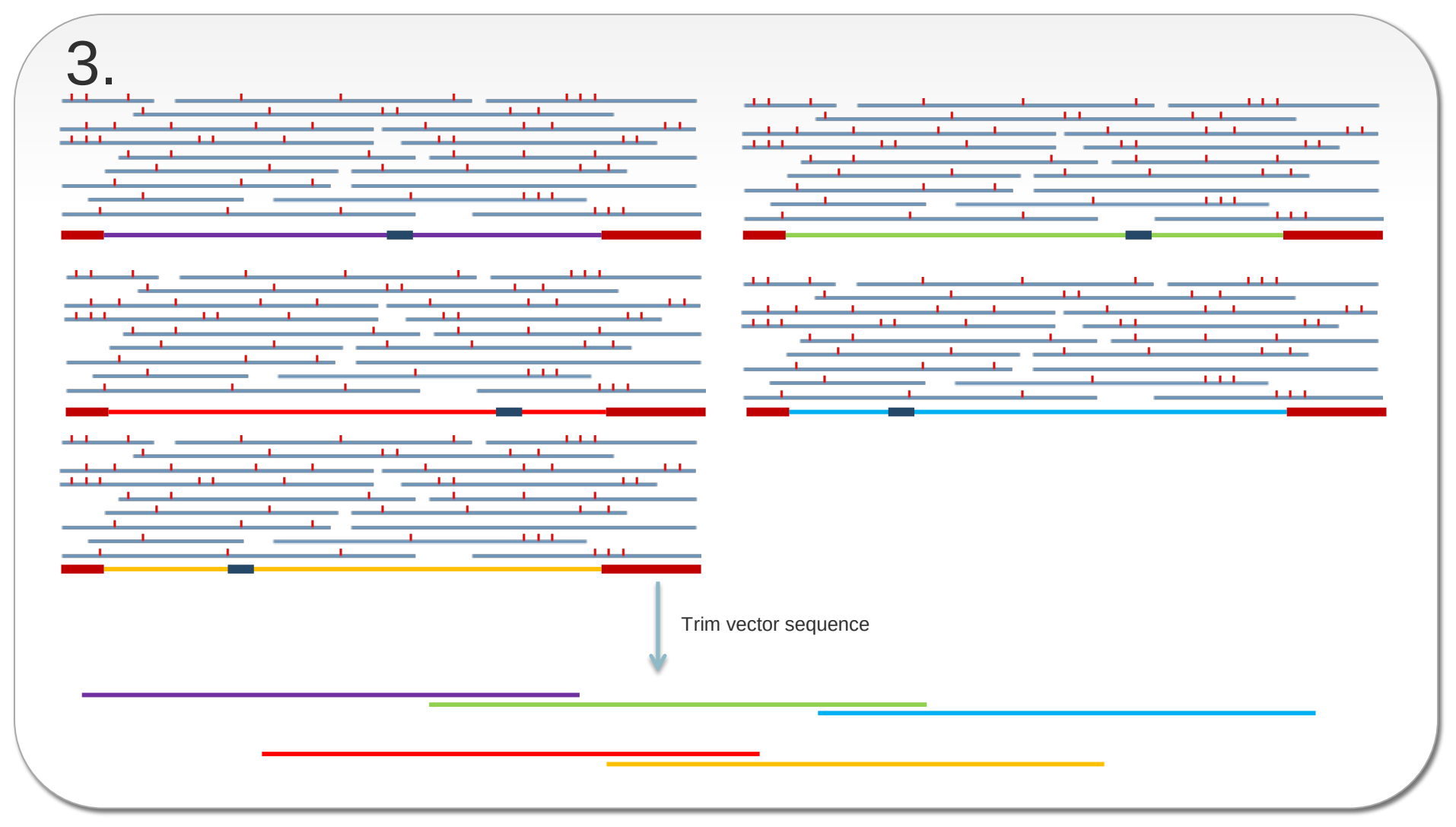


1. Error correction of subreads that possibly contain the sequence of a full-length fosmid. Subreads >15 kb are mapped to seed reads >25 kb and a consensus is calculated. Note selection of seeds is based on length and subreads that do not span a complete fosmid are also error corrected.

2a. Filtering for a set of full-length fosmids, i.e. an error corrected read spanning fosmid vector sequences, including the amp+ selection vector. Representatives are selected for redundant fosmid sequences.

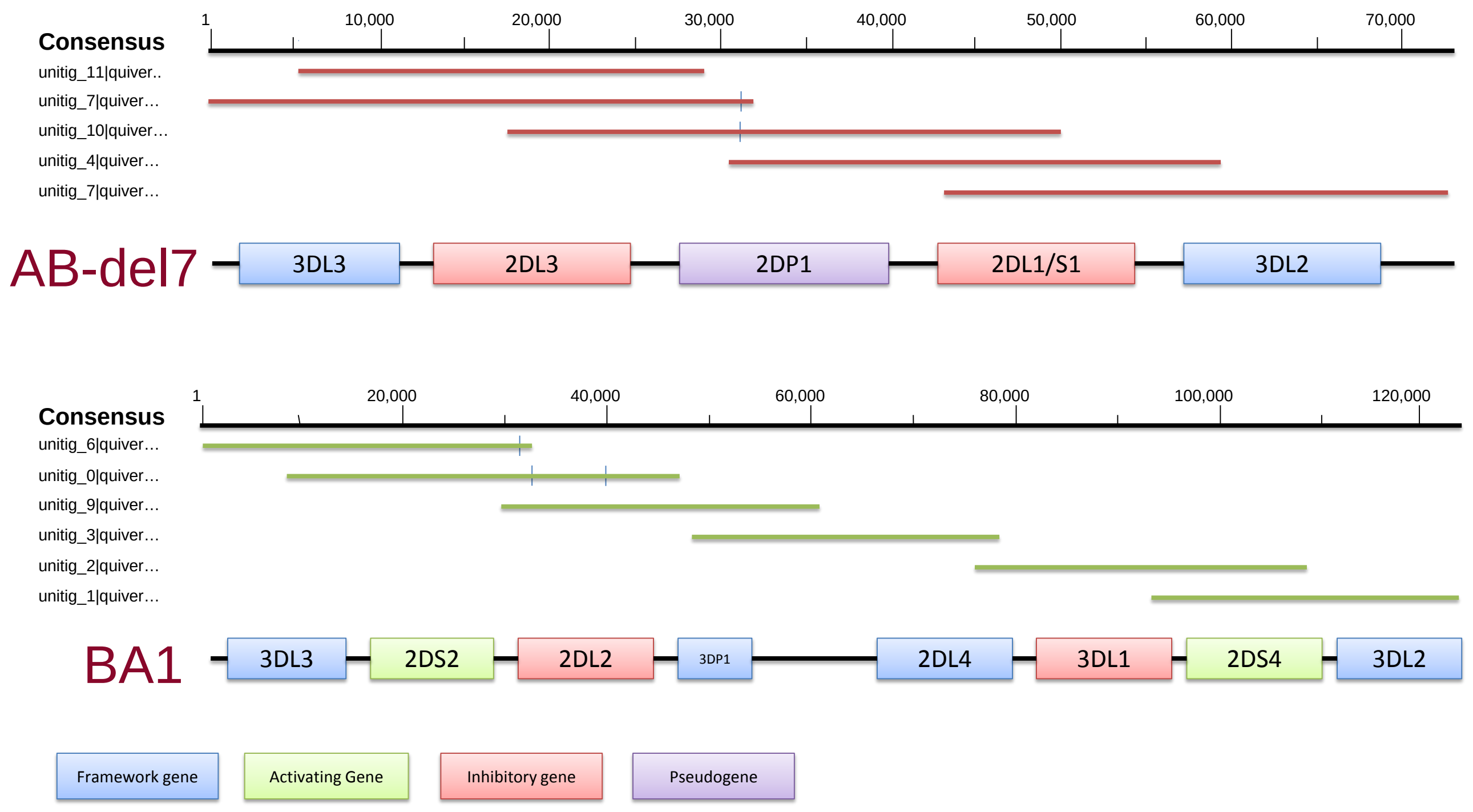
2b. If a full-length fosmid read is not obtained, an overlapping of half-length fosmid sequences can be used to generate complete sequence for the remaining fosmids.





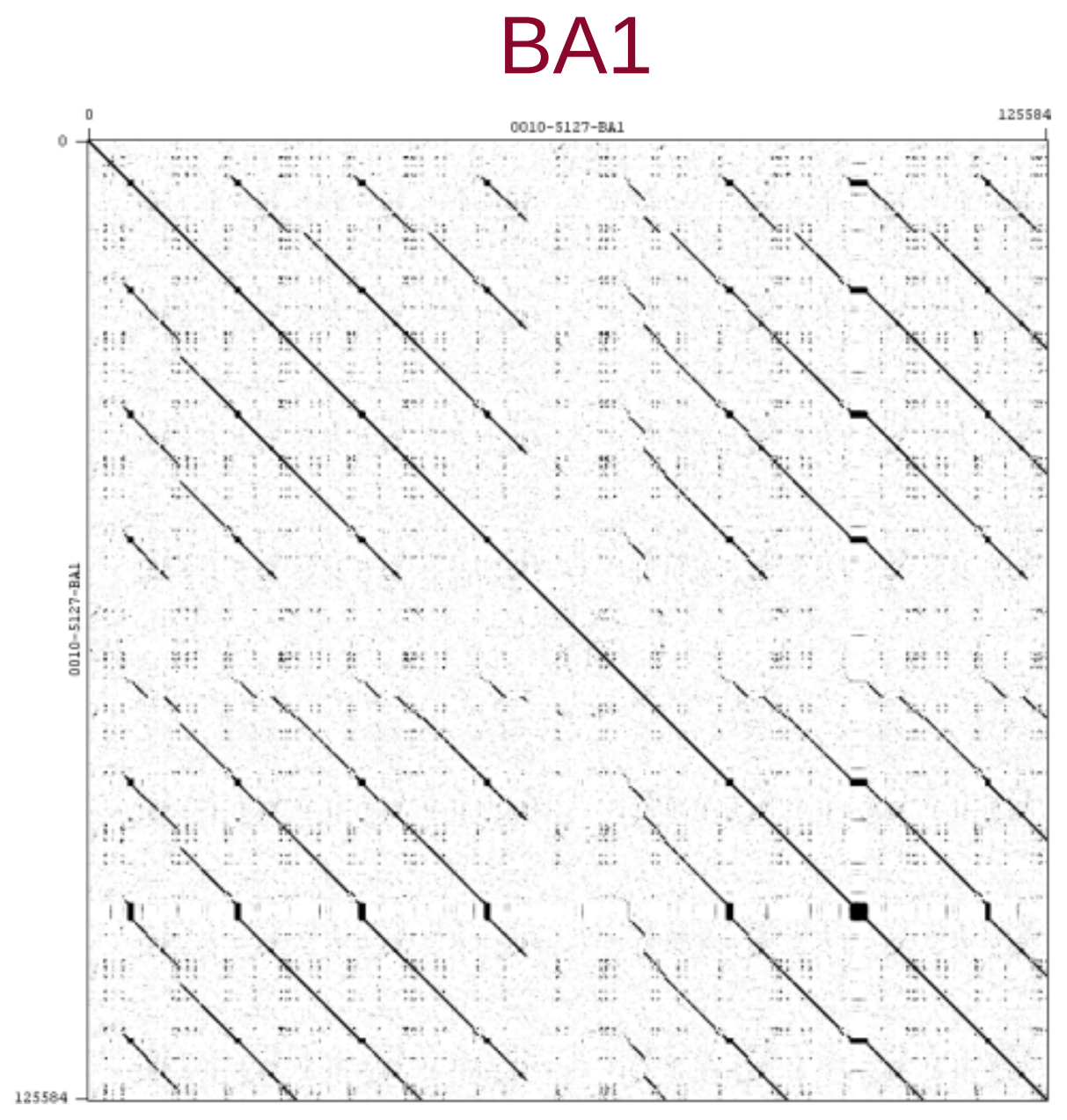
3. Re-mapping of all subreads to the error-corrected, full-length fosmids and Quiver consensus[1] generation results in >QV50 (99.999%) fosmid sequences. After vector trimming, high-quality fosmid sequences can be overlapped into continuous haplotypes.

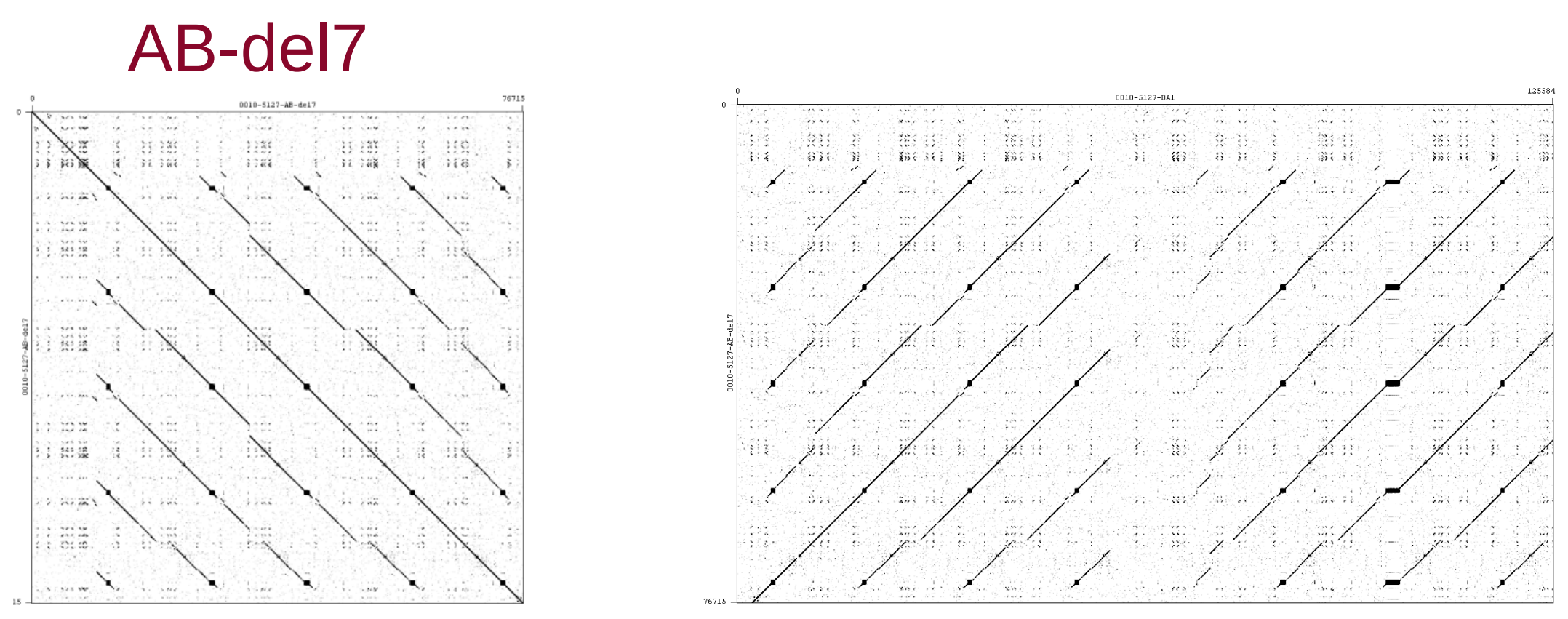
Results



Results of sequencing 11 fosmids covering both haplotypes for a single individual in 2 SMRT Cells. The Quiver-corrected, full-length fosmid sequences can be tiled to form continuous sequence for both haplotypes, with the majority of cases having 100% identity overlap. Single base mismatches in the overlap are indicated. Genes are labeled, classified, and haplotypes are named accordingly.[2,3]

Overlap (base pairs)	Sequence identity
18,406	95.67
8,646	95.09
9,114	95.04
2,958	96.82
23,370	95.46
16,030	98.23
6,905	99.97





In-depth look at the repetitive content of the two haplotypes sequenced. The table shows sequence identity for overlaps >6 kb within both haplotypes. The three dot plots show comparisons within haplotype sequence for both haplotypes (BA1 & AB-del7), and a dot plot comparing haplotypes.

Conclusion

We demonstrate it is possible to pool and sequence fosmids covering a highly repetitive immunogenetic region and fully resolve the sequence of both haplotypes. This is possible because of the very long reads in SMRT Sequencing. We prepared pooled libraries of full-length fosmids and using the extremely long reads, generated high-quality consensus sequences without the need for assembly.

References

1. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data., Chin, CS, Alexander D. A, Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W. & Korlach J., Nature Methods 10, 563–569 (2013) .

2. The killer cell immunoglobulin-like receptor (KIR) genomic region: gene-order, haplotypes and allelic polymorphism., Hsu KC, Chida S, Geraghty DE, Dupont B., Immunological Review (2002) 190:40-52

3. Different patterns of evolution in the centromeric and telomeric regions of group a and B haplotypes of the human killer cell Ig-like receptor locus., Pyo CW, Guethlein LA, Vu Q, Wang R, Abi-Rached L, Norman, PJ, Marsh SGE, Miller JS, Parham P, Geraghty DE. PLoS One (2010)