Development of Copy Number Variation Detection Algorithms and Their Application to Genome Diversity Studies

by

Feichen Shen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in The University of Michigan
2019

Doctoral Committee:

Associate Professor Jeffrey Kidd, Chair
Professor David Burke
Assistant Professor Jacob Kitzman
Assistant Professor Jacob Mueller
Professor Kerby Shedden
Professor Thomas Wilson

Feichen Shen

feichens@umich.edu

ORCID 0000-0001-9689-0375

DEDICATION

This dissertation is dedicated to my family, especially to my father, Yu Shen, who has devoted invaluable spiritual support to me.

ACKNOWLEDGEMENT

I would like to thank all my lab members for making this research possible, especially Amanda Pendleton, Sarah Emery and my mentor Jeffrey Kidd on their dedication. In addition, my thesis committee gave me valuable guidance throughout my research. Finally, I'd like to pay special tribute to my family members especially my father who provided spiritual support to me over the years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Copy number variation (CNV) is an important class of variation that contributes to genome evolution and disease. CNVs that become fixed in a species give rise to segmental duplications; and already duplicated sequence is prone to subsequent gain and loss leading to additional copy-number variation. Multiple methods exist for defining CNV based on high-throughput sequencing data, including analysis of mapped read-depth. However, accurately assessing CNV can be computationally costly and multi-mapping-based approaches may not specifically distinguish among paralogs or gene families.

We present two rapid CNV estimation algorithms, QuicK-mer and fastCN, for second generation short sequencing data. The QuicK-mer program is a paralog sensitive CNV detector which relies on enumerating unique k-mers from a pre-tabulated reference genome. The latest version of QuicK-mer 2.0 utilizes a newly constructed k-mer counting core based on the DJB hash function and permits multithreaded CNV counting of a large input file. As a result, QuicK-mer 2.0 can produce copy-number profiles form a 10X coverage mammalian genome in less than 5 minutes. The second CNV estimator, fastCN, is based on sequence mapping and has tolerance for mismatches. The pipeline is built

around the mrsFAST read mapper and does not use additional time compared to the mrsFAST mapping process. We validated the accuracy of both approaches with existing data on human paralogous regions from the 1000 Genomes Project. We also employed QuicK-mer to perform an assessment of copy number variation on chimpanzee and human Y chromosomes.

CNV has also been associated with phenotypic changes that occur also during animal domestication. Large scale CNVs were observed previously in cattle, pigs and chicken domestication. We assessed the role of CNV in dog domestication though a comparison of semi-feral village dogs and a global collection of wolfs. Our CNV selection scan uncovered many previously confirmed duplications and deletions but did not identify fixed variants that may have contributed to the initial domestication process. During this selection study, we uncovered CNVs that are errors in the existing canine reference assembly. We attempted to the complement the current CanFam3.1 reference with the de novo genome assembly of a Great Dane breed dog named Zoey. A 50x PacBio long reads sequencing with median insert size of 7.8kbp was conducted. The resulting assembly shows significant improvement with 20x increased continuity and two third reductions of unplaced contigs. The Zoey Great Dane assembly closes 80% of CanFam3.1 gaps where high GC content was the major culprit in the original assembly.

Using unique k-mers assigned in these closed gaps, QuicK-mer was able to find many of these regions are fixed across dogs while small proportion shows variability.

Chapter 1: Introduction

This introduction chapter will provide background on genomic polymorphism and especially copy number variation. The important roles of copy number variation in evolution and the domestication process will be discussed. I will also review the bioinformatic approaches necessary to analyze the massive data produced by next generation sequencing technology. Finally, I will cover the recent progress on *de novo* genome assembly approaches and potential improvements provided for resequencing projects.

1.1. Genome variation

At the turn of last century, the human genome project completed the first reference genome assembly of our species (Richards and Scott Hawley 2005). Comparative sequencing analysis using multiple species improved our understanding to the structure and variation across the species. One important discovery shows that large differences lie not in the conserved coding regions but in interspersed non-coding regions. Potential variations in these regions include DNA sequences domains responsible for promoters and enhancing elements, as well as repeats affecting the distances between interacting

elements. Variations between species can be quantified as distances for constructing phylogenetic trees and provide evolutionary history.

Another important benefit of reference assembly is that it provides a map for looking for variations across populations within a species. Several techniques that rely on a reference exist. Oligo-microarrays utilize probe hybridization to DNA samples and require probes designed specifically to a target region. A binary signal can be generated whether the sample under query is identical to the reference in the probe region. Later, massively parallel sequencing enabled the discovery of novel variants when combined with a reference assembly.

Generally, two types of variants exist within a sample based on their size. The smaller point mutations, ranging from a single base pair change to a few base pairs of insertion or deletion, are the predominant type of mutations. The second type, structural variations, is on a much larger scale. These include chromosomal inversion, translocation, duplication and deletion. The duplication and deletion variants change the number of occurrences of these sequences and are defined as copy number variation (CNV). Genes with copy number differences could show differences in expression level which in turn affects phenotypes (Geistlinger et al. 2018; Gamazon and Stranger 2015). Chromosomal inversion and reciprocal translocations are copy number neutral, meaning that the genes

within the inversion or on the adjacent chromosomal arm do not change in their number of copies. However, genes intersecting with the breakpoint will be disrupted. Such process can occur in somatic tissues, leading to tumorigenesis. Another example is a fusion between a highly active promoter of one tissue specific gene and a cell proliferation factor (Annala et al. 2013; Kloosterman et al. 2017).

1.1.1 Point mutations

Base pair substitution and small scale insertion/deletion (indels) are some forms of point mutation. Indels within a coding region usually generate a premature stop codon from coding frameshift, which may further lead to nonsense mediated decay of RNA. These indels are usually detrimental and generally evolve under purifying selection. Substitutes can be more tolerated for three reasons. First, the codons are degenerate, with 64 triplets coding for only 20 amino acids. Thus, many single base pair changes do not alter the encoded amino acid. Second, amino acids form similar groups based on their charge and affinity to water. Should a similar amino acid replace the original one, the effect on the protein might be small. Lastly, proteins contain multiple functional domains and the effects for each change can vary greatly. Missense hits in functional domains that are less important, such as linkers, would have a subtle effect.

For evolutionary studies looking at time history, the distance can be drawn based on the number of mutation changes and a mutation rate. Thus, these studies require mutations to be preserved instead of being selected against.

1.1.2 Structural Variation

Structural changes range from short segmental duplication/deletion to much larger chromosomal scales like inversion, translocation or even duplication of entire chromosomes. Although rare compared to point mutations, these changes affect far more genes. Over the past ten years, various discoveries have identified diverse polymorphism of copy number among the ethnic groups in humans (J. Li et al. 2009; Lou et al. 2011). Structural changes are also a major driver in cancer.

1.2. Role of copy number variation in evolution

Copy number variation is a major component during evolution. Numerous methods have been developed to detect such genetic divergence. Before the advent of high-throughput sequencing, we had quantitative PCR by comparing the relative abundance of DNA to a standard. Later, array CGH came along. This approach comes in various shapes and forms. One of the most popular is a genome tiling array with probes uniformly tiling across a reference assembly. The signal difference reveals the copy number ratio between

two samples, labeled with different dyes, hybridized on the array. Other indirect methods could also be used to identify copy number variations and larger structural variations by their effect on other markers, such as altered recombination rates due to inversions or gain or loss of genotype signal from single nucleotide variants due to duplication or deletion (Conrad and Hurles 2007). Various bioinformatic algorithm also utilize these signals to indirectly detect structural variation and breakpoints (Becker et al. 2018; Zöllner and Teslovich 2009).

By employing both aCGH and qPCR as a validation approach, the human CNV map immediately revealed the abundance of copy number changes as a major source of variation across populations (Redon et al. 2006). Many of the CNVs discovered were adjacent to assembly gaps. Even though the tiling aCGH has limited resolution, the same study was able to associate a list of Mendelian diseases to copy number variation loci. Some genes that were deleted stood out in the CNV map as well. Apparently, large scale duplication has a huge effect on genes on those regions. For example, an expanded research effort draws a link between autism and schizophrenia to CNVs (McCarroll 2008). In 2005, the first chimpanzee genome showed that activity of retrotransposable elements was a major driver for small insertions in the primate evolution (Chimpanzee Sequencing and Analysis Consortium 2005). Another form of CNV is due to large

segmental duplication. One mechanism for copy number variation among duplications is nonallelic homologous recombination (NAHR), as suggested by numerous CNV studies(Chimpanzee Sequencing and Analysis Consortium 2005)(Graubert et al. 2007; Redon et al. 2006; Lee et al. 2008)Higher resolution aCGH maps showed that these duplication regions clustered in "hotspots" common between the chimpanzee and human genome. Many of these regions are still not fixed in the species and have variable copy number between unrelated individuals (Graubert et al. 2007; Redon et al. 2006; Lee et al. 2008)(Perry et al. 2008). Higher resolution achieved by sequencing further implicated NAHR as major mechanism of duplication and deletion between humans (Kidd et al. 2008)The presence of these duplication across the genome creates seed for further structural rearrangements and leads to rapid evolution (Figure 1) (Zhou et al. 2013).

Comparative genome analysis shows the mechanism of insertion and deletions, as well as their effect on the recent evolutionary history in human and other mammals. Yet on a longer time scale, segmental duplication gives opportunities for shaping new genes. The same gene in different species is called an ortholog. Orthologous genes usually share the same function and similar sequence content due to conservation. The process underlying speciation might be much more complicated than accumulating mutations. For example, copy number increases can give rise to additional copies of genes. Over time, mutations

might randomly disable one copy (Figure 1). The functional version might be carried to a different chromosomal location due to inversion or translocation. This process may create mating incompatibility and generate species (Lynch 2002). In other cases, the additional copy of the gene might accumulate changes for a new function or simply the increased dosage of such genes can be beneficial. Other genetic mechanisms have been proposed for speciation, and this is an active area of much research (Shapiro, Leducq, and Mallet 2016; Noor and Feder 2006).

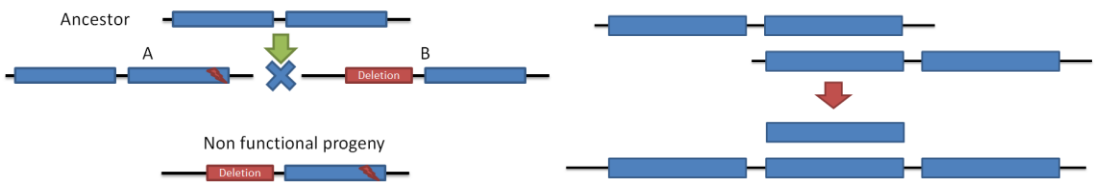


Figure 1 Mechanisms of duplication driven speciation

Two important roles of CNV in evolution. Left: Genes are duplicated and randomly deactivated afterwards. Mating between progenies could create hybrid incompatibility and thus drive speciation. Right: Nonallelic homologous recombination within segmental duplications drives deletion or expansion of gene copies. Presence of these duplication sites seeds more complex rearrangements in the future.

1.2.1 CNV and Adaptation

In some cases, copy number expansion can be directly related to a higher transcription level, which in turn affects protein translation (Orozco et al. 2009). This increase in copy number dosage could be an advantage should the protein expressed be advantageous to

survival. Several examples in the human genome have been shown using sequencing studies (Perry 2008). For example, a higher copy number of the *CCL3L1* gene has been shown related to lower risk of HIV infection in certain African populations. Expansion of copy number in olfactory receptors is also observed (Nozawa, Kawahara, and Nei 2007). Additional copy number of *P53* is associated with increased cancer resistance and longevity in animals (Sulak et al. 2016; Donehower 2009).

1.2.2 Duplication and neofunctionalization

A major evolutionary importance of gene duplication is providing an additional copy for new function to evolve. The red opsin gene on the X chromosome is an example (Hunt et al. 2009). The later accumulated mutations that altered the structure of the protein and shifts the peak of spectra sensitivity. The tricolor vision in primates enables easy distinction for fruits as a source of food (Melin et al. 2013). This process of duplication followed by beneficial mutation is sometimes called neofunctionalization. To a greater extent, changes in domain binding factors could have a cascading effect on evolution. Zinc-finger factor (ZNF) is a DNA binding protein. The rapid neofunctionalization of new copies of ZNFs give rise to diverse regulatory network (Nowick, Carneiro, and Faria 2013).

1.2.3 Examples of CNVs in domestication

Copy number expansion and contraction is also observed during the domestication process (Clop, Vidal, and Amills 2012). The pea-comb phenotype in domesticated chicken is related to SOX5 duplications (Wright et al. 2009). CNV is also shown in pigs linked to certain traits (Chen et al. 2012). A complex CNV landscape has also been suggested in numerous studies of cattle and sheep (Chen et al. 2012; Keel, Lindholm-Perry, and Snelling 2016) with some variants associated with particular breeds. Expansion of olfactory receptor and immune system related genes were a common observation between cattle and pigs (Chen et al. 2012; Keel, Lindholm-Perry, and Snelling 2016).

1.2.4 Limitation of reliance on a good / contiguous genome assembly

Calling structural variation requires decent reference assemblies. Collapsed duplications during the *de novo* assembly process lack the location information for additional sets of copies. Regions not well assembled, such as unplaced contigs, limit the continuity of CNVs. Segmental duplication can also alter the local linkage disequilibrium in many cases and lead to misidentification of single nucleotide variants. Yet CNV are often flanked with low complexity sequences leading to misassembled regions. These misplaced CNV would show incorrect linkage disequilibrium in the vicinity. Mapping

tools relies on unique regions to find the best matching locations for sequencing reads. In doing so, these mappers will generate a mapping score based on the uniqueness of the query relative to the reference. For example, the MAPQ score is the log10 probability of observing such placement by random chance. Collapsed duplications remove the unique signature for each copy, further limits the mapping quality of a read. This is also the reason that repeat regions have limited genotype calling accuracy and are often excluded from use.

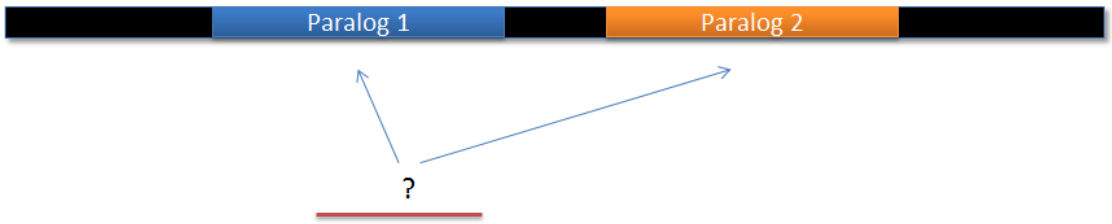


Figure 2 Random assignment of reads in duplication

Most sequencing read mapping algorithms will have difficulty assigning reads in repeated sequences due to reduced content complexity. Here is an example of read matching sequences within a pair duplicated paralogs. To quantify the correctness of read placement, a mapping score is devised based on the sequence complexity inside the reference and the number of base pairs matched. This score is calculated by most mapping algorithms and embedded in the alignment file.

1.3. Advent of high throughput sequencing

To identify these variants, determining the sequence of each base pair is a standard approach. Biologists are long used to probing and cloning to narrow down a region of

interest from whole genomic DNA. Later, this technique was greatly improved with the PCR approach. By designing primers in the region of interest, one can amplify the DNA segment without cloning steps. The amplified region can then be studied with sequencing approach. The invention of DNA sequencing permits direct readout of nucleotides in order. In this section I will briefly review the sequencing techniques as they are the fundamental methods in CNV detection and analysis.

1.3.1 The Sanger approach to DNA sequencing

The first-generation approach, called Sanger sequencing, is similar to PCR. It requires a concentrated DNA template from a genomic region. The sequence extension is randomly terminated and electrophoresis separates fragments into single base pair resolution. If terminal oligonucleotides are fluorescently labeled, the sequence can be read automatically. As sequencing reads get longer, the co-mingling of large linear molecules causes their migration speed to deviate from its molecular weight. This is observed as peaks get wider after 800 bp (Dovichi 1997). This technique also requires PCR primers and thus a portion of the underlying sequence has to be partially known. In genome projects this problem can be solved by sequencing the ends of fragments inserted into a vector of known sequence.

At the inception of human genome project, Sanger sequencing was the only feasible technology to get actual reads. In order to assemble sequence into chromosomes, genomists at that time resorted to a layered hierarchical approach by constructing bacterial artificial chromosomes (BAC), fosmids and plasmids, as well as primer walking methods. A more detailed review of genome assembly will be given later in this chapter at section 4.

1.3.2 Massively parallel sequencing

Almost at the time when the human genome project was completed, we saw the first proliferation of cost-effective second-generation sequencing. These methods aim at reducing cost and increasing throughput. Instead of PCR in a bulk tube, each of the technologies aimed to isolate each and every single DNA molecular randomly and amplify them in parallel without mutual interference. This step is called library preparation. Two major techniques dominated the market, solid surface based and bead based (Goodwin, McPherson, and McCombie 2016). The first method involved using a bead covered by an oil droplet, which was employed by 454 pyrosequencing (Margulies et al. 2005) and its related technology - Ion Torrent (Rothberg et al. 2011). Also limited by input concentration, each bead is expected to have only one DNA molecule and hence the PCR amplification is on that DNA only. The beads are then loaded onto a predefined

microwell array and the sequence determined, usually with sequencing-by-synthesis or sequencing-by-ligation. The second involves a substrate, usually transparent glass for imaging, that contains binding primers to allow a PCR reaction in a limited radius defined by the length of the DNA fragment. Solexa/Illumina and Solid sequencing employed this approach. As long as the input DNA concentration and length is constrained, the PCR products would form an isolated cluster from each other, and the sequence of each cluster could be obtained using labeled oligos in a stepwise fashion. These amplification steps achieve the necessary signal to noise ratio since at that time photons from a single molecule were hard to detect.

During the actual sequencing step, base signal can be generated by a labeled dye on a reversible terminated nucleotide in a sequencing-by-synthesis approach, or short probe in sequencing-by-ligation approach. For contiguous sequencing methods, four types of nucleotides have to be added and washed consecutively and are prone to homopolymer errors. Signal can be detected by a secondary reaction giving of light or by directly measuring the proton generation through an ion sensitive diode (Rothberg et al. 2011). This step happens in parallel for each DNA molecule. Even though the entire process might take up to a few days and be expensive for a single run, the time and financial cost can quickly be offset by the large number of DNA reads obtained.

The second-generation sequencing immediately opened up numerous resequencing projects aimed at identifying variation across the genome. Since each DNA fragment is randomly selected from the genome, the second-generation sequencing will have less bias and enable discovery of any variants not likely covered in previous studies. However, on the other side, without knowing where each fragment is located on the reference assembly, the resequencing studies require significantly more computational analysis compared to targeted approaches.

1.4. *de novo* genome assembly

Genome assembly is the process of piecing together a complete sequence for each chromosome from basic sequencing reads. A simple analogy is like detective work by piecing together shredded paper. When Sanger sequencing was developed in 1970's, biologist immediately tried to map the genome of various organisms. These attempts were focused on small DNA like vectors and plasmids from bacteria by sequencing one segment at a time. Segments were then further extended with a primer designed at previously resolved sequence. This linear iterative approach could feasibly resolve small genomes in the size range of kilobase to hundreds of kilobases. But it quickly became impossible for bacteria genomes or even vertebrate genomes.

1.4.1 Methods of genome assembly

To solve this problem, cloning approaches came into view by breaking the genome into large fragments and cloning them into fosmids or bacterial artificial chromosomes (BACs). One can sequence from the common, known backbone sequence on these clones simultaneously without knowing the exact sequencing for the next starting point. But, isolation and purification of individual clones could be a tedious process.

Another parallel process is called whole genome shotgun sequencing (Figure 3). By randomly shearing the DNA into small fragments, hence "shotgun", each DNA fragment can be tackled by attaching a common adapter or cloning into a known vector. Individual reads from each are obtained which later can be pieced together bioinformatically. The underlying assumption is that unique sequences shared by adjacent and overlapping reads will provide a unique tiling path.

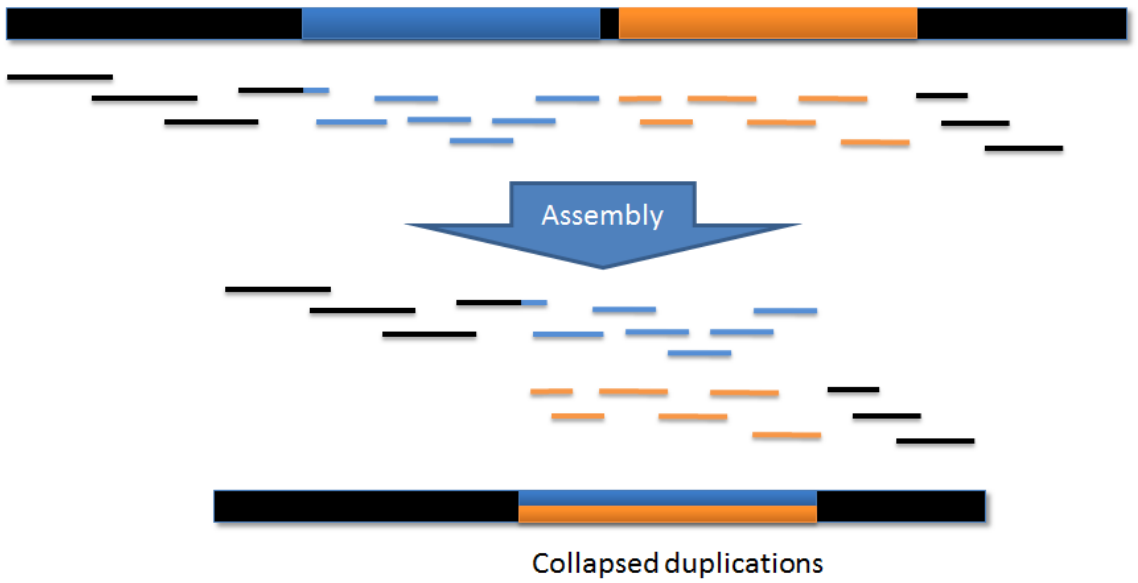


Figure 3 WGS and collapsed duplications

Whole genome shotgun sequencing could not separate reads in repeats and duplications and could result in collapsed sequences in the assembly.

However, for a vertebrate and especially mammalian genome harboring repetitive elements, this assumption is no longer true. For example, the human LINE-1 element has a full length of 6kb let alone the microsatellite and simple repeats and large segmental duplications, which easily exceed the longest sequencing read available at that time. The initial human genome project found 17% of our DNA content is made up by these elements (International Human Genome Sequencing Consortium 2001). When the whole genome shotgun approach is used on such an assembly, the continuity of the tiling path

will be broken upon encountering repetitive regions. Highly repetitive regions and duplications would typically result in missing sequences or linking related DNA by a common shared region (Figure 3).

To resolve these issues, the public human genome project employed a hierarchical approach which combines shotgun methods and BAC clones (Figure 4). Each BAC is first mapped using markers forming a rough order on the chromosomes. Then each BAC is resolved using shotgun sequencing. The chances of multiple duplications inducing mis-assembly within a BAC are greatly reduced. Comparison between the hierarchical approach against the private whole genome shotgun sequencing clearly demonstrates that having localized information greatly increases contig length (Waterston, Lander, and Sulston 2002).

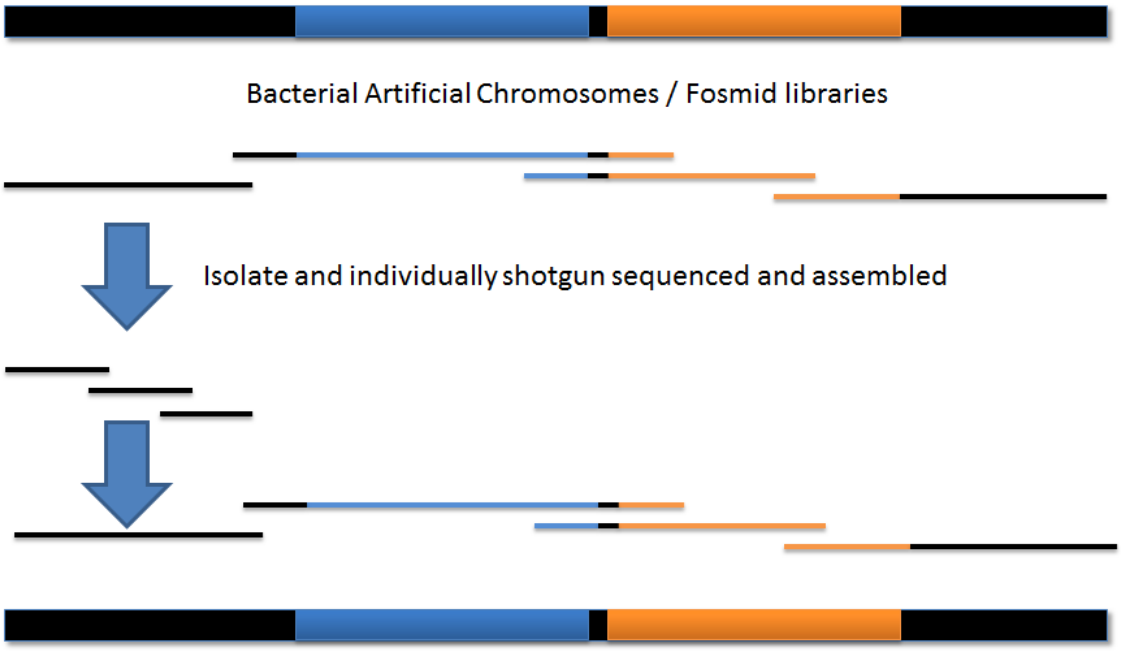


Figure 4 Hierarchical shotgun assembly

Hierarchical shotgun assembly isolates each duplicated region into individual DNA clones through an overlapping tiling path. The ensuing random shearing of each clone reduces the chances of collapsed repeats.

The second-generation short read platforms promised much lower cost for the initial assembly process. But the long continuity of a gold standard reference genome would still require long range information to link short contigs together.

1.4.2 Benefits of an improved reference assembly

The second generation short-read sequencing approach provides easy access to whole genome SNP variant analysis as well as some structural variation calling. But to generate a reference assembly from this technology can have numerous limitations. Mainly due to the limits in sequence read length and DNA insert size, short read sequencing has difficulties in repeated regions and segmental duplications. Without sufficient unique identifiers in each repeat, the *de novo* assembler is unable to assign the read to its true location. To make this process more complicated, the flanking regions of structural variants and duplications are occasionally enriched with repeat sequences (Satyanarayana and Strominger 1992; Bacolla et al. 2016). Together, short reads and *de novo* assembly usually result in short contigs or contigs mixed with assembly errors.

A metric to measure the quality of assembly continuity is the N_{50} value. Calculation of N_{50} is done by sorting the assembled contigs from the shortest to the longest. Then combining the total genome coverage from the shortest contigs until it reaches 50% of genome size. The length of the last contig added to the sum is N_{50}. For example, the first draft assembly of the giant panda obtained with Illumina resulted in an N_{50} of 40kb even with the use of large insert jumping libraries (R. Li et al. 2010). Even with the whole genome shotgun method, the increased read length of older Sanger sequencing based

assembly preserves more segmental duplications compared to that from a short-read data (Alkan, Sajjadian, and Eichler 2010). To further improve, a scaffold backbone is necessary in order to reorganize these into meaningful representation on chromosomes. Alternatively, the method has to be improved with long read sequencing technologies.

Moving onward from the second-generation short reads, recently the third-generation sequencing technology has led to a decline in cost and an increase in read length. This method, initially lead by PacBio single molecule real time method, was quickly followed by even longer reads from Oxford Nanopore. The read length in such technology can be quantified by L_{50}, which is the read length at the 50% of summed length from the longest reads. The L_{50} of PacBio raw sequencing reads could easily exceed 10kb. When a long insert library is carefully prepared, we can expect the majority of repeat elements in a genome to be correctly linked to their flanking unique sequences.

1.4.2.1 Identifying duplication by comparative genome analysis

Copy number variation on a reference can be annotated with multiway alignment methods. By comparing against another reference assembly of a species sufficiently close on the phylogenetic tree, one could find duplication and deletion still under evolution between the species. Another approach is to do a self to self alignment to reveal regions by constructing dot-plots. Regions successfully assembled into each of their respective

copies will show up as diagonal match pairs across the genome. But due to the limitation of most de novo assembly projects illustrated before, these duplications are usually limited in length.

1.4.2.2 Existing duplications interfering with unique sequence identification

Searching for CNV across a reference with correctly assembled duplications is also challenging. Should duplications be represented multiple times on a reference assembly, the mapping software needs to determine which copy to correctly assign the sequencing read (Treangen and Salzberg 2011). Since the majority of base pairs are identical within the duplicated copies, mapping algorithms must choose a strategy to deal with the read placement and determine the mapping quality score accordingly.

1.5. Summary

Here we described the importance of copy number variations in evolution and animal domestication, as well as methods and difficulties to detect CNVs using high throughput sequencing data. In the following chapters of my dissertation, I will further dive into methodological development for CNV calling and its application on studies of domestication in dogs. I will also explore the benefits of a newly improved assembly of the canine genome for these CNV methodologies.

1.6 References

Alkan, Can, Saba Sajjadian, and Evan E. Eichler. 2010. “Limitations of next-Generation Genome Sequence Assembly.” Nature Methods 8 (1): 61–65.

Annala, M. J., B. C. Parker, W. Zhang, and M. Nykter. 2013. “Fusion Genes and Their Discovery Using High Throughput Sequencing.” Cancer Letters 340 (2): 192–200.

Bacolla, Albino, John A. Tainer, Karen M. Vasquez, and David N. Cooper. 2016. “Translocation and Deletion Breakpoints in Cancer Genomes Are Associated with Potential Non-B DNA-Forming Sequences.” Nucleic Acids Research 44 (12): 5673–88.

Becker, Timothy, Wan-Ping Lee, Joseph Leone, Qihui Zhu, Chengsheng Zhang, Silvia Liu, Jack Sargent, et al. 2018. “FusorSV: An Algorithm for Optimally Combining Data from Multiple Structural Variation Detection Methods.” Genome Biology 19 (1): 38.

Chen, Congying, Ruimin Qiao, Rongxing Wei, Yuanmei Guo, Huashui Ai, Junwu Ma, Jun Ren, and Lusheng Huang. 2012. “A Comprehensive Survey of Copy Number Variation in 18 Diverse Pig Populations and Identification of Candidate Copy Number Variable Genes Associated with Complex Traits.” BMC Genomics 13 (December): 733.

Chimpanzee Sequencing and Analysis Consortium. 2005. “Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome.” Nature 437 (7055): 69–87.

Clop, A., O. Vidal, and M. Amills. 2012. “Copy Number Variation in the Genomes of Domestic Animals.” Animal Genetics 43 (5): 503–17.

Conrad, Donald F., and Matthew E. Hurles. 2007. “The Population Genetics of Structural Variation.” Nature Genetics 39 (7s): S30–36.

Donehower, Lawrence A. 2009. “Using Mice to Examine p53 Functions in Cancer, Aging, and Longevity.” Cold Spring Harbor Perspectives in Biology 1 (6): a001081.

Dovichi, Norman J. 1997. “DNA Sequencing by Capillary Electrophoresis.” Electrophoresis 18 (12-13): 2393–99.

Gamazon, Eric R., and Barbara E. Stranger. 2015. “The Impact of Human Copy Number Variation on Gene Expression.” Briefings in Functional Genomics 14 (5): 352–57.

Geistlinger, Ludwig, Vinicius Henrique da Silva, Aline Silva Mello Cesar, Polyana Cristine Tizioto, Levi Waldron, Ralf Zimmer, Luciana Correia de Almeida Regitano, and Luiz Lehmann Coutinho. 2018. “Widespread Modulation of Gene Expression by Copy Number Variation in Skeletal Muscle.” Scientific Reports 8 (1): 1399.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. “Coming of Age: Ten Years of next-Generation Sequencing Technologies.” Nature Reviews. Genetics 17 (6): 333–51.

Graubert, Timothy A., Patrick Cahan, Deepa Edwin, Rebecca R. Selzer, Todd A. Richmond, Peggy S. Eis, William D. Shannon, et al. 2007. “A High-Resolution Map of Segmental DNA Copy Number Variation in the Mouse Genome.” PLoS Genetics 3 (1): e3.

Hunt, David M., Livia S. Carvalho, Jill A. Cowing, and Wayne L. Davies. 2009. “Evolution and Spectral Tuning of Visual Pigments in Birds and Mammals.” Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 364 (1531): 2941–55.

International Human Genome Sequencing Consortium. 2001. “Initial Sequencing and Analysis of the Human Genome.” Nature 409 (6822): 860–921.

Keel, Brittney N., Amanda K. Lindholm-Perry, and Warren M. Snelling. 2016. “Evolutionary and Functional Features of Copy Number Variation in the Cattle Genome.” Frontiers in Genetics 7 (November): 207.

Kidd, Jeffrey M., Gregory M. Cooper, William F. Donahue, Hillary S. Hayden, Nick Sampas, Tina Graves, Nancy Hansen, et al. 2008. “Mapping and Sequencing of Structural Variation from Eight Human Genomes.” Nature 453 (7191): 56–64.

Kloosterman, Wigard P., Robert R. J. Coebergh van den Braak, Mark Pieterse, Markus J. van Roosmalen, Anieta M. Sieuwerts, Christina Stangl, Ronne Brunekreef, et al. 2017. “A Systematic Analysis of Oncogenic Gene Fusions in Primary Colon Cancer.” Cancer Research 77 (14): 3814–22.

Lee, Arthur S., Mar á Guti érrez-Arcelus, George H. Perry, Eric J. Vallender, Welkin E. Johnson, Gregory M. Miller, Jan O. Korbel, and Charles Lee. 2008. “Analysis of Copy Number Variation in the Rhesus Macaque Genome Identifies Candidate Loci for Evolutionary and Human Disease Studies.” Human Molecular Genetics 17 (8): 1127–36.

Li, Jian, Tielin Yang, Liang Wang, Han Yan, Yinping Zhang, Yan Guo, Feng Pan, et al. 2009. “Whole Genome Distribution and Ethnic Differentiation of Copy Number Variation in Caucasian and Asian Populations.” PloS One 4 (11): e7958.

Li, Ruiqiang, Wei Fan, Geng Tian, Hongmei Zhu, Lin He, Jing Cai, Quanfei Huang, et al. 2010. “The Sequence and de Novo Assembly of the Giant Panda Genome.” Nature 463 (7279): 311–17.

Lou, Haiyi, Shilin Li, Yajun Yang, Longli Kang, Xin Zhang, Wenfei Jin, Bailin Wu, Li Jin, and Shuhua Xu. 2011. “A Map of Copy Number Variations in Chinese Populations.” PloS One 6 (11): e27341.

Lynch, Michael. 2002. “Genomics. Gene Duplication and Evolution.” Science 297 (5583): 945–47.

Margulies, Marcel, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, et al. 2005. “Genome Sequencing in Microfabricated High-Density Picolitre Reactors.” Nature 437 (7057): 376–80.

McCarroll, Steven A. 2008. “Extending Genome-Wide Association Studies to Copy-Number Variation.” Human Molecular Genetics 17 (R2): R135–42.

Melin, A. D., C. Hiramatsu, N. A. Parr, Y. Matsushita, S. Kawamura, and L. M. Fedigan. 2013. “The Behavioral Ecology of Color Vision: Considering Fruit Conspicuity, Detection Distance and Dietary Importance.” International Journal of Primatology 35 (1): 258–87.

Noor, Mohamed A. F., and Jeffrey L. Feder. 2006. “Speciation Genetics: Evolving Approaches.” Nature Reviews. Genetics 7 (11): 851–61.

Nowick, Katja, Miguel Carneiro, and Rui Faria. 2013. “A Prominent Role of KRAB-ZNF Transcription Factors in Mammalian Speciation?” Trends in Genetics: TIG 29 (3): 130–39.

Nozawa, Masafumi, Yoshihiro Kawahara, and Masatoshi Nei. 2007. “Genomic Drift and Copy Number Variation of Sensory Receptor Genes in Humans.” Proceedings of the National Academy of Sciences of the United States of America 104 (51): 20421–26.

Orozco, Luz D., Shawn J. Cokus, Anatole Ghazalpour, Leslie Ingram-Drake, Susanna Wang, Atila van Nas, Nam Che, Jesus A. Araujo, Matteo Pellegrini, and Aldons J. Lusis. 2009. “Copy Number Variation Influences Gene Expression and Metabolic Traits in Mice.” Human Molecular Genetics 18 (21): 4118–29.

Perry, G. H. 2008. “The Evolutionary Significance of Copy Number Variation in the Human Genome.” Cytogenetic and Genome Research 123 (1-4): 283–87.

Perry, G. H., F. Yang, T. Marques-Bonet, C. Murphy, T. Fitzgerald, A. S. Lee, C. Hyland, et al. 2008. “Copy Number Variation and Evolution in Humans and Chimpanzees.” Genome Research 18 (11): 1698–1710.

Redon, Richard, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, et al. 2006. “Global Variation in Copy Number in the Human Genome.” Nature 444 (7118): 444–54.

Richards, Julia E., and R. Scott Hawley. 2005. “The Human Genome Project.” In The Human Genome, 279–86.

Rothberg, Jonathan M., Wolfgang Hinz, Todd M. Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H. Leamon, et al. 2011. “An Integrated Semiconductor Device Enabling Non-Optical Genome Sequencing.” Nature 475 (7356): 348–52.

Satyanarayana, Karuturi, and Jack L. Strominger. 1992. “DNA Sequences near a

Meiotic Recombinational Breakpoint within the Human HLA-DQ Region." Immunogenetics 35 (4): 235–40.

Shapiro, B. Jesse, Jean-Baptiste Leducq, and James Mallet. 2016. "What Is Speciation?" PLoS Genetics 12 (3): e1005860.

Sulak, Michael, Lindsey Fong, Katelyn Mika, Sravanthi Chigurupati, Lisa Yon, Nigel P. Mongan, Richard D. Emes, and Vincent J. Lynch. 2016. "TP53 Copy Number Expansion Is Associated with the Evolution of Increased Body Size and an Enhanced DNA Damage Response in Elephants." eLife 5. https://doi.org/10.7554/elife.11994.

Treangen, Todd J., and Steven L. Salzberg. 2011. "Repetitive DNA and next-Generation Sequencing: Computational Challenges and Solutions." Nature Reviews. Genetics 13 (1): 36–46.

Waterston, Robert H., Eric S. Lander, and John E. Sulston. 2002. "On the Sequencing of the Human Genome." Proceedings of the National Academy of Sciences 99 (6): 3712–16.

Wright, Dominic, Henrik Boije, Jennifer R. S. Meadows, Bertrand Bed'hom, David Gourichon, Agathe Vieaud, Michèle Tixier-Boichard, et al. 2009. "Copy Number Variation in Intron 1 of SOX5 Causes the Pea-Comb Phenotype in Chickens." PLoS Genetics 5 (6): e1000512.

Zhou, Weichen, Feng Zhang, Xiaoli Chen, Yiping Shen, James R. Lupski, and Li Jin. 2013. "Increased Genome Instability in Human DNA Segments with Self-Chains: Homology-Induced Structural Variations via Replicative Mechanisms." Human Molecular Genetics 22 (13): 2642–51.

Zöllner, Sebastian, and Tanya M. Teslovich. 2009. "Using GWAS Data to Identify Copy Number Variants Contributing to Common Complex Diseases." Statistical Science: A Review Journal of the Institute of Mathematical Statistics 24 (4): 530–46.

Chapter 2: Copy number estimation through a depth of coverage approach

In this chapter, I will introduce two copy number variation (CNV) detection algorithms based on the depth of coverage approach using 2nd generation short read sequencing data. The first pipeline, QuicK-mer, is a CNV detection pipeline with unique paralog sensitivity that uses k-mer counting. This pipeline is later updated to a full-fledged, stand-alone approach that does not rely on external k-mer counting programs. A second pipeline, called fastCN, is an efficient approach based on multi-mapping of reads. It achieves similar efficiency with mismatch tolerance. The described programs are publically available at https://github.com/KiddLab/ I have used these methods to survey copy-number variation in dogs, wolves, chimpanzees, and bonobos. Portions of this chapter appeared in these previously published manuscripts (Oetjens et al. 2016; Pendleton et al. 2018).

2.1. Background

2.1.1 Definition of paralog

A paralog gene is defined as a new gene which arises due to gene duplication within a species. This is in contrast to an ortholog, which are genes in different species related due to speciation events. The duplicated paralogs can undergo neofunctionalization or subfunctionalization and create large gene families responsible for diverse functionality, or a similar function with different targets. Examples include various DNA binding domains, such as zinc finger DNA binding regions, which are responsible for regulating diverse processes. The latter includes examples olfactory receptors with each binding to a different chemical ligand, and duplication between red and green rhodopsin binding protein that gave primates tri-color vision. Together through diversification over evolution time scale, both processes shape the function and phenotypes we see across species today.

To correctly detect copy number variation in a reference genome, we not only need to classify the copies for each DNA segment, but also accurately account for the mutations accumulated within unique paralogs.

2.1.2 CNV detection methodologies

The copy number of a sequence is defined as the number of times it occurs in a sample's genome. For a unique region in a diploid genome, this value is equal to two. Through duplication and deletion process the copy number of a region can increase or decrease. In the case of deletion, copy number can decrease resulting in the loss of sequence or gene. Other repeated regions, such as mobile elements have greatly expanded copy number through other biological processes and are typically analyzed using specialized tools.

Comparative sequence analysis can be employed to identify copy number variation between two or more species when high quality reference assemblies are available. However, to detect a CNV event across population of one species, or subspecies where reference is lacking, we must resort to other means. With the advent of short read sequencing technologies in the last 20 years, CNV information could be readily extracted. Several bioinformatic methods had been proposed to infer copy number using short read data.

2.1.2.1 CNV detection through de novo genome assembly

One method to delineate copy number variations is de novo genome assembly. By constructing the sample into fully assembled sequences as accurately as possible, a pairwise comparison can be performed against a reference in search of duplications and

deletions. The apparent drawback for such an approach is usually the prohibitive cost required for an accurate assembly of duplicated regions. Optimistically, an ideal de novo assembly should yield megabase level scaffolds with preserved duplications. However, in reality, high depth assembly from next-generation sequencing is limited by read length. As a result, these assemblies are usually fragmented with regions of duplication collapsed into a false copy number of two (Hartasánchez et al. 2018). A good quality de novo assembly requires long read sequencing at high depth to overcome repeats and duplications within a genome. Even with long reads of current instruments (20-50kb length), approaches such as BAC clone sequencing remain required to accurately reconstruct the duplicated segments of typical mammalian genomes (Hoeppner et al. 2014; Chaisson et al. 2015). These long-read sequencing technologies are far from being mature compared to short reads and often require customized software and lengthy parameter tuning for analysis.

2.1.2.2 Duplication/deletion detection using pair-end or jumping libraries

A second approach to discover copy number variation is to seek various structural variation signatures in data aligned to an existing reference genome. In particular, sequencing reads spanning a breakpoint junction, called a split-read, can be detected. However the chances of finding such split-reads can be low given the short read length

and low depth in many resequencing projects. To alleviate such issues, pair-end read and specially constructed jumping libraries can improve the physical coverage obtained with the same read depth. Through statistical filtration, a candidate duplication sites can be detected with based on reads not mapping in a concordant fashion. Duplications (or insertions of any sequence) smaller than the fragment size of the library can be directly detected via aberrant read-pair signatures. Tandem duplication of otherwise unique sequence can also be identified. Other types of duplications may be predicted based on aberrant anchoring of read-pairs. In the case of deletion, the associated signatures are read pairs from fragments with apparently long insert.

Though it is possible to accurately determine the breakpoints for such an event, these methods will struggle to accurately survey the precise copy number of a region. Secondly, since many breakpoints are flanked by repeats, the mapping quality for such a read is low resulting in a lack of statistical power for calling a breakpoint.

2.1.2.3 CNV detection with depth of coverage approach

Lastly, copy number variation can be assessed using coverage and depth information. Assuming a random DNA shearing and sequencing process when short sequencing reads

are generated, the number of reads at each location should follow a Poisson distribution with a mean proportional to the copy number of that segment in the sampled genomes. If a segment of DNA is duplicated, we'd expect the number of read observation to increase and vice versa for deletion. The power of detection increases dramatically when the length of duplication is increased. At a megabase level, CNV can be accurately quantified even with less than 1x of sequencing coverage. On the other hand, this approach requires a good reference. Should a region be misassembled or even missing from the reference, there will be no chance of finding it based on sequencing depth.

2.1.3 Summary

All three methods of CNV detection can be used. In term of cost effectiveness, depth of coverage clearly stands out due to number of reads available given a region. In the following sections, we are going employ this method with varying degree of paralog sensitivity in constructing two bioinformatic algorithms.

2.2 QuicK-mer: A paralog sensitive rapid CNV estimator

In order to achieve paralog sensitivity, we choose an approach to consider only unique sequences within a reference assembly. This problem can then be simplified with a combination of the depth of coverage method described previously and enumeration of

predefined k-mers based on an existing reference assembly. In this section I will describe the concept and realization of the QuicK-mer CNV estimator.

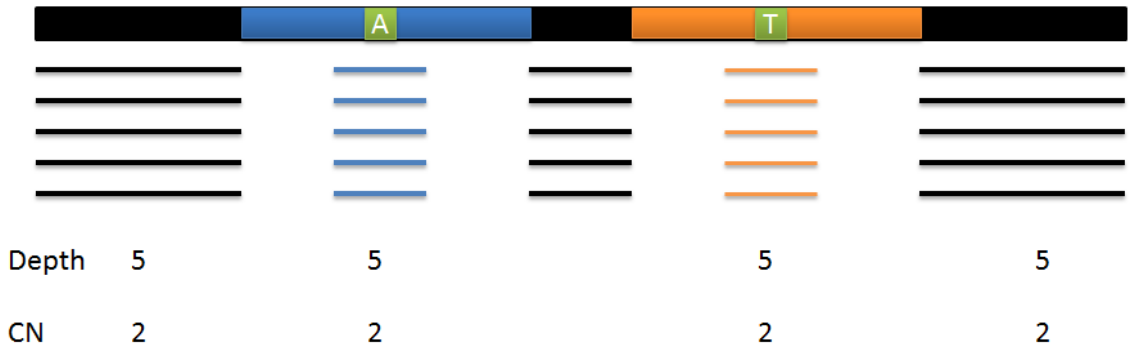


Figure 5 QuicK-mer working principle

QuicK-mer achieves paralog sensitivity by counting the sequencing depth only within unique regions of a reference genome. Here, a simplified example showing unique regions of each paralogous gene is correctly normalized to copy number of two.

2.2.1 Paralog sensitivity through unique k-mer counting

To achieve efficient and paralog-specific CNV estimation, we focused on counting specific k-mer sequences rather than aligning reads to a reference, an approach that has also been proposed for analysis of RNA-Seq data (Zhang and Wang 2014; Patro, Mount, and Kingsford 2014). The QuicK-mer pipeline was designed to utilize the existing Jellyfish k-mer counting application (Marçais and Kingsford 2011). Accepting both

FASTQ and BAM files as input, QuicK-mer is designed for sequences generated by the Illumina platform.

To speed up copy number estimation, QuicK-mer requires two major pre-processing steps for each genome assembly. These two steps are essential to generate binary files for efficient access within the core copy number estimation pipeline. These two binary files are described further in the next section and are used to estimate the copy number in each sample. For detailed operation, refer to the QuicK-mer operation manual in the software package.

2.2.2 QuicK-mer 1.0 Implementation

2.2.2.1 Tabulate a catalog of unique k-mer

Defining a catalog of unique k-mers requires seven individual steps. In practice, we utilize a size of k=30 for consistency with previous studies. (Alkan et al. 2009; Sudmant et al. 2010) An example of the command lines for each of the following steps can be found in the QuicK-mer User Manual v1.0 Section 9.

1. List all unique 30-mers: All 30-mers in the reference genome are enumerated with Jellyfish by setting k-mer size equal to 30 and using the reference genome FASTA

sequence as the input. A k-mer and its reverse complement are considered equal (Jellyfish option –C). The 30-mers with a count of 1 are exported into text format.

2. Determine unique 30-mer locations: Unique 30-mers are mapped to the genome reference using mrsFAST (Hach et al. 2010) with an edit distance setting of 0. This step serves to map the location of each unique 30-mer and is used in the following steps for region overlapping and exclusion.

3. Enumerate highly repetitive 15-mers: The same procedure for Step 1 is repeated for the reference assembly except now k is set equal to 15 and all 15-mers with counts ≥ 1,000 are exported.

4. Determine repetitive 15-mer locations and filter 30-mers: Step 2 is repeated with the 15-mers determined in Step 3. Here, each k-mer will have multiple genome locations. Finally, locations of the 15 and 30-mers are merged together, and all 30-mers (from Step 2) that overlap with the high frequency 15-mer track are removed from subsequent analyses.

5. Remove highly similar 30-mers: The 30-mers that pass Step 4 are mapped onto the reference genome using mrsFAST with an edit distance of 2. All 30-mers with ≥ 100 mapped positions are removed. Note that mrsFAST only considers substitutions.

6. Considering indels, remove highly similar 30-mers: The k-mers that pass Step 5 are mapped again using mrFAST with an edit distance of 2. All 30-mers with ≥ 100 mapped positions are removed. The mrsFAST search is performed prior to mrFAST due to the speed advantage of mrsFAST only considering mismatches. Steps 5 and 6 serve to reduce the chances of matching k-mers with sequencing errors into unintended locations.

7. Combine final k-mer catalog: The final list of highly unique 30-mers is sorted based on chromosome location and outputted in BED format. This output file will then be used by QuicK-mer and for the generation of required auxiliary files.

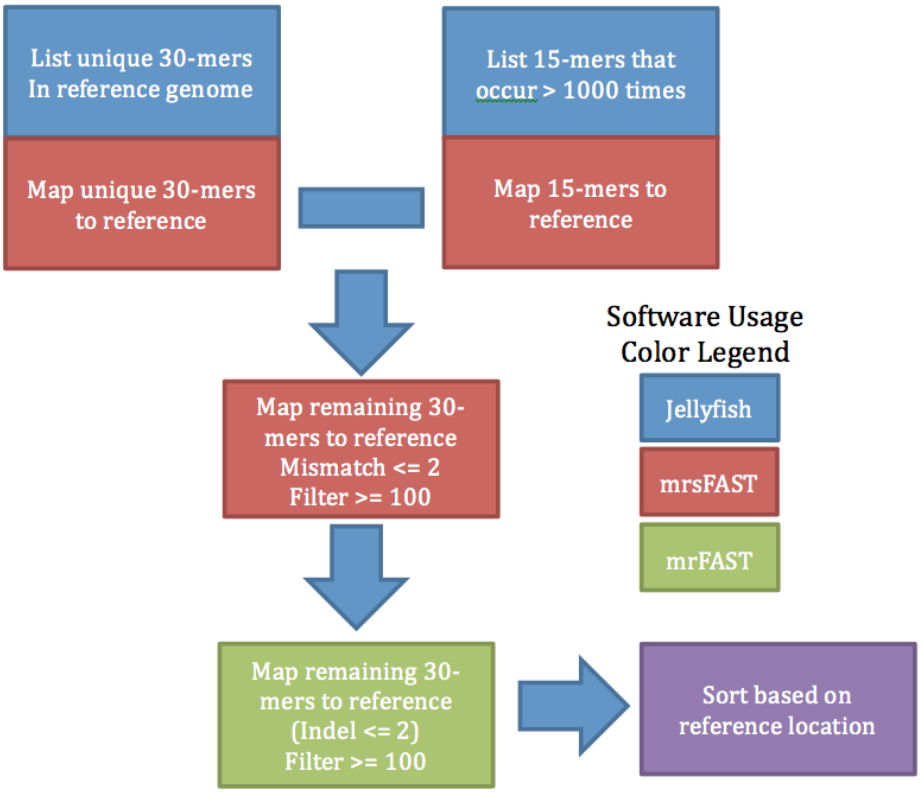


Figure 6 Reference 30-mers generation flow

Diagram depicting QuicK-mer workflow for 30-mer generation from a genome reference assembly. The color for each box corresponds to each of the three software packages used. Multiple rounds of filtration steps guarantee sequencing errors won't map to highly represented regions.

2.2.2.2 Definition of control regions

The resulting file is encoded in a binary format for convenient access during the GC correction and copy number normalization step.

2.2.2.3 PCR Bias and GC correction

To counter the amplification bias during library preparation and flow cell bridge amplification, a moving window of local GC content is calculated for a reference genome assembly. For each K-mer, this window is taken by extending from the central base pair by half of the window length. In our study, the GC window is set to a value of 400bp, which is typical for a WGS library. In the same manner as the previous step, the values are stored in binary file for rapid access.

2.2.2.4 Realization of pipeline

2.2.2.4.1 Depth estimation and GC correction

The QuicK-mer core program is written in C++ and Object Pascal and wrapped with Python for control flow. The control flow consists of calling Jellyfish-2 (Marçais and Kingsford 2011) for building the 30-mer hash library followed by the k-mer query step. At the beginning of the query step, two axillary binary files are preloaded and memory space for count values is allocated. QuicK-mer then interrogates the Jellyfish hash library with the sorted 30-mer list, storing each raw count value in memory. The core program verifies each 30-mer's status as a normalization control and, if indicated, the 400 bp GC-content value is fetched from the associated binary file and incorporated into the GC bias curve. Once the process is finished, the core program builds the GC curve based on the

average counts obtained from each GC percentage bin and uses the lowess smoothing algorithm to generate a correction curve. The targeted average depth is calculated using a weighted average based on GC content of 25~75%. A 0.3x minimum and 3x maximum correction factor is also enforced to reduce over-correcting extreme GC regions due to a lack of representative k-mers. The GC bias curve is output in a text format and, along with correction curve, is represented in a PNG image (example in Figure 7). Lastly, the correction factor is applied to each k-mer count value based on its GC content and the resulting GC-corrected k-mer counts are outputted in binary format.

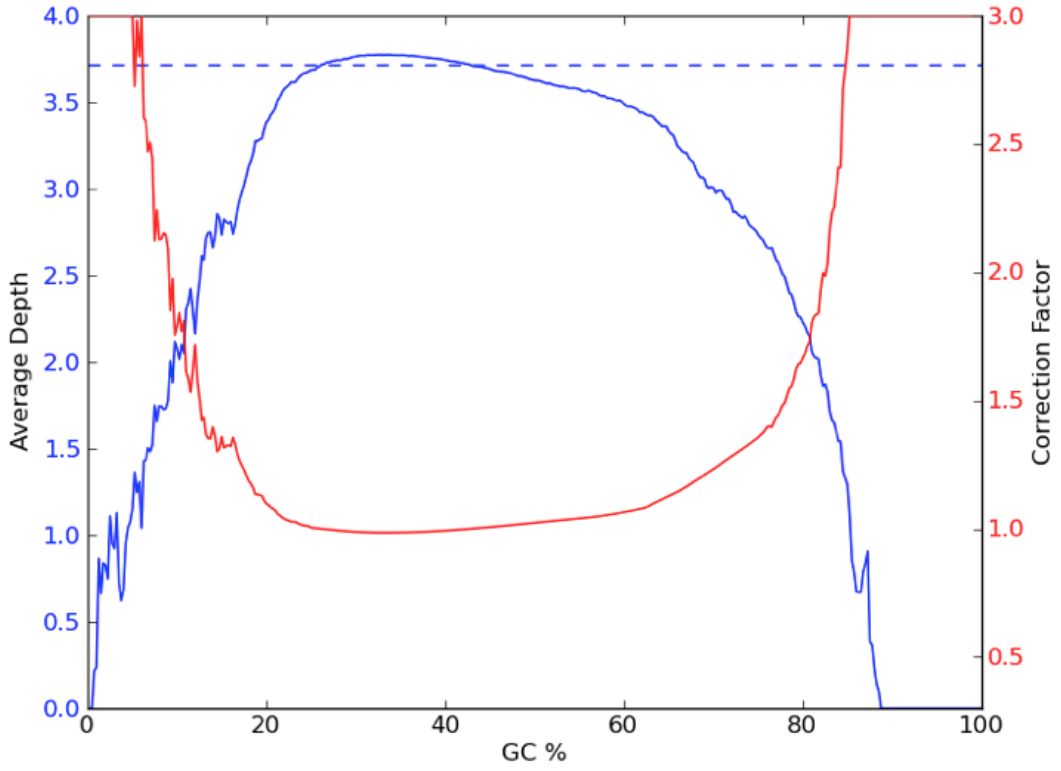


Figure 7 GC correction curve

GC Correction and Bias The majority of sequencing coverage bias is related to local GC content. The blue curve indicates the average depth for the 30-mers with the same GC content in 400bp surrounding the center of each k-mer location, rounded in steps of 0.25%. The red curve is the lowess smoothed correction factor, targeted for the average depth indicated by the dashed line. A 3x max correction value is enforced. The GC curve represents QuicK-mer run from WGS experiment SRX734522.

Due to different GC biases within sequencing libraries and across flow cell lanes, the user is encouraged to apply QuicK-mer GC normalization separately for each sequencing lane.

Resulting GC-corrected k-mer counts can then be merged together for each sample using the CorDepthCombine command.

2.2.2.4.2 Normalization and CNV estimation

Another program in the QuicK-mer package (kmer2window) converts counts to copy number estimates. The normalization program loads the same binary control region file then, using the corrected depth for the control 30-mers, calculates a scaling factor based on an assumed copy number of two for these regions. Normalization is performed on windows of equal number of k-mers (default = 500 k-mers per window, but is adjustable by the user). The median k-mer count for each window is used for the normalization, and only windows where all k-mers are in the defined control intervals are used in subsequent steps. The resulting normalization is then applied to all windows.

2.2.2.5 Performance

To assess the efficiency of QuicK-mer, we randomly sampled subsets of reads from the human genome sequence for sample HG02799, which was sequenced to a depth of 17x. The selected fractions were individually analyzed using 35GB memory and 4 cores during library construction and 2 cores during querying on an empty compute node with 4 Xeon E7 4850 2GHz processors and 1TB of total memory. Wall clock-time statistics

indicate a constant time cost for the querying step once the average sequencing depth exceeds 1x. The nature of counting predefined k-mers also means the memory usage is unlikely to be affected by the sequencing depth. The library building time is linearly correlated with the input read counts.

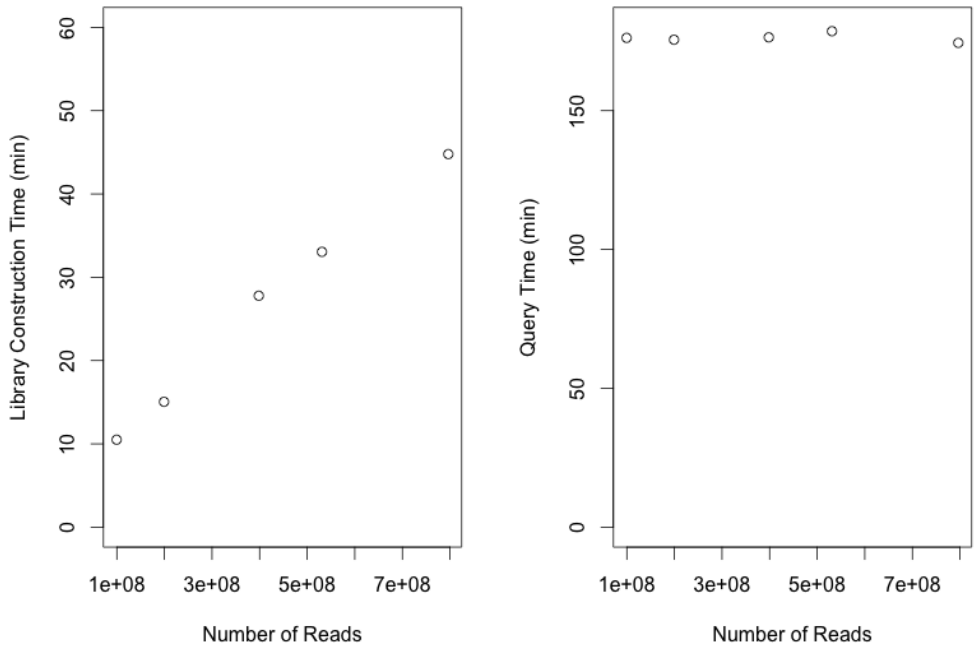


Figure 8 QuicK-mer 1.0 CPU wall time

Wall time statistics of QuicK-mer based on random read sampling of a HG02799 Illumina library. The left panel shows the time cost for Jellyfish 2 to construct the k-mer array from a FASTQ file with varying number of reads. The right panel shows constant query time for each k-mer using prebuilt k-mer list.

2.2.2.6 Validation with 1000 genome dataset

For comparison, we reanalyzed the dataset from the 1000 Genome Project and other sources using QuicK-mer and compared the estimated copy number profiles with supplementary data from the (Sudmant et al. 2010) study. The dataset was downloaded from 1000 Genome Project Pilot, Phase 1, and Phase 3 studies (1000 Genomes Project Consortium et al. 2010, 2012, 2015). Sequencing files were individually run through the QuicK-mer pipeline and GC corrections were performed for each sequencing lane. Corrected data is combined and normalized into copy number estimates. Table 1 contains the details of samples used to assess the accuracy of QuicK-mer in known CNV regions.

Table 1 List of samples used for QuicK-mer validation

Samples used for QuicK-mer validation. The mean depth is calculated based on the median depth obtained from windows of 500 30-mers that fully overlap a defined control region. For sample NA19240, the SRA accession identifiers are provided.

Sample Name	**Data Source**	**Mean 30-mer Depth in Control**
NA12156	1000 Genome Phase 3	4.40
NA12878	1000 Genome Phase 1, Pilot 1/2	7.23

NA18507	(Bentley et al. 2008)	23.05
NA18508	(Bentley et al. 2008)	7.30
NA18517	1000 Genome Phase 3	3.82
NA18555	1000 Genome Phase 3	4.09
NA18956	1000 Genome Phase 3	3.63
NA19129	1000 Genome Phase 1, Pilot 1/2	0.87
NA19240	SRX574476, SRX582073 (Song et al. 2017)	13.26

We evaluated the genome regions depicted in S52 and S60 – S71 of (Sudmant et al. 2010). QuicK-mer accurately estimated copy number for many highly paralogous gene families, such as the *UGT2* gene family (Figure 9), for each sampled human genome. Other regions in which QuicK-mer CNV estimations are consistent with the original study, further demonstrating the accuracy of QuicK-mer in detecting copy number of unique paralogs.

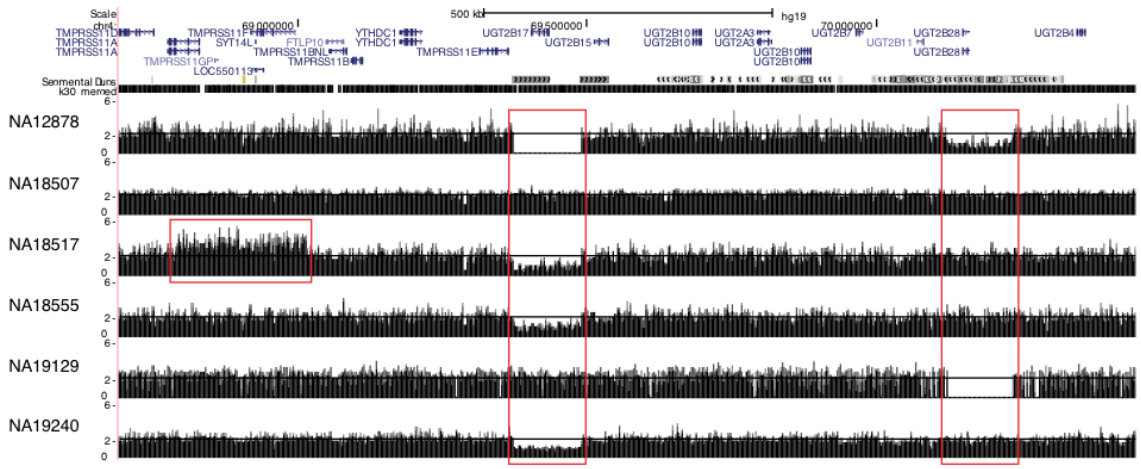


Figure 9 Validation using 1000 Genome data

Diverse and paralog-specific CNV detected by QuicK-mer at the *UGT2* family locus for numerous gene models (top track) at chr4q13.2. Red boxes indicate regions of detected CNV. *UGT2B17* is hemizygously deleted in NA19240, NA18555 and NA18517. TMPRSS11F and SYT14L are duplicated in NA18517, resulting in a copy number of 3. This figure corresponds to the region shown in Figure S65 in (Sudmant et al. 2010). The k30_merged track indicates the locations with unique 30-mers.

2.2.4 Application of QuicK-mer

2.2.4.1 Chimpanzee Y-chromosomal amplicon copy number detection

2.2.4.1.1 Methods

In addition to validation on the human data, we employed QuicK-mer to amplicon copy number on Y chromosomes on primates. Since these amplicons is already duplicated on chrY, we specifically isolate one copy for each family as a separate FASTA file. We then generated the k-mer unique to each amplicon and only kept the ones distinct from other

ampicon copies. This ensures paralog sensitivity. The additional k-mers were then appended to the ordinary k-mers generated based on PanTro4 reference. QuicK-mer is then ran on samples using human as a comparison.

2.2.5 QuicK-mer 2.0: Speed improvement and user-friendly interface

The original QuicK-mer was published with four sets of prebuilt 30-mer indexes. These include human HG19, mouse MM10, chimpanzee PanTro4 and dog CanFam3.1. These files are large and hard to distribute to end users. If users are dealing with a different reference version or completely distinct organism, they are required to spend four days and hundreds of independent tasks for processing. These processing steps uses mrsFAST and mrFAST for repeated mapping, merging, sorting and intersection described in section 2.2.1. These steps require constant human intervention. We feel like a unified and user-friendly improvement is necessary to make QuicK-mer suitable for general public.

To further improve performance and simplify the pipeline, a new version of QuicK-mer V2.0 with a novel internal core was designed and released. In this new version, a single application contains all three functionalities. 1. Searching unique k-mer list from reference assembly 2. Counting k-mer from sample FASTQ files. 3. GC correction and copy number estimation. This results in a more user-friendly program. In addition, a

single sample scan combining the previous enumeration/query steps results in a massive run time reduction.

2.2.5.1 K-mer encoding

Each base pair can be represented using two bits. Based on the standard ASCII table encoding, bit position 1 - 2 happens to be unique across A, T, G and C. More importantly, A and T, or G and C happen to differ by two. Thus, reverse complement conversion can be realized with subtraction of value of two in a two-bit unsigned integer space. The use of both encoding tricks reduces CPU instruction cycles and avoids branching instruction execution with simple bit manipulation.

Base	2-bit Binary
A	00
T	10
C	01
G	11

Table 2 Bit encoding of QuicK-mer 2.0

Binary representation of four nucleotides using bits 1-2 from ASCII encoding of characters. Efficient complementary conversion can be achieved by addition of two (2'b10) in two-bit space. Overflow will cause the value to flip back.

A 64-bit unsigned integer can store a maximum length of 32 base pairs. This is sufficient in most of the applications. In QuicK-mer 2, the 3'-end base pair is encoded in the least

significant bit in the integer. During stream processing, previous values can be shifted to the left every two bit at a time. Since nucleotide ‘N’ found in sequencing data is ambiguous, k-mers containing ‘N’ characters are discarded. Each k-mer can be represented using both the forward or reverse strand. This redundancy is resolved by taking the smaller value after encoding for hashing process. Thus, a k-mer and its reverse complement are taken to be identical.

2.2.5.2 Data structure

A reference genome with length of N can in the worst-case scenario generate N-k+1 k-mer should no repeats occur. With N in the range of billions, storing such an array efficiently while also enabling rapid access presents a challenge. To quickly access the depth information of each k-mer, several mapping strategies have been proposed in the literature. In the end, there is usually a tradeoff between computing time or memory space usage. Burrow wheeler transformation encoding with a suffix array is a typical example of trading computing time for space. For example, BWA and BOWTIE used this approach.(Li and Durbin 2009; Langmead et al. 2009) Hash functions on the other hand trades space for time and provide constant O(1) access time independent of the size of the array. We chose this method since the number of k-mer can be easily reduced without sacrificing depth estimation precision. On the other hand, with multithreading, each

thread on average does not use much memory. Lastly, the time cost at large datasets is typically more important compared to memory usage and QuicK-mer 2.0 can finish a 20x high depth analysis in less than 20 minutes.

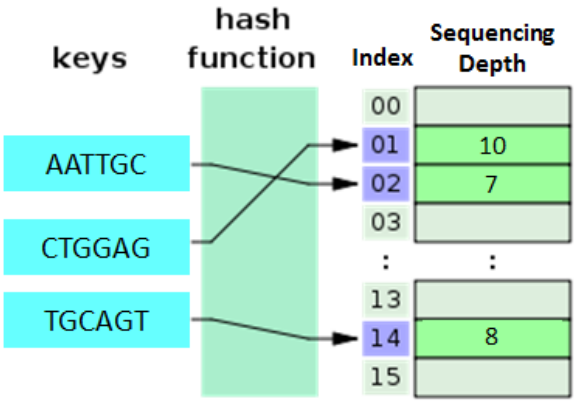


Figure 10 Hash array and data structure

Hash data structure used in QuicK-mer 2.0. Encoded k-mers are transformed into a random array index through a hash function. This randomness is determined by the hash function. The array slots can store information such as the sequencing depth, the original encoded k-mer itself and next array index on chromosomal order.

In a hash array, the index value is determined by converting the item value using a hash function. An ideal hash function will generate a distinct index value for each distinct input. However, the same value from different inputs can occur in reality. This is called a hash collision. A good hash function should have a low collision rate and a uniform distribution of index value given a fixed space. To resolve collisions, QuicK-mer 2

employs the linear probing approach, where the value colliding is appended in the adjacent array cell. In QuicK-mer 2.0 the appending direction is flipped between the upper and lower half of the array. This way the array size will not exceed allocated memory. In such a collision resolving scheme, a k-mer search scan will start at the hash index, follow the indicated direction until the encoded k-mer value cell is found, or stop when an empty cell is reached where the k-mer is absent.

Additional arrays with the same indexing strategy are allocated for other purposes. During k-mer enumeration from a reference genome assembly, two integer arrays are used to store the occurrence of each k-mer and to store the number of repeats during edit distance search. Additionally, a linked list stores the exact index of the next k-mer. This is used to reorganize the depth information into chromosomal order.

2.2.5.3 K-mer hashing with DJB2

In QuicK-mer 2, I chose DJB2 due to its speed and efficiency. Each 64-bit encoded k-mer is considered a string length of eight characters. In total DJB2 goes through each character for hashing. The final index is calculated by taking modulo of the hashed value.

2.2.5.4 Edit distance search

During the search step, each k-mer can be permuted with up to two substitutions. The additional time cost can be calculated in the following equation. Since this step only requires a shared memory for depth access, multithreading is implemented to facilitate the process. Compared to the 1.0 version, 256 CPU hours is sufficient. Previously the complex filter procedure requires hundreds of mapping jobs using external tools and intersecting calculations which take multiple days.

$$Permutations = C\ {}^{2}_{k} \times 3\ {}^{2} + C\ {}^{1}_{k} \times 3$$

2.2.5.5 Multithreading implementation

QuicK-mer 2 also implemented multithreading in the count step. The multithreading is designed as a feeder - consumer scheme, where the feeder thread fetches sequencing reads and generates encoded k-mer values while each consumer thread hashes the encoded k-mer, search and accumulate the depth in corresponding index location using the lock-add CPU instruction. The process is thread safe and scales according to input read count and thread number. Eventually the process becomes I/O bounded at more than 6 consumer threads. The process is able to finish for a 10x genome in well under 10 minutes.

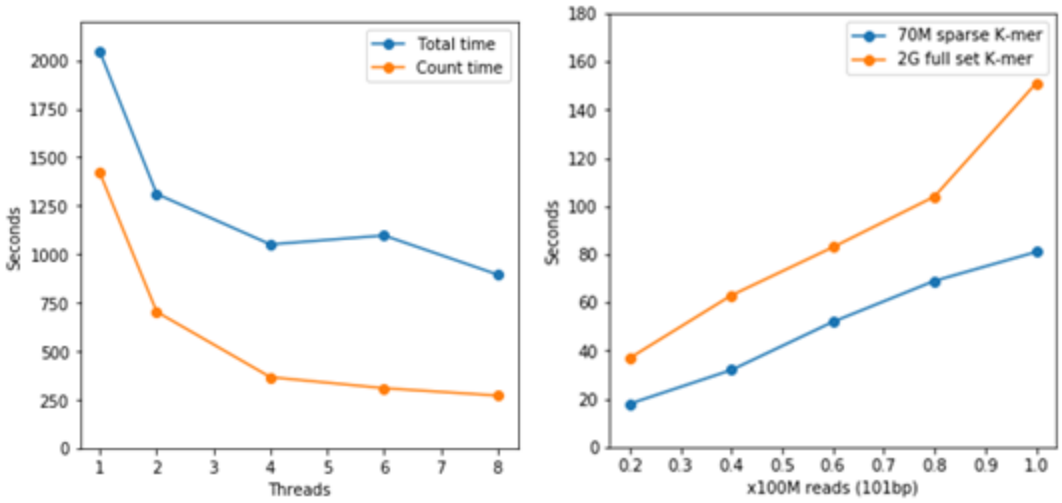


Figure 11 QuicK-mer 2.0 CPU time and multithreading

Time cost for QuicK-mer 2 counting step with varying number of thread and input data

2.2.5.6 K-mer sparsing and memory reduction

QuicK-mer 2 requires 10 bytes per k-mer of memory space during the counting step. Thus, for a typical mammalian genome two billions of k-mers are enumerated. Since hash table should exceed an 80% fill rate to avoid excessive collision, 40GB of memory is required for index. Here I explored the reduction of control region in order to conserve memory space. Naturally, we do not have to count every k-mer by shifting 1bp. Instead, we can skip k-1 k-mers for every k as sequencing reads are contiguous. To demonstrate, the reference k-mer list is progressively reduced to 1/100 for k=30. We demonstrate that the GC correction accuracy is not affected (Figure 12). This approach drastically reduces

the memory consumption. In the meantime, there's a small time reduction in CPU time as well, possibly due to lower chances of cache miss when smaller memory is used.

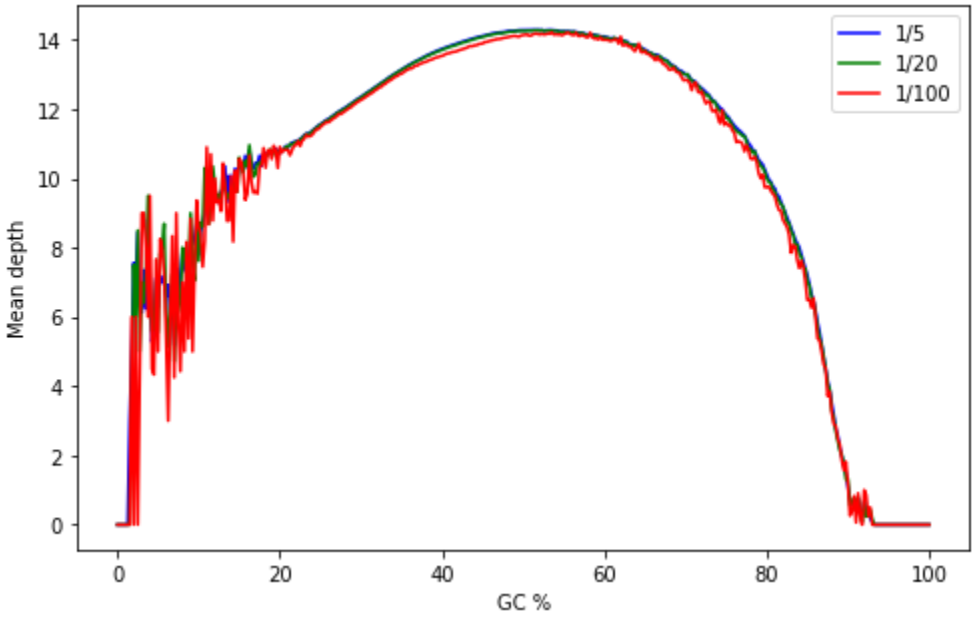


Figure 12 GC correction using sparse function

Successive reduction of k-mers in control region does not affect GC bias curve. Accurate GC correction can be achieved with smaller memory footprint

2.2.5.7 Summary

Here I presented a paralog specific CNV tool with high efficiency. The new version provides all in one k-mer search and filter along with count and estimation in a single user-friendly application. The new QuicK-mer 2 stores accumulated uncorrected GC

information as intermedia file. This method enable user to iteratively refine control region based a population of samples to find regions with copy number two for GC correction. New version of QuicK-mer 2 is available on GitHub at https://github.com/jackshencn/QuicK-mer2

2.3 fastCN: A multi-mapping CNV detection pipeline

2.3.1 Introduction

Multiple approaches that utilize read depth to identify regions of copy number variation have been developed. One successful set of approaches utilize the mrFAST and mrsFAST aligners, tools which efficiently return all matching locations for short sequencing reads within a specified edit distance. These tools have been used to analyze CNV patterns in multiple studies of humans and non-human primates (Sudmant et al. 2015, 2013; Alkan et al. 2009; Sudmant et al. 2010). However, this estimation required separate steps including mapping, BAM file sorting based on location, and read pileup followed by GC corrections, requiring the storage and manipulation of several large files. Since the total time for disk I/O and the use of multiple intermediate files is a serious bottleneck for large scale analyses, we developed fastCN to efficiently estimate genome copy number from short read data. This program utilizes the data output from the short

read mapper mrsFAST (Hach et al. 2010), and reports per-bp read depth in an efficient compressed binary format. The fastCN software package is available on the Kidd Lab GitHub.

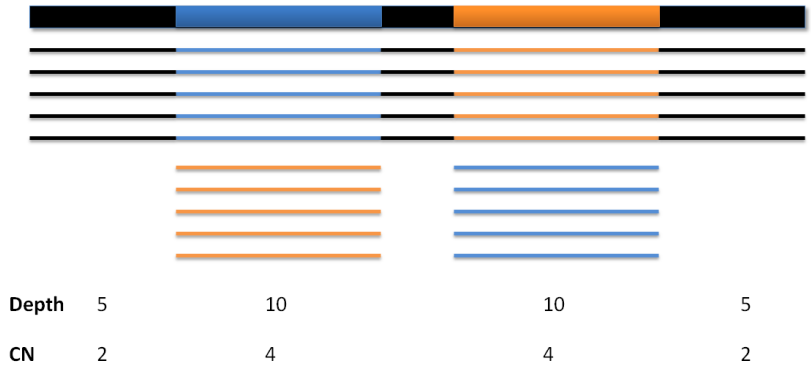


Figure 13 fastCN working principle

fastCN is implemented using a paralog insensitive approach. This example shows both paralog copies have the other's reads mapping to it. Thus, the final copy number is four in a sample matching the original reference assembly.

2.3.2 Implementation and optimization

The fastCN core pipeline consists of two major applications responsible for generating reference files and depth pile up respectively.

2.3.2.1 Reference file

The first program, GC_control_gen, generates a control region file for the next stage of the pipeline based on (1) the reference genome and user supplied files indicating (2)

regions of the genome assumed to not be copy number variable and (3) regions of the genome which have been masked prior to read mapping. To avoid excessive depth pile ups due to repetitive regions, we utilize a version of the genome reference where all elements defined by RepeatMasker, elements defined by tandem repeat finder (TRF), and 50-mers with at least 20 genome matches within an edit distance of two are masked to 'N' prior to short read mapping. To avoid the shadow effect of mapping against a masked genome, the coordinates of the masked segments are extended by the length of the utilized reads. For compatibility with previous work, we utilize a read length of 36 bp, and divide longer Illumina reads into disjoint 36 bp long sequences.

The control region file is encoded as a 32-bit float per base pair in a pure binary representation, and the value for each float corresponds to the local GC content at that base pair. The window length for GC content calculation is defined by the user (typically 400 bp), and GC content values are assigned to the centers of each sliding window. Signed values are given for each base-pair with expected values between -1.0 and 1, where a negative bit value indicates a base pair that should not be used as a control for normalization. Masked genomic regions are identified by a value of negative infinity and are omitted from processing. This encoding scheme allows rapid access during the

subsequent normalization stage. For each reference assembly, the above step should be executed once.

2.3.2.2 Depth pileup

The second application, SAM_GC_correction, processes the data output from the mrsFAST mapper (an unsorted SAM file format). As such, a memory space proportional to the size of the haploid reference genome is required for this random access. Once the mapping input is processed, GC normalization ensues with the aid of the binary file from the previous step. GC normalization utilizes a multiplicative correction factor determined by lowess fitting, as utilized in QuicK-mer. The end result is corrected depth preserved with half floating-point precision, which contains sufficient dynamic range and precision while significantly saving disk space. Depths at regions masked out in the reference are assigned a fixed depth value of -1.0. The resulting binary normalized depth files are subsequently compressed using gzip. Mean or median depth values in predetermined windows can then be efficiently calculated from these files, and converted to estimates of genome copy number.

2.3.3.3 Depth combine

A utility application is included for the user to convert between half float and single precision float point. Please refer to readme file from the fastCN software package on GitHub for additional instructions.

2.3.3 Performance

The fastCN pipeline achieves excellent performance. The core pipeline consumes negligible additional time compared to mrsFAST mapping. The control region files can be constructed less than 3 minutes for a typical 3Gb mammalian genome.

2.4 References

1000 Genomes Project Consortium, Gonçalo R. Abecasis, David Altshuler, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Richard A. Gibbs, Matt E. Hurles, and Gil A. McVean. 2010. "A Map of Human Genome Variation from Population-Scale Sequencing." Nature 467 (7319): 1061–73.

1000 Genomes Project Consortium, Goncalo R. Abecasis, Adam Auton, Lisa D. Brooks, Mark A. DePristo, Richard M. Durbin, Robert E. Handsaker, Hyun Min Kang, Gabor T. Marth, and Gil A. McVean. 2012. "An Integrated Map of Genetic Variation from 1,092 Human Genomes." Nature 491 (7422): 56–65.

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. "A Global Reference for Human Genetic Variation." Nature 526 (7571): 68–74.

Alkan, Can, Jeffrey M. Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O. Kitzman, et al. 2009. "Personalized Copy Number and Segmental Duplication Maps Using next-Generation Sequencing."

Nature Genetics 41 (10): 1061–67.

Bentley, David R., Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, et al. 2008. “Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry.” Nature 456 (7218): 53–59.

Chaisson, Mark J. P., John Huddleston, Megan Y. Dennis, Peter H. Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, et al. 2015. “Resolving the Complexity of the Human Genome Using Single-Molecule Sequencing.” Nature 517 (7536): 608–11.

Hach, Faraz, Fereydoun Hormozdiari, Can Alkan, Farhad Hormozdiari, Inanc Birol, Evan E. Eichler, and S. Cenk Sahinalp. 2010. “mrsFAST: A Cache-Oblivious Algorithm for Short-Read Mapping.” Nature Methods 7 (8): 576–77.

Hartasánchez, Diego A., Marina Brasó-Vives, Jose Maria Heredia-Genestar, Marc Pybus, and Arcadi Navarro. 2018. “Effect of Collapsed Duplications on Diversity Estimates: What to Expect.” Genome Biology and Evolution 10 (11): 2899–2905.

Hoeppner, Marc P., Andrew Lundquist, Mono Pirun, Jennifer R. S. Meadows, Neda Zamani, Jeremy Johnson, Görel Sundström, et al. 2014. “An Improved Canine Genome and a Comprehensive Catalogue of Coding Genes and Non-Coding Transcripts.” PloS One 9 (3): e91172.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” Genome Biology 10 (3): R25.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” Bioinformatics 25 (14): 1754–60.

Marçais, Guillaume, and Carl Kingsford. 2011. “A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers.” Bioinformatics 27 (6): 764–70.

Oetjens, Matthew T., Feichen Shen, Sarah B. Emery, Zhengting Zou, and Jeffrey M.

Kidd. 2016. “Y-Chromosome Structural Diversity in the Bonobo and Chimpanzee Lineages.” Genome Biology and Evolution 8 (7): 2231–40.

Patro, Rob, Stephen M. Mount, and Carl Kingsford. 2014. “Sailfish Enables Alignment-Free Isoform Quantification from RNA-Seq Reads Using Lightweight Algorithms.” Nature Biotechnology 32 (5): 462–64.

Pendleton, Amanda L., Feichen Shen, Angela M. Taravella, Sarah Emery, Krishna R. Veeramah, Adam R. Boyko, and Jeffrey M. Kidd. 2018. “Comparison of Village Dog and Wolf Genomes Highlights the Role of the Neural Crest in Dog Domestication.” BMC Biology 16 (1): 64.

Song, Shiya, Elzbieta Sliwerska, Sarah Emery, and Jeffrey M. Kidd. 2017. “Modeling Human Population Separation History Using Physically Phased Genomes.” Genetics 205 (1): 385–95.

Sudmant, Peter H., John Huddleston, Claudia R. Catacchio, Maika Malig, Ladeana W. Hillier, Carl Baker, Kiana Mohajeri, et al. 2013. “Evolution and Diversity of Copy Number Variation in the Great Ape Lineage.” Genome Research 23 (9): 1373–82.

Sudmant, Peter H., Jacob O. Kitzman, Francesca Antonacci, Can Alkan, Maika Malig, Anya Tsalenko, Nick Sampas, et al. 2010. “Diversity of Human Copy Number Variation and Multicopy Genes.” Science 330 (6004): 641–46.

Sudmant, Peter H., Swapan Mallick, Bradley J. Nelson, Fereydoun Hormozdiari, Niklas Krumm, John Huddleston, Bradley P. Coe, et al. 2015. “Global Diversity, Population Stratification, and Selection of Human Copy-Number Variation.” Science 349 (6253): aab3761.

Zhang, Zhaojun, and Wei Wang. 2014. “RNA-Skim: A Rapid Method for RNA-Seq Quantification at Transcript Level.” Bioinformatics 30 (12): i283–92.

Chapter 3: Detection of copy number variation associated with dog domestication

Portions of this chapter including figures, tables and text was published and described in (Pendleton et al. 2018). This chapter details the application of QuicK-mer and fastCN to the study of canine domestication.

3.1 Introduction to domestication

Domestication of animals and plants is a complex process that accompanies human evolution for at least the last ten thousand years (Larson and Fuller 2014). Generally, domestication is defined as an evolutionary process that gradually transforms organisms to suit human needs with some form of human intervention. Domestication was once viewed as humans actively selecting traits and phenotypes. This was the core concept of selective breeding in recent times. Another process might be non-deliberate actions and subtle influences from humans that alter an organism's environment, which further leads to adaptation by certain organisms to human presence. As a result, some common

characteristics from domestic animals are consistently observed. This includes increased tameness and docility, changes of coat color, presence of floppy ears, changes of reproduction frequency and finally, body shapes across a variety of animals. This observation is known as the "Domestication Syndrome" noticed first by Darwin (Wilkins, Wrangham, and Fitch 2014; Larson and Fuller 2014). Debates on how and why similar morphological traits arise are still ongoing. Evidence from Belyaev's wild fox experiment refuted the earlier hypothesis that the traits observed in dog domestication were the result of hybridization of a myriad of breeds (Trut, Plyusnina, and Oskina 2004). The result of this study leads to another line of thought, that some common upstream regulatory network governs the genes responsible for all the phenotypes observed in domestic syndrome.

With multiple genome assemblies largely complete and well annotated, studies on regulatory pathways and gene networks suggests this might also be infeasible. The sheer number of pathways involved for a common regulatory factor would be too much for only a few traits observed. Recently, Wilkins et al proposed that initial selection of tameness might drive the "domestication syndrome" through affecting the neural crest migration during the early embryo development (Wilkins, Wrangham, and Fitch 2014). The implication of the neural crest pathway in the "domestication syndrome" is that

selection by tameness alone could affect multiple pathways due to common cell lineage during the early embryo development.

3.1.1 CNV in animal domestication

As a large-scale variant, CNV represents an important aspect during domestication. Numerous studies on farm and companion animals have revealed diverse changes of the genomic landscape due to CNVs. In pig domestication, a high resolution CNV map based on genome sequencing data revealed large expansion of olfactory receptors as the primary source of copy number increase (Paudel et al. 2015; C. Chen et al. 2012). Genes related to immune defense were identified as copy number variable across the pig population. Several genes related to starch digestion including amylase are variable as well. The same study deduced that copy number evolves at a much faster rate than SNPs. The sheer number of these CNV regions might contribute to rapid speciation and adaptation through interspecific hybridization. Other CNV surveys highlighted CNV regions overlapping with QTL linked to meat quality and fat deposition (C. Chen et al. 2012). Diseases associated genes due to copy number changes are also discovered (Long et al. 2016; Wang et al. 2015).

The CNV landscape in the cattle genome shows some similarity to that from domesticated pigs. CNV is enriched in the immune system related pathway and olfactory receptors (Jia et al. 2013; Keel, Lindholm-Perry, and Snelling 2016), with segmental tandem duplication as the primary cause for copy number expansion. Domestication related CNV is also observed in chickens. Pea-comb shape was one of the earliest discovered traits related to CNV in chicken (Wright et al. 2009). Feather features such as timing and coat color are also shown driven by CNV induced variations (Elferink et al. 2008; Dorshorst et al. 2011). Whole genome tiling array analysis revealed pathways enriched with CNV expansion (Jia et al. 2013). Yet the implications of such large-scale change are difficult to draw because phenotypes like meat growth are usually polymorphic.

In all, copy number variation represents an essential component in animal evolution and domestication process. To further gain insights on the regions evolved during dog domestication, it remains essential we also look into the copy number variations.

3.2 Dog domestication history

The domestication history of dog stands in a unique place among all other animals. Foremost, the dog is the first animal to be domesticated, which shares at least ten

thousand years alongside humans. Secondly, the recent breed formation has been carefully recorded, allowing breed associated genetic differences to be easily compared against the breed tree. Thirdly, dogs exhibit a wide range of phenotypic and morphological diversity within a species unlike any other species.

Two important aspects of domestication are its timing and location. This timing is usually determined by two methods either archaeologically or genetically. Excavated fossils maintained the bone shape at that time. By comparing bone shape to the sample of living animals, one can determine the relative distance of ancient fossils to the modern domesticated animal. The age of a fossil can be determined using radioisotope dating by residual carbon-14. Similarly, such comparisons can also be drawn at the molecular level by comparing the number of mutations between the wild and domesticated samples. The relative time can be inferred by assuming a constant mutation rate in the genome. The physical evidence provided by fossils can be compelling. But early dogs would be rare and are also difficult to distinguish from the wolf morphology (Freedman and Wayne 2017). Another problem would be divergence of modern wild species compared to the ancient individuals from where were domesticated. In the case of dog domestication, the ancient wolf progenitor is now extinct (Freedman et al. 2014). Thus, by combining molecular timing, a more accurate picture can be drawn especially when ancient dog

DNA is available. Based on distance clustering, domestication location can be inferred based on proximity to older ancient samples.

At present, the timing and location for dog domestication is still under debate, with each method drawing a different conclusion. Factors affecting the result including the region of DNA used (mitochondria, Y chromosome or autosomes), filtration and clustering algorithm and criteria, as well as the most important, the samples selected and used (Freedman and Wayne 2017). Yet general consensus place the domestication split time around 10-40 thousand years ago and a single place of origin based on shared common haplotype (Freedman and Wayne 2017). Recently, our research group surveyed more than 5,000 dogs along with two ancient dog DNA samples. The result supported single origin theory and narrowed the domestication time frame to 2-40,000 years ago (Botigué et al. 2017). After this ancient, long enduring domestication process, the modern breed dogs were the result of selective breeding for certain desirable traits in the last 250 years. This breed formation creates the second population bottleneck, where the first being the origin domestication event tens of thousands of years back.

3.3 Results

To truly capture the domestication signature that distinguishes between dogs and wolves, we need to isolate the origin of mutations in the 40,000 years' history. Previous studies on many of such selection scans focused on using breed dogs as a resource (Freedman et al. 2014; Marsden et al. 2016). Village dogs were the relative wild dogs detached from human selective breeding. Due to this property, village dogs are more pristine and unlikely harbors regions affected due to breeding process.

3.3.1 Copy number estimation using Illumina sequencing data

Copy number was estimated using Illumina whole genome sequencing data with both the fastCN and QuicK-mer methods described above. Input sequencing data for both approaches was derived from BAM files with duplicated reads removed. Non-CNV autosomal control regions for depth normalization were predefined for the CanFam3.1 reference by excluding regions previously reported to be duplicated or copy number variable (Nicholas et al. 2009, 2011; Freedman et al. 2014; W.-K. Chen et al. 2009). Copy number estimates were created in windows of 3,000 unmasked bp (fastCN) or 3,000 unique k-mers (QuicK-mer) for the autosomes and chromosome X. Unplaced contigs were merged into one chrUn for copy number estimation.

3.3.1.1 QuicK-mer CN estimation

The canine reference assembly was divided into consecutive windows that each have 3,000 k-mers. Since k-mer locations are not uniform or consecutive, the actual genomic span (or length) of each window varies depending on the local sequence complexity. Window definition is constructed using an utility application in the QuicK-mer pipeline (Chapter 2). Copy number estimates were calculated using the kmer2window program, which requires the 3,000 k-mer windows, as well as the normalized binary files for each sample (available for download at http://kiddlabshare.umms.med.umich.edu/public-data/QuicK-mer/Ref/).

3.3.1.2 fastCN CN estimation

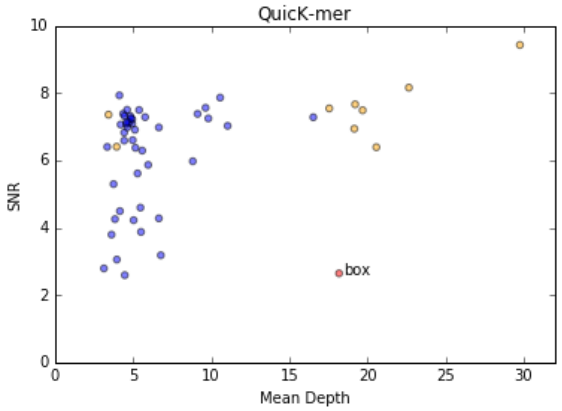
Similar to QuicK-mer CNV estimation, we divided the canine genome into consecutive 3kb windows, with the exclusion of masked regions defined in the fastCN pipeline. The depth for each window is first assigned the mean normalized depth of all intersecting unmasked base pairs.

This value is then scaled to copy number estimate per window by dividing the average depth in all control windows assumed to be copy number of two.

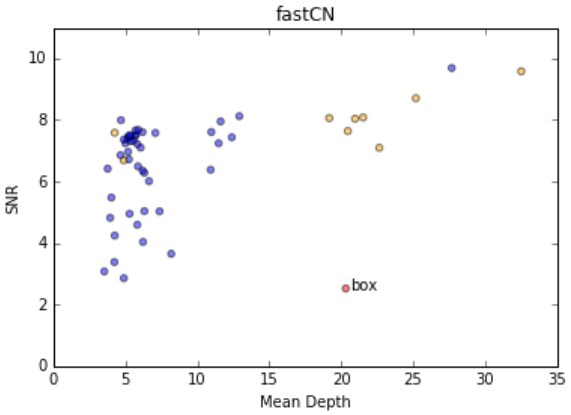
3.3.2 Comparison of noise across samples

The signal-to-noise ratio (SNR), defined as the mean depth in autosomal control windows divided by the standard deviation, was calculated for the 53 dogs and wolves that were processed through the F_{ST} pipeline. Because the wolf samples were typically sequenced to a higher depth than the village dog samples, wolves display larger SNR than dogs (Figure 14 A and B). However, many village dogs with lower average sequence depth (~4-10x) exhibit comparable SNRs with wolves that have higher depths. The correlation of noise in control regions between both pipelines indicates a consistent noise originating from the sequencing data (Figure 14 C). Results from a boxer breed dog (box), which is used in subsequent QuicK-mer and fastCN validations, are also included in these plots. The SNR of this sample indicates that the boxer sequencing data is unusually noisy, an observation accounted for in later analysis.

A



B



C

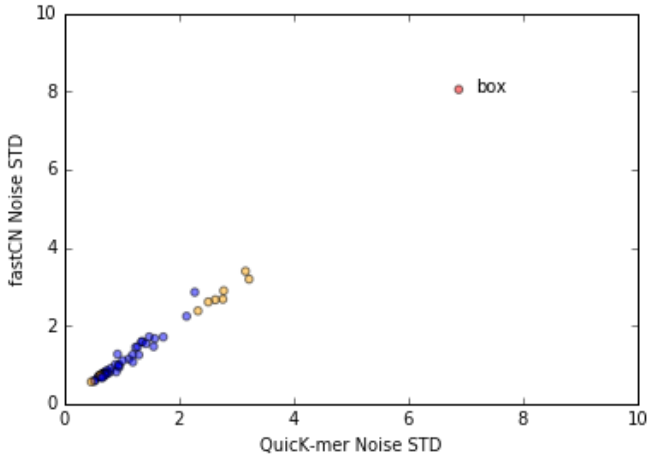


Figure 14 Sample coverage noise

SNR values based on (A) QuicK-mer and (B) fastCN (upper right) analyses are plotted against genome sequence depth for all samples used in the study. The SNR values were obtained from 3kb control region windows. (C) Correlation of noise standard deviations between QuicK-mer and fastCN. Village dogs are in blue and wolves in orange, while the aCGH reference sample (box) is red.

3.3.3 Comparison with CGH array data

Comparative genomic hybridization array (aCGH) data from a previous study (Ramirez et al. 2014) was downloaded from Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih/gov/geo/info/linking.html) under the accession number GSE58195. This study utilized a NimbleGen aCGH chip that contained 598,733 probes with average spacing of oligonucleotide probes at 157 bp, and tested the comparative binding of DNA to estimate CNV between dogs and wolves at sites incorporated into the aCGH design (Ramirez et al. 2014).

Table 3 aCGH data used for wolf CNV validation

Sample information for aCGH data deposited under GEO accession number GSE58195 from (Ramirez et al. 2014). The sample identifiers used in this study, sample descriptions, GEO accession for the aCGH data, SRA data accession for whole genome sequence, and sample sex is provided.

Sample ID	**Sample Description**	**aCGH Data Accession ID**	**SRA Data Accession ID(s)**	**Sex**
chw	Chinese Wolf	GSM1402955	SRX1137190, SRX1137189, SRX1137188	Female
glw	Great Lakes Wolf	GSM1402952	SRX655630, SRX655629	Male
inw	Indian Wolf	GSM1402956	SRX655632, SRX655631	Male

irw	Iranian Wolf	GSM1402953	SRX655634, SRX655633	Female
mxa	Mexican Wolf	GSM1402954	SRX655637, SRX655636	Female
ptw	Portuguese Wolf	GSM1402949	SRX655640, SRX655639	Female
ysa	Yellowstone Wolf	GSM1402951	SRX655648, SRX655647, SRX655646	Female
box	Boxer	GSM1402940	SRX655611, SRX655610	Female

In (Ramirez et al. 2014), DNA from a Boxer breed dog (box) is used in the aCGH control channel. However, the sequencing data from the same sample shows poor quality due to extremely uneven coverage. To circumvent the impact of this noisy sample in our aCGH validation, we employed a simple log difference transformation (Equation 3.3.1). Assuming the hybridization for the boxer sample performs equivalently across experiments, this approach effectively cancels out the boxer as the aCGH reference sample and instead directly compares copy number between samples 1 and 2. To make validation based on sequencing depth comparable to relative estimates from array CGH, we employed a in silico transformation on the copy number estimates using Equation

3.3.2, where the numerator and denominator are the normalized copy number state in each 3kb window for samples 1 and 2, respectively.

Equation 3.3.1 $$log_2 \frac{Probe\ Intensity\ \ of\ Sample\ 1}{Probe\ Intensity\ of\ box} - log_2 \frac{Probe\ Intensity\ of\ Sample\ 2}{Probe\ Intensity\ of\ box}$$

Equation 3.3.2 $$log_2 \frac{Copy\ Number\ of\ Sample\ 1}{Copy\ Number\ of\ Sample\ 2}$$

To compare the result between Equation 3.3.1 and 3.3.2, we lifted over the aCGH probe location to CanFam3.1 reference coordinate. Next, the values from Equation 3.3.1 for all the probes those intersect with a 3kb fastCN or QuicK-mer window were averaged and then assigned to the window. Figure 15 illustrates the probe count distribution for 3kb windows, a similar distribution was found for 3,000 k-mer windows, and we observed that QuicK-mer and fastCN had similar distributions. In total, 12,584 and 17,989 3kb windows intersect with at least one aCGH probe for fastCN and QuicK-mer, respectively. We further filtered these windows to only include those containing at least three aCGH probes, thus reducing the window counts to 11,876 and 17,774 for fastCN and QuicK-mer, respectively.

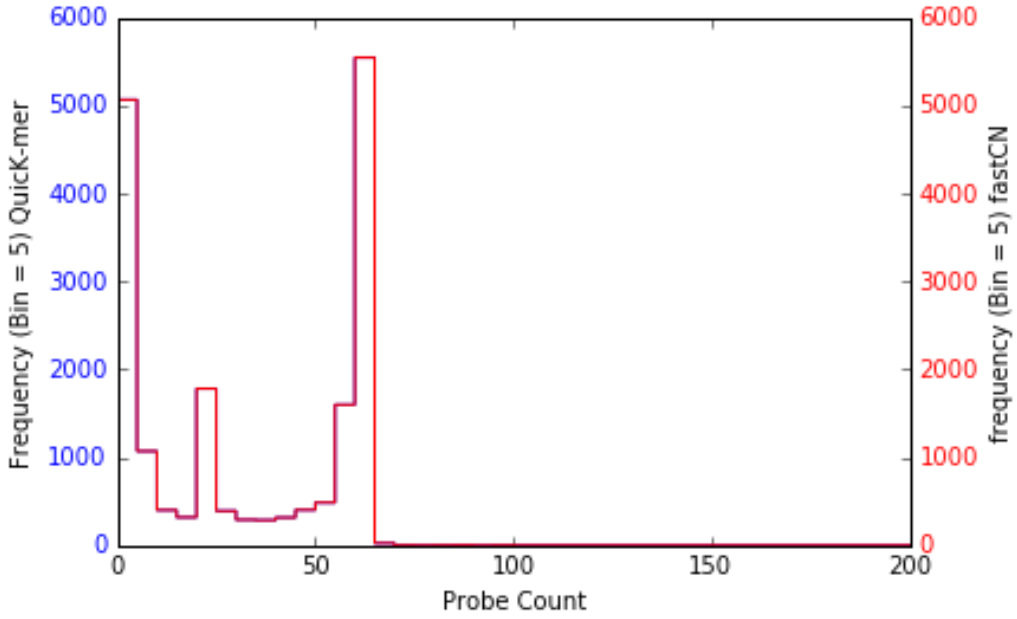


Figure 15 aCGH probe count distribution

Distribution of probe in fastCN and QuicK-mer window

From the previous two steps, each QuicK-mer and fastCN window was assigned two log-ratio values, one from the mean of log-ratios from aCGH probes overlapping the window (Equation 3.3.1) and another from the *in silico* copy number estimates (Equation 3.3.2). We filtered windows that contained less than three probes and whose in silico vs aCGH log ratio values fell within a circle around the origin on the plot, defined to be $x^2 + y^2 < R^2$ (Figure 16). We set the radius R equal to 0.4 for both fastCN and QuicK-mer, which

corresponds to 1.4x change in probe intensity or copy number. This filtration step is necessary because linear regression will be skewed toward a cluster of noise which has no meaningful correlation near the plot origin, since most probes in the aCGH are in regions that are not variable between the two samples being compared. The remainder of the data points is used for linear regression.

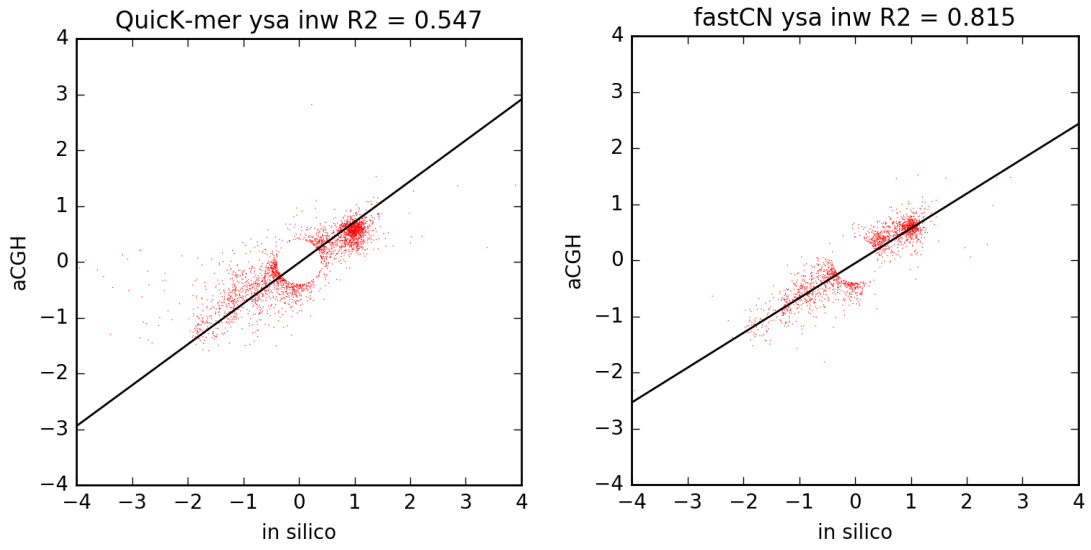


Figure 16 aCGH validation by correlation

Example of scatter plot displaying correlation between our *in silico* copy number estimations and the actual aGCH CN for a Yellowstone wolf (ysa) and an Indian wolf (inw). R^2 equals 0.55 and 0.81, respectively for QuicK-mer (left) and fastCN (right).

We calculated pairwise correlation coefficient among all seven available wolf samples and the resulting R^2 values for 21 comparisons are illustrated in a heatmap in Figure 17. The average R^2 value is 0.71 for fastCN and 0.55 for QuicK-mer. Based on the correlation coefficients, we observed that fastCN typically scores higher than QuicK-mer. This could reflect the probe binding chemistry acting in a paralog-insensitive fashion (permitting a certain number of mismatches to hybridize) or that paralog uniqueness was not considered during the probe design. It is also evident that sample pairs with higher coefficient values from QuicK-mer usually have a higher value in fastCN as well, indicating that data from certain samples harbor less noise. In summary, greater than 50% of variance can be explained by this correlation and we employed both methods in the following V_{ST} analysis.

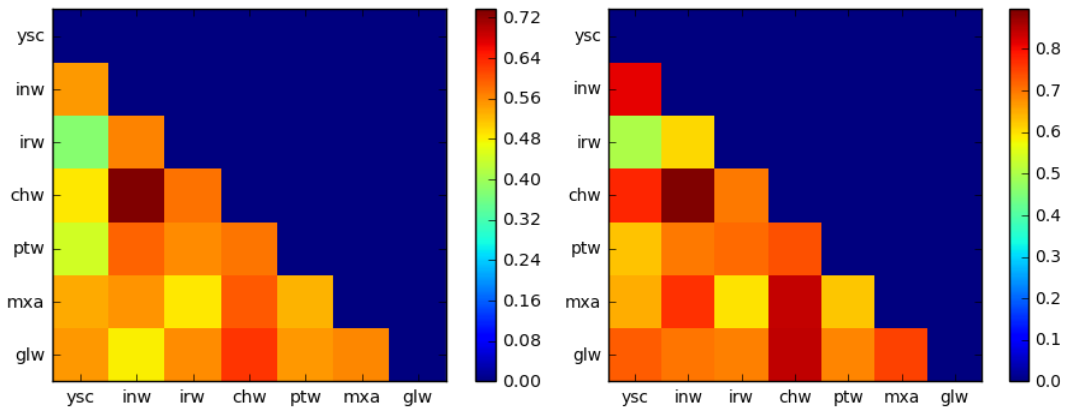


Figure 17 aCGH correlation heatmap

Heatmap showing pairwise correlation coefficients (R^2) between the log ratio of aCGH probe intensities and that of the *in silico* methods from QuicK-mer (left) and fastCN (right)

3.3.4 Detection of CN sweeps through V_{ST} analysis

To screen for CNV regions under selection between village dogs and wolves, we evaluated a metric called V_{ST}. This value is similar to the fixation index used to look for divergence in genotype between populations, except V_{ST} can be applied to copy number. A CNV under selection in one population will lead to an increased V_{ST} value. Point mutations within CNV regions might also show increased F_{ST}.

3.3.3.4.1 Filtration of genomic regions based on CN estimates

To determine genomic regions with differentiated copy number states between dogs and wolves, we first selected a subset of 3kb windows from the canine genome that showed evidence of copy number variation among the studied samples (Figure 18). We selected windows with a copy number range greater than 1.5 (copy number estimates on the non-PAR region of the X chromosome in males were doubled). This filtration step is necessary to limit the subsequent analysis to variable regions, rather than to noisy estimates derived from a large number of invariable windows. Window selection for fastCN (92,037) and QuicK-mer (37,626) were determined independently.

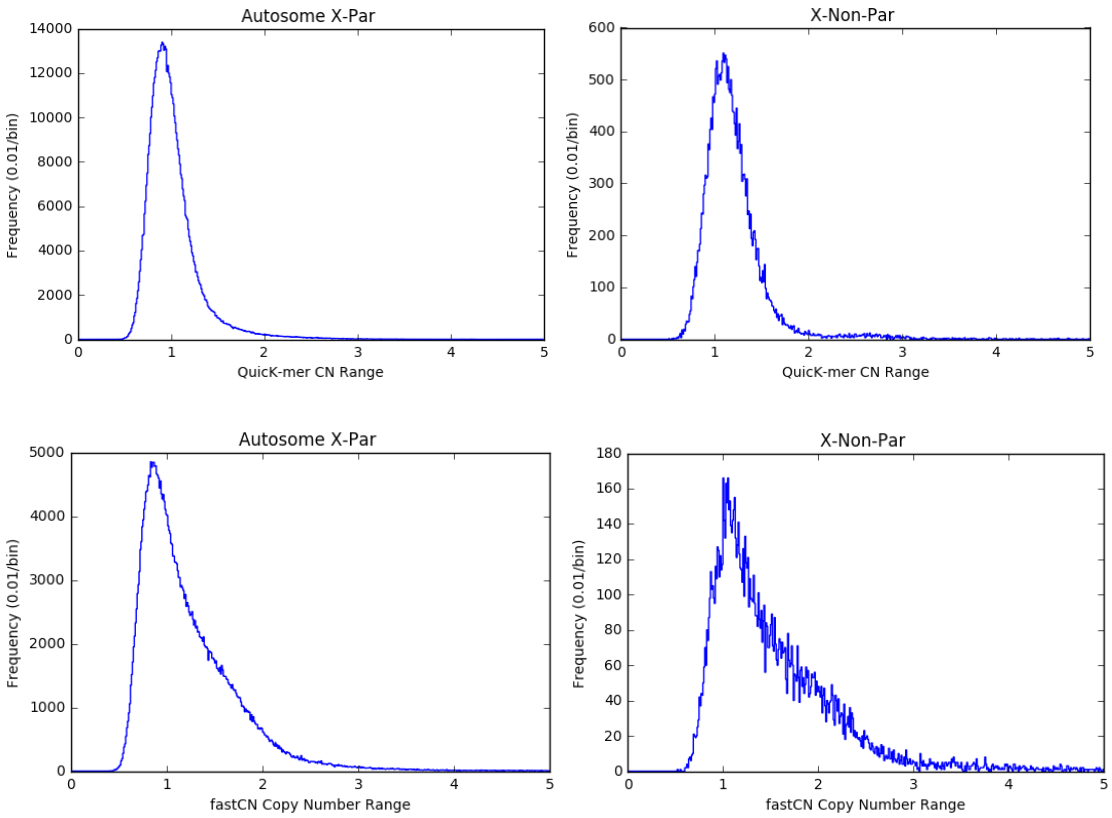


Figure 18 Copy number range distribution

Distribution of copy number estimates of autosomes plus chrX-PAR (left) and chrX-NonPAR (right) for QuicK-mer (top) and fastCN (bottom) pipelines, respectively, across all samples.

3.3.3.4.2 Calculation of V_{ST} values

A VST value for each of the selected 3 kb windows (CN range > 1.5) is then calculated with the following equation (Equation 3.4.2) according to (Redon et al. 2006), where VT

and VS denote variance of copy number in each window across the total or sub-population. Calculations were performed separately for QuicK-mer and fastCN estimates.

Equation 3.4.2 $$V_{ST} = \frac{V_T - V_S}{V_T}$$

The VST value is similar to a FST value where a higher VST value indicates a greater divergence in copy number between wolves and village dogs. However, the value alone will not indicate which population has increased or decreased copy number. We therefore calculated the average copy number in each window for wolves and village dogs separately. The VST value distribution for each pipeline is illustrated in Figure 19. We observed a narrower distribution for QuicK-mer VST values, likely because duplicated regions are prone to additional copy number variation and QuicK-mer only interrogates unique regions in the genome assembly.

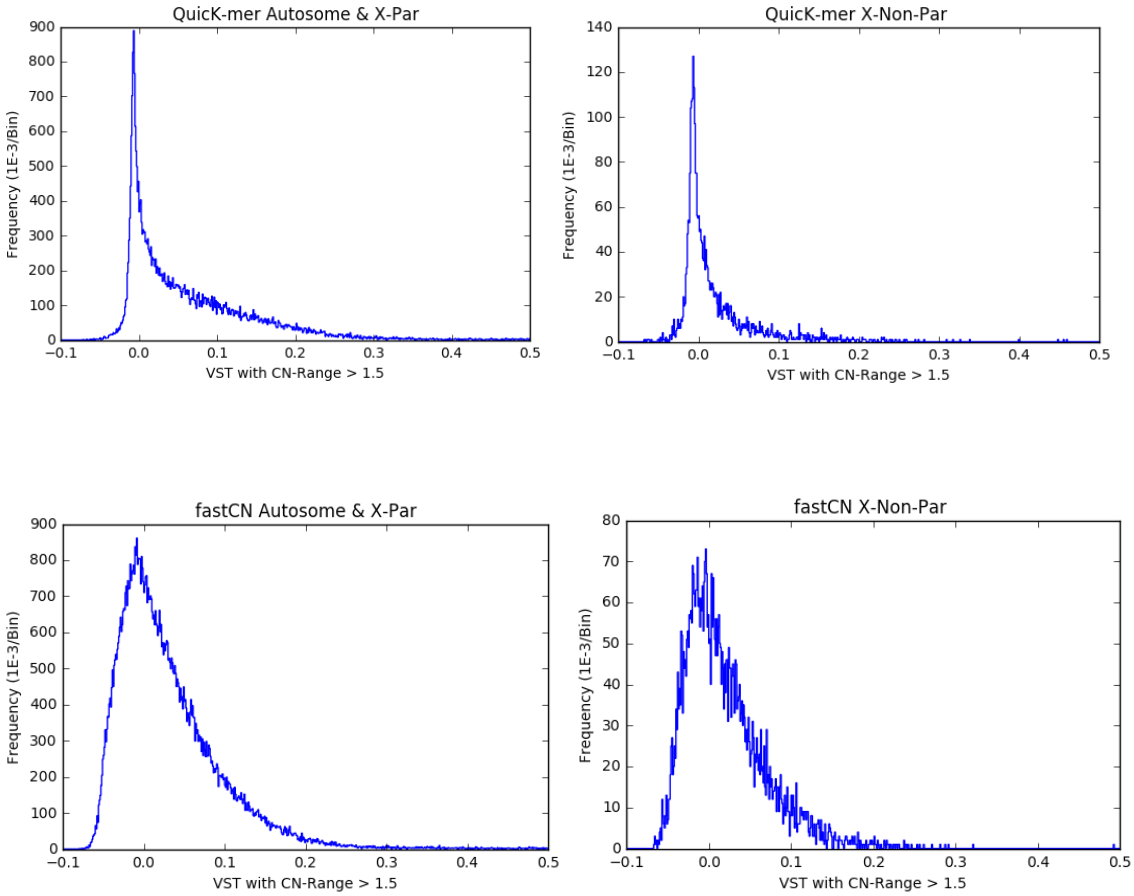


Figure 19 VST distribution

V_{ST} distribution of subsetted windows with copy number range greater than 1.5 for both QuicK-mer (top) and fastCN (bottom), in autosomes + chrX-PAR (left), and chrX-NonPAR (right).

3.3.3.4.3 Z-score normalization of V_{ST} distribution

The VST distributions from the windows with CN > 1.5 across all samples were Z-transformed to generate ZVST scores per window. This transformation was separately

completed for the autosomes + chrX-PAR, and the X-nonPAR. Similar to the FST filtrations, autosomal and chrX-PAR windows with greater than five standard deviations (or ZVST > 5) were selected as significant VST outliers, while significant chrX-NonPAR windows included all those that achieved ZVST > 3. The numbers of windows following each filtration step are detailed in Table 4.

Table 4 VST scan summary

The number of windows at each filtration stage including total windows analyzed in the V_{ST} pipeline, the number of windows that had >1.5 CN across samples, and the final windows with significant ZV_{ST} scores (greater than 5 for autosomes and chrX-PAR, and >3 for chrX-NonPAR).

Source	**Whole Genome Windows**	**> 1.5 CN Range**	**Significant ZV_{ST} Score**
QuicK-mer Autosomes + chrX-PAR	614,143	34,682	182
QuicK-mer chrX-NonPAR	28,060	2,944	11
fastCN Autosomes + chrX-PAR	366,945	86,276	513
fastCN chrX-NonPAR	13,786	5,761	6

3.3.3.4.4 Generation of candidate domestication regions from V_{ST} results

Windows that met significance thresholds set above were selected from the fastCN and QuicK-mer analysis, and within a given pipeline's dataset, adjacent significant windows

were merged into larger windows. From the original 519 windows with significant ZVST scores from the fastCN pipeline, 120 windows were generated from merging with other adjacent, significant windows. Similarly, of the 120 windows remained following merging of the 193 significant QuicK-mer windows.

To designate candidate domestication regions (CDRs) from the VST data, we intersected the significant windows determined from fastCN and QuicK-mer with one another using bedtools (Quinlan and Hall 2010). For all intersections, the minimum start coordinate and the maximum end coordinate of the intersecting window(s) were selected to define a VST candidate domestication region (VCDR). Any window unique to fastCN or QuicK-mer was automatically classified as a VCDR. For all resulting VCDRs, the maximum Z-score was extracted from the fastCN and QuicK-mer dataset to evaluate the level of significance of the region from each set or determine if the region was even evaluated in the opposite analysis set. Due to the stringency of requiring unique k-mer sequence in the QuicK-mer pipeline, it is foreseeable that a window analyzed by fastCN would not be present in the final QuicK-mer dataset.

Upon intersection, we identified 202 regions of copy number deviation between dogs and wolves through our VST pipeline. Of these final windows, 121 regions were found to be significant only by fastCN while 46 windows were identified only by QuicK-mer. 35

windows had significant ZVST scores from both fastCN and QuicK-mer. Again, QuicK-mer is a much more conservative and restrictive copy number estimator based on its reliance for sufficient unique k-mers in a region. For this reason, we observe considerably fewer windows with QuicK-mer support. However, the 35 windows with support from both pipelines are noteworthy, having significant CNV between village dogs and wolves.

Due to low genome coverage of some dog samples, confident detection of small CNVs using read depth is difficult (Sudmant et al. 2010). Therefore, the 202 outlier windows from above were further filtered to require at least two adjacent CN windows from either fastCN or QuicK-mer, or combined. Comparable to the nomenclature of such regions undergoing significant sequence deviation found through FST analysis, we distinguish these filtered regions as VCDRs, or VST Candidate Domestication Regions. Therefore, the filtration step resulted in 67 VCDRs that were either supported by both pipelines (N=35) or fastCN only (N=32), but no region was identified by QuicK-mer alone. In total, four VCDRs intersected with a FST CDR. This includes VCDR20 (chr6: 46945638-46957719), which intersects with a CDR8, a window that corresponds to previously published sweep loci (Cagan and Blass 2016; Axelsson et al. 2013), and harbors the copy number variable AMY2B gene. Next, a cluster of intersections are observed on chromosome 9 that include VCDRs 27 and 28 overlapping with CDR10, and VCDR31

co-localizing with CDR11. Detailed analysis of this region is provided in Supplemental Note 8.2.

3.3.3.4.5 Chromosome unknown analysis

In addition to the autosomes and X chromosome, we also calculated the level of copy number differentiation of unplaced contigs in the CanFam3.1 reference assembly. FST was not calculated on these contigs because the redundancy of these sequences reduces quality mappability, thus affecting accurate SNP calling. However, copy number changes through VST analysis could still be assessed with the fastCN and QuicK-mer pipelines. CN estimation with mrsFAST (and therefore also fastCN) is limited to a certain number of input chromosomes, so to facilitate this analysis, we merged all 3,228 unplaced contigs (chrUn's) into a single, continuous chromosome with 200 'N' bases inserted between each contig. Contig-specific 3kb windows were generated for both fastCN and QuicK-mer pipelines requiring that the last window of each contig does not extend into the next or contain 'N' bases. Next, the coordinates of each 3kb window were lifted over to the original unplaced contig with its corresponding location in order to assign CN to a single unplaced contig following processing with both fastCN and QuicK-mer. Finally, the

combined chromosome unknown was incorporated into the genome reference during the copy number estimation.

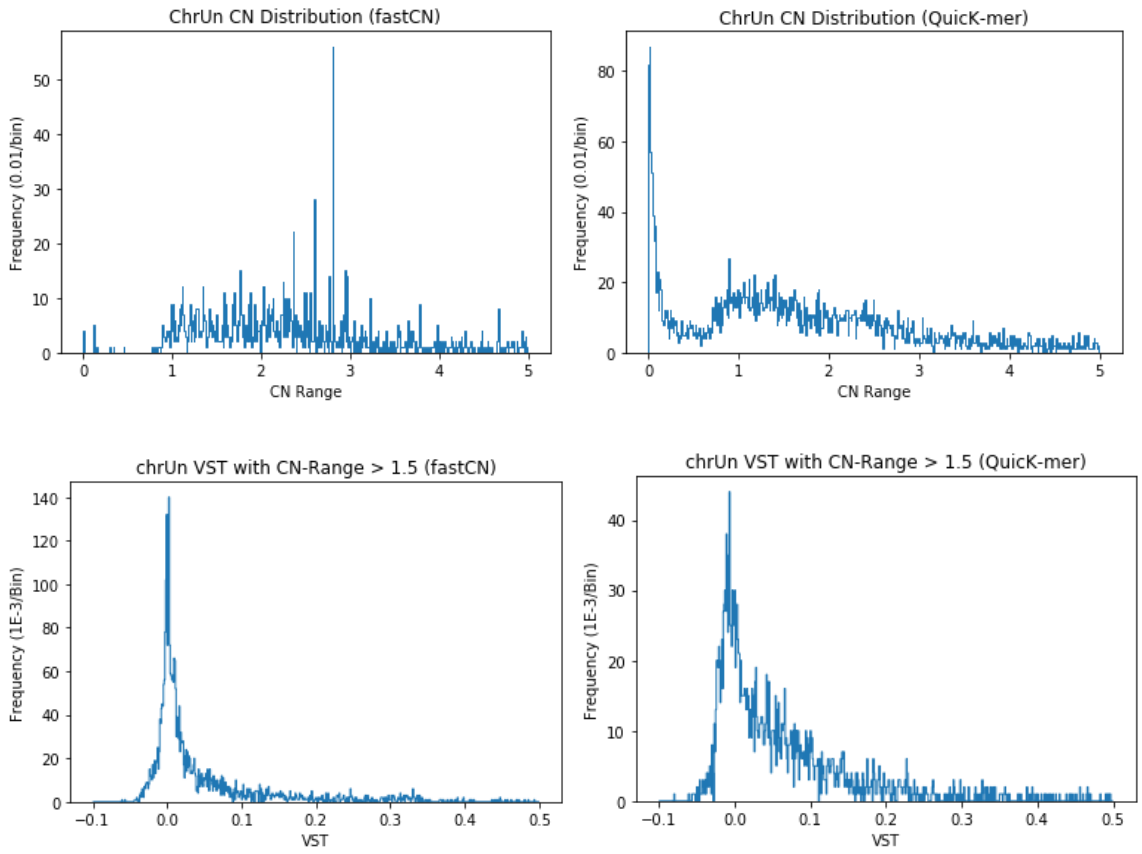


Figure 20 VST distribution for chrUs

The copy number range distribution for chromosome unknown based on fastCN (top left) and QuicK-mer (top right) distincts themselves from that of autosomes or chromosome X. The histogram for VST were shown after filtering windows that have less than 1.5 copy number range among all samples.

VST selection scans were completed on the merged unknown chromosome using methods previously implemented for autosomal VST scans (Figure 21). Windows with copy number ranges greater than 1.5 were selected from each pipeline, which included 3,370 windows from fastCN and 2,346 windows from QuicK-mer. Following Z-transformation of these subset windows, only windows with Z-scores greater than 5 were selected as candidate VST sweeps. Initially, we identified 21 fastCN and 10 QuicK-mer windows with Z-scores greater than five. After merging adjacent significant windows, the reduced to 8 fastCN and 9 QuicK-mer windows (Table 5), however no overlapping region was called by both fastCN and QuicK-mer. Upon further filtration, five fastCN and one QuicK-mer windows remained that consisted of at least two adjacent significant windows, yielding 6 additional candidate VCDRs. The largest of these is found on chrUn_AAEX03020568 which contains the pancreatic alpha amylase-2b (AMY2B) gene, a known copy number variable gene (Botigue et al. 2017; Axelsson et al. 2013; Arendt et al. 2016; Ollivier et al. 2016). Most unmerged windows achieving the VST threshold discovered by QuicK-mer contain micro-satellites interrupted by unique sequence queried by QuicK-mer. However, the role of this variation in domestication is unclear.

Table 5 chrU VST scan summary

Regions on chromosome unknown revealed by both fastCN and QuicK-mer. Segments greater than one window in size would meet the criteria to be VCDRs.

fastCN							
Chromosome	Start	End	No. 3kb window	Max VST	Mean CN in dogs	Mean CN in wolves	Mean CN range
chrUn_AAEX03020568	433	38543	7	0.63487	10.7707	1.86921	17.5122
chrUn_AAEX03024353	36	8845	2	0.634171	11.5489	1.91201	18.3228
chrUn_AAEX03024600	7409	7889	1	0.623222	4.85336	12.0129	14.6209
chrUn_AAEX03025786	95	4932	1	0.639081	8.4489	1.96336	12.1835
chrUn_JH373575	14842	33896	4	0.66059	5.52413	16.3035	21.2698
chrUn_JH373917	683	22389	3	0.633881	5.3227	13.4491	15.8141
chrUn_JH374030	36	15233	2	0.648995	5.38054	13.9423	16.6026
chrUn_JH374046	4411	11006	1	0.618721	4.34593	10.8845	13.2208
QuicK-mer							
chrUn_AAEX03021660	560	25346	1	0.626874	0.169814	1.4167	2.463
chrUn_AAEX03022211	1381	20389	1	0.689001	0.127372	1.391	2.236
chrUn_AAEX03022212	25	20386	1	0.618471	2.18914	9.3152	13.311
chrUn_AAEX03024092	132	7895	1	0.656465	0.127163	1.6749	2.692
chrUn_AAEX03026048	678	3035	1	0.648537	0.0331163	1.2794	1.933
chrUn_JH373233	199637	2045139	2	0.619935	1.79078	0.42245	2.446

	9						
chrUn_JH373337	582	77223	1	0.634374	0.419674	1.3641	2.006
chrUn_JH373343	15	87451	1	0.608101	0.358233	1.1457	1.725
chrUn_JH373779	15	21473	1	0.585038	0.378628	1.3411	1.974

3.3.5 Chromosome 9 Regions

Co-localization analysis indicated a clustering of VCDR and CDR windows within the first 25Mb of chromosome 9. Upon closer inspection of the copy-number and FST data at this region, we observed anomalous patterns not found elsewhere in the genome for our datasets. Average CN values from fastCN and QuicK-mer both indicate significantly higher CN in wolves within 19 VCDR windows here. Notably, boundaries of the VCDRs are directly adjacent to regions undergoing significant allele frequency differentiation, as highlighted in per site ZFST peaks in Figure 21. Such a pattern of extended divergence is reminiscent of inverted haplotypes which have been characterized in several species (Kirkpatrick and Barton 2006; Yeaman 2013; Jones et al. 2012). To further characterize this locus, we identified candidate inversions in the dogs and wolves separately using inveRsion (Cáceres et al. 2012) which relies on SNP genotypes to locally phase alleles and determine haplotype blocks for inversion breakpoint estimations. Although no

inversions were detected in the wolves on chromosome 9, five potential inversions were identified in village dogs clustered within this region of interest. Interestingly, predicted breakpoints of two inversions are situated at the transition point between the elevated FST region (chr9: ~9.0-16.7 Mb) and major copy-number peaks. Correlations between copy number states of VCDRs (per fastCN and QuicK-mer) and SNP genotypes of the 53 samples on chromosome 9 indicates two loci 8 Mb apart in the reference genome share elevated R2 value, patterns consistent with genome rearrangements at this region (Figure 22B). More specifically, the copy number states of VCDR 31 and VCDR 48 (Figure 23) share similar correlation even though the locations are separated by 8 Mb.

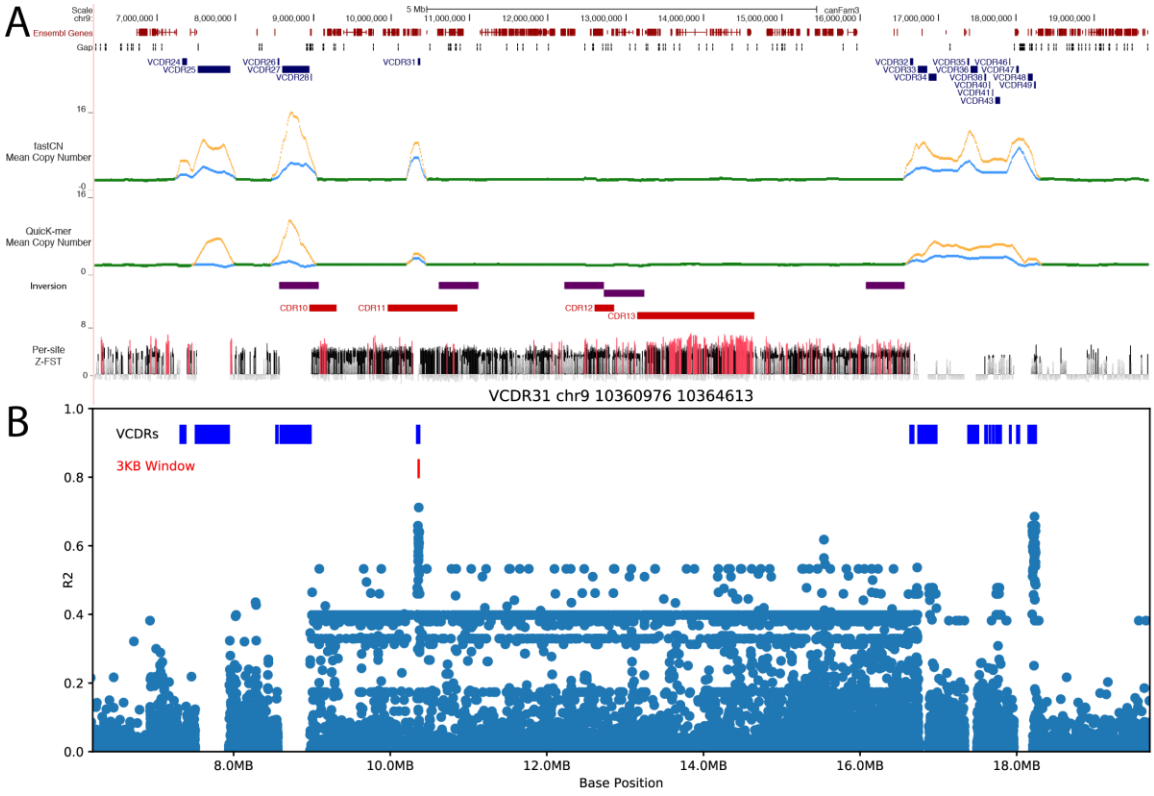


Figure 21 Chromosome 9 region

Region of complex structural variation on chromosome 9.

(A) Relative to Ensembl gene models and reference assembly gaps (top two tracks), the co-localization of VCDRs (dark blue) with regions of copy number expansion can be observed. Tracks 3 and 4 display the average copy number states of wolf (orange) and village dog (blue) populations as determined by fastCN and QuicK-mer, with regions of consistent CN between the populations as green. Putative inversions (purple bars), FST CDRs (red bars), and per site ZF ST values from the total SNP set (red = ZFST > 5), are also provided in tracks 5-7. (B) SNP genotypes are correlated with CNV states for each VCDR. The horizontal axis indicates the genome position of each VCDR and SNP site, while the vertical axis indicates R2 value between each SNP and the QuicK-mer based copy number state of VCDR31. The resulting scatterplot demonstrates the R 2 value between the individual SNP genotype and the copy number of the 3Kb window indicated in red (chr9:10,367,629-10,370,795). VCDRs positions are shown along the top of the scatterplot.

A **B**

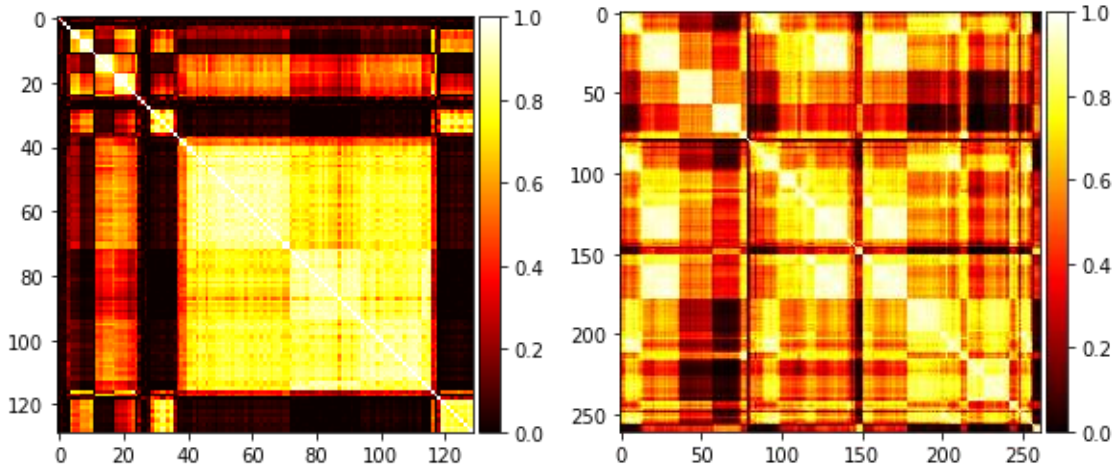


Figure 22 CNV pairwise correlation

(A) Mutual correlation of copy number state between 3KB QuicK-mer windows reveal duplicating/deleting segments. Intensity indicates R^2 values. VCDR 26 and 27 (axis index 11~23) shows good correlation in copy number state with VCDR 32~46 (axis index 37 - 115). Each segment is indicated as square block along the diagonal line. (B) Mutual correlation of copy number state in 3KB window from fastCN pipeline reveals individual blocks of segmental duplications and their relative correlation states.

3.4 Summary

QuicK-mer and fastCN were successfully applied to copy number related domestication scan of village dogs and modern wolves. Changes in copy number could be validated with aCGH showing valid change signal. However, the numerous VCDRs discovered through this approach did not achieve statistical significance of in gene enrichment analysis. This suggests CNV might not played a significant role in early canine

domestication. This conclusion should be limited to the scope of current reference assembly and sample sets. Another interesting discovery was the clustering of CNV in a 8Mbp locus on chromosome 9. Correlation of CNV and genotypes far apart plus previous studies pointed to an error during genome assembly.

3.5 References

Arendt, M., K. M. Cairns, J. W. O. Ballard, P. Savolainen, and E. Axelsson. 2016. "Diet Adaptation in Dog Reflects Spread of Prehistoric Agriculture." Heredity 117 (5): 301–6.

Axelsson, Erik, Abhirami Ratnakumar, Maja-Louise Arendt, Khurram Maqbool, Matthew T. Webster, Michele Perloski, Olof Liberg, Jon M. Arnemo, Ake Hedhammar, and Kerstin Lindblad-Toh. 2013. "The Genomic Signature of Dog Domestication Reveals Adaptation to a Starch-Rich Diet." Nature 495 (7441): 360–64.

Botigue, Laura, Botigue Laura, Song Shiya, Scheu Amelie, Gopalan Shyamalika, Pendleton Amanda, Oetjens Matthew, et al. 2017. "Ancient European Dog Genomes Reveal Continuity since the Early Neolithic." https://doi.org/10.1101/068189.

Botigué, Laura R., Shiya Song, Amelie Scheu, Shyamalika Gopalan, Amanda L. Pendleton, Matthew Oetjens, Angela M. Taravella, et al. 2017. "Ancient European Dog Genomes Reveal Continuity since the Early Neolithic." Nature Communications 8 (July): 16082.

Cáceres, Alejandro, Suzanne S. Sindi, Benjamin J. Raphael, Mario Cáceres, and Juan R. González. 2012. "Identification of Polymorphic Inversions from Genotypes."

BMC Bioinformatics 13 (1): 28.

Cagan, Alex, and Torsten Blass. 2016. “Identification of Genomic Variants Putatively Targeted by Selection during Dog Domestication.” BMC Evolutionary Biology 16 (January): 10.

Chen, Congying, Ruimin Qiao, Rongxing Wei, Yuanmei Guo, Huashui Ai, Junwu Ma, Jun Ren, and Lusheng Huang. 2012. “A Comprehensive Survey of Copy Number Variation in 18 Diverse Pig Populations and Identification of Candidate Copy Number Variable Genes Associated with Complex Traits.” BMC Genomics 13 (December): 733.

Chen, Wei-Kang, Joshua D. Swartz, Laura J. Rush, and Carlos E. Alvarez. 2009. “Mapping DNA Structural Variation in Dogs.” Genome Research 19 (3): 500–509.

Dorshorst, Ben, Anna-Maja Molin, Carl-Johan Rubin, Anna M. Johansson, Lina Strömstedt, Manh-Hung Pham, Chih-Feng Chen, Finn Hallböök, Chris Ashwell, and Leif Andersson. 2011. “A Complex Genomic Rearrangement Involving the Endothelin 3 Locus Causes Dermal Hyperpigmentation in the Chicken.” PLoS Genetics 7 (12): e1002412.

Elferink, Martin G., Amélie A. A. Vallée, Annemieke P. Jungerius, Richard P. M. A. Crooijmans, and Martien A. M. Groenen. 2008. “Partial Duplication of the PRLR and SPEF2 Genes at the Late Feathering Locus in Chicken.” BMC Genomics 9 (August): 391.

Freedman, Adam H., Ilan Gronau, Rena M. Schweizer, Diego Ortega-Del Vecchyo, Eunjung Han, Pedro M. Silva, Marco Galaverni, et al. 2014. “Genome Sequencing Highlights the Dynamic Early History of Dogs.” PLoS Genetics 10 (1): e1004016.

Freedman, Adam H., and Robert K. Wayne. 2017. “Deciphering the Origin of Dogs: From Fossils to Genomes.” Annual Review of Animal Biosciences 5 (February): 281–307.

Jia, X., S. Chen, H. Zhou, D. Li, W. Liu, and N. Yang. 2013. “Copy Number Variations Identified in the Chicken Using a 60K SNP BeadChip.” Animal Genetics 44 (3): 276–84.

Jones, Felicity C., Manfred G. Grabherr, Yingguang Frank Chan, Pamela Russell, Evan Mauceli, Jeremy Johnson, Ross Swofford, et al. 2012. “The Genomic Basis of Adaptive Evolution in Threespine Sticklebacks.” Nature 484 (7392): 55–61.

Keel, Brittney N., Amanda K. Lindholm-Perry, and Warren M. Snelling. 2016. “Evolutionary and Functional Features of Copy Number Variation in the Cattle Genome.” Frontiers in Genetics 7 (November): 207.

Kirkpatrick, Mark, and Nick Barton. 2006. “Chromosome Inversions, Local Adaptation and Speciation.” Genetics 173 (1): 419–34.

Larson, Greger, and Dorian Q. Fuller. 2014. “The Evolution of Animal Domestication.” Annual Review of Ecology, Evolution, and Systematics 45 (1): 115–36.

Long, Yi, Ying Su, Huashui Ai, Zhiyan Zhang, Bin Yang, Guorong Ruan, Shijun Xiao, et al. 2016. “A Genome-Wide Association Study of Copy Number Variations with Umbilical Hernia in Swine.” Animal Genetics 47 (3): 298–305.

Marsden, Clare D., Diego Ortega-Del Vecchyo, Dennis P. O’Brien, Jeremy F. Taylor, Oscar Ramirez, Carles Vilà, Tomas Marques-Bonet, Robert D. Schnabel, Robert K. Wayne, and Kirk E. Lohmueller. 2016. “Bottlenecks and Selective Sweeps during Domestication Have Increased Deleterious Genetic Variation in Dogs.” Proceedings of the National Academy of Sciences of the United States of America 113 (1): 152–57.

Nicholas, Thomas J., Carl Baker, Evan E. Eichler, and Joshua M. Akey. 2011. “A High-Resolution Integrated Map of Copy Number Polymorphisms within and between Breeds of the Modern Domesticated Dog.” BMC Genomics 12 (August): 414.

Nicholas, Thomas J., Ze Cheng, Mario Ventura, Katrina Mealey, Evan E. Eichler, and Joshua M. Akey. 2009. “The Genomic Architecture of Segmental Duplications and Associated Copy Number Variants in Dogs.” Genome Research 19 (3): 491–99.

Ollivier, Morgane, Anne Tresset, Fabiola Bastian, Laetitia Lagoutte, Erik Axelsson, Maja-Louise Arendt, Adrian Bălăşescu, et al. 2016. “Amy2B Copy Number

Variation Reveals Starch Diet Adaptations in Ancient European Dogs." Royal Society Open Science 3 (11): 160449.

Paudel, Yogesh, Ole Madsen, Hendrik-Jan Megens, Laurent A. F. Frantz, Mirte Bosse, Richard P. M. A. Crooijmans, and Martien A. M. Groenen. 2015. "Copy Number Variation in the Speciation of Pigs: A Possible Prominent Role for Olfactory Receptors." BMC Genomics 16 (April): 330.

Pendleton, Amanda L., Feichen Shen, Angela M. Taravella, Sarah Emery, Krishna R. Veeramah, Adam R. Boyko, and Jeffrey M. Kidd. 2018. "Comparison of Village Dog and Wolf Genomes Highlights the Role of the Neural Crest in Dog Domestication." BMC Biology 16 (1). https://doi.org/10.1186/s12915-018-0535-2.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." Bioinformatics 26 (6): 841–42.

Ramirez, Oscar, Iñigo Olalde, Jonas Berglund, Belen Lorente-Galdos, Jessica Hernandez-Rodriguez, Javier Quilez, Matthew T. Webster, et al. 2014. "Analysis of Structural Diversity in Wolf-like Canids Reveals Post-Domestication Variants." BMC Genomics 15 (June): 465.

Redon, Richard, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, et al. 2006. "Global Variation in Copy Number in the Human Genome." Nature 444 (7118): 444–54.

Sudmant, Peter H., Jacob O. Kitzman, Francesca Antonacci, Can Alkan, Maika Malig, Anya Tsalenko, Nick Sampas, et al. 2010. "Diversity of Human Copy Number Variation and Multicopy Genes." Science 330 (6004): 641–46.

Trut, L. N., I. Z. Plyusnina, and I. N. Oskina. 2004. "An Experiment on Fox Domestication and Debatable Issues of Evolution of the Dog." Russian Journal of Genetics 40 (6): 644–55.

Wang, Jiying, Jicai Jiang, Haifei Wang, Huimin Kang, Qin Zhang, and Jian-Feng Liu. 2015. "Improved Detection and Characterization of Copy Number Variations Among Diverse Pig Breeds by Array CGH." G3 5 (6): 1253–61.

Wilkins, Adam S., Richard W. Wrangham, and W. Tecumseh Fitch. 2014. “The ‘Domestication Syndrome’ in Mammals: A Unified Explanation Based on Neural Crest Cell Behavior and Genetics.” Genetics 197 (3): 795–808.

Wright, Dominic, Henrik Boije, Jennifer R. S. Meadows, Bertrand Bed’hom, David Gourichon, Agathe Vieaud, Michèle Tixier-Boichard, et al. 2009. “Copy Number Variation in Intron 1 of SOX5 Causes the Pea-Comb Phenotype in Chickens.” PLoS Genetics 5 (6): e1000512.

Yeaman, Sam. 2013. “Genomic Rearrangements and the Evolution of Clusters of Locally Adaptive Loci.” Proceedings of the National Academy of Sciences of the United States of America 110 (19): E1743–51.

Chapter 4: A new canine genome assembly using long read sequencing

4.1 Limitations of the current canine assembly

After the successful completion of the human genome project, biologists tried to push for assemblies of additional species in order to gain an understanding of the evolution, biological diversity and phenotypical differences between many of other species. As our closest friend, dogs share much of their evolutionary history alongside human and creating a dog genome assembly became a priority. In 2004, the Broad Institute released the first reference assembly derived from a female boxer (Lindblad-Toh et al. 2005). This assembly used Sanger sequencing and was performed using the whole genome shotgun methodology. It utilized end-sequences from bacterial artificial chromosomes, fosmids, and plasmids constructed from genomic DNA. Individual clones were isolated, purified and end sequenced using known primer in the BAC or Fosmid backbone with Sanger approach. The CanFam 1.0 *de novo* assembly solely relied on 7.5x of read coverage. The

same study improved the reference into 2.0 by addressing some of the errors using FISH technique (Lindblad-Toh et al. 2005; Breen et al. 2001) and sequencing of some individually isolated BAC clones. Like many previous initial assembly projects, it lacks a male specific Y chromosome due to use of a female sample. In 2006 a partial assembly of the dog Y chromosome was released (Mustafa and Yuen 1991). This was further improved in 2013 with pooled BAC sequencing on a 454 platform (G. Li et al. 2013), although the dog Y sequence remains incomplete. The latest effort of canine reference improvement was facilitated with primer walking and resequencing a few hundreds of selected BACs (Hoeppner et al. 2014). The same study also annotated the CanFam 3.1 with additional RNA sequencing data, providing a more thorough view of the transcriptome landscape.

4.1.1 Unplaced contigs and assembly gaps

Even after several rounds of patch and correction, the latest version dog reference genome assembly, CanFam 3.1, is still not without flaws. For example, 15,800 autosomal and chromosome X gaps dot the genomic landscape. Based on the size estimates, which might be completely inaccurate, the combined missing gap sequence amounts to 18.2 million base pairs in total. In addition to gaps, there is a long catalog of 3227 contigs and scaffolds with their location on the chromosome unknown. These unplaced contigs have

a combined length of 83.3 MB including the hypothetical gap length. Compared to the latest human assembly GRCh38 (Guo et al. 2017) of 2,512 unplaced contigs totaling only 10 MB, there remains huge potential for further improvement in continuity and missing sequences.

4.1.2 Indication of mis-assemblies

Besides the apparent base content of the CanFam3.1 reference, evidence from numerous recent studies suggests potential mis-assemblies. A notable example is the chromosome 9 regions described in the previous chapter. A strong correlation of copy number variation and SNP genotype for regions placed 8M B apart is unlikely due to true linkage disequilibrium, suggesting a large inversion in the assembly. Further evidence is provided by the Rossi, et al study of the Sox9 gene, where FISH probing of the chr9 regions related to non-SRY sex reversal in dogs also support a missassemby (Rossi et al. 2015).

These missing sequence and mis-assemblies also manifest as incomplete gene models. For example, Holden et al points out some dog genes contain gaps or are missing known alternative splicing variants (Holden et al. 2018). In other cases, critical mammalian gene models are entirely missing from the assembly.

The tool sets such as QuicK-mer and fastCN described in Chapter Two as well as many general mapping tools like BWA, Bowtie, Star and Rsam all rely on a reference. If a sequence is missing or mis-assembled, errors will result in CNV, SNP calling and expression analysis. An improved genome would well address these issues.

4.1.3 Expectation of improvement

Second generation sequencing has limited read length and is unlikely to improve repeat regions and genome continuity compared to classic BAC assembly methodologies. However, the sequencing technologies have been improving over time. Modifications of sequencing library construction can overcome the insert length limitation of 1kb in Illumina bridge amplification. For example, mate-pair libraries swap the insert sequencing direction with the help of a common backbone sequence (Vasmatzis et al. 2012). This method extends the insert size to 10kb and greatly increases the physical coverage, which show promising in structural breakpoint discoveries and scaffolding in *de novo* assembly (Love et al. 2016; Williams LJ n.d.). Other technologies such as 10x Genomics attempt to fragment a single very long DNA insert while attaching the same barcode from a random pool, followed by sequencing on a regular short read platform. Combined with sophisticated bioinformatic analysis to assign reads from each region, this library construction emulates some of the benifits old BAC selection and purification in a

massive parallel manner, permitting the anchoring of sequences in a genomic context (Mostovoy et al. 2016).

Finally, the PacBio sequencing approaching combines both PCR-free library and very long read length. With the fragment sizes of 15kb and greater, it's easy to expect correct sequencing through LINE and ERV repeats, which are the longest common repetitive sequences in a mammalian genome. Resolving large segmental duplications however remains a challenge. Using a combination of such techniques, we would expect the majorities of small gaps to be completely closed and greatly improve continuity. It would also reduce the number of unplaced contigs by bridging the repeats between them and their anchoring chromosomes.

4.2 Single Molecular Real Time Sequencing from PacBio

Most sequencing technology uses fluorescence signal for readout. But each fluorescent label can only emit ~10k photons before being photobleached (Luchowski et al. 2008) and meanwhile the DNA extension under polymerase is rapid. This makes it impossible to observe in real time. To combat these problems, second generation sequencing methods rely on local isolation and amplification of individual DNA insert fragment by a variation of PCR, thus increases the optical signal for base pair calling. These existing

short-read methods either step the reaction one at a time, or integrate the signal by cycling through each of the four nucleotides one at a time. Due to incomplete reaction, the lagging strands inside a cluster generates mixed signal. This effect limits the sequencing length and usually observed as systematic errors at particular trinucleotides or homopolymers (Nakamura et al. 2011). In order to increase read length, the third generation employed a completely different approach with technological innovations.

4.2.1 Real time single molecule sequencing

Recently, improvements in laser and sensor technologies have been brought by research in nanotechnology and photonics. Electron-multiplying charge coupled devices (EMCCD) and scientific CMOS sensors allow single photon detection at a very high frame rate (Saurabh, Maji, and Bruchez 2012). This enabled the observation of a single polymerase-DNA pair in action with just one nucleotide label. In the meantime, nano-fabricated holes called Zero Mode Waveguide (ZMW) limited the background fluorescence noise from the unincorporated free nucleotides in solution. The fluorescence light is then dissected into a micro spectrum with an optical prism in the microscope (Lundquist et al. 2008). The distribution of each spectrum indicates the dye color of the current base in the polymerase reaction pocket and hence the base pair. By taking a movie of thousands of such holes in parallel and analyze their color spectrum in real time, one could deduce the

DNA sequence uninterrupted (Eid et al. 2009). In this manner, PacBio could enable PCR-free sequencing of original DNA from cellular extract with very long insert.

However, such real-time single photon detection and base calling is not without cost. First, the error rate is much higher due to the rapid action of polymerase and the still low signal-to-noise ratio. The error mode is also different than technologies such as Illumina. Instead of substitutions, PacBio errors are predominantly insertions and deletions, making sequence alignment much harder (Eid et al. 2009). The insertion is mostly due to transiting nucleotides that are not actually incorporated while deletion is mostly due to very fast extension event. Secondly, the per base pair cost is much higher due to the limitation of throughput by the number of pixels available on the camera. Several methods have been proposed to circumvent the problem and will be discussed in the section 2.2.

Another recent progress on single molecular sequencing is the Oxford Nanopore technology (Jain et al. 2016). In this method, the underlying sequence is resolved directly like reading a magnetic tape. A nanoscale hole made from specifically engineered transmembrane protein limits the passage of a single denatured strand one nucleotide at a time. A voltage gradient is then applied across the membrane and the current is determined by the electro-chemical property of a few base pairs inside the nanopore. By

measuring the current passing through the hole and deconvolute the signal, the software could call the base pair directly. Nanopore sequencing can achieve very long read only limited by DNA length during extraction. Recent improvement decreased the error rate to that similar in PacBio (Weirather et al. 2017) and could be a promising approach to achieve genome scaffolds with even better continuity.

4.2.2 *de novo* assembly process using long reads

Because the third-generation sequencing is costly and error prone, researcher has proposed various methods to increase the efficiency. Most such techniques, known as hybrid assembly, combining high depth short-read data with longer reads at a much lower depth.

4.2.2.1 Hybrid assembly

Owing to the long insert sizes, PacBio sequencing can detect structural variations using only a limited sequencing depth (Merker et al. 2016). By using it as a structural backbone, one can correct the error in short read assemblies using these reads and improve the reference with second generation short sequencing reads. The first hybrid assembly algorithms supporting PacBio long reads are ALLORA and ALLPATH-LG which performs correction after PacBio *de novo* assembly using short reads (Gnerre et al. 2011; Schaeffer 2012). Later, pre-assembly correction of raw PacBio reads was also realized

(Koren et al. 2012). These studies suggest inclusion of short read data could improve the assembly by increasing N50 length and base pair accuracy. But both PacBio-only and hybrid assembly would greatly exceed the continuity of a pure short-read *de novo* process.

However, many of these algorithms are benchmarked against bacteria genomes, which are far less complex compared to a vertebrate genome on a multitude of levels. First, eukaryotic genomes contain far more repeat elements. Second, most of eukaryotic genes are multi-exons interrupted by lengthy introns which may harbor these repeats. Third, mammalian genomes are diploid and regions of heterozygous deletion and duplication will form bubbles and forks in a *de novo* assembly graph. In such correction scheme, one might mis-align a short read to the wrong paralog and hence incorrectly introduce a true variant private to that specific copy. Afterall, the *de novo* assembly requires much more investments comparing to a typical resequencing project. Careful planning and weighing the benefits of each sequencing technique is essential to yield the maximum quality given limited resources.

4.3 Library construction and sequencing

In our interest of seeking for copy number changes and gene duplications, we appreciate a more contiguous genome reference. As such, we desire the sequencing reads as long as possible to improve the scaffold for our dog genome project.

4.3.1 Sample origin and breed information

The DNA was extracted from the blood of a female Great Dane breed dog named Zoey. We choose this sample for several reasons. A breed dog is more homozygous compared to a village dog which has not gone through the recent selective breeding process. This property makes it simpler for *de novo* assembly process by reducing chances bubble formations due to heterozygosity. Next, Great Dane is in a separate breed group than the boxer but also sufficiently close on breed phylogeny tree to utilize existing resources (Parker et al. 2017). Having a different breed dog reference would provide insight into regions possibly lost due to breed selection. This additional genome resource would help the canine research community. Finally, we had previously invested a portfolio of sequencing project on Zoey, which could later well serve to improve the genome with error correction and hybrid assembly. These resources include a 14x coverage whole genome shotgun sequencing using Illumina HiSeq, a mate-pair library consisting of 4kb insert size, and finally 96 whole genome fosmid pools all from the same blood sample.

The mate-pair is a library preparation technique to circumvent insert size limitation on short-read sequencing. The DNA is circularized into a common backbone, restriction digested and circularized to reduce insert fragment. Besides these DNA resources, we also kept a low quantity stock of RNA from Zoey for potential gene models or expression validation.

4.3.2 Depth, insert length and coverage

Three PacBio libraries were constructed to satisfy the DNA quantity required for 90 SMRT sequencing runs. The actual sequencing was done at the University of Michigan Sequencing Core on a PacBio RS II platform using 6th generation polymerase and 4th generation chemistry. For initial assessment, we mapped the PacBio raw reads using BLASR (Chaisson and Tesler 2012) to the original CanFam3.1 reference assembly. We then filter the mapping with requiring it to be primary mapping MAPQ of 20. Due to the presence of circular consensus reads, we looked at each mapping from the same ZMW and picked the longest read and assigning it as the DNA insert. In all, we achieved closed to 50x sequencing depth for all the reads assuming a 2.4 billion haploid genome, and a 28x when only consider unique DNA insert. The L50 is around 10kb and median is at 7.5kb (Figure 23).

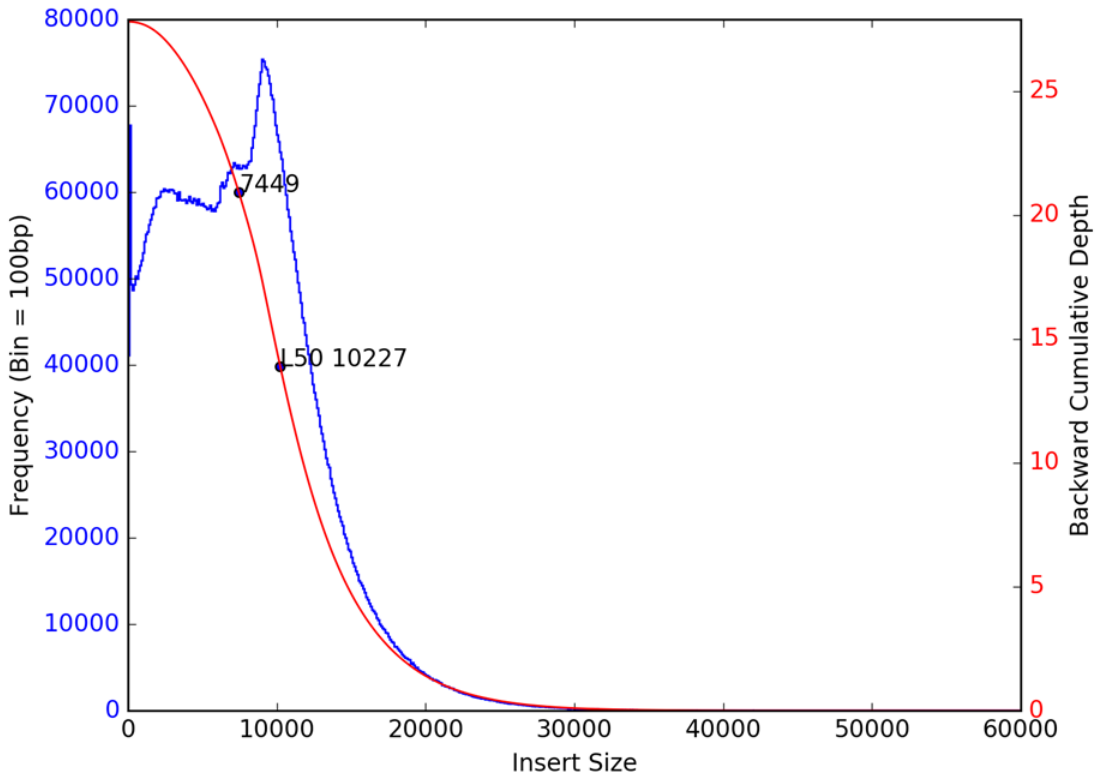


Figure 23 PacBio insert size distribution

PacBio sequencing insert and coverage statistics. The blue curve shows the distribution of unique insert under filtration criteria described above. The red curve is cumulative depth by adding the longest reads in the pool first.

In addition, we also observed around 80% of existing gaps in the CanFam3.1 has been spanned by at least one of above filtered reads.

4.3.3 *de novo* assembly with FALCON-unzip

Due to limitation of computing resource, we outsourced the initial assembly project to the DNANexus. First, 50-fold whole-genome, single-molecule, real-time sequencing (SMRT) data was passed through the TANmask and REPmask modules from the Damasker suite. This data was then used as input to the traditional FALCON pipeline (Chin et al. 2016), using a length cut-off of 1,775 bp during the initial error-correcting stage. This resulted in 15 million error corrected reads with an N50 read length equal to 8.7 kbp covering 38x of the dog genome.

Second, the error-corrected reads again passed through the TANmask and REPmask modules, followed by the overlap portion of the FALCON pipeline. For the overlap portion, a length cut-off of 5,087 bp was used. The aligned reads were assembled in the third stage of FALCON into 2,688 primary contigs containing 2.3Gbp with an N50 contig length of 4.4 Mbp.Finally, the assembly was polished through PacBio's Quiver algorithm from SMRT Link 3.1, using the original raw-reads. This assembly approach yielded 2688 primary contigs with N50 at 4.4Mb and N90 at 1Mb. The largest single contig is 28.8Mb long.

4.4 The Zoey reference assembly

4.4.1 Quality control

Two steps of initial quality assessment were done on contigs. On the small scale it's the base level quality and on the large scale to eliminate the chimeric contigs.

4.4.1.2 Base quality assessment

To assess the base level accuracy of the Zoey assembly, we used full length sequences from 59 fosmids using PacBio deep sequencing data as a gold standard. The fosmid library from which each fosmid was isolated was generated using whole blood DNA from Zoey. Most of these fosmids achieved 1000x sequencing depth. To verify that these 59 fosmids were truly gold standard, we mapped Illumina paired-end read data of seven fosmids onto their respective assembled full length fasta using BWA-mem 0.7.15. (H. Li 2013) Resulting BAM files were then sorted, marked for duplicates using Picard tools 2.3, and processed through GATK 3.5 HaplotypeCaller SNP caller. (McKenna et al. 2010) All three steps were run with default parameters. In total, 50 variants were called from the combined length of 265,191 bp, yielding a combined accuracy of QV37 across all seven fosmids. However, we observed both reference and alternative alleles achieving significant depth. It is unclear if these variants are related to PCR or systematic bias

associated with the Illumina platform, because these fosmid clones were individually isolated and purified, therefore only one haploid allele should be represented.

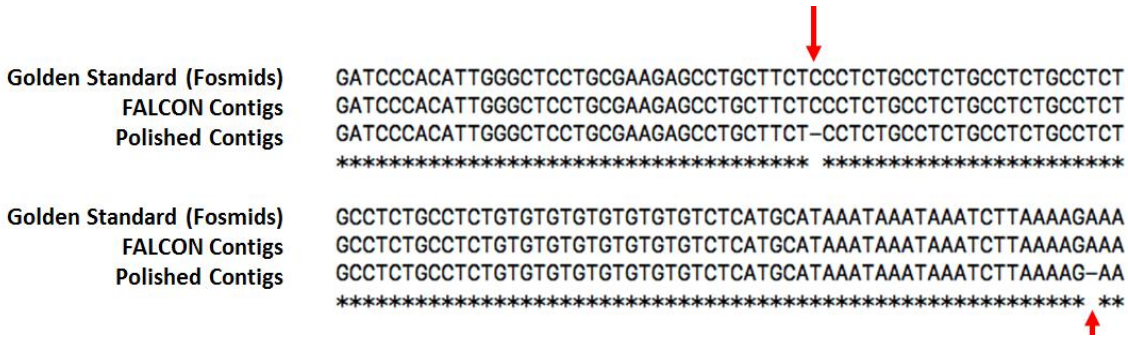


Figure 24 Base level error verification

Gold standard showing second round of polishing actually induce more deletion errors. Red arrow points to the second round of polishing induces deletion into contigs. Such polishing is abandoned.

Next, we mapped the original PacBio reads used in the genome assembly onto Zoey 2.2 reference and ran SMRTanalysis 2.3.0 pipeline VariantCaller with quiver polishing algorithm to generate a polished reference Zoey 2.3. The above 59 fosmids were then mapped to both 2.1 and 2.3 references using BWA-mem. The mapped fosmids were then filtered based on the following criteria: 1. the mapping should not have hard or soft clipping on neither end. And 2. the fosmid must be a primary mapping. This resulted in 37 valid mapping on both references. We then extracted the reference backbone sequence

and performed a three-way multiple sequence alignment using Clustal-Omega 1.2.4 and EMBOSS Stretcher 6.6.0 with default parameters to verify the base level accuracy. We found that Zoey 2.1 achieved higher similarity scores and lower gap open rates overall. Thus, we concluded that Zoey 2.3 was over-corrected by Quiver, and we opted to continue with misassembly assessments with Zoey 2.2 as the draft assembly.

4.4.1.2 Chimeric contigs and mis-assembly

Due to existence of large-scale duplications on different chromosomes, chimeric contigs can form even with long PacBio reads (Figure 25). In another word, pieces of DNA from different chromosomes are joined together as an assembly artifact. To resolve this, we make the primary contigs into a reference genome and mapped three different sequencing libraries onto it: 1. Mate pair library with insert size of 4kb from Zoey; 2. Tasha BAC end sequences; 3. Zoey short Illumina library pools. To look for chimeric candidate sites, we require that a region lack continuous mate pair insert coverage and showed translocation of BAC end sequence to a different primary contig. Further, this candidate should also show deletion or duplication in Illumina read coverage. MUMMER plots are generated for each primary contigs with canFam3.1 reference. Strong chimeric candidate should bear hallmarks of diagonal alignment to multiple chromosomes as distinctive segments (Figure 26).

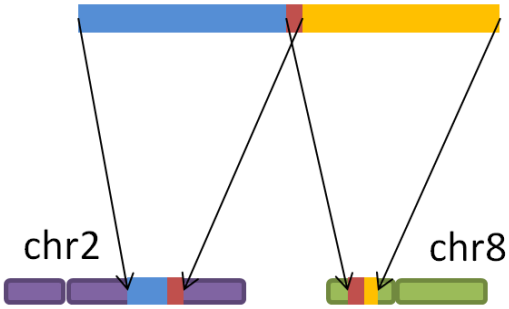
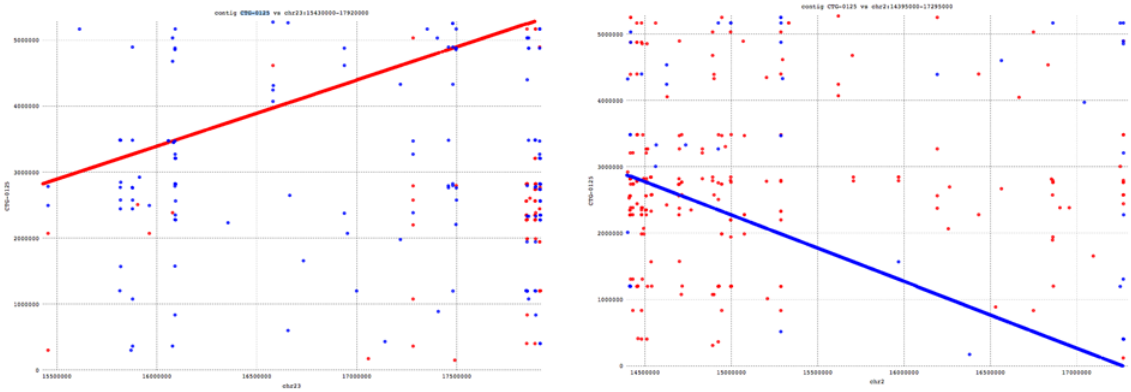


Figure 25 Formation of chimeric contig

Formation of chimeric contigs is usually induced by large scale interchromosomal duplications. The red segments indicate a duplication in this example.



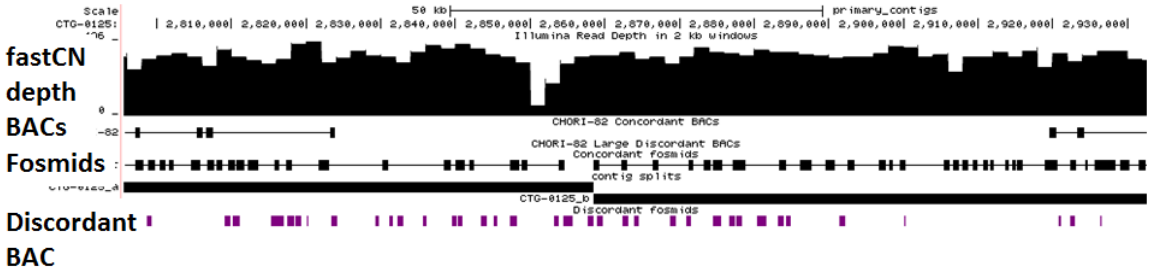


Figure 26 Chimeric contig

Example of a chimeric contigs with two ends mapping to two chromosomes. Top two graphs show MUMMER plot between the same contigs and chromosome 2 and 23. Bottom panel shows the same contig and supporting evidence including mapping of Fosmid pools, discordant and concordant BAC end sequences and CNV calling based fastCN.

In total, 19 of these primary contigs are split into two contigs for scaffolding. The split sites are selected based on inner most boundary on continuous mate pair read coverage and BAC end sequence alignment.

4.4.2 Local assembly and gap filling

To increase continuity, we supplement the contig set with local *de novo* assembly. Raw PacBio reads and primary contigs are aligned to CanFam3.1. Gaps between contigs are defined with actual location extended in both directions by 10kb. Raw PacBio reads intersect with these predefined regions are then pulled into individual FASTQ file for each gap region. For each of these FASTQ files, we employed Canu 1.3 (Koren et al. 2016) to assemble the extracted PacBio reads. The best assembly contig is chosen to

align against the flanking primary contigs using BLAT. When sufficient score (Need further elaboration, >90% identity, no edge effect) is achieved for both contig ends, contigs are kept for next stage of scaffolding.

4.4.3 Draft assembly and scaffolds

The collection local assembled contigs and primary contigs are used as input for scaffolding process. The long-range linking information was selected with Zoey mate pair sequencing reads and original Tasha BAC end sequences. We employed BEEST scaffolding algorithm (Sahlin et al. 2014) to generate chromosomal layout.

4.4.4 Error correction with Illumina data

To further improve base pair accuracy, we aligned the 14x Illumina pair end short reads to the scaffold assembly of Zoey using BWA-MEM (H. Li 2013). These reads were sorted and masked for duplicates using standard Picard procedure. Variants were then called using GATK HaplotypeCaller with the same parameter in section 4.1.2. For base correction, we only select sites with homozygous sites with > 90% alternative allele frequency and mapping quality MAPQ of at least 20. Polished assembly was then output as final Zoey assembly along with unplaced contigs and scaffold during the assembly process.

4.5 Genome improvement by Zoey *de novo* assembly

We next analyzed the improvement of the final Zoey assembly relative to the existing CanFam3.1 reference genome.

4.5.1 Continuity improvement and gap reduction

The scaffolding of the Zoey final assembly was compared to CanFam3.1 assembly using LiftOver, which generated a list of rough coordinates for further refinement. We extracted 5kb of the end sequence for each contig defined in the LiftOver output and aligned them against the corresponding chromosome of CanFam3.1 reference using BLAT. The exact coordinate was then calculated with a custom script based on the best alignment and filtration criteria. A total of 15,800 CanFam3.1 autosome and chrX gaps were successfully aligned with 1bp resolution on the Zoey assembly. Based on these precise alignments, we hereby define two regions of interest. 1. “Assembled gaps” are defined as gaps in CanFam3.1 reference filled with actual sequence without “N” base in the Zoey assembly. 2. “Novel regions” are segments of sequence presence in the Zoey assembly but absent in the CanFam3.1. In summary, we found 2,489 novel regions and 14187 assembled gaps in the Zoey assembly relative to CanFam3.1.

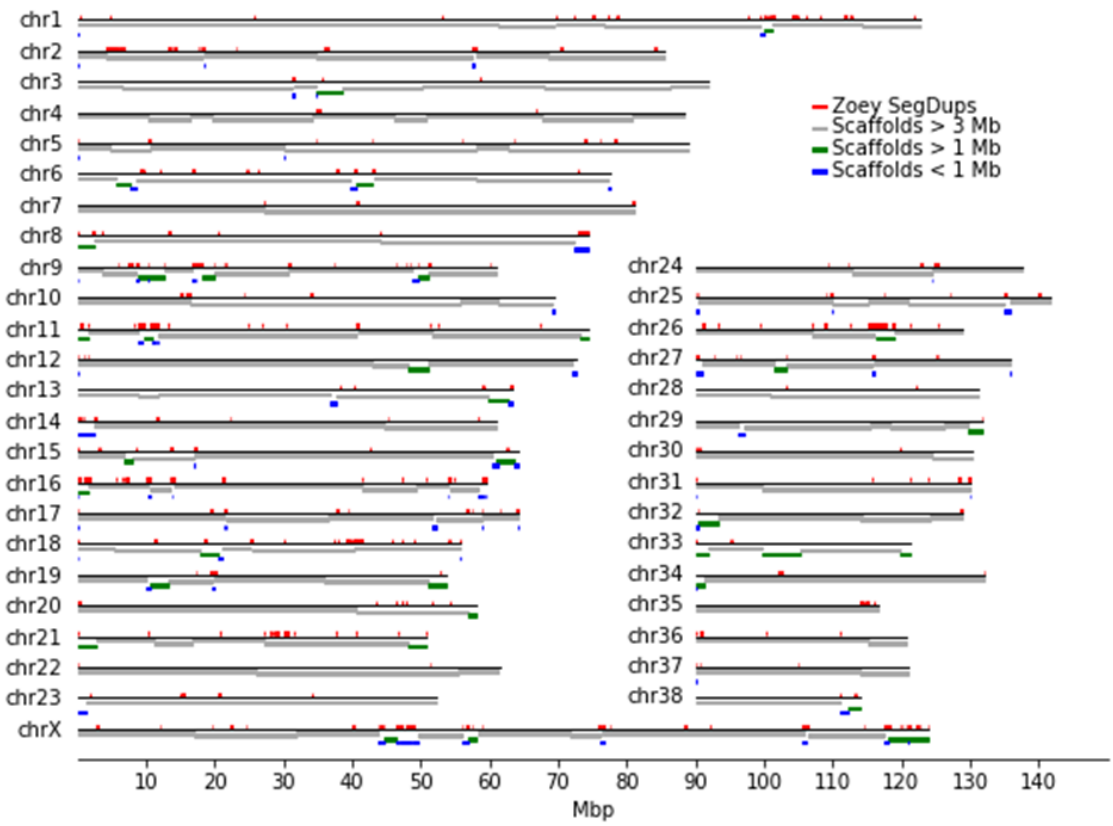


Figure 27 Zoey scaffolds

Scaffold layout of Zoey onto CanFam3.1 reference. Large scale chromosomal scaffolds are only interrupted by presence of large segmental duplications.

The mean continuity improved with assembled gaps to 2.3Mbp from a mean of 124kbp in CanFam3.1.

4.5.2 Improvement of coverage in high-GC regions

4.5.2.1 QuicK-mer Copy Number Interrogation

Next, we investigated the degree of polymorphism of these gap and novel regions among dogs. Since repetitive sequences often flank gaps, we interrogated copy number using QuicK-mer (Pendleton et al. 2018) which utilizes unique sequences within these regions. To do this, we first constructed a list of unique 30-mers using methods described in our previous study (Pendleton et al. 2018). To speed up the process, we next lifted over the previously determined CanFam3.1 control regions to our Zoey assembly and reduced the control 30-mers by 100 fold. These thinned control 30-mers were then merged and sorted with 30-mers that intersected with assembled gap regions and novel regions. We then ran QuicK-mer on 58 samples generated copy number estimates for each region.

4.5.2.2 High GC Content is Primary Cause for CanFam3.1 Gaps

We generated heat maps for the copy number of these regions (Figure 28) across all 58 samples for the autosomal chromosomes and the X chromosome separately. One very apparent pattern emerged, where a major proportion of these novel and gap regions displayed copy numbers of zero across the samples. Further investigation of BWA read alignment patterns at these regions showed no sign of large discordant read pairs near defined gap junctions, which would be indicative of deletions or other structural variation.

Instead, the likely cause of this discrepancy points to errors in the Illumina sequencing process when either DNA inserts could not form clusters during bridge amplification or failure during library PCR.

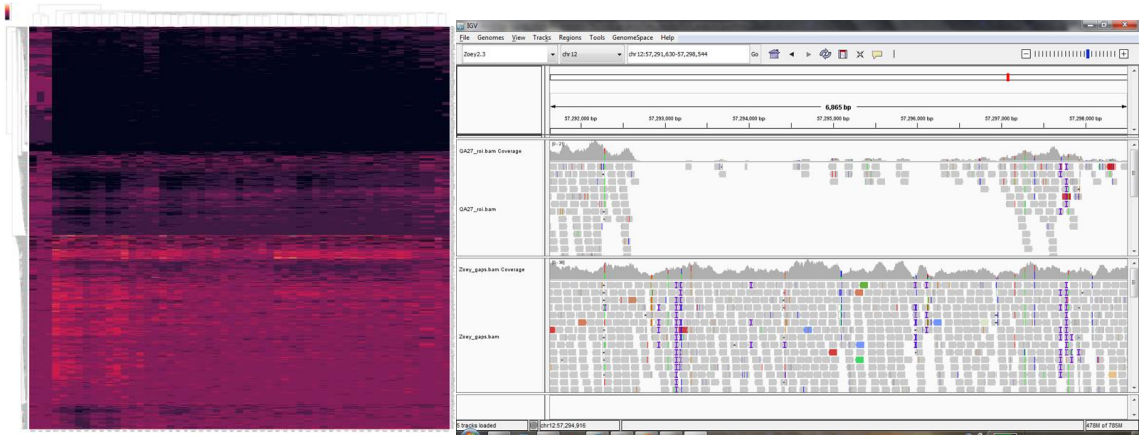


Figure 28 Artifact of false positive deletion

Left: QuicK-mer survey of all assembled gaps and novel regions. Right: Lack of discordant reads inconsistent with deletion signal.

We suspect local repeat or GC content might hold the answer, which could explain either of the two previous possible errors. To address this, we performed a permutation test by randomly shuffling the novel and gap regions on the autosome and X chromosome 1000 times. Empirical p-values were calculated as the chance of observing GC percentage

greater in our regions of interest compared to that of the permutations. The distribution of p-values of these regions is shown in the Figure 29.

We concluded that 55.2% of these regions are highly enriched in GC and thus failed the sequencing. This highlighted a serious limitation of previous sequencing technology which necessitates longer reads and a PCR free approach to improve the reference. We then regenerated the heat plot using only 6174 regions with empirical p-value having greater than 0.05 empirical p-value.

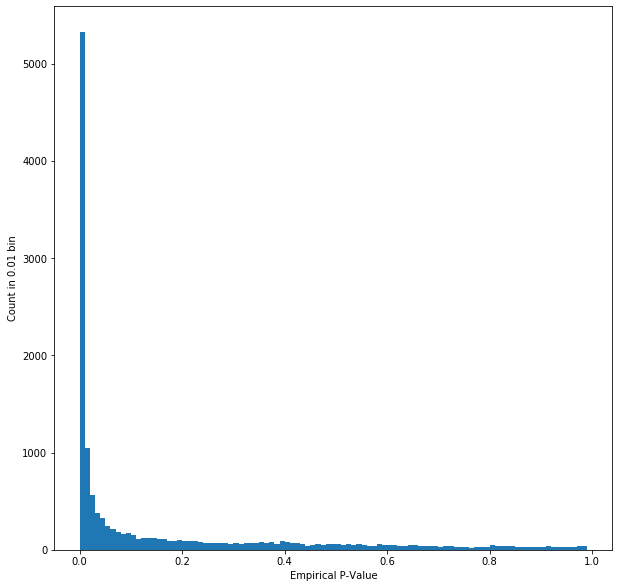


Figure 29 Empirical p-value distribution

Distribution of empirical p-value based on number of shuffled region with GC percentage greater than observed

4.5.2.3 Array CGH validation

To make sure the copy number in these regions is correct, we employed an aCGH data for a previous dog project. The probes were designed using CanFam3.1 thus we remapped the probe sequence against our Zoey assembly. Gaps containing at least three fully bonded aCGH probes and have empirical p-value greater than 0.1 are selected for analysis. We compared the log-ratio of probe intensity against the in silico CNV ratio based on all the unique 30-mers inside the window using method described in Chapter 3.

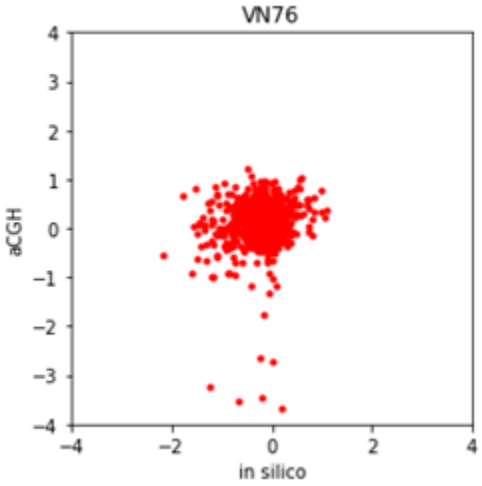


Figure 30 aCGH validation in novel sequences

Example of an aCGH validation of the copy number of gapped regions.

Results in figure 30 demonstrated most of these regions showed fixed copy number with log ratio within +/-1, consistent with the heat plot.

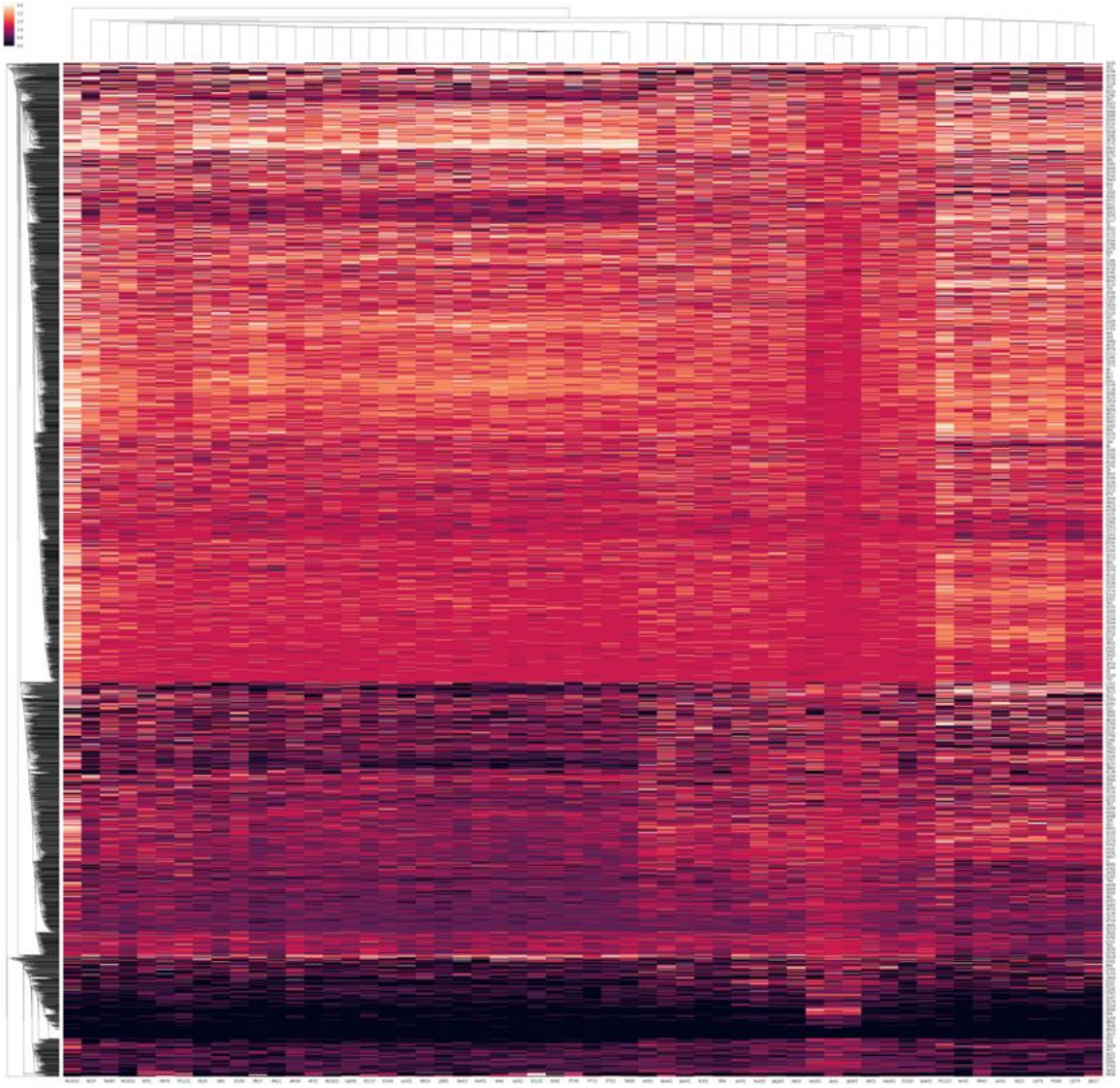


Figure 31 Copy number heatmap of novel sequences

Regenerated heatmap showing copy number of novel and assembled gap regions with p-value greater than 0.1 on autosomes.

4.6 Conclusion

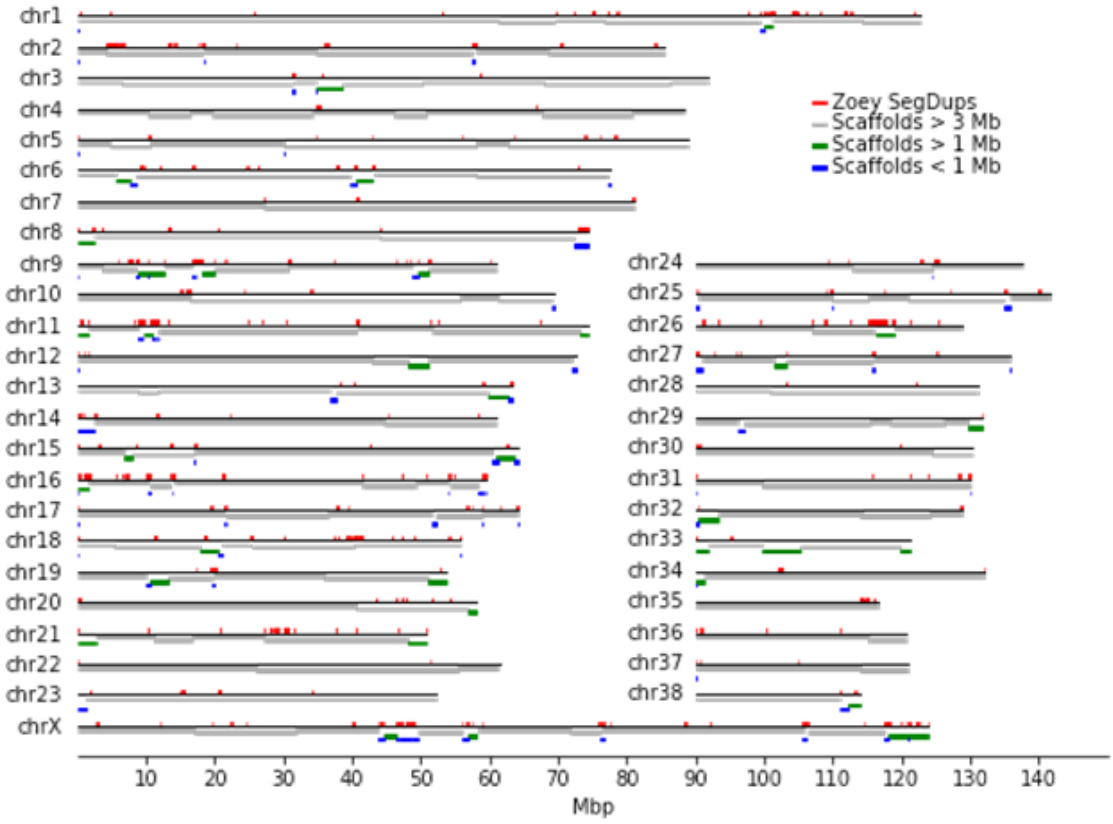


Figure 32 Zoey Scaffolds

Scaffolds of Zoey assembly. Continuity is greatly improved by using mate-pair and BAC libraries. Most discontinuity only happened at large-scale segmental duplications such as chromosome 9 regions.

In conclusion, we improved the CanFam3.1 reference with *de novo* assembly of another breed dog. The new assembly had superior continuity and vastly reduced number of gaps.

4.7 References

Breen, M., S. Jouquand, C. Renier, C. S. Mellersh, C. Hitte, N. G. Holmes, A. Chéron, et al. 2001. "Chromosome-Specific Single-Locus FISH Probes Allow Anchorage of an 1800-Marker Integrated Radiation-Hybrid/linkage Map of the Domestic Dog Genome to All Chromosomes." *Genome Research* 11 (10): 1784–95.

Chaisson, Mark J., and Glenn Tesler. 2012. "Mapping Single Molecule Sequencing Reads Using Basic Local Alignment with Successive Refinement (BLASR): Application and Theory." BMC Bioinformatics 13 (September): 238.

Chin, Chen-Shan, Paul Peluso, Fritz J. Sedlazeck, Maria Nattestad, Gregory T. Concepcion, Alicia Clum, Christopher Dunn, et al. 2016. "Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing." Nature Methods 13 (12): 1050–54.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009. "Real-Time DNA Sequencing from Single Polymerase Molecules." Science 323 (5910): 133–38.

Gnerre, Sante, Iain Maccallum, Dariusz Przybylski, Filipe J. Ribeiro, Joshua N. Burton, Bruce J. Walker, Ted Sharpe, et al. 2011. "High-Quality Draft Assemblies of Mammalian Genomes from Massively Parallel Sequence Data." Proceedings of the National Academy of Sciences of the United States of America 108 (4): 1513–18.

Guo, Yan, Yulin Dai, Hui Yu, Shilin Zhao, David C. Samuels, and Yu Shyr. 2017. "Improvements and Impacts of GRCh38 Human Reference on High Throughput Sequencing Data Analysis." Genomics 109 (2): 83–90.

Hoeppner, Marc P., Andrew Lundquist, Mono Pirun, Jennifer R. S. Meadows, Neda Zamani, Jeremy Johnson, Görel Sundström, et al. 2014. "An Improved Canine

Genome and a Comprehensive Catalogue of Coding Genes and Non-Coding Transcripts." PloS One 9 (3): e91172.

Holden, Lindsay A., Meharji Arumilli, Marjo K. Hytönen, Sruthi Hundi, Jarkko Salojärvi, Kim H. Brown, and Hannes Lohi. 2018. "Author Correction: Assembly and Analysis of Unmapped Genome Sequence Reads Reveal Novel Sequence and Variation in Dogs." Scientific Reports 8 (1): 11853.

Jain, Miten, Hugh E. Olsen, Benedict Paten, and Mark Akeson. 2016. "The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community." Genome Biology 17 (1): 239.

Koren, Sergey, Michael C. Schatz, Brian P. Walenz, Jeffrey Martin, Jason T. Howard, Ganeshkumar Ganapathy, Zhong Wang, et al. 2012. "Hybrid Error Correction and de novo Assembly of Single-Molecule Sequencing Reads." Nature Biotechnology 30 (7): 693–700.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2016. "Canu: Scalable and Accurate Long-Read Assembly via Adaptive K-Mer Weighting and Repeat Separation." https://doi.org/10.1101/071282.

Li, Gang, Brian W. Davis, Terje Raudsepp, Alison J. Pearks Wilkerson, Victor C. Mason, Malcolm Ferguson-Smith, Patricia C. O'Brien, Paul D. Waters, and William J. Murphy. 2013. "Comparative Analysis of Mammalian Y Chromosomes Illuminates Ancestral Structure and Lineage-Specific Evolution." Genome Research 23 (9): 1486–95.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." http://arxiv.org/abs/1303.3997.

Lindblad-Toh, Kerstin, Claire M. Wade, Tarjei S. Mikkelsen, Elinor K. Karlsson, David B. Jaffe, Michael Kamal, Michele Clamp, et al. 2005. "Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog." Nature 438 (7069): 803–19.

Love, R. Rebecca, Neil I. Weisenfeld, David B. Jaffe, Nora J. Besansky, and Daniel

E. Neafsey. 2016. “Evaluation of DISCOVAR de novo Using a Mosquito Sample for Cost-Effective Short-Read Genome Assembly.” BMC Genomics 17 (March): 187.

Luchowski, Rafal, Evgenia G. Matveeva, Ignacy Gryczynski, Ewald A. Terpetschnig, Leonid Patsenker, Gabor Laczko, Julian Borejdo, and Zygmunt Gryczynski. 2008. “Single Molecule Studies of Multiple-Fluorophore Labeled Antibodies. Effect of Homo-FRET on the Number of Photons Available before Photobleaching.” Current Pharmaceutical Biotechnology 9 (5): 411–20.

Lundquist, Paul M., Cheng F. Zhong, Peiqian Zhao, Austin B. Tomaney, Paul S. Peluso, John Dixon, Brad Bettman, et al. 2008. “Parallel Confocal Detection of Single Molecules in Real Time.” Optics Letters 33 (9): 1026–28.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” Genome Research 20 (9): 1297–1303.

Merker, Jason, Aaron M. Wenger, Tam Sneddon, Megan Grove, Daryl Waggott, Sowmi Utiramerur, Yanli Hou, et al. 2016. “Long-Read Whole Genome Sequencing Identifies Causal Structural Variation in a Mendelian Disease.” https://doi.org/10.1101/090985.

Mostovoy, Yulia, Michal Levy-Sakin, Jessica Lam, Ernest T. Lam, Alex R. Hastie, Patrick Marks, Joyce Lee, et al. 2016. “A Hybrid Approach for de novo Human Genome Sequence Assembly and Phasing.” Nature Methods 13 (7): 587–90.

Mustafa, Amir, and Leonard Yuen. 1991. “Identification and Sequencing of theChoristoneura Biennisentomopoxvirus DNA Polymerase Gene.” DNA Sequence: The Journal of DNA Sequencing and Mapping 2 (1): 39–45.

Nakamura, Kensuke, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, et al. 2011. “Sequence-Specific Error Profile of Illumina Sequencers.” Nucleic Acids Research 39 (13): e90.

Parker, Heidi G., Dayna L. Dreger, Maud Rimbault, Brian W. Davis, Alexandra B. Mullen, Gretchen Carpintero-Ramirez, and Elaine A. Ostrander. 2017. “Genomic

Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development." Cell Reports 19 (4): 697–708.

Pendleton, Amanda L., Feichen Shen, Angela M. Taravella, Sarah Emery, Krishna R. Veeramah, Adam R. Boyko, and Jeffrey M. Kidd. 2018. "Comparison of Village Dog and Wolf Genomes Highlights the Role of the Neural Crest in Dog Domestication." BMC Biology 16 (1): 64.

Rossi, Elena, Orietta Radi, Lisa De Lorenzi, Alessandra Iannuzzi, Giovanna Camerino, Orsetta Zuffardi, and Pietro Parma. 2015. "A Revised Genome Assembly of the Region 5' to Canine SOX9 Includes the RevSex Orthologous Region." Sexual Development: Genetics, Molecular Biology, Evolution, Endocrinology, Embryology, and Pathology of Sex Determination and Differentiation 9 (3): 155–61.

Sahlin, Kristoffer, Francesco Vezzi, Björn Nystedt, Joakim Lundeberg, and Lars Arvestad. 2014. "BESST--Efficient Scaffolding of Large Fragmented Assemblies." BMC Bioinformatics 15 (August): 281.

Saurabh, Saumya, Suvrajit Maji, and Marcel P. Bruchez. 2012. "Evaluation of sCMOS Cameras for Detection and Localization of Single Cy5 Molecules." Optics Express 20 (7): 7338–49.

Schaeffer, Edward M. 2012. "Re: Origins of the E. Coli Strain Causing an Outbreak of Hemolytic-Uremic Syndrome in Germany." The Journal of Urology 187 (2): 514–15.

Vasmatzis, George, Sarah H. Johnson, Ryan A. Knudson, Rhett P. Ketterling, Esteban Braggio, Rafael Fonseca, David S. Viswanatha, et al. 2012. "Genome-Wide Analysis Reveals Recurrent Structural Abnormalities of TP63 and Other p53-Related Genes in Peripheral T-Cell Lymphomas." Blood 120 (11): 2280–89.

Weirather, Jason L., Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. 2017. "Comprehensive Comparison of Pacific Biosciences and Oxford Nanopore Technologies and Their Applications to Transcriptome Analysis." F1000Research 6 (February): 100.

Williams LJ, Et al. n.d. "Paired-End Sequencing of Fosmid Libraries by Illumina. -

PubMed - NCBI.” Accessed January 14, 2019.
https://www.ncbi.nlm.nih.gov/pubmed/22800726.

Chapter 5: Conclusions and Future Directions

5.1. CNV detection algorithms

The performance of QuicK-mer 2.0 greatly exceeds that of the first version. To assess the potential for additional improvements I looked into the time cost for the feeder thread on I/O. The latest build shows 80-90% CPU usage for the feeder thread. I suspect a majority of CPU time is still spent on encoding k-mers or looping through the FASTQ input. Two strategies could further improve the performance: 1) feed the working thread with raw reads and unload the k-mer encoding to multiple threads and 2) read the file as binary blocks into memory and scan through the reads directly without prior parsing. Either approach should accelerate the process further on a compute node that has a solid-state drive. However, since we had observed network drive delay in practice and uncompressing the gzip format had a significant impact these improvements are likely to have a marginal return. Thus, they were currently not implemented yet.

We have observed that the fastCN algorithm, which is based on read mapping, has better correlation with results from aCGH validation. This observation can be explained by similarity with the chemical process of probe hybridization. Occasionally a single mismatch will not necessarily yield no-signal whereas QuicK-mer is much more stringent where a single mismatch leads to dropout. Such mismatches could be the result of recent polymorphism in the population.

5.2. *de novo* assembly for dogs

Our *de novo* assembly has achieved vast improvement relative to CanFam3.1. The vast number of assembled gaps is consistent with our observation of gap spanning PacBio reads during initial mapping. The origin and content of these gapped regions is of great interest. We had already observed that they majority of such opening gaps have enriched GC content. In the process of reaching such a conclusion, we had also performed analysis of repeated elements. In previous genome improvement projects, tandem repeats were a primary cause for gap sequences (Chaisson et al. 2015; Pendleton et al. 2015; Seo et al. 2016) or enriched in missing sequences such as centromere and telomeres (Alkan et al. 2011). Using similar permutations as described in Chapter 4 Section 5.2.3, we generated the empirical p-value for tandem repeats as well. (Figure 33L) We found that tandem

repeats are also enriched but to a much lesser extent compared to GC content. The joint distribution shows sequences are enriched with 20% of tandem repeats, yet a majority failed during the classical assembly process was most likely due to GC.

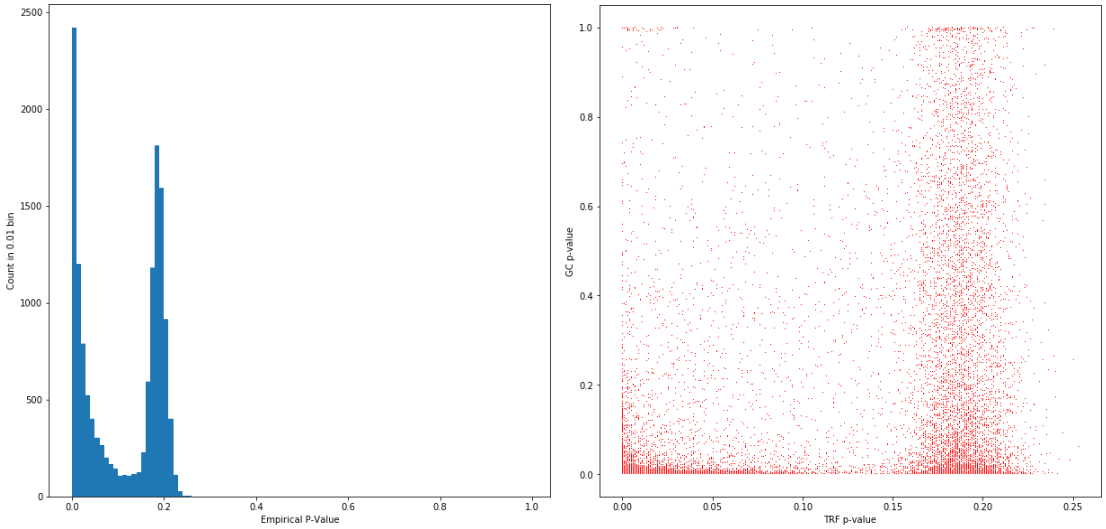


Figure 33 Empirical p-values of TRFs

Left: Empirical p-value based on 1000 times permutation of closed gaps and novel regions with intersection of tandem repeats greater than actual observed. Right: Joint p-value plot between GC content and tandem repeats

We had initially suspected the failure is probably due to bad library or sequencing in many of those Illumina samples. Yet the presence of these gaps in the first place

indicated the Sanger end sequencing of clones also failed for such regions in the original CanFam1/2 assembly (Lindblad-Toh et al. 2005). When we considered that the Sanger sequencing is a variation of primer-based PCR, it might inherit some of its drawback at these regions not only repetitive but also highly GC rich. The denature-annealing cycle could generate secondary structure in the single strand template and further produce sequence slippage during the extension process. Such PCR related phenomenon has been experimentally shown in previous studies (Hommelsheim et al. 2014) Further evidence supporting a failed sequencing-through than an actual discontinuity in clone is supported by resolved gap size in our *de novo* Zoey assembly (Figure 34). For the majority of smaller gaps, the actual resolved size in Zoey correlates well against the hypothetic gap size in CanFam3.1. This concludes that the size estimation from original BAC during the CamFam1/2 project is generally correct. GC rich and repeat elements can be resolved in PacBio due to use of single molecule readout, which completely eliminated PCR, and good processivity of improved phi29 polymerase (Korlach et al. 2008).

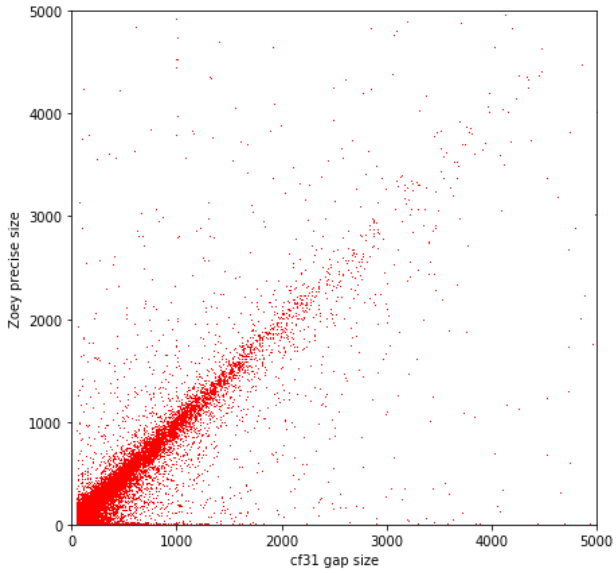


Figure 34 Correlation between estimated and true gap size

Assembled gap size of Zoey compared to original proposed gap size in CanFam3.1

The GC content of such regions is also intriguing given CpG islands usually had a functional role in gene expression and regulation. Due to lack of an equivalent to the ENCODE project on dogs, we have to rely on comparative genome analysis and gene annotation. We further explored this idea by annotating our genome with RNA sequencing data and intersect these regions against the annotated gene models. It turns out more than 2000 of such regions intersect with the first exon (Figure 35). This

indicates that our reference assembly greatly improves the dog annotation where many promoter regions were previously missing in CanFam3.1.

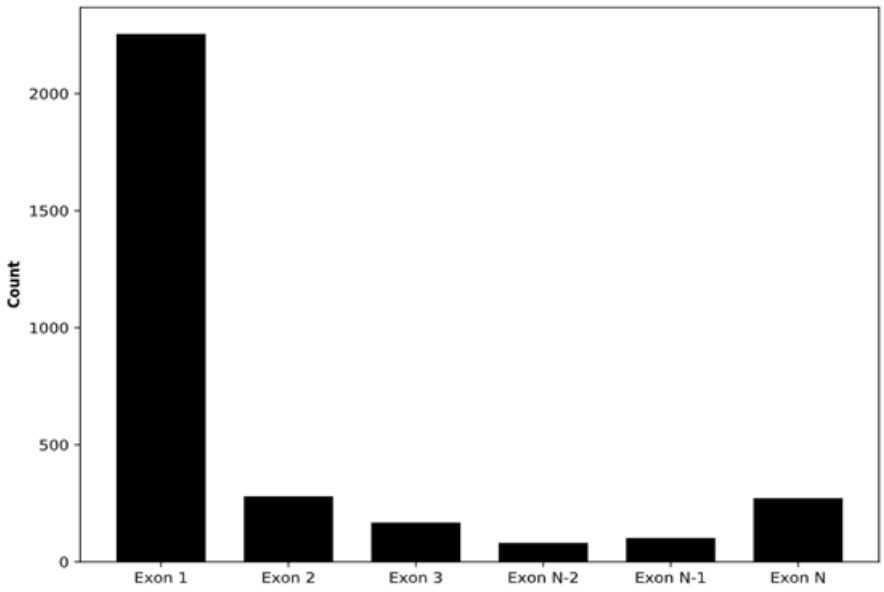


Figure 35 Intersect between exons and novel sequences

Large proportion of gap regions intersect with the first exons of many genes.

Some of these assembled gap and Zoey novel regions shows variable copy number based on QuicK-mer data. Since aCGH data is sparse and still noisy within +/-1 log-ratio, it would be near impossible to validate with high accuracy. Identification and sequencing in

individual canine samples might be required for those were potentially variable. Whether the copy number observed in Chapter 4 Section 5.2 plays a role in expression regulation will be an interesting question to further explore.

The improvement to small gaps is apparent. We then shift our focus to large scale structural variation and especially the mis-assembled chromosome 9 region identified in the selection scan. Based on the MUMMER plot, we found one of the small inversions was correctly resolved in our Zoey assembly. However, due to the existence of segmental duplication approximately one megabase in size, even long PacBio reads could not resolve these regions. In fact, Figure 32 clearly shows discontinuity of scaffold concentrates on segmental duplications. To further resolve these regions, we attempt to sequence the original BACs used in CanFam3.1. A tiling path near chromosome 9 for these BACs are chosen and selected (Figure 37). Due to presence of mobile elements, individual BACs were still fragmented using short-reads based assembly even though each BAC is barcoded. To further improve, we decide to resolve these BAC using PacBio sequencing.

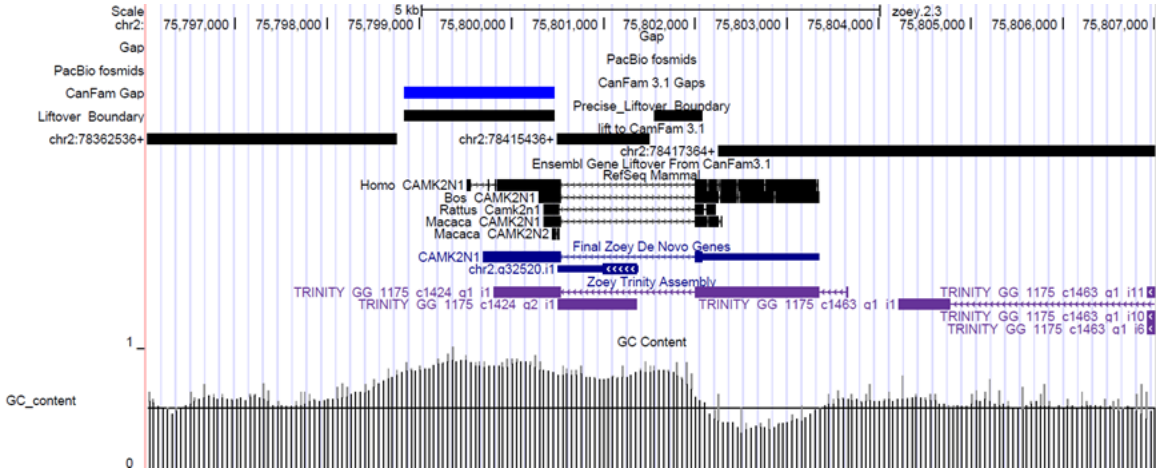


Figure 36 Example of an assembled promoter

Example of assembled CanFam3.1 gap showing extreme GC content and missing promoter region of gene CAMK2N1.

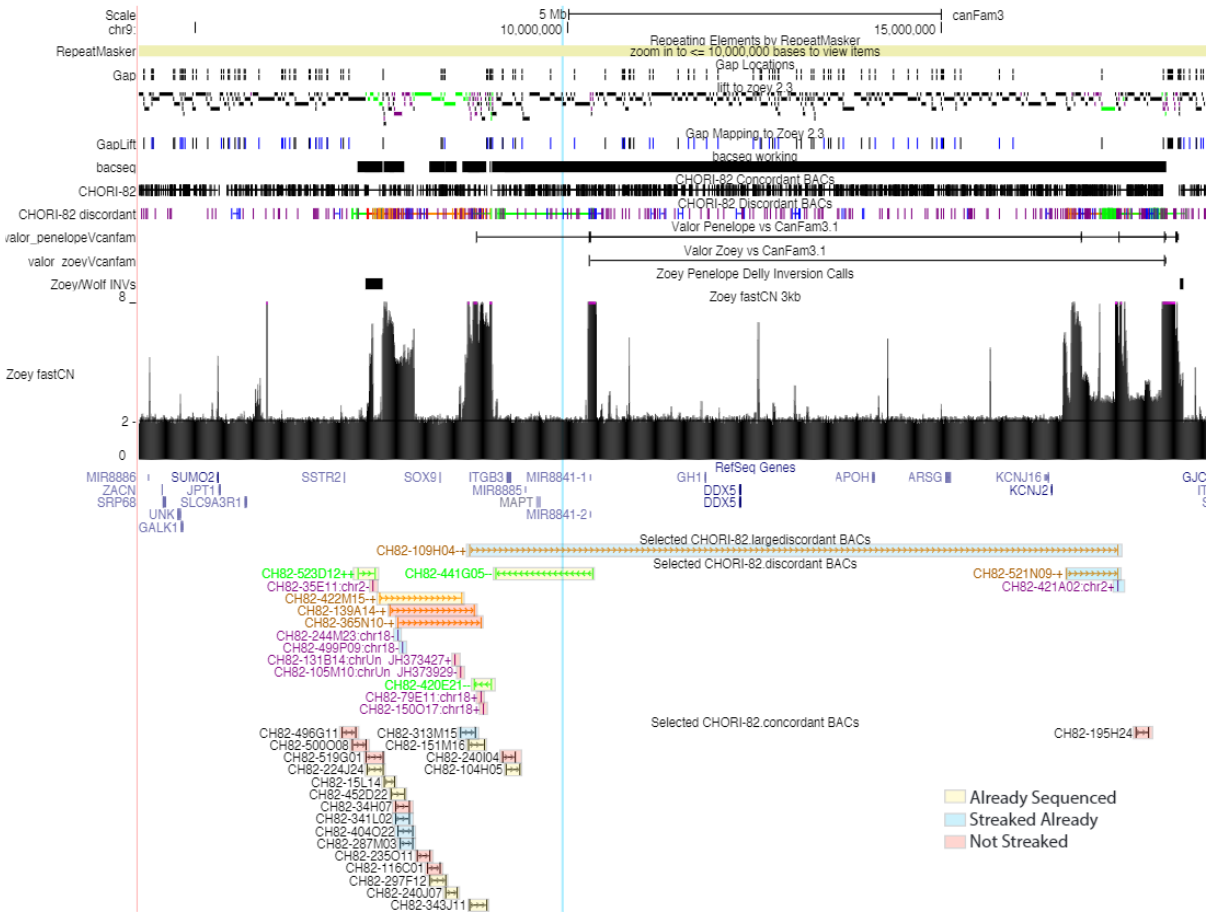


Figure 37 Alignment of selected BACs

Mapping of selected BACs on CanFam3.1 reference around the 8Mbp on chromosome 9. CNVs from fastCN of Zoey is shown.

Since PacBio employed a proprietary barcode inside the dumbbell adapter, individual library cost is high. To reduce the cost, we choose to mix four BACs into each barcode. A total of 16 BACs will be allocated into four libraries. An optimal solution for such mixing

should minimize the pairwise alignment for the BACs within a library. The ideal solution should also restrict length of matching to below PacBio insert size. Such a knapsack or combinatorial optimization problem is NP-hard. However, since the search space is sufficiently small, optimal solution is found with a script less than 6 hours. More analysis would be done in the future to improve continuity in this region.

5.3. Future Directions

Our current *de novo* reference assembly has greatly increased the continuity and improved the gene annotation relative to the existing assembly. This implies that the existing CanFam3.1 may poses a bias towards regions which are easy to correctly assemble. Using the chromosome 9 regions as an example, we clearly observed an inversion signature induced by misrepresentation of the underlying genome structure. It is also unlikely that V_{ST} analysis will correctly reflect regions missing in assembly gaps or uncovered sequences. Another bias in these analyses was introduced by using dog as the genome reference. It is probable that some regions were deleted during the domestication process relative to the wolves. Due to absence of such sequences in the dog reference assembly, wolf reads would remain unmapped or uncovered by k-mers based on that genome reference. This is supported by evidence in our previous study. We had assembled the unmapped reads from a wolf Illumina library and compiled a list of wolf

specific contigs distinct from CanFam3.1 and dog novel sequences. A wolf reference assembly could potentially address this bias. At present, a wolf *de novo* reference based on mate-pair and short-reads is available (Gopalakrishnan et al. 2017). However, its quality remains poor relative to our long read based approach.

To further survey the CNVs and point mutations during the dog domestication process, an improved wolf assembly is highly desirable. A similar contiguous assembly using long reads could better capture wolf novel regions relative to the domesticated dogs. Thirdly, we should also focus on the genome diversity within the village dogs. Our current Zoey reference and the CanFam3.1 are both based on purebred dogs. The use of relatively homozygous breeds improves the assembly but has the potential to miss important variation within the village dog populations. It is evident that even at shorter DNA insert size we could achieve a decent N_{50} compared to the improved gorilla genome (Gordon et al. 2016) simply due to homozygosity of the breed dog sample. In the near future, low depth PacBio sequencing of village dogs can capture these structural variations and provide a supplement to the Zoey reference. Finally, we demonstrated the limitation of PCR-based Sanger and short-read sequencing technologies in GC rich regions. The proximity of such GC rich regions to the first gene exon hints at the potential regulatory

function from CpG islands. Since these cost-effective approaches still dominate resequencing studies, other methods have to be devised to overcome these biases. Hybridization based approaches such as array CGH can be designed to survey CG rich regions. But stringent requirements in the probe melting temperature and synthesis might also limit the outcome. Capture kits can be constructed to limit the input for a long read PacBio sequencing in these extreme GC regions in addition to short-reads whole genome data.

Should potential CNV regions be uncovered using an improved reference and techniques, experiments can be designed to verify their biological impacts. Specially designed CRISPR cas9 mutagenesis can disable additional copies in the dog cell lines to observe biological outcome related to corresponding CNVs.

The speed and efficiency of QuicK-mer 2.0 opens up research opportunities related to paralog specific sequences. One of those interesting areas is the evolution and divergence of gene families. Pertaining to my topic of CNVs, new copies of genes from duplication could lead to future neo-functions and sub-functions. One of the most important are the genes encoding the DNA binding domains controlling gene expression regulation. Previous research has surveyed such transcription factors using similar sequence

matching (Shen et al. 2018). Since the counting step in QuicK-mer 2.0 takes almost no time, a even wider survey of sequence frequency in a paralogous specific manner can be conducted. For example, we can first define unique k-mers inside mutated regions within each paralog. These k-mers can be organized into groups by ordering them using the QuicK-mer 2.0 linking information. Once the k-mer hash index is built, multiple samples can be queried efficiently for paralog abundance. This approach can also be expanded across different genomes and organism, or even include mutated k-mers in specific domains to search for variations.

In summary, this dissertation reports the construction of two CNV detection algorithms and their successful application to genome diversity studies. Various comparisons demonstrated the precision and efficiency of these approaches. Combining these tools with an improved *de novo* canine assembly, we showed that a quality reference could reduce bias in genome variation studies. In the future, more genomes will be improved using single molecule long reads or even resort to traditional clone selection technique to reveal the true extent structural rearrangements.

5.4. References

Alkan, Can, Maria Francesca Cardone, Claudia Rita Catacchio, Francesca Antonacci, Stephen J. O'Brien, Oliver A. Ryder, Stefania Purgato, et al. 2011. "Genome-Wide Characterization of Centromeric Satellites from Multiple Mammalian Genomes." Genome Research 21 (1): 137–45.

Chaisson, Mark J. P., John Huddleston, Megan Y. Dennis, Peter H. Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, et al. 2015. "Resolving the Complexity of the Human Genome Using Single-Molecule Sequencing." Nature 517 (7536): 608–11.

Gopalakrishnan, Shyam, Jose A. Samaniego Castruita, Mikkel-Holger S. Sinding, Lukas F. K. Kuderna, Jannikke Räikkönen, Bent Petersen, Thomas Sicheritz-Ponten, et al. 2017. "The Wolf Reference Genome Sequence (Canis Lupus Lupus) and Its Implications for Canis Spp. Population Genomics." BMC Genomics 18 (1): 495.

Gordon, David, John Huddleston, Mark J. P. Chaisson, Christopher M. Hill, Zev N. Kronenberg, Katherine M. Munson, Maika Malig, et al. 2016. "Long-Read Sequence Assembly of the Gorilla Genome." Science 352 (6281): aae0344.

Hommelsheim, Carl Maximilian, Lamprinos Frantzeskakis, Mengmeng Huang, and Bekir Ülker. 2014. "PCR Amplification of Repetitive DNA: A Limitation to Genome Editing Technologies and Many Other Applications." Scientific Reports 4 (May): 5052.

Korlach, Jonas, Arek Bibillo, Jeffrey Wegener, Paul Peluso, Thang T. Pham, Insil Park, Sonya Clark, Geoff A. Otto, and Stephen W. Turner. 2008. "Long, Processive Enzymatic DNA Synthesis Using 100% Dye-Labeled Terminal Phosphate-Linked Nucleotides." Nucleosides, Nucleotides & Nucleic Acids 27 (9): 1072–83.

Lindblad-Toh, Kerstin, Claire M. Wade, Tarjei S. Mikkelsen, Elinor K. Karlsson, David B. Jaffe, Michael Kamal, Michele Clamp, et al. 2005. "Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog." Nature 438 (7069): 803–19.

Pendleton, Matthew, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar

Franzen, Tobias Rausch, Adrian M. Stütz, et al. 2015. “Assembly and Diploid Architecture of an Individual Human Genome via Single-Molecule Technologies.” Nature Methods 12 (8): 780–86.

Seo, Jeong-Sun, Arang Rhie, Junsoo Kim, Sangjin Lee, Min-Hwan Sohn, Chang-Uk Kim, Alex Hastie, et al. 2016. “De novo Assembly and Phasing of a Korean Human Genome.” Nature 538 (7624): 243–47.

Shen, Ning, Jingkang Zhao, Joshua L. Schipper, Yuning Zhang, Tristan Bepler, Dan Leehr, John Bradley, John Horton, Hilmar Lapp, and Raluca Gordan. 2018. “Divergence in DNA Specificity among Paralogous Transcription Factors Contributes to Their Differential In Vivo Binding.” Cell Systems 6 (4): 470–83.e8.