이학박사학위논문

비교유전체학을 이용한 선충의 서브텔로미어 진화와 표현형 변이 연구

Comparative genomic studies on subtelomere evolution and phenotypic variation in nematodes

2019년 12월

서울대학교 대학원

생명과학부

김 준

ABSTRACT

Comparative genomic studies on subtelomere evolution and phenotypic variation in nematodes

Jun Kim

Department of Biological Sciences

The Graduate School

Seoul National University

Long-read sequencing technologies have contributed greatly to comparative genomics among species and can also be applied to study genomics within a species. In this study, to determine how substantial genomic changes are generated and tolerated within a species, a *C. elegans* strain, CB4856, was sequenced which is one of the most genetically divergent strains compared to the N2 reference strain. For this comparison, the Pacific Biosciences (PacBio) RSII platform (80×, N50 read length 11.8 kb) was used and de novo genome assembly were generated to the level of pseudochromosomes containing 76 contigs (N50 contig = 2.8 Mb). I identified structural variations that affected as many as 2,694 genes, most of which are at chromosome arms. Subtelomeric regions contained the most extensive genomic rearrangements, which even created new subtelomeres in some cases. The subtelomere structure of Chromosome VR implies that ancestral telomere damage was repaired by alternative lengthening of telomeres even in the presence of a functional telomerase gene and that a new subtelomere was formed by break-induced replication. My study demonstrates that substantial genomic changes including structural variations and new subtelomeres can be tolerated within a species, and that these changes may accumulate genetic diversity within a species. Secondly, I

also assembled draft genomes of two *C. elegans* relative species, *Auanema freiburgensis* and *Auanema* sp. APS14, which have and a distinct reproductive (three genders; male, female, and hermaphrodite) and behavioral repertoire (tube-nictation). *A. freiburgensis* and *Auanema* sp. APS14 were sequenced using the PacBio RSII (270×, N50 read length 12.5 kb) and the Oxford Nanopore Technologies (ONT) MinION platforms (113×, N50 read length 3.6 kb), respectively, and their reads were assembled as smaller genomes (55 and 69 Mb, respectively) compared to that of *C. elegans* (~100 Mb). Comparative genomic studies of these genomes will help understand how genomic changes in close relative species affect evolution of novel traits.

Keywords: *C. elegans*, nematode, long-read sequencing, *de novo* genome assembly, telomere, subtelomere, alternative lengthening of telomreres (ALT)

Student Number: 2015-20422

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter I

Introduction

Long-read sequencing and *de novo* genome assembly

All species have a variety of heritable phenotypic variations. Studying the genetic factors contributing to these differences is one of the major challenges in genetics. Geneticists have studied the relationship between genotype and phenotype using various methodologies including heritability estimation (Johnson and Wood 1982), mutant study (Brenner 1974; Nigon and Dougherty 1950), and natural variation study (Fatt and Dougherty 1963). In addition, recent long-read sequencing technologies have enabled genome-level comparisons, which further advances investigating chromosome-scale changes in the same species as a novel phenotype, or comparative genomics using *de novo* genome assembly of closely related species (Rödelsperger et al. 2017; Yin et al. 2018).

De novo genome assembly is a process of genome reconstruction from sequencing data, and long-read sequencing technologies produce high quality genome assemblies with high throughput. Since it is currently impossible to sequence a chromosome from end to end at once, genome reconstruction should be performed by reading the same position many times and assembling overlapped reads, similar to jigsaw puzzle matching (Figure 1). Just as jigsaw puzzles are difficult to match with a wide, white sky, it is difficult to assemble consecutive, long repetitive sequences in the genome. In particular, the current short-read sequencing technologies produce accurate, but very short reads of 100-500 bp, which make it almost impossible to resolve such repetitive regions. Long-read sequencing technologies, however, can generate up to 100 kb reads, which can sometimes fully cover very long repetitive sequences at once. It enhances qualities and throughputs of *de novo* genome assemblies.

Caenorhabditis and *Caenorhabditis elegans* as a model system for comparative genomics

The genus *Caenorhabditis* is a great resource and example for comparative genomic studies using *de novo* genome assemblies. Species diversity and small genome sizes have made the *Caenorhabditis* genus a subject for molecular dissection of the genome and of trait evolution (Stein et al. 2003; Slos et al. 2017; Yin et al. 2018). Over 50 species of the *Caenorhabditis* genus have been

collected, and the genomes of 25 of them have been sequenced (Stevens et al. 2018). Although inter-species comparisons have found many genomic differences that have provided insights into genome evolution, different species have already undergone numerous changes. Little is known about where and how genomic changes within a species have accumulated. To understand genomic changes within a species, I compared the genome of the reference N2 strain with that of CB4856, a highly divergent *C. elegans* wild strain (Koch et al. 2000; Wicks et al. 2001).

N2 and CB4856 have numerous heritable phenotypic differences. The recombinant inbred lines and the recombinant inbred advanced intercross lines produced by crossing the two strains have revealed several genetic loci that cause phenotypic variations such as aggregation behavior, mating, nictation behavior, pathogen response, and genetic incompatibility (de Bono and Bargmann 1998; Tijsterman et al. 2002; Schulenburg and Müller 2004; Kammenga et al. 2007; Palopoli et al. 2008; Seidel et al. 2008, 2011; Kim et al. 2017; Lee et al. 2017). Attempts have been made to obtain the CB4856 genome that accurately represents these genetic variants, but the currently available CB4856 reference genome has the limitation that it has been assembled from sequences that were obtained using short-read sequencing (Thompson et al. 2015). These sequences may underrepresent genomic rearrangements that are longer than the insert length and may miss insertions and repetitive sequences.

Repetitive nature of subtelomere and the trace of alternative lengthening of telomeres (ALT) in subtelomeric regions

The occurrence of repetitive sequences is generally highest near the ends of chromosomes. A subtelomere is a hypervariable region adjacent to the telomere and has various repeats including segmental duplicated blocks. The repetitive nature of subtelomeric and telomeric regions can impair their assembly by short-read sequencing. For example, in the human genome hg19 version released in 2009, telomeric repeats directly linked to subtelomere sequences appear in only 17 out of 46 chromosome ends (Rudd 2014). In addition, in the *C. elegans* VC2010 de novo assembly by Nanopore long-read sequencing, telomeric repeats directly linked to subtelomere sequences appear in only six out of 12 chromosome ends (Tyson et al. 2018). Therefore, these regions could be underrepresented in de

novo assembled genomes. The high variability of subtelomeres over generations facilitates the emergence of new genes and may help to increase the fitness of organisms. The possibility of the involvement of subtelomeres in chromosome evolution has not been extensively studied because of the difficulty in the genome assembly near subtelomeres.

Telomeres are the ends of linear chromosomes of eukaryotic cells. In most cases, telomeres are composed of specific sequence repeats to form highly ordered structures. Critically shortened telomeres can lead to chromosome dysfunction, so all eukaryotic cells must maintain appropriate telomere length (Harley et al. 1990; O'Sullivan and Karlseder 2010). Organisms that fail to maintain the telomere in the germline cells eventually become sterile (Blackburn 1991; Blasco et al. 1997; Meier et al. 2006). The telomere lengthening is mainly fulfilled by using telomerase and telomeric repeats, but in some cases alternative lengthening of telomeres (ALT) can be used to lengthen telomeres without utilizing telomerase (Lundblad and Blackburn 1993; Nakamura et al. 1998).

ALT is defined as telomere lengthening in the absence of functional telomerase activity. ALT occurs in certain cancer cells in humans and in organisms in nature; for example, Drosophila uses retrotransposon rDNA sequences and onions use minisatellite rDNA sequences to maintain telomeres. This ALT process uses sequences other than canonical telomeric repeats (Bryan et al. 1997; Pich and Schubert 1998; Cesare and Reddel 2010; Garavís et al. 2013; Mason et al. 2016). In *C. elegans*, the telomerase-deficient animals survived telomere attrition by replicating template for ALT (TALT) at the end of every chromosome (Seo et al. 2015; Kim et al. 2016). Break-induced replication (BIR) is another major mechanism to maintain telomeres without the action of telomerase, as reported in human cancer cells and yeasts (Lydeard et al. 2007; Dilley et al. 2016). During BIR, homologous templates from either the same chromosome or even a nonallelic region can be used for replication of the templates, up to the size of 200 kb, which can establish new subtelomeres (Costantino et al. 2014; Mason and McEachern 2018).

Phenotypic diversity in the genus *Auanema*

Although *Caenorhabditis* is a good model for studying trait evolution, this genus does not

have all the phenotypic diversity in nematodes, and it limits the research scope to specific *Caenorhabditis* phenotypes. For this reason, various nematodes have been investigated to study novel traits that do not appear in *Caenorhabditis*. For example, nematodes of *Oscheius* and *Pristionchus* genus are used as important satellite model systems for evolutionary developmental research as they have different vulva (Félix 2006) and buccal cavity structure (Sommer 2006). *Strongyloides* clade is among the most suitable models to study the evolution of parasitism (Viney and Lok 2007). A comparison of the genomes of free-living, facultative parasitic, and obligatory parasitic nematodes included in the clade has shown changes in genome and gene contents as parasitism evolves (Hunt et al. 2016). Many interesting phenotypes have not yet been studied because no model systems or just few resources are available for such phenotypes. To expand the nematode collection to study the trait evolution deeper and even broader, I have collected ~20 nematode species from rotten fruits in South Korea and focused on the *Auanema* genus, which has been closely related to *Caenorhabditis* but has not been studied in detail.

Almost all *Auanema* species have distinct reproductive and behavioral repertoire compared to *Caenorhabditis*. Their sex is mainly determined by the number of sex chromosome X, that is, XX worms develop into females and XO worms into males. However, young XX larvae also can develop into hermaphrodites under harsh conditions including high population density (Félix 2004). This interesting three-gender phenotype was considered as an unstable intermediate stage between female/male and hermaphrodite/male reproductive modes, but the genus *Auanema* shares the phenotype in a quite stable fashion (Kanzaki et al. 2017). It has also a novel behavioral phenotype, tube-nictation. In *C. elegans* and close relative species, a dispersal behavior, nictation, facilitates the migration from an old habitat to a new one. Worms use rough surface such as fungi to support their tails and wave their heads into the air, which may increase the possibility to attach the worms to their carrier animals including isopods and snails. However, *Auanema* worms can use their own cuticles (tube) of the previous molt for tail support instead of fungi or any rough surface. It was hypothesized that they can recognize and respond quickly to their carriers to hitchhike using this tube-waving behavior.

In addition, one uncharacterized *Auanema* species, *Auanema* sp. APS14, has another

interesting behavior – group nictation. Tens to thousands of *C. elegans* worms aggregate and hold each other to form a huge rod using a complex surface to intensify the dispersal probability. On the other hand, *Auanema* sp. APS14 worms can show this group nictation behavior without any surface, simply by crawling to a specific site and aggregate together. This interesting behavior has not yet been found even in the same genus except the species, thus this genus can serve as a model system to study how novel reproductive and behavioral phenotypes have evolved using close relative species.

Purposes of the study

The purposes of my thesis researches were two-fold. First, I aimed at obtaining a complete CB4856 genome by long-read sequencing and report the identification and characterization of structural variations (SVs) within the genome and structural changes in the subtelomeric regions. I also discuss the significance of new subtelomere formation in generating new genetic materials for evolution of new traits. Secondly, I aimed at assembling two genomes of the genus *Auanema*, which is close to *Caenorhabditis*, to lay the groundwork for studying novel traits in non-standard model organisms.

Materials and methods

Worm maintenance

Worms were cultured at 20°C under standard culture conditions.

gDNA extraction and long-read sequencing

Mixed stage worms were collected and washed 5× in M9 buffer. Worms were lysed in lysis buffer for 8 h (100 μg mL^{-1} ProteinaseK, 50 mM KCl, 10 mM Tris (pH 8.3), 2.5 mM $MgCl_2$, 0.45% NP-40, 0.45% Tween 20, and 1% beta-mercaptoethanol). DNA was extracted using phenol-chloroform extraction and ethanol precipitation. To minimize DNA shearing, I used phase-lock gel and minimized pipetting. DNA in TE buffer was treated with RNase (10 μg mL^{-1}) for 2 h and re-extracted, before being dissolved in TE buffer. Macrogen performed library preparation and sequencing using the PacBio Single Molecule, Real-Time (SMRT) DNA sequencing technology (platform: PacBio RS II; chemistry: P6-C4) for the *C. elegans* CB4856 and *A. freiburgensis* APS7 strains. DNA library of the *Auanema* sp. APS14 strain was prepared using the SQK-LSK109 kit and sequenced using the FLO-MIN106 flowcell of the ONT MinION platform.

Total RNA extraction and RNA sequencing

Mixed stage worms of the CB4856 strain were harvested in the M9 buffer and TRIzol. To disrupt worms, I performed flash-freeze/thaw cycles 10×. RNA was extracted using chloroform and isopropanol precipitation. Macrogen performed library preparation and sequencing using HiSeq 4000 (Illumina) with 101-bp paired-end reads. Technical duplicate samples were sequenced in this study.

Genome assembly and polishing

For the CB4856 strain, d*e novo* genome assembly was generated with 80× coverage PacBio reads using Canu (Koren et al. 2017) (version 1.6; *canu minReadLength=1000 correctedErrorRate=0.040 genomeSize=100m -pacbio-raw ∗.pacbio.subreads.fastq.gz*). To increase base quality, the assembly was corrected using PacBio raw reads with Quiver (Chin et al. 2013) and

HiSeq raw reads with Pilon (Walker et al. 2014). First, I converted PacBio raw reads to BAM files using bax2-bam (version 0.0.8; *bax2bam --subread --pulsefeatures=DeletionQV,DeletionTag,InsertionQV,IPD,MergeQV,SubstitutionQV,PulseWidth,SubstitutionTag*), aligned PacBio raw reads to the Canu-only assembly using pbalign (version 0.3.1; default option), merged BAM files using BamTools (version 2.4.1; *bamtools merge*), and polished it using Quiver (version 2.2.1; *variantCaller --algorithm quiver*). Quiver, bax2bam, BamTools, and pbalign were from the Genomic-Consensus package (https://github.com/PacificBiosciences/GenomicConsensus). I repeated this process with the Quiver polished assembly instead of the Canu-only one. Next, to remove bacterial sequence contamination, I aligned the contigs with 3,000 bacterial genomes downloaded from European Nucleotide Archive (ENA) (on March 30, 2018) from ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/embl_genomes/genomes/Bacteria using BLAST + (Camacho et al. 2009) (version 2.7.1; *makeblastdb -input_type fasta -dbtype nucl and blastn -task megablast -evalue 1e-06 -outfmt 6 -perc_identity 50*). Nine contigs were excluded that contain bacterial homology sequences longer than 50% in contig length. Lastly, homopolymers were corrected with mapping CB4856 short reads downloaded from NCBI (accession numbers: SRR3440952, SRR3441150, SRR3441428, and SRR3441550; 73× coverage) (Cook et al. 2017) to 128 contigs using BWA-MEM (Li 2013) (version 0.7.17) and Pilon (version 1.22). The following rounds of Pilon polishing were performed with the same parameters except using the previous round Pilon-polished contigs as a reference. I repeated the polishing using Pilon 4× in total.

A *de novo* genome assembly of the APS7 strain was performed with Canu (version 1.6; *canu minReadLength=2500 correctedErrorRate=0.030 genomeSize=55m -pacbio-raw*) to obtain the most contiguous assembly and polished with Quiver 1× and Pilon 4× in total. For the APS14 strain, the genome was assembled with Canu (version 1.8; *canu -correct genomeSize=67m corOutCoverage=500 corMinCoverage=0 corMhapSensitivity=high minReadLength=4000 -nanopore-raw* and *canu genomeSize=67m correctedErrorRate=0.3 utgGraphDeviation=50 -nanopore-corrected*). Its polishing was conducted by two steps: first, mapping, sorting, and indexing ONT reads to the Canu assembly using minimap2 (Li 2018) and samtools (*minimap2 -ax map-ont -t 20 assembly.fa reads.fq | samtools*

sort -o reads.sorted.bam -T reads.tmp && samtools index reads.sorted.bam), then calling consensus variants and polishing the assembly using Nanopolish (https://github.com/jts/nanopolish) and GNU Parallel (Tange 2011) (Nanopolish version 0.11.2 and GNU Parallel 20161222; *nanopolish_makerange.py assembly.fa | parallel --results nanopolish.results -P 8 nanopolish variants --consensus -o polished.{1}.vcf -w {1} -r reads.fq -b reads.sorted.bam -g assembly.fa -t 4 --min-candidate-frequency 0.1 && nanopolish vcf2fasta -g assembly.fa polished.*.vcf > polished.fa*).

Scaffolding contigs

To determine a subset of CB4856 genome assembly that aligned syntenically onto the N2 genome, I used NUCmer and showtiling from the MUMmer package (Kurtz et al. 2004; Marçais et al. 2018) (version 4.0.0 beta). The final 128 polished contigs were aligned onto the N2 genome (Ensembl WBcel235/ce11) using NUCmer (*nucmer --mum -l 100 -c 300*). The most well-aligned 74 contigs were obtained using show-tiling (*show-tiling -l 1 -g -1 -i 80.0 -v 1.0 -V 0.0*). The right end contigs of Chr I and Chr V had no telomeric repeats, so I manually selected telomere-containing contigs among not placed ones. I judged whether these contigs showed similarity to either end using NUCmer (*nucmer --mum -l 100 -c 300*), then assessed linkage data from recombinant inbred lines between N2 and CB4856. First, reads aligned onto the N2 genome were extracted using Picard (http://broadinstitute.github.io/picard/) (version 2.18.6; *picard SamToFastq*), realigned to the CB4856 genome using BWA-MEM, and sorted using SAMtools (Li et al. 2009) (version 1.6; *samtools sort*). Duplicated reads were removed using *picard MarkDuplicatesWithMateCigar, REMOVE_DUPLICATES=true, read groups were added using picard AddOrReplaceReadGroups*, and indexed using *samtools index*. Variants were called using GATK (Poplin et al. 2017) (version 4.0.5.1; *HaplotypeCaller -ERC GVCF --use-new-qual-calculator,GenomicsDBImport*, and *Genotype GVCFs --founder-id 'CB4856' --use-new-qual-calculator --max-alternate-alleles 2*). I further analyzed whether leftover telomere-containing contigs have linkage with the ends of Chr I and Chr V, then placed the remaining right end contigs for Chr I and Chr V (Supplemental Code). Lastly, the initial version of mitochondrial contig was aligned to the N2 mitochondrial genome using progressiveMauve (Darling et

al. 2010). The CB4856 mitochondrial contig was repeated twice as compared with the N2's, so the ends were trimmed to make a linearized-circular genome. Placed and not-placed contigs were compared for their length, lower-quality nucleotide ratio based on Quiver, and repetitive element ratio using RepeatMasker (Smit et al. 2016) (version open-4.0.7; http://www.repeatmasker.org). Scatterplots were created using an excel template (Weissgerber et al. 2015). All gaps between contigs were filled with 1000 Ns to generate a chromosome-level assembly. Assembly statistics were measured using *nucmer --maxmatch -l 100 -c 300* and *dnadiff*, and the numbers of SNPs were counted using *show-snps -C*. Fosmids were also used to scaffold contigs. I used 15,360 fosmids, removed <500 bp, and mapped them using BWA-MEM. Only fosmids that had both ends mapped were used to check mapping regions, and two contigs were scaffolded if they had at least the same mapped fosmid. Unless otherwise specified, this assembly was used for all following analyses.

Genome quality assessment

BUSCO (Simão et al. 2015) and BWA-MEM were used to verify the completeness of the CB4856 genome. First, the N2 and CB4856 genomes were assessed using BUSCO OrthoDB v9 (*-l eukaryota_odb9 -m geno -sp caenorhabditis*). Next, PacBio raw reads were aligned to the CB4856 genome by using pbalign from the GenomicConsensus package, and its average coverage was calculated by SAMtools depth. Finally, CB4856 HiSeq reads were aligned to two genomes, variants were called using BCFtools (Li 2011) (version 1.6; *bcftools mpileup -Ou -f | bcftools call -vmO z -o* and *bcftools filter -O v -o -s LOWQUAL -i '%QUAL>10'*), and positions with allele frequency of 40%–60% were extracted to visualize them.

Gene annotation transfer and gene prediction

The EMBL-formatted gene annotation (Ensembl 91) was transferred to the CB4856 genome using the Rapid Annotation Transfer Tool (Otto et al. 2011) (RATT; version 24-Dec-2011). I optimized parameters of start.ratt.sh (*Strain, -c 400 -l 20 -g 500, and -o 75*), and reformatted the resulting EMBL file to the GFF format. N2-specific genes were defined as genes whose exons were not transferred at

all using RATT. According to the canonical gene set of WS266 version downloaded from WormBase (c_elegans.PRJNA13758.WS266.canonical_geneset.gtf), the annotations for 45,457 of 46,742 N2 genes (including 1655 genes of total 1891 N2 pseudogenes) were transferred to the CB4856 genome. I also confirmed that 19,355 of the total of 20,039 N2 protein-coding genes were transferred into the CB4856 genome. To further confirm that 684 N2-specific genes (including 661 protein-coding genes) are not found in the CB4856 genome, I searched the sequence of those genes in the CB4856 genome using BLAST+ (*blastn -outfmt 7 -html -perc_identity 95.0 -qcov_hsp_perc 9*5). I identified five genes with copy-number changes only. I repeated this same procedure for the Thompson genome and finally identified 619 genes that are specific to N2.

I then used the MAKER annotation pipeline (Cantarel et al. 2008) (version 2.31.9) to further annotate the CB4856 genome and generated ab initio gene prediction with several tools, including AUGUSTUS (Stanke et al. 2006) (version 3.2.3), SNAP (Korf 2004) (version 2006-07-28), and BUSCO, referred to the pipeline posted on a GitHub website (https://gist.github.com/darencard/bb1001ac1532dd4225b030cf0cd61ce2). Data analyzed in the MAKER pipeline included (1) *de novo* assembled transcripts from CB4856 RNA-seq data with two biological replicates, (2) N2 strain proteome sequences for protein homology evidence (Caenorhabditis_elegans.WBcel235.pep.all.fa; download from the WBcel235 release of WormBase), (3) trained ab initio prediction data set from the SNAP gene prediction tool, and (4) another trained ab initio AUGUSTUS data set optimized by BUSCO. De novo assembled transcripts of CB4856 RNA-seq data were generated using STAR (Dobin et al. 2013) (version 020201; *STAR --readFilesIn --readFilesCommand gzip -cd*) and Trinity (Haas et al. 2013) (version 2.6.6; *Trinity --genome_guided_bam --genome_guided_max_intron 100920*). Before running the first-round MAKER, I masked repeat sequences in the CB4856 genome using RepeatMasker (*RepeatMasker --engine ncbi -lib celrep.Repbase.ref –pa 60*) and Repbase data (Bao et al. 2015) (https://www.girinst.org/repbase/). Complex repeats were isolated and reformatted. Taken together, the gene annotation using MAKER was guided by hints from *de novo* assembled transcript, known protein sequences, and complex repeat and transposable element protein sequences bundled in RepeatMasker (*maker -base*

cb4856_rnd1_RM_trinity_mixed_published round1_maker_opts.Repbase.repeat.trinity.mixed.published.ctl maker_bopts.ctl maker_exe.ctl; st2genome=1, protein2genome=1 in the maker_exe.ctl file). I combined the resulting FASTA files and GFF files using fasta-merge and gffmerge in the MAKER package. I then predicted genes in the CB4856 genome with ab initio gene prediction tools to improve my gene annotation. For training AUGUSTUS, I used nematode-specific BUSCO gene models (nematode_odb9) and the sequence with mRNA annotations based on the initial MAKER result containing 1 kb on each side. At the end, I refined training parameters for AUGUSTUS using BUSCO (*BUSCO.py -i maker.all.maker.transcripts1000.fasta –o rnd1_maker -l nematode_ odb9/ -m genome -c 8 --long -sp worm -z --augustus_parameters="--progress=true"*). For training SNAP, I used maker2zff, fathom, forge, and hmm-assembler in the MAKER package to filter the initial MAKER result (*maker2zff -x0.25 -l 50*) and extracted the annotation and sequences containing 1 kb on each side for the training (*fathom -gene-stats*; *fathom -validate; fathom -categorize 1000; fathom -export 1000 -plus*). Based on this information, I generated training parameters for SNAP (*forge*; *hmm-assembler.pl -params*). Then, the second round of MAKER was run to predict genes with the AUGUSTUS and SNAP training data set (*maker -base cb4856_RM_trinity_mixed_published_rnd2round2_maker_opts.Repbase.repeat.trinity.mixed.published.ctlmaker_bopts.ctl maker_exe.ctl*). Parameters were changed for ab initio gene prediction (*est2genome=0, protein2genome=0*).

After running two rounds of the MAKER ab initio gene prediction pipeline, I filtered out less reliable genes by using the following criteria (Stanley et al. 2018): (1) Discard MAKER gene models that overlap regions that are covered by genes annotated in RATT gene-transfer pipeline; (2) discard genes that encode proteins of shorter than 30 amino acids (as 90 bp); (3) if two or more different MAKER gene models overlap in their coding sequence, discard the model that has the lower eAED score. After these steps, I predicted 781 MAKER gene models and integrated them into the previous gene lists to make the complete set of 46,238 genes. The resulting FASTA files and GFF files were merged using *fasta-merge* and *gffmerge* in the MAKER package as well (Supplemental Material).

Structural variations and GO analysis

I used NUCmer (*nucmer --maxmatch -l 100 -c 500*) to align the final Quiver-Pilon-polished 128 contigs to the N2 genome, or to a CB4856 genome that had been assembled from short reads (Thompson et al. 2015), then called SVs by using the NUCmer output file and Assemblytics (Nattestad and Schatz 2016) (http://assemblytics.com/). Large rearrangements on the chromosome scale were analyzed using progressiveMauve. To assess the effects of genetic variations, I reformatted the Assemblytics result and annotated effects of SVs using SnpEff (Cingolani et al. 2012) (version 4.3t; *java -jar snpEff.jar*). The SnpEff result was summarized based on size and impact categories (modifier, low, moderate, and high) on genes and visualized using Circos version 0.69-6 (Krzywinski et al. 2009) (http://circos.ca/software/download/circos). To evaluate the functional effects of high-impact SVs on genes, I further identified genes which have "lethal" or "sterile" phenotypic evidence reported by RNAi depletion experiment or allelic deletion mutation experiments using the SimpleMine web tool (Lee et al. 2018) (https://www.wormbase.org/tools/mine/simplemine.cgi) and also predicted Gene Ontology (GO) terms for gene functions with the gene set enrichment analysis web tool (Angeles-Albores et al. 2016) (https://www.wormbase.org/tools/enrichment/tea/tea.cgi). N2/CB4856 local recombination data were obtained from https://github.com/AndersenLab/linkagemapping.

Comparison of SVs and the determination of coverage of specific SVs

Each set of SVs was further analyzed (*nucmer --maxmatch -l 100 -c 500*, Assemblytics SV minimum length: 50 bp). First, I extracted the coordinations of SVs on each chromosome from the SV files by using NUCmer and Assemblytics. On each chromosome of the Thompson genome or the Kim genome, I collected the SV region and the additional left side 500 bp (start position −500) and right side 500 bp (end position +500 bp) of the coordinates. The widening of the region was done to prevent mistakes that may occur due to trivial coordination errors. I determined genome-specific SVs and their corresponding genomic positions as a BED file. Finally, from the BAM file that aligned the Canu corrected reads to the genomes using pbalign, I extracted the depth information using mosdepth

(mosdepth 0.2.4; *mosdepth --by v1.novel.snps.v1_coordination.bed cb4856.v1.only.sv.pbalign.depth v1.correctedReads.pbalign.sorted.bam* and *mosdepth --by v2.cb4856_contig_scaffold_novel_sv_v2.v2_coord.bed cb4856.v2.only.sv.pbalign.depth v2.correctedReads.pbalign.sorted.bam*) (Pedersen and Quinlan 2018).

SNP and indel calling by use of GATK

The calling was performed using the FASTQ files downloaded from NCBI (accession numbers: SRR3440952, SRR3441150, SRR344 1428, and SRR3441550) (Cook et al. 2017). The FASTQ files are aligned to the reference genome by BWA-MEM (*bwa mem -M -R*). Aligned SAM files were processed with Picard SortSam and MarkDuplicates to remove PCR duplicates and were converted to BAM files (*picard SortSam SORT_ORDER=coordinate picard MarkDuplicates*). Four BAM files were used for SNP and indel calling with GATK (McKenna et al. 2010) HaplotypeCaller (Poplin et al. 2017) (*GenomeAnalysisTK -T HaplotypeCaller*) against the reference genome. I then distinguished SNPs (*GenomeAnalysisTK SelectVariants -selectType SNP*) and indels by using GATK Select Variants (*GenomeAnalysisTK SelectVariants -selectType INDEL*). I then filtered SNPs and indels using GATK VariantFiltration with a standard filter option (*GenomeAnalysisTK -T VariantFiltration –filterExpression 'QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 4.0'', GenomeAnalysisTK -T VariantFiltration –filterExpression 'QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0 || SOR > 10.0'*). Base calibration was done for each BAM file using the first round SNPs and indels, with GATK BaseRecalibrator (*GenomeAnalysisTK -T Base Recalibrator -BQSR recal_data.table*). Correction was performed with GATK PrintReads (*GenomeAnalysisTK -T PrintReads -BQSR recal_data.table*). Finally, I integrated each BAM file into a single file by using Picard MergeSamFile (*Picard MergeSamFiles*), and the second round of calling for SNPs and indels was done with GATK HaplotypeCaller (*GenomeAnalysisTK -T HaplotypeCaller*). GATK SelectVariants was again used to distinguish SNPs and INDELs (*GenomeAnalysisTK SelectVariants -selectType SNP and GenomeAnalysisTK SelectVariants -selectType INDEL*). SNP results from the second round were processed with corresponding filters, and only the SNPs common to all four short-

read sequencing data (accession numbers: SRR3440952, SRR3441150, SRR3441428, and SRR3441550) were collected with GATK SelectVariants (*GenomeAnalysisTK -T VariantFiltration –filterExpression 'QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 4.0'*) and maxNOCALLnumber (*GenomeAnalysisTK -T SelectVariants -ef –maxNOCALLnumber 0*).

Subtelomere analysis

The subtelomere was defined as the 200-kb end of each chromosome. All subtelomere pairs of N2 and CB4856 strains were aligned usingNUCmer and progressiveMauve, and unaligned regions were obtained. These regions were searched using BLAST+ (*blastn -task megablast -evalue 1e-06 -outfmt 6 -perc_identity 50*) to identify any homology in the N2 genome. To analyze the extreme difference of Chr VR, internal and duplicated sequences were extracted and aligned to each other using *nucmer --maxmatch*, and the alignment was visualized using *mummerplot*. Lastly, short reads of 14 strains were aligned to the CB4856 genome using BWA-MEM, and the positional depth of the last contig was parsed using *samtools depth -a -r*. The short reads were downloaded from NCBI (accession numbers: CB4856: SRR3440952, SRR3441150, SRR3441428, SRR3441 550; CX11262: SRR3441573, SRR3441359; CX11264: SRR3452 248, SRR3452255, SRR3441549; CX11314: SRR3441488, SRR3441191, SRR3440991; CX11315: SRR3441659, SRR3441435, SRR3441151; DL226: SRR3441461, SRR3441168, SRR3440967; DL238: SRR3452231, SRR3452104, SRR3452184; LKC34: SRR3452180, SRR3441481, SRR3441206; MY16: SRR3452112, SRR3441454, SRR3441180; MY23: SRR3452187, SRR3452234, SRR3441433; N2: SRR3441391, SRR3452263, SRR3441113; QX1791: SRR3452145, SRR3452136, SRR3441468; QX1794: SRR3441473, SRR3441189, SRR3440987; QX1793: SRR3452168, SRR3452175, SRR3441470) (Cook et al. 2017). This depth was normalized by the average whole genome depth of each strain.

TALT copy number estimation and phylogenetic analysis

TALT copy number was estimated by calculating the normalized coverage of putative TALT

regions. Normalized coverage was calculated by dividing the depth of coverage within TALT regions by the mean depth of coverage of the nuclear genome. Depth of coverage calculations were performed using VCF-kit (Cook and Andersen 2017) across sequence-alignment files for 150 wild isolates. Variant data for dendrogram comparisons were assembled by constructing a FASTA file with the genome-wide variant positions across all strains and subsetting by regions as described (Cook et al. 2016). MUSCLE (Edgar 2004) (version v3.8.31) was used to construct neighbor-joining trees. The R packages APE (Paradis et al. 2004) (version 3.4) and phyloseq (McMurdie and Holmes 2013) (version 1.12.2) were used for data processing and plotting. Haplotype block analysis was conducted as previously described (Lee et al. 2019).

Chapter II

Part I. *De novo* genome assembly of the CB4856 genome and structural variants compared to the reference strain, N2

Long-read sequencing and de novo assembly of the CB4856 genome

To compare the N2 and CB4856 genomes, I used the Pacific Biosciences (PacBio) RSII platform to construct a nearly complete, chromosome-scale, high-quality genome of CB4856. The genome of CB4856 was assembled with Canu (Koren et al. 2017) using 80× coverage raw reads and was composed of 137 contigs of 104 Mb in total length (Figure 2). Elimination of bacterial contamination, followed by base corrections using PacBio and HiSeq raw reads (Chin et al. 2013; Walker et al. 2014), left an assembled genome of 128 contigs, which were assembled to the level of pseudochromosomes by using fosmid, linkage information, and tiling to the N2 genome (Figure 3; Table 1; Figure 4, Figure 5A,B,E–G). The final assembled genome of CB4856 was 103 Mb in total, 99.4% identical to the N2 genome, and contained 0.2% SNPs between N2 and CB4856 (Tables 1, 2). BUSCO analysis based on gene content information showed that the completeness of the CB4856 genome was comparable to that of the N2 genome (Figure 5C; Simão et al. 2015). In addition, all of the chromosome ends had assembled telomeres longer than 2 kb; this observation suggests that the genome assembly toward the chromosome ends is of high quality (Figure 5D). Most of the genome regions are covered by PacBio raw reads, an average of 60× (Figure 3B). To further evaluate the quality of my genome assembly, I measured the quality of alignment among CB4856 HiSeq reads, a reference genome (N2 genome), a CB4856 genome assembly obtained using short reads (Thompson genome) (Thompson et al. 2015), and a CB4856 genome obtained in this study (Kim genome). I aligned the CB4856 HiSeq reads to the genomes (72.2×, 74.6×, 72.5×, respectively) and tried to call SNPs, indels, and heterozygous variants. The CB4856 HiSeq reads were used for alignment, so I expected to get few SNPs, indels, or heterozygous variants from a well assembled genome of CB4856 and a large number from N2. The number of SNPs and indels found in the Kim genome here was only about 5% of that detected in the Thompson genome (Figure 6). I also found that the numbers of heterozygous variants were 21,432 in the N2 genome, 13,412 in the Thompson genome, and 562 in the Kim genome (Figure

7).

To further analyze the two CB4856 genomes, I aligned them to the N2 genome and determined the numbers of SNPs, indels, and SVs larger than 50 bp. The number of SNPs was similar in the Thompson and Kim genomes, but the Kim genome had substantially more indels and SVs (Table 2). The patterns of hypervariable regions in which SNPs are densely distributed was similar in the Thompson and Kim genomes (Figure 8). Taken together, these results indicate that my (Kim) CB4856 genome was of sufficiently high quality.

Long-read sequencing identified new structural variations

With the newly de novo assembled genome of CB4856, I assessed SVs between the N2 reference genome and my CB4856 genome at fine-scale resolution. SVs longer than 50 nucleotides altered more nucleotides than did SNPs. On the chromosomal scale, a 170-kb sequence block from the N2 Chr V: 1,105,418–1,274,268 was located at the CB4856 Chr II: 4,153,071–4,323,030, and a 90-kb sequence block from N2 Chr IV: 9,413,332–9,503,493 was inverted in CB4856 Chr IV: 9,614,155–9,523,953. Furthermore, the Chr V right arm in CB4856 contained numerous small rearrangements that ranged from 10 to 100 kb in size (Figure 3C; Figure 9). SVs also caused substantial changes in the two genomes (Figure 10): They included 3349 SVs, which together add up to more than 4.95 Mb (Figure 10A).

I then further analyzed the properties of the SVs that I identified using the Kim genome based on long-read sequencing, compared with those from the Thompson genome. The Kim genome detected an additional 1.6 Mb of SVs, including insertions, tandem expansions, and repeat expansions (Figure 11A). The Kim genome also included ~4M bases unaligned to the N2 gnome that were not present in the Thompson genome (Figure 11B). Unaligned bases occurred in 264,580 regions, which were mostly near the ends of chromosomes (Figure 11C–H). The Kim genome included 467 unaligned regions of >1 kb; of these, 293 regions contained repeat sequences. Over 90% of the SVs found in the Thompson genome were also found in the Kim genome (Figure 11I). I found SVs that had not been found in the short-read-based assembly, so the Kim genome is larger than the Thompson genome.

To examine the consequence of the SVs in the context of the genes affected, I inspected the genome-wide gene annotations of the CB4856 genome based on synteny with N2 or RNAseq data (Figure 12). The SVs of 2694 genes in CB4856 generated predicted effects on gene function, including start-codon losses, stop-codon losses, frameshifts, or exon losses (Figure 10B-D; Figure 13). In addition, more than 600 genes are specific to one strain or the other (Figure 12). Half of the completely missing genes were identified in previous CGH data; the other half are identified here for the first time (Figure14; Maydan et al. 2007, 2010). Among the genes that are missing in CB4856, 31 are reported to cause sterile or lethal phenotypes by RNAi, and six of them, including the incompatibility gene zeel-1, showed sterile or lethal phenotypes when deleted in the N2 background (Figure 15; Supplemental Table S2; Seidel et al. 2008, 2011). High impact SVs as defined by SnpEff (Cingolani et al. 2012) and strain-specific genes are more concentrated on autosome arms than centers. These results show that chromosomes are changing more rapidly on the arms than at the center (The *C. elegans* Sequencing Consortium 1998) and show another example of how variants on the chromosome center regions, where recombination frequency is relatively low (Figure 10B), have been eliminated, together with other deleterious mutations, by background selection (Rockman and Kruglyak 2009; Cutter and Choi 2010; Cutter and Payseur 2013). My analysis also implies that substantial genetic changes including gene gain or loss have been tolerated during genetic differentiation within these two strains without decreasing brood size (Andersen et al. 2014; Lee et al. 2017).

Part II. Subtelomere evolution via past ALT events in *C. elegans*

Long-read sequencing revealed the hypervariable nature of subtelomeres

The subtelomeric regions, which I arbitrarily defined as the 200-kb ends of each chromosome, have many regions without alignment. I used high-coverage long-read sequencing to construct contigs of an average size of 700 kb including telomeres on all chromosomes (Figure 5D). The assembled telomere length of each chromosome end is ～40% of mean telomere length (Figure 17). This information allows a direct comparison of the subtelomeres of the CB4856 genome with those of the reference genome. Only 76% of the sequences from the N2 subtelomeric regions and 74% of those from CB4856 were aligned with those of the other strain. These numbers of aligned nucleotides are relatively small when compared with that of the entire genome, which is 95% of the N2 genome and 93% of the CB4856 genome (Figure 16A).

The subtelomere sequences show large insertions, deletions, or inversions at more than half of the chromosome ends (Figure 17); these changes suggest that half of subtelomeric regions have undergone substantial changes. These subtelomeres showed complex structures composed of sequences with homology to preexisting subtelomeres, sequences with partial homology from internal regions, and sequences with no homology at all (Figure 16B–G).

The structure of Chr VR subtelomere is unique, in consequence of past ALT and BIR events

Among the subtelomeres, Chr VR is unique in that new sequences of more than 200 kb are inserted, and these regions are derived from an internal Chr V region with high homology (71% aligned, 91% identity) (Figure 18A; Figure 19A). I analyzed the right end of Chr V in more detail to provide an insight into the possible mechanism of new subtelomere formation in the ancestor of CB4856. I found that the right subtelomere of Chr V of CB4856 contained telomere sequences (Figure 13B; marked as ‘N2 end’ in Figure 13C) 10-fold shorter than the estimated mean telomere, which were followed by 200 kb of extra sequences (Figure 18C). This extra region contains five tandemly

duplicated copies of the TALT sequence (marked as red bars in Figure 18C; Figure 19B,C), flanked by telomeric repeats of lengths ranging from 780 to 1182 nt (marked as blue bars in Figure 18C). The TALT sequence was previously identified and defined as the replication template for ALT in *C. elegans* animals that survived telomere shortening caused by telomerase deficiency (Seo et al. 2015). These TALT copies were followed by sequences that have 91% identity with an internal 200-kb sequence block next to the internal TALT (Figure 18C). The real end of the Chr VR in CB4856 contained at least 3-kb-long telomeric repeats. The features of Chr VR are consistent with the hypothesis that the new subtelomere was formed by telomere attrition followed by two sequential telomere damage repair events using ALT and BIR (see Discussion; Figure 25, below).

New genes in the subtelomeric region

The internal region and the newly duplicated subtelomeric regions shared many, but not all, genes (Figure 20; Figure 21). Sixteen common genes are predicted in both regions, and more than 10 genes are predicted to be specific to each region. The duplicated new subtelomere also contains genes copied from different chromosomes. In addition, the analysis of short-read whole-genome sequence data from 151 wild strains (Cook et al. 2016) revealed that seven of them showed a high copy number of TALT sequences and also contained the same unique sequences of the duplicated 200-kb region seen in the CB4856 subtelomere (Figure 22). To investigate the seven strains in more detail, I identified which exons of the duplicated genes have SNPs. QX1793 and QX1794 had no SNP in exon, DL226, CX11262, CX11264, and CX11315 had one SNP in the C14B4.2 gene, and CX11262, CX11264, and CX11315 had also common SNPs on other five genes (T26H2.3, F21D9.1, F55C9.3, Y43F8A.4, C25F9.8). CX11264 had one SNP in the F55C9.4 gene, and CX11315 had one SNP in T26H2.12 and two more SNPs in C14B4.2, uniquely. Since the common SNPs were all at the same position, they were probably diverged from a same common ancestor. To examine whether the TALT duplication that was observed in the seven strains had arisen independently during evolution, I constructed a phylogenetic tree of the haplotype block (~400 kb) that is closely linked to the chromosome arms that bear the TALT duplication. The seven strains that have high TALT copy numbers shared the same

TALT-linked haplotype block, and these seven strains are grouped alone into a single cluster (Figure 23; Figure 24). Genomic regions that are subject to duplication and changes may act as genetic resources by providing redundant gene sets that can facilitate adaptation to new environments during evolution (Zhang 2003; Leister 2004).

Chapter III

Phenotypic characterization of Korean nematodes and draft genome assembly of two *Auanema* species

Korean nematode collection

To investigate phenotypic diversity of nematodes in more detail, I collected nematodes from rotten fruits in orchards in South Korea. A total of twenty species belonging to eight genera of nematodes were obtained, including *Caenorhabditis*, *Panagrellus*, and *Panagrolaimus*. *Caenorhabditis* was confirmed to be new species based on the difference in rDNA sequence and hybrid incompatability with related species, *C. sinica* (JU1201) and *C. zanzibari* (JU2190). The collected nematodes showed a phenotypic difference in various traits, especially the vulva position (at 50%, 70-80%, or 90-100% length of the body) and the reproductive mode (oviparity or ovoviviparity). One of these species showed a very unique nictation behavior on smooth agar media without any physical support, which was more interesting because it seemed to be independent on mechanical sensing. I am currently working on preparing inbred lines to obtain their draft genomes.

Phenotypic diversification in the genus *Auanema*

To understand genetic and phenotypic variation between different species, I examined the phenotype of an *Auanema* species, *Auanema* sp. APS14, closely related to *Caenorhabditis*. Three species belonging to *Auanema* are currently reported, and they share several phenotypes such as three-gender and arsenic resistance (Shih et al. 2019). However, the *Auanema* species collected from Mono Lake, unlike the other two species (*A. rhodensis* and *A. freburgensis*), does not show tube-nictation and is ovoviviparous rather than oviparous. These results show the phenotypic diversity among the genus *Auanema*.

To determine whether *Auanema* species have other new unknown phenotypes in *Auanema*, I compared the phenotypes of an unreported *Auanema* species, *Auanema* sp. APS14 (APS14 strain), with those of a typical *Auanema* species, *A. freiburgensis* (APS7 strain). As a result, I confirmed that both

strains have three-gender, tube-nictation, and hermaphroditic spot pattern phenotypes in common. Interestingly, APS14 exhibited a previously unreported phenotype, group nictation without any external support (Table 3; Figure 26). While both APS7 and APS14 could perform nictation with their cuticle (tube) of the previous molt, and APS14 has a novel group nictation phenotype where dozens of worms shake their heads together using each other's bodies as a physical support. All of these features are absent in *C. elegans*. *C. elegans* has a two-gender (female and male) and can do nictation or group nictation, but they all require external support such as rough surfaces to shake its head in the air. In order to gain a deeper understanding of these interesting phenotypic differences between the two species and within the genus *Auanema*, I assembled their genomes.

Highly contiguous genome assembly using two long-read sequencing technologies

I used two different long-read sequencing technologies, PacBio RSII and ONT MinION, to obtain draft genomes of APS7 and APS14 strains, respectively, while optimizing the sequencing strategies required for phylogeny-scale genome assembly of nematodes in the future. PacBio RSII produced a total of 14.8 Gb (268×, N50 read length 12.5 kb), and ONT MinION produced 7.8 Gb (113×, N50 read length 3.6 kb). ONT MinION produced much longer reads than PacBio RSII (longest read length: 153.4 kb vs. 55.3 kb), but the read ratio of longer than 5 kb was nearly 10 times lower in the MinION (7% vs. 69%), and its reads are enriched in shorter reads, especially near ~3.5 kb. Since this distribution was different from the known ONT MinION's, a related species *Rhabditella axei* was also sequenced using ONT MinION to confirm that the distribution was reproduced. Sequencing results of *R. axei* showed the elimination of enrichment of a certain read length. Therefore, it is presumed that the library preparation process for *Auanema* sp. APS14 suffered from an unknown problem.

After that, I used similar pipelines to assemble their genomes, and polished them using different methods optimized for PacBio and ONT, respectively. In case of APS14, the reads produced from ONT MinION were not properly assembled when the enriched ~3.5 kb reads were included. Therefore, only reads longer than 4 kb were used. PacBio RSII did not suffer from the read length distribution problem and achieved much better assembly results compared to ONT, where the most

contiguous assembly was obtained with reads longer than 2.5 kb. As a result, draft genomes of APS7 (55 Mb, N50 contig 3.0 Mb) and APS14 (70 Mb, N50 contig 0.6 Mb) were obtained (Table 4). ONT MinION, however, was 2- to 10-fold cheaper than PacBio RSII, so it has a much greater price advantage.

Chapter IV

Discussion

Since the first collection of *C. elegans* (Maupas 1901; Nigon and Felix 2017), 330 isotypes comprising more than 750 strains have been collected from all over the world (Cook et al. 2017). Among them, the reference strain N2, collected in the Bristol area of England, and the CB4856 strain, collected in Hawaii, are the best-known and most extensively studied strains. In this study, I constructed a highly contiguous genome of the CB4856 strain by de novo assembly using long-read sequencing. Because of chromosome-scale selective sweeps in *C. elegans* wild strains, some strains, including CB4856, exhibit distinct polymorphism patterns from most other wild strains (Andersen et al. 2012). For this reason, my completed CB4856 genome will serve as a better reference genome for those wild strains distinct from most other wild strains including N2. In addition, the numerous SVs between N2 and CB4856, identified based on my long-read sequencing, will also help to better understand the effect of SVs on traits by association studies using these strains.

Enrichment of genetic variations in chromosome arms and subtelomeres by background selection and error-prone recombination

Due to background selection, the polymorphism level and the recombination rate are correlated in most species (Kern and Hahn 2018); genetic variations are enriched in chromosome arms, which also show a high recombination rate in many nematodes such as in the genera Pristionchus and Caenorhabditis (Rockman and Kruglyak 2009; Andersen et al. 2012; Rödelsperger et al. 2017; Yin et al. 2018). In particular, repeat sequences are enriched and essential genes are sparsely distributed in chromosome arms (The *C. elegans* Sequencing Consortium 1998; Kamath et al. 2003). Comparison of Pristionchus species has shown that the more conserved, old genes are present in chromosome centers, whereas newly generated orphan genes are preferentially found in chromosome arms (Prabh et al. 2018; Werner et al. 2018). Similar patterns are shown in *C. elegans*. Among the *C. elegans* chromosomes, the largest one, Chr V, contains the fewest essential genes but the highest density of gene families (The *C.*

elegans Sequencing Consortium 1998; Kamath et al. 2003). The right arm of Chr V has the lowest homology gene ratio compared to other closely related species (Stein et al. 2003); it is the region in which mutations accumulate more rapidly than in other chromosome regions, and many deletions are accumulated.

The hypervariable features of the subtelomeric and telomeric regions also contribute to variation enrichment in chromosome arms. Subtelomeres and telomeres are fragile regions that are prone to double-strand breaks (DSBs) during replication, and accurate repair of the DSBs is critical to maintaining genomic integrity (Glover and Stein 1987; Sfeir et al. 2009; Vannier et al. 2012). Most DSBs are repaired by nonhomologous end joining or homologous recombination (Ceccaldi et al. 2016). However, DSBs at subtelomeric and telomeric regions often lead to one-ended DSBs that lose telomeric parts, and thus are repaired by BIR, which finds a homologous sequence instead of missing ends (Bosco and Haber 1998; McEachern and Haber 2006; Kramara et al. 2018). DSBs in telomeres can use remote homologous sequences for repair by executing a searching process (Cho et al. 2014). Repeat sequences are enriched in subtelomeric and telomeric regions, so templates located elsewhere are likely to be used in the homology searching process; their use may increase the variations in the subtelomeric regions. Indeed, each subtelomeric region of CB4856 contains a complex subsequence from a homologous sequence elsewhere in the genome, so they have a new subtelomere that differs from the corresponding one of N2.

New subtelomere formation by ALT and BIR

Among the newly formed subtelomeres, Chr VR shows a unique feature that is reminiscent of telomere damage, ALT, and BIR. My hypothesis for the Chr VR subtelomere formation in the ancestor of CB4856 is that the telomere underwent attrition followed by two sequential telomere-damage repair events, one using ALT and the other using BIR (Fig. 5).

The presence of short telomeric repeats within the subtelomeric region of CB4856 Chr VR implies that telomere attrition and repair had occurred. The multiple copies of TALT sequences next to the telomeric sequences suggests that the repair of telomere attrition was not performed by the canonical telomerase-mediated lengthening mechanism but by an ALT mechanism, even in the presence of the telomerase gene. TALT copies were not the end of the Chr VR: TALT copies were followed by sequences very similar to the region next to the internal TALT, probably by segmental duplication of a 200-kb internal sequence block. The last TALT sequences may have acted as a homology template for BIR in this process. The chromosome ends with a fewTALT copies may have been recognized as a breakage, which in turn could induce the BIR mechanism. Searching for homologous sequences with that of the TALT homology template must have found the internal TALT, resulting in the duplication of sequences next to the internal TALT up to 200 kb via BIR. I postulate that harsh environmental stimuli or stresses, yet to be identified, may have induced Chr VR–specific DSBs in CB4856 ancestors and that these stimuli activated the intrinsic subtelomeric recombination mechanisms by which a new subtelomere was formed by ALT and BIR.We did not fail to notice that telomerase also had an important, though limited, function in the new subtelomere formation. Short traces of telomeric repeats between the tandem TALT copies suggest that telomerase was briefly activated on each end of TALT but was not enough to produce long telomeric repeats. In addition, the duplicated block end was repaired by the action of telomerase, as the real end of the Chr VR contains at least 3-kb-long telomeric repeats.

My analysis of the genomic feature in the CB4856 subtelomere of Chr VR shows that an ALT template can repair telomere attrition even when the telomerase gene is intact. Consistent with this inference, mouse embryonic stem cells or mouse somatic cells may have ALT features when telomerase is present, and ALTand telomerase coexist to perform their unique functions in cells (Zalzman et al. 2010; Neumann et al. 2013). Currently, little is known about the normal function of ALT, and my analysis of the genomic features of CB4856 shows for the first time that ALT activity may be present in the germline to repair abrupt telomere attrition of an individual that already has telomerase activity. My analysis also shows that BIR can induce subtelomere evolution by replicating internal genetic materials. Subtelomeres are enriched with ‘contingency genes,’ which are critical for adaptation to novel

or stressful environments, and the gene families located in subtelomeres tend to expand rapidly (Barry et al. 2003; Brown et al. 2010). By this process, the subtelomere and telomere DSB-induced BIR can operate as a mechanism in the evolutionary process. To summarize, my findings suggest that a species can tolerate substantial structural changes in the genome without losing integrity as the same species and that new subtelomeres, and eventually new chromosomal contents, can evolve by the ALT and BIR mechanisms.

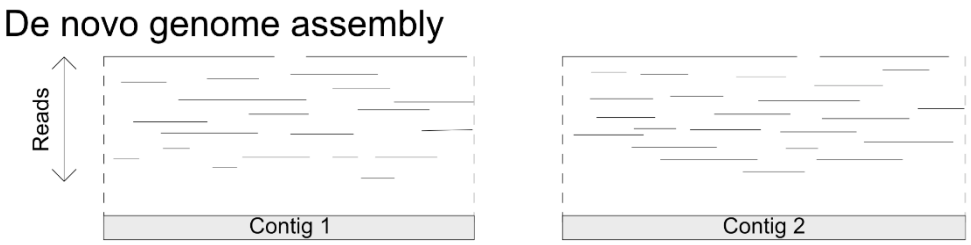
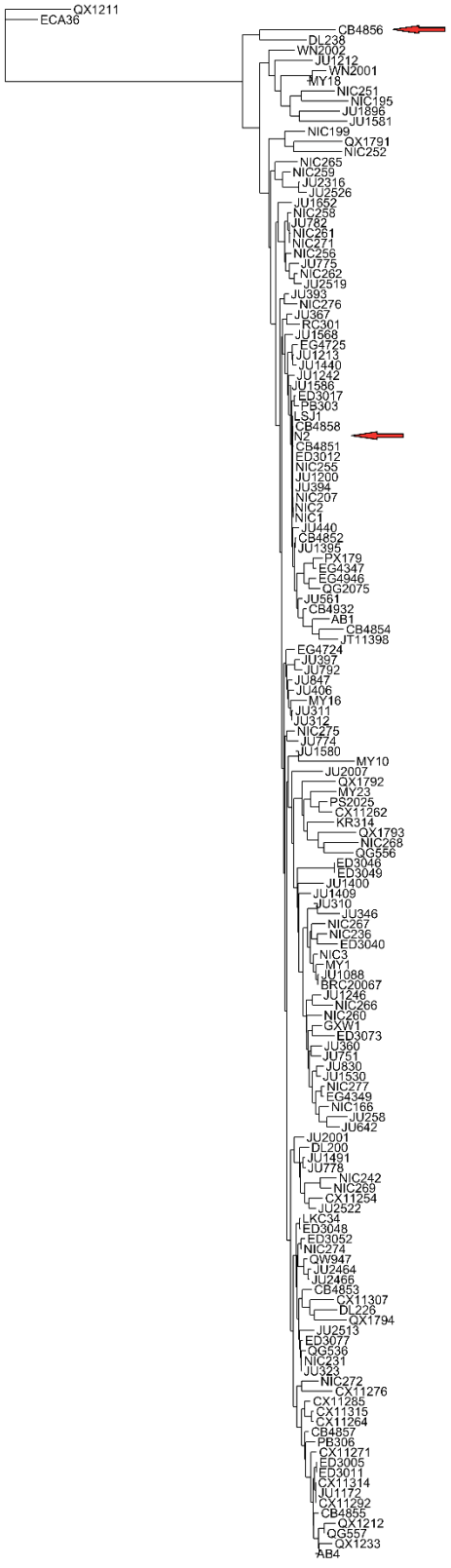
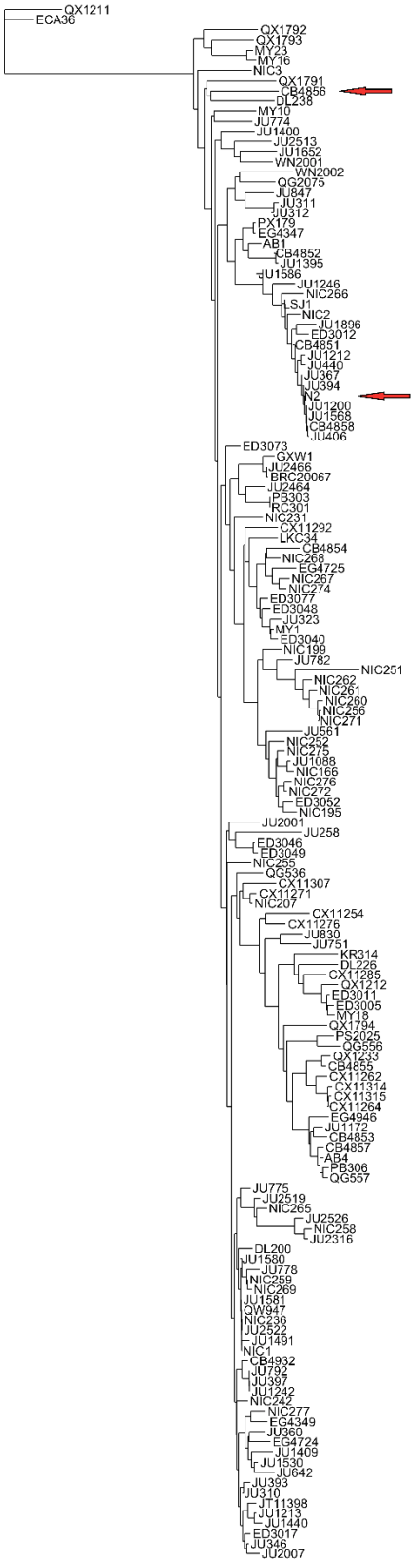


Figure 1. Schematic representation of *de novo* genome assembly. DNA sequencing produces reads (horizontal gray lines), which are DNA fragments read at one time. Finding and assembling these overlapping parts yields a larger chunk called contigs (gray bars), and adding linkage or physical map information can improve the assembly up to chromosome-level assembly.

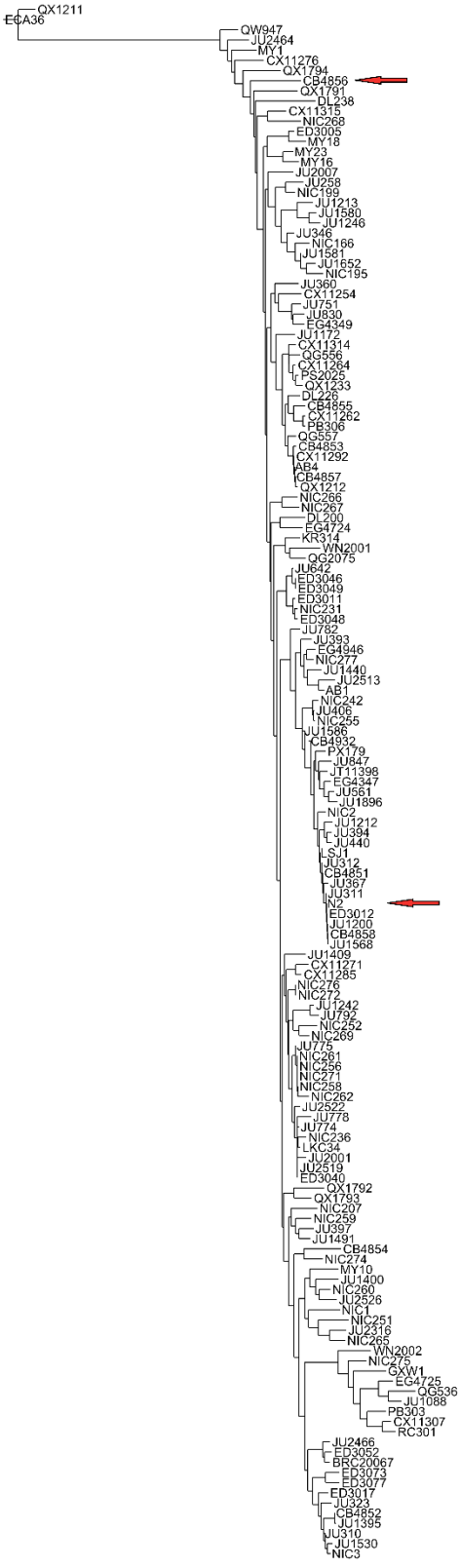
I tree
NA


QX1211
ECA36
CB4856
DL238
WN2002
JU1212
WN2001
MY18
NIC251
NIC195
JU1896
JU1581
NIC199
QX1791
NIC252
NIC265
NIC259
JU2316
JU2526
JU1652
NIC258
JU782
NIC261
NIC271
NIC256
JU775
NIC262
JU2519
JU393
NIC276
JU367
RC301
JU1568
EG4725
JU1213
JU1440
JU1242
JU1586
ED3017
PB303
LSJ1
CB4858
N2
CB4851
ED3012
NIC255
JU1200
JU394
NIC207
NIC2
NIC1
JU440
CB4852
JU1395
PX179
EG4347
EG4946
QG2075
JU561
CB4932
AB1
CB4854
JT11398
EG4724
JU387
JU792
JU847
JU406
MY16
JU311
JU312
NIC275
JU774
JU1580
MY10
JU2007
QX1792
MY23
PS2025
CX11262
KR314
QX1793
NIC268
QG556
ED3046
ED3049
JU1400
JU1409
JU310
JU346
NIC267
NIC236
ED3040
NIC3
MY1
JU1088
BRC20067
JU1246
NIC266
NIC260
GXW1
ED3073
JU360
JU751
JU830
JU1530
NIC277
EG4349
NIC166
JU258
JU642
JU2001
DL200
JU1491
JU778
NIC242
NIC269
CX11254
JU2522
LKC34
ED3048
ED3052
NIC274
QW947
JU2464
JU2466
CB4853
CX11307
DL226
QX1794
JU2513
ED3077
QG536
NIC231
JU323
NIC272
CX11276
CX11285
CX11315
CX11264
CB4857
PB306
CX11271
ED3005
ED3011
CX11314
JU1172
CX11292
CB4855
QX1212
QG557
QX1233
AB4

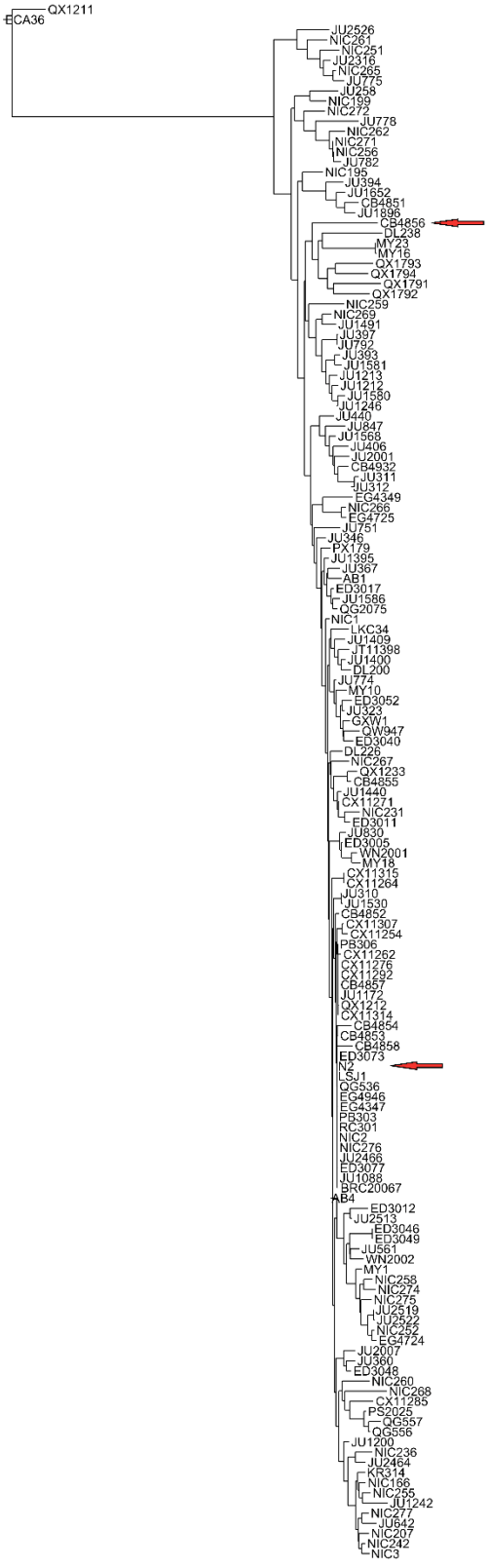
II tree
NA


QX1211
ECA36
QX1792
QX1793
MY23
MY16
NIC3
QX1791
CB4856
DL238
MY10
JU774
JU1400
JU2513
JU1652
WN2001
WN2002
QG2075
JU847
JU311
JU312
PX179
EG4347
AB1
CB4852
JU1395
JU1586
JU1246
NIC266
LSJ1
NIC2
JU1896
ED3012
CB4851
JU1212
JU440
JU367
JU394
N2
JU1200
JU1568
CB4858
JU406
ED3073
GXW1
JU2466
BRC20067
JU2464
PB303
PB301
NIC231
CX11292
LKC34
CB4854
NIC268
EG4725
NIC267
NIC274
ED3077
ED3048
JU323
MY1
ED3040
NIC199
JU782
NIC251
NIC262
NIC261
NIC260
NIC256
NIC271
JU561
NIC252
NIC275
JU1088
NIC166
NIC276
NIC272
ED3052
NIC195
JU2001
JU258
ED3046
ED3049
NIC255
QG536
CX11307
CX11271
NIC207
CX11254
CX11276
JU830
JU751
KR314
DL226
CX11285
QX1212
ED3011
ED3005
MY18
QX1794
PS2025
QG556
QX1233
CB4855
CX11262
CX11314
CX11315
CX11264
EG4946
JU1172
CB4853
CB4857
AB4
PB306
QG557
JU775
JU2519
NIC265
JU2526
NIC258
JU2316
DL200
JU1580
JU778
NIC259
NIC269
JU1581
QW947
NIC236
JU2522
JU1491
NIC1
CB4932
JU792
JU397
JU1242
NIC242
NIC277
EG4349
JU360
EG4724
JU1409
JU1530
JU642
JU393
JU310
JT11398
JU1213
JU1440
ED3017
JU346
JU2007

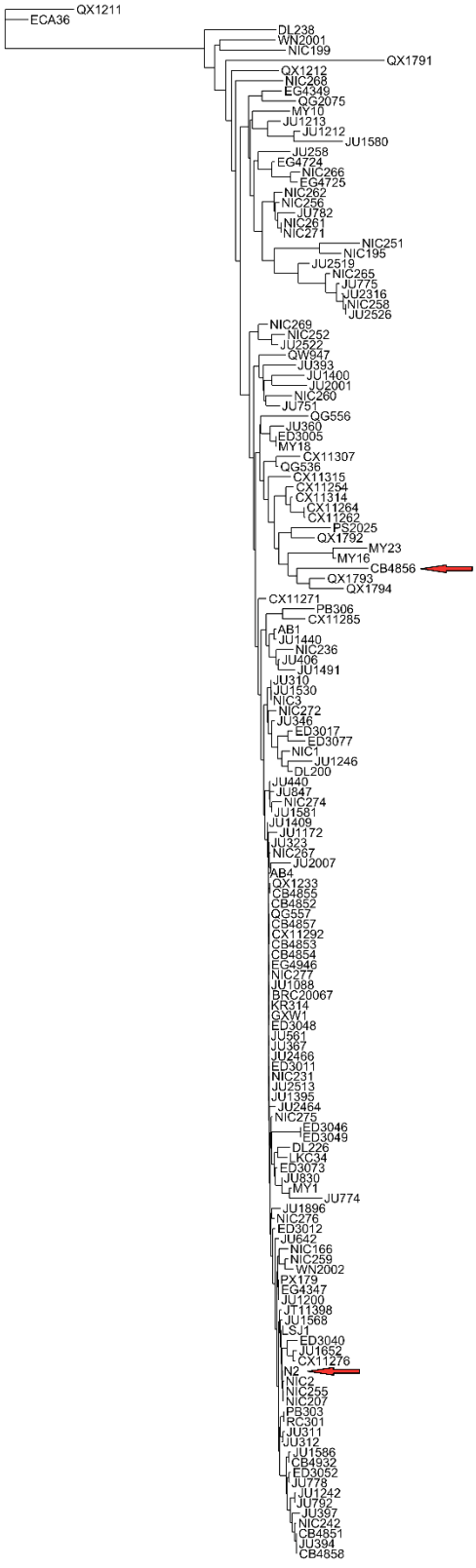
III tree
NA


QX1211
ECA36
QW947
JU2464
MY1
CX11276
QX1794
CB4856
QX1791
DL238
CX11315
NIC268
ED3005
MY18
MY23
MY16
JU2007
JU258
NIC199
JU1213
JU1580
JU1246
JU346
NIC166
JU1581
JU1652
NIC195
JU360
CX11254
JU751
JU830
EG4349
JU1172
CX11514
QG556
CX11264
PS2025
QX1233
DL226
CB4855
CX11262
PB306
QG557
CB4853
CX11292
AB4
CB4857
QX1212
NIC266
NIC267
DL200
EG4724
KR314
WN2001
QG2075
JU642
ED3046
ED3049
ED3011
NIC231
ED3048
JU782
JU393
EG4946
NIC277
JU1440
JU2513
AB1
NIC242
JU406
NIC255
JU1586
CB4932
PX179
JU847
JT11398
EG4347
JU561
JU1896
NIC2
JU1212
JU394
JU440
LSJ1
JU312
CB4851
JU367
JU311
N2
ED3012
JU1200
CB4858
JU1568
JU1409
CX11271
CX11285
NIC276
NIC272
JU1242
JU792
NIC252
NIC269
JU775
NIC261
NIC256
NIC271
NIC258
NIC262
JU2522
JU778
JU774
NIC236
LKC34
JU2001
JU2519
ED3040
QX1792
QX1793
NIC207
NIC259
JU397
JU1491
CB4854
NIC274
MY10
JU1400
NIC260
JU2526
NIC1
NIC251
JU2316
NIC265
WN2002
NIC275
GXW1
EG4725
QG536
JU1088
PB303
CX11307
RC301
JU2466
ED3052
BRC20067
ED3073
ED3077
ED3017
JU323
CB4852
JU1395
JU310
JU1530
NIC3

IV tree
NA


QX1211
ECA36
JU2526
NIC261
NIC251
JU2316
NIC265
JU775
JU258
NIC199
NIC272
JU778
NIC262
NIC271
NIC256
JU782
NIC195
JU394
JU1652
CB4851
JU1896
CB4856
DL238
MY23
MY16
QX1793
QX1794
QX1791
QX1792
NIC259
NIC269
JU1491
JU397
JU792
JU393
JU1581
JU1213
JU1212
JU1580
JU1246
JU440
JU1847
JU1568
JU406
JU2001
CB4932
JU311
JU312
EG4349
NIC266
EG4725
JU751
JU346
PX179
JU1395
JU367
AB1
ED3017
JU1586
QG2075
NIC1
LKC34
JU1409
JT11398
JU1400
DL200
JU774
MY10
ED3052
JU323
GXW1
QW947
ED3040
DL226
NIC267
QX1233
CB4855
JU1440
CX11271
NIC231
ED3011
JU830
ED3005
WN2001
MY18
CX11315
CX11264
JU310
JU1530
CB4852
CX11307
CX11254
PB306
CX11262
CX11276
CX11292
CB4857
JU1172
QX1212
CX11314
CB4854
CB4853
CB4858
ED3073
N2
LSJ1
QG536
EG4946
EG4347
PB303
RC301
NIC2
NIC276
JU2466
ED3077
JU1088
BRC20067
AB4
ED3012
JU2513
ED3046
ED3049
JU561
WN2002
MY1
NIC258
NIC274
NIC275
JU2519
JU2522
NIC252
EG4724
JU2007
JU360
ED3048
NIC260
NIC268
CX11285
PS2025
QG557
QG556
JU1200
NIC236
JU2464
KR314
NIC166
NIC255
JU1242
NIC277
JU642
NIC207
NIC242
NIC3

V tree
NA


QX1211
ECA36
DL238
WN2001
NIC199
QX1791
QX1212
NIC268
EG4349
QG2075
MY18
JU1213
JU1212
JU1580
JU258
EG4724
NIC266
EG4725
NIC262
NIC256
JU782
NIC261
NIC271
NIC251
NIC195
JU2519
NIC265
JU775
JU2316
NIC258
JU2526
NIC269
NIC252
JU2522
QW947
JU393
JU1400
JU2001
NIC260
JU751
QG556
JU360
ED3005
MY18
CX11307
QG538
CX11315
CX11254
CX11314
CX11264
CX11262
PS2025
QX1792
MY23
MY16
CB4856
QX1793
QX1794
CX11271
PB306
CX11285
AB1
JU1440
NIC236
JU406
JU1491
JU310
JU1530
NIC3
NIC272
JU346
ED3017
ED3077
NIC1
JU1246
DL200
JU440
JU847
NIC274
JU1581
JU1409
JU1172
JU323
NIC267
JU2007
AB4
QX1233
CB4855
CB4852
QG557
CB4857
CX11292
CB4853
CB4854
EG4946
NIC277
JU1088
BRC20067
KR314
GXW1
ED3048
JU561
JU367
JU2466
ED3011
NIC231
JU2513
JU1395
JU2464
NIC275
ED3046
ED3049
DL226
LKC34
ED3073
JU830
MY1
JU774
JU1896
NIC276
ED3012
JU642
NIC166
NIC259
WN2002
PX179
EG4347
JU1200
JT11398
JU1568
LSJ1
ED3040
JU1652
CX11276
N2
NIC2
NIC255
NIC207
PB303
BRC301
JU311
JU312
JU1586
CB4932
ED3052
JU778
JU1242
JU792
JU397
NIC242
CB4851
JU394
CB4858

X tree
NA


QX1211
ECA36
NIC262
NIC261
NIC260
NIC265
JU2526
JU2519
NIC258
JU775
JU2316
WN2001
NIC251
NIC199
CX11264
QX1793
MY23
QX1794
QX1791
CB4856
DL238
JU1581
JU1580
JU1246
JU346
JU2001
JU397
JU1242
JU792
JU847
DL200
JU323
QX1233
PB306
CB4855
DL226
CX11285
PS2025
QG557
QG556
WN2002
NIC195
JU1409
CX11276
JU1491
ED3011
CX11307
CX11254
JU2464
MY10
JU360
EG4349
QX1792
ED3017
CB4857
CB4853
NIC1
JU1586
JU440
NIC166
NIC231
ED3073
NIC255
JU1088
JU2466
NIC2
NIC207
ED3048
BRC20067
PX179
ED3077
JU751
JU2513
JU1400
ED3046
ED3049
JU1652
KR314
EG4347
JU1200
CX11315
GXW1
JU1896
PB303
RC301
ED3012
QW947
AB1
JU830
JT11398
JU561
CB4852
JU1395
JU1440
CB4854
JU774
ED3040
CX11271
ED3005
CX11262
AB4
JU2007
QG2075
CX11292
EG4946
QX1212
JU1172
CX11314
QG536
EG4725
ED3052
NIC275
NIC236
JU2522
JU782
NIC256
NIC271
JU778
NIC259
LKC34
EG4724
MY1
NIC268
NIC252
NIC267
NIC266
JU258
NIC272
NIC274
NIC269
NIC276
JU642
JU1213
NIC277
JU310
JU1530
NIC3
CB4932
JU393
JU367
MY18
MY16
JU311
JU312
NIC242
JU1568
LSJ1
CB4858
N2
CB4851
JU394
JU1212
JU406

Figure 2. Phylogenetic tree of *C. elegans* wild isolates. Each chromosomes were used to analyze the relationship among wild isolates.

A

	Sub-read	Sub-read (> 20 kb)
Number of reads	968,475	32,180
Number of bases (bp)	8,289,680,042	764,689,469
Coverage	79.71	7.35
Maximum read length (bp)	51,996	
N50 (bp)	11,828	

B

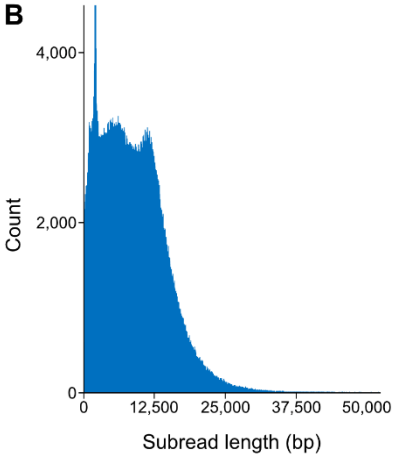


Figure 3. Stats of PacBio raw reads. (A) PacBio raw sub-reads statistics. (B) PacBio raw reads length distribution.

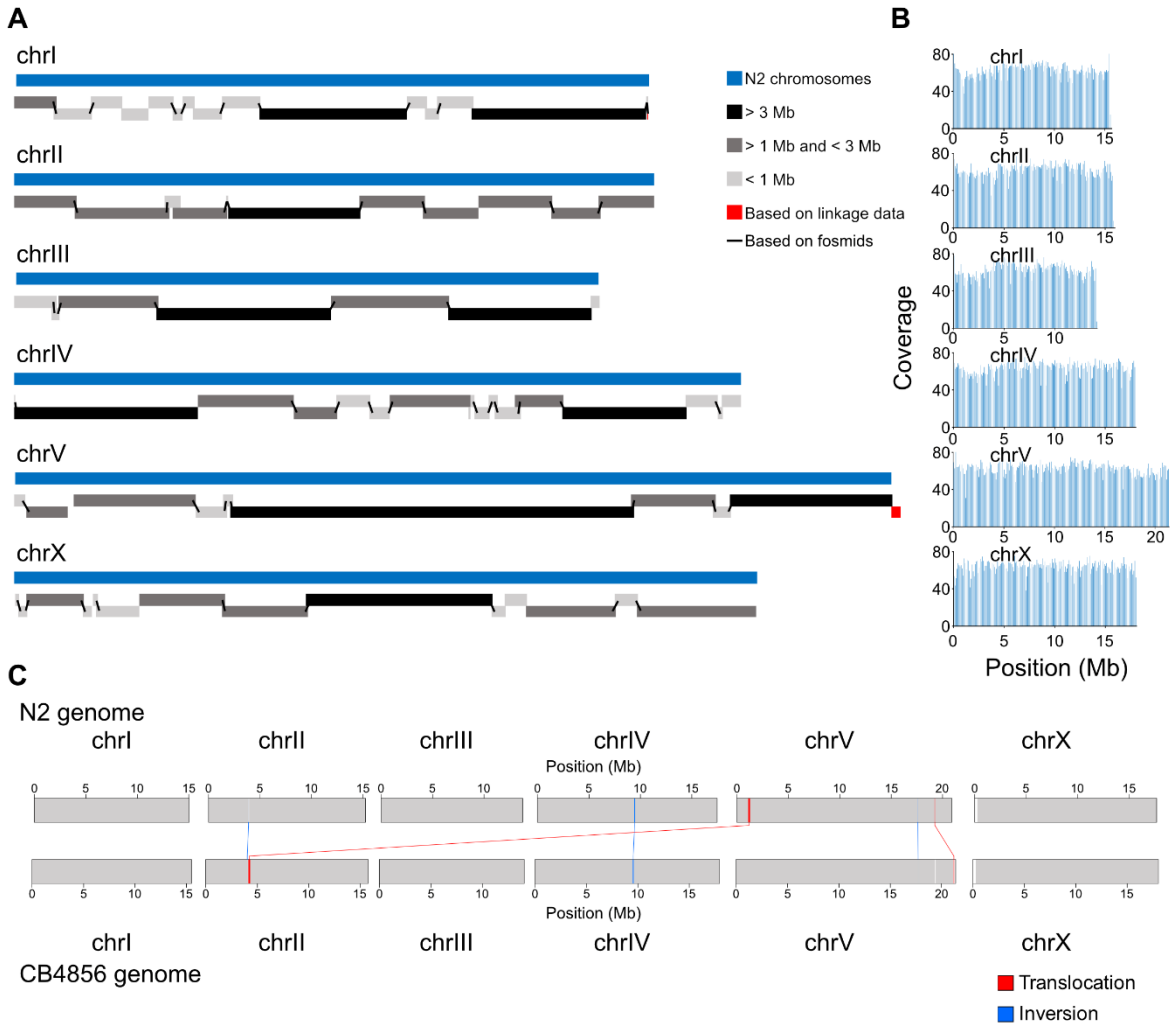


Figure 4. CB4856 genome assembly and comparison with the N2 genome at a chromosome level. (A) Schematic representation of CB4856 contig lengths mapped to N2 WBcel235 chromosomes. (B) PacBio raw read coverage, mapped on CB4856 chromosomes (100-kb binned). Reads were distributed at average 60× coverage. (C) Schematic of large chromosomal rearrangement between N2 and CB4856 genomes identified using progressiveMauve. The blue box and line indicate inversion; the red box and line, translocation; and the white box indicates the unaligned block. Chr VR has several small rearrangements and unaligned blocks. Chr II: 3,896,126–3,900,949 in N2 was inverted in CB4856 (Chr II: 4,045,653–4,040,823), Chr V: 17,616,880–17,623,484 in N2 was inverted in CB4856 (Chr V: 17,734,209–17,728,873), and Chr V: 19,258,912–19,289,935 in N2 was located at Chr V: 21,193,104–21,237,336 in CB4856.

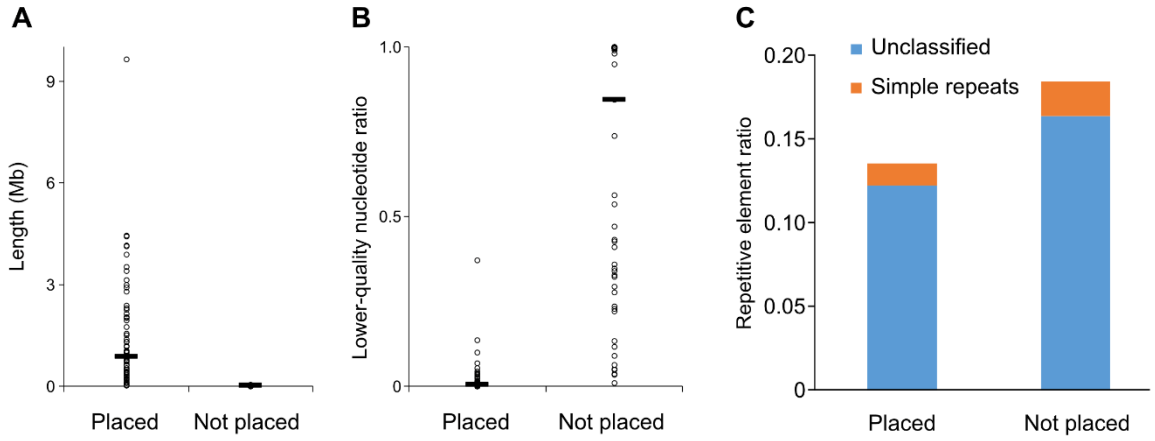
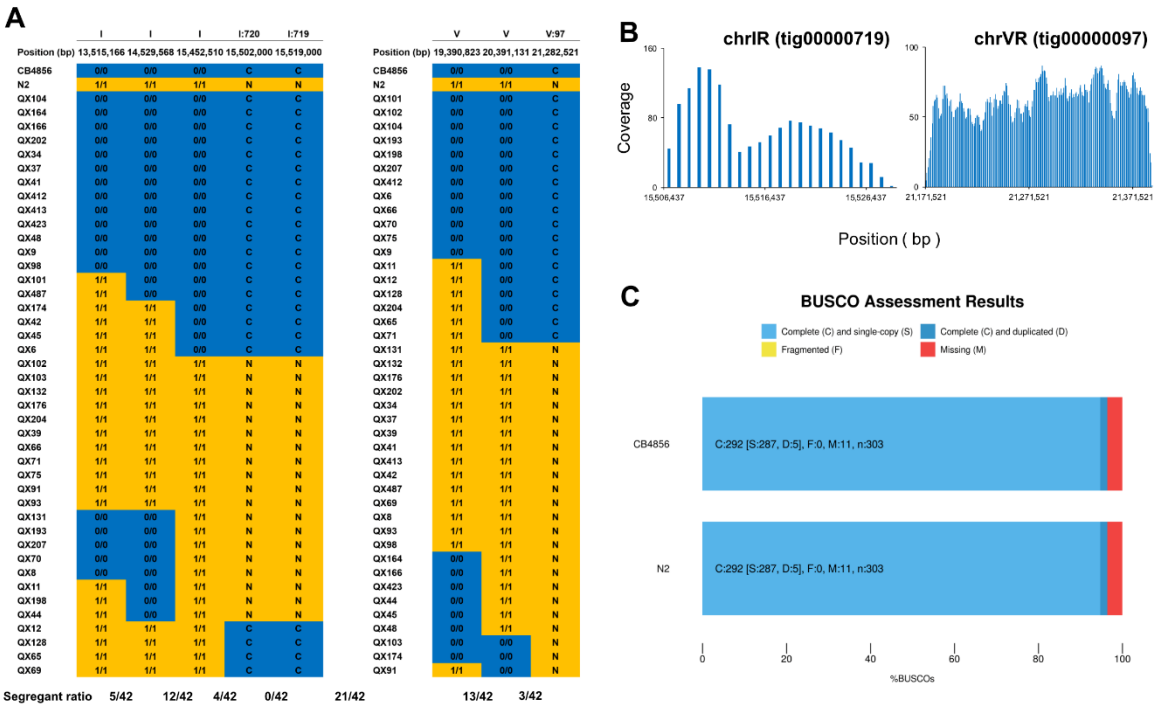


Figure 5. Stats of PacBio not placed contigs. (A) length, (B) lower-quality nucleotides ratio, and (C) repetitive element types of placed and not placed contigs on the N2 genome. Black horizontal bars represent median values.



D

	Contig name	Contig size (bp)	Start of telomeric array	End of telomeric array	Telomeric repeat sequence	Telomere length (bp)	Position of telomere on contig	Number of canonical telomeric repeats
IL	tig00000193	1,021,041	1,018,151	1,021,041	TTAGGC	2,886	End	481
IR	tig00000719	22,460	1	3,348	GCCTAA	3,348	Start	508
IIL	tig00000150	1,498,322	1	3,979	GCCTAA	3,979	Start	587
IIR	tig00000156	1,307,113	1,303,937	1,307,113	TTAGGC	3,177	End	473
IIIL	tig00000201	999,175	1	3,121	GCCTAA	3,121	Start	480
IIIR	tig00000291	205,719	1	2,524	GCCTAA	2,524	Start	390
IVL	tig00000324	22,756	19,410	22,755	TTAGGC	3,347	End	532
IVR	tig00000212	451,830	1	2,772	GCCTAA	2,772	Start	399
VL	tig00000764	282,952	280,057	282,952	TTAGGC	2,896	End	401
VR	tig00000097	218,346	216,009	218,346	TTAGGC	2,338	End	344
XL	tig00000307	97,378	1	3,259	GCCTAA	3,259	Start	516
XR	tig00000073	2,855,796	2,853,115	2,855,796	TTAGGC	2,682	End	386

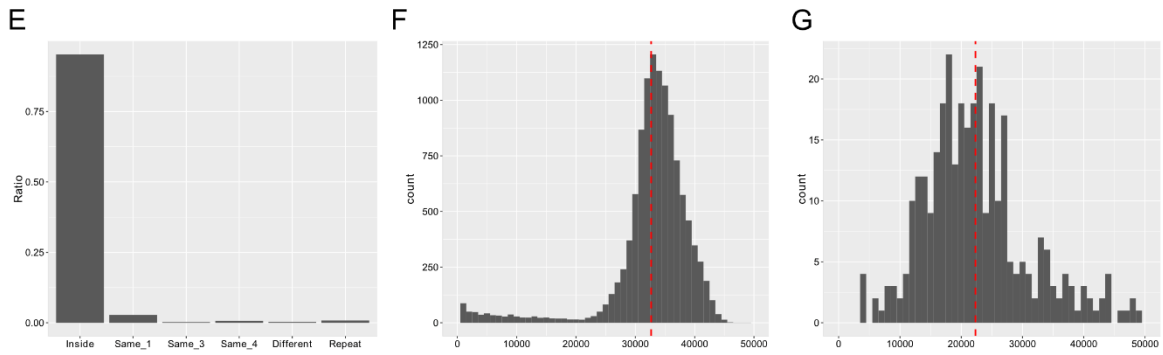
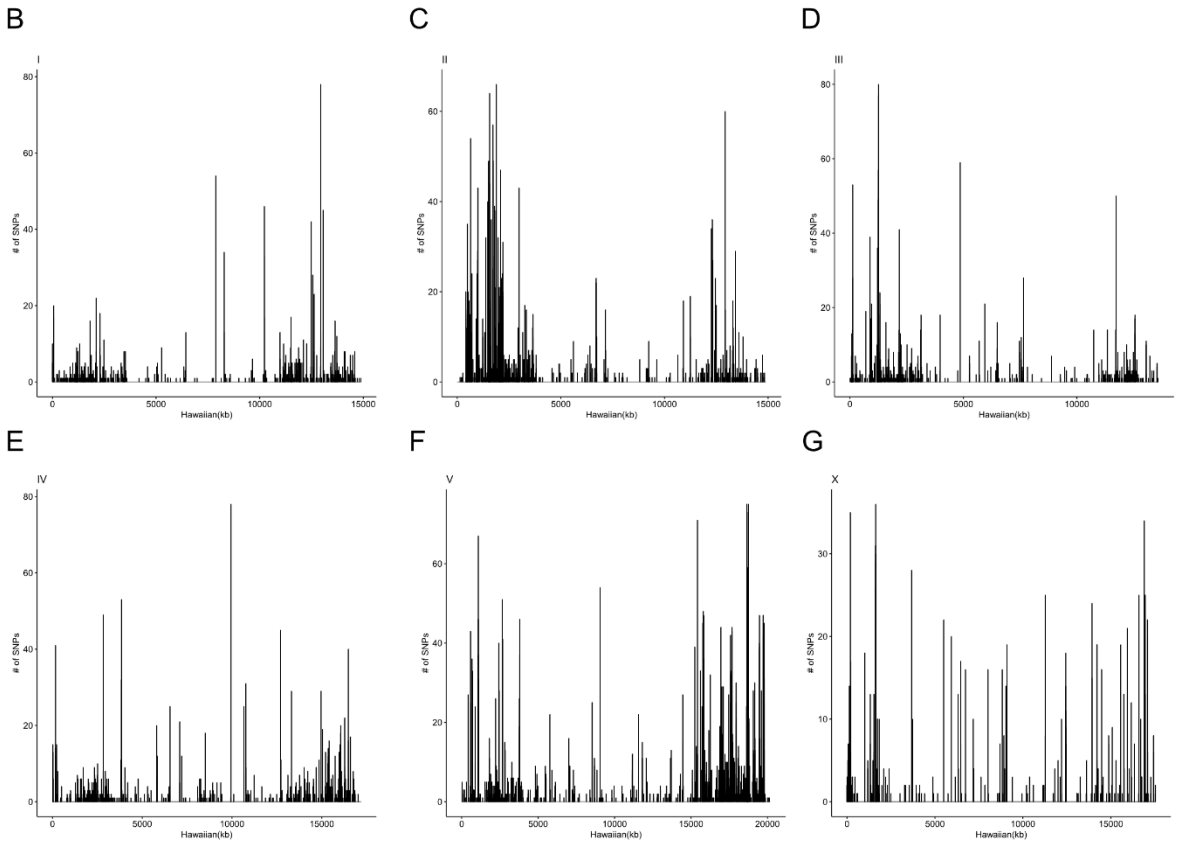


Figure 6. Linkage map and genome quality. (A) Linkage map of Chr IR and Chr VR ends. Blue: CB4856 alleles; yellow: N2 alleles. 0/0: homozygous for CB4856 alleles; 1/1 homozygous for N2 alleles as a result of variant calling with GATK. C, homozygous for CB4856; or N, homozygous for N2 alleles, based on whether each strain contains CB4856-specific sequences (N2 has zero coverage in this region). Segregant ratio represents the number of strains that have different allele types between positions. For example, 21/42 in the middle of the graph indicates that 21 strains have N2 and CB4856 allele types (or vice versa) in Chr I and Chr V, respectively, but not N2 and N2 or CB4856 and CB4856 alleles. tig00000719 and tig00000097 were not placed on the N2 genome, but they are highly linked with chrIR and chrVR ends, respectively. (B) Coverage plot of both contigs. (C) BUSCO assessment of CB4856 and N2 genomes. (D) Telomere lengths and telomere-containing contigs. (E) Fosmid alignment to the Kim et al., 2019 genome. Inside, fosmids aligned within a single contig ; Same_1, fosmid aligned to two adjacent contigs; Same_3, the distance between two ends of the fosmid was longer than 3 contigs; Same_4, longer than 4 contigs; different, one fosmid aligned to two different contigs; repeat, fosmid sequence from repeat sequences. (F) length distribution of 'inside' fosmids, which are contained within a single contig. (G) length distribution of fosmids that are aligned to two adjacent contigs.

A

	Thompson et al., 2015	Kim et al., 2019
SNP	19827	987
Indel	8336	352



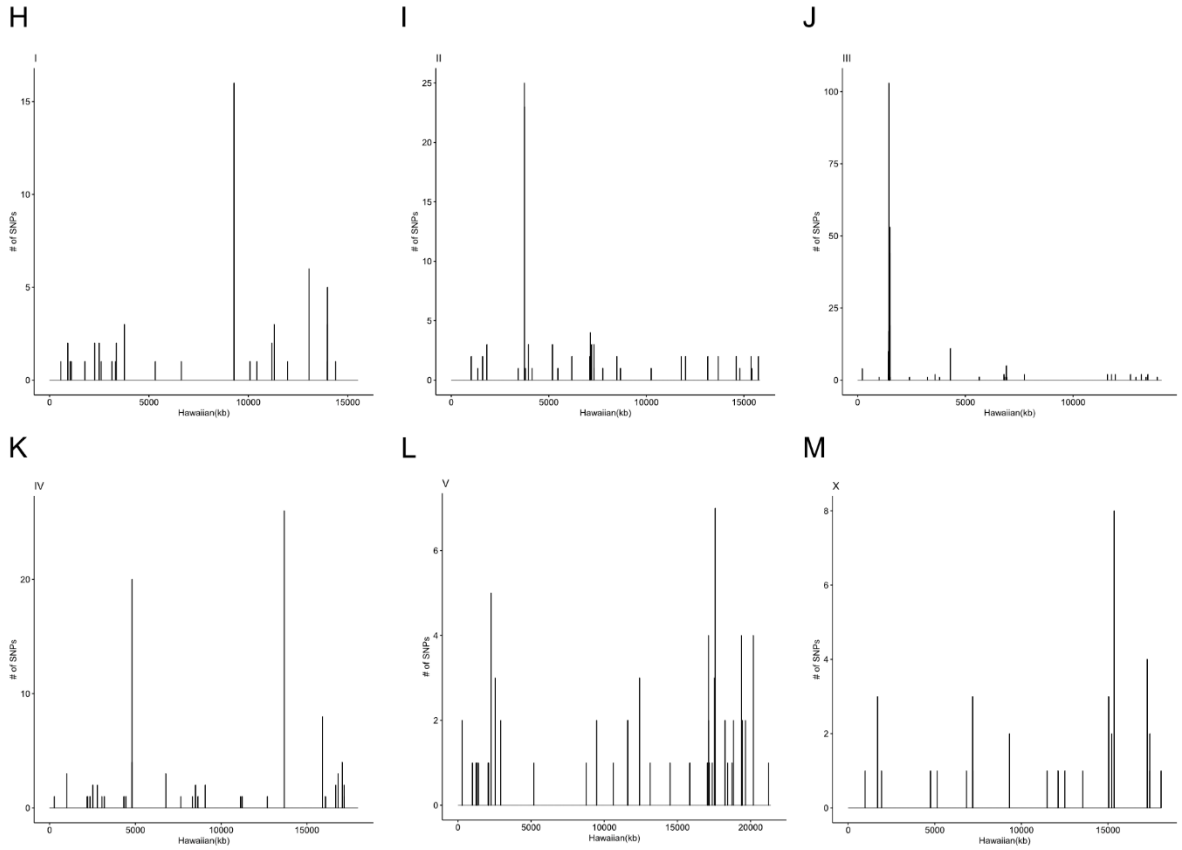


Figure 7. Variant call by CB4856 HiSeq short read at both genomes. (A) Number of SNP and Indel in both genomes. (B-M) The chromosomal distributions of short read sequencing align against the Thompson et al., 2015 genome (B-G), or the Kim et al., 2019 genome (H-M).

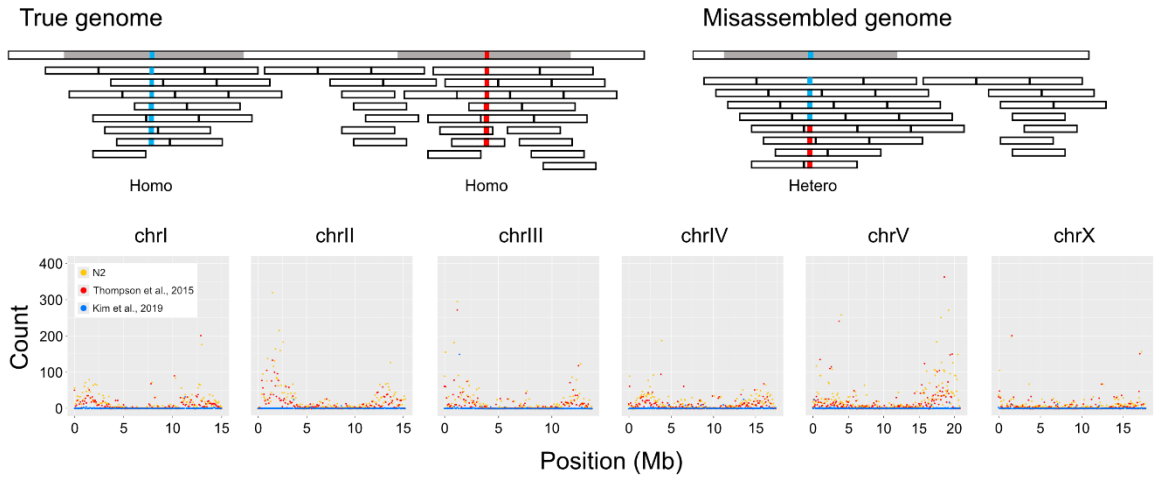
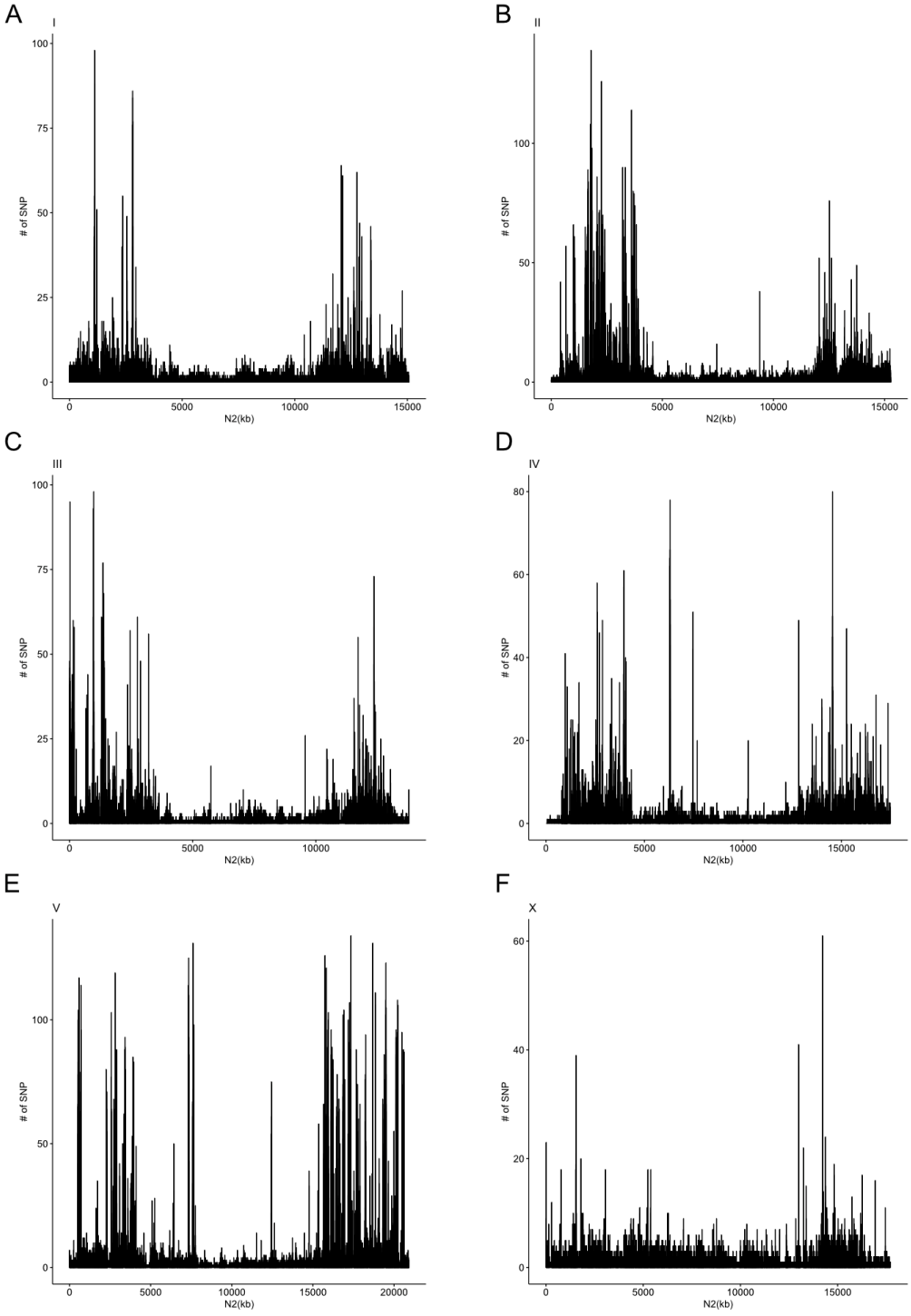


Figure 8. Schematic representation of CB4856 HiSeq reads mapped on the CB4856 genome (blue) or the N2 genome (yellow). Each dot shows the heterozygous base count (100-kb interval) from Chr I to Chr X.


A
I
B
II
C
III
D
IV
E
V
F
X
# of SNP
N2(kb)

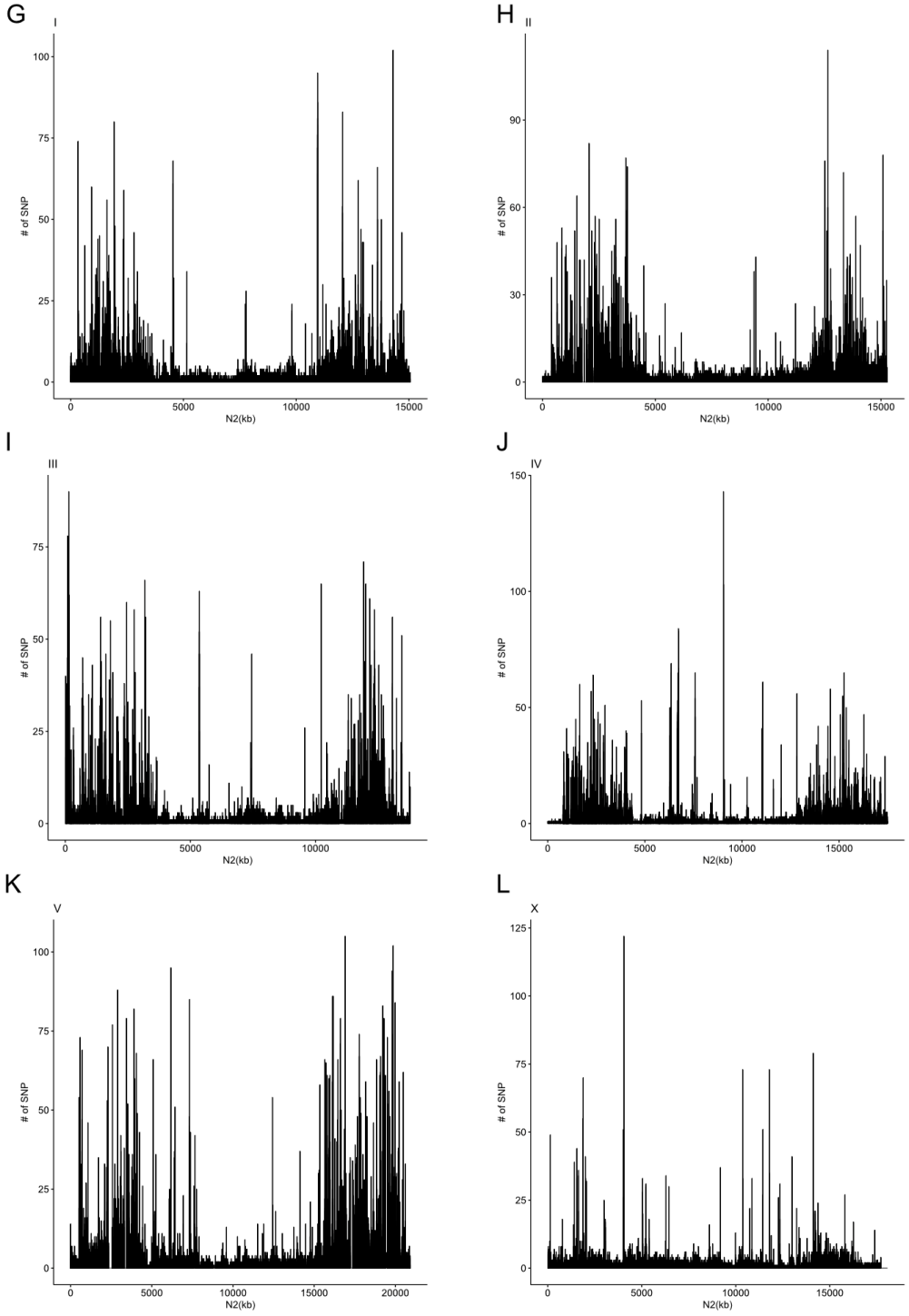


Figure 9. Density of SNP variant sites across chromosomes. Density was calculated in 9 kb windows moving in 1 kb steps. Density of SNP at the Thompson et al., 2015 genome (A-F), Density of SNP at the Kim et al., 2019 genome (G-L).

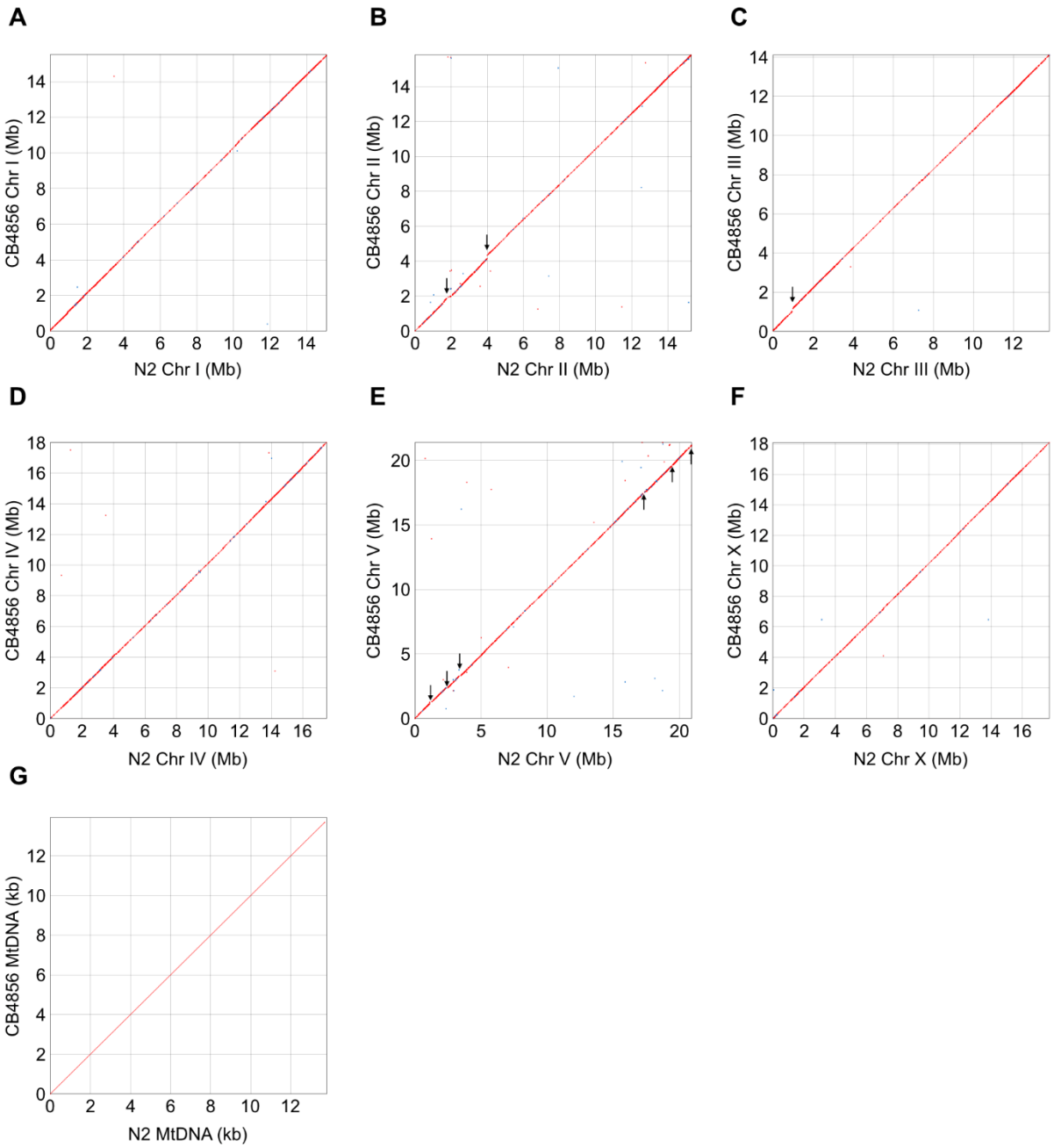


Figure 10. Alignment and structural variations between N2 and CB4856 chromosomes. (A-G) Dot plots showing alignment between Chr Is (A), Chr IIs (B), Chr IIIs (C), Chr IVs (D), Chr Vs (E), Chr Xs (F), and mitochondrial genomes (G) of N2 and CB4856 strains. Red: forward strand matches; Blue: reverse strand matches; Arrows: extreme break points.

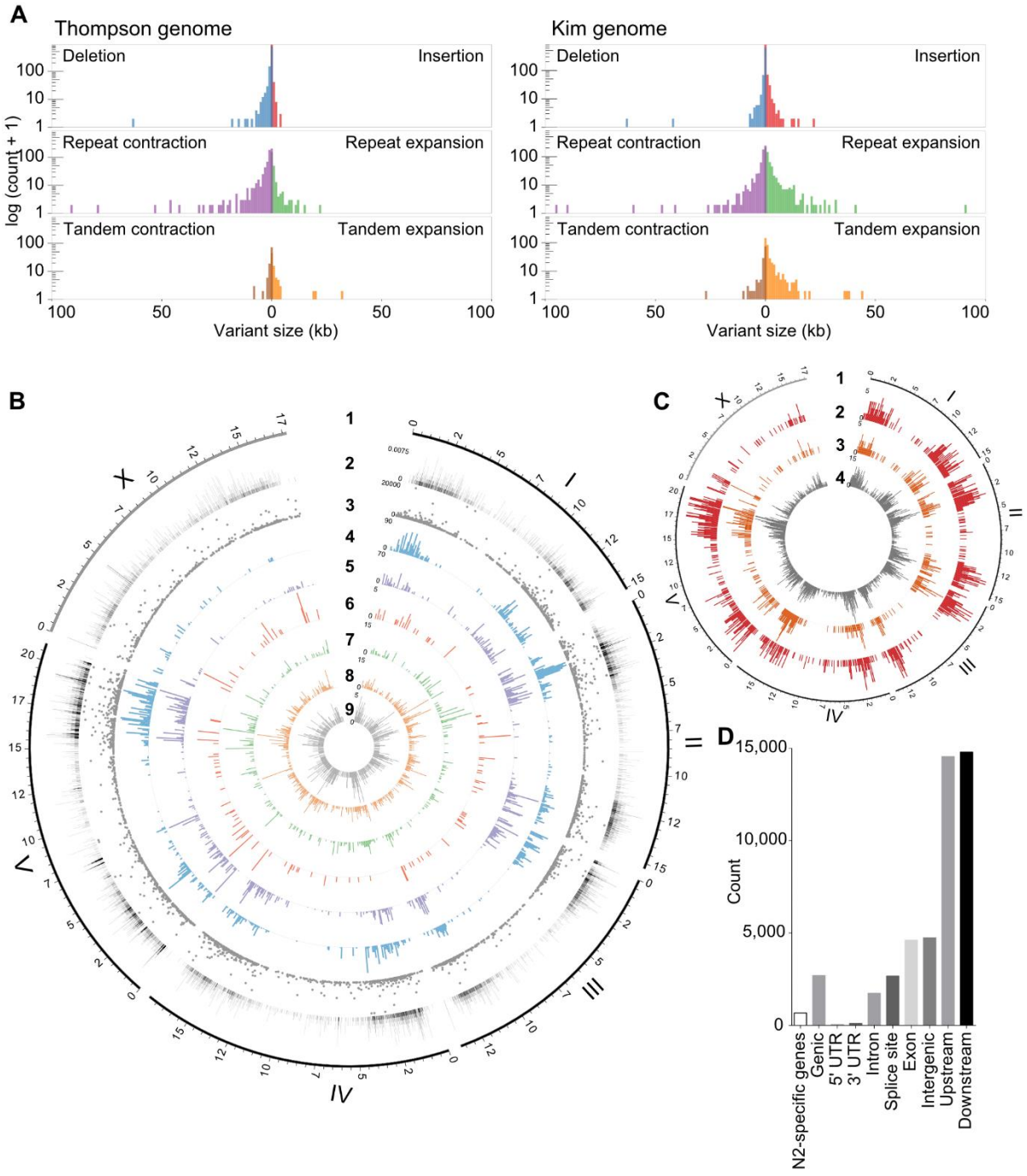
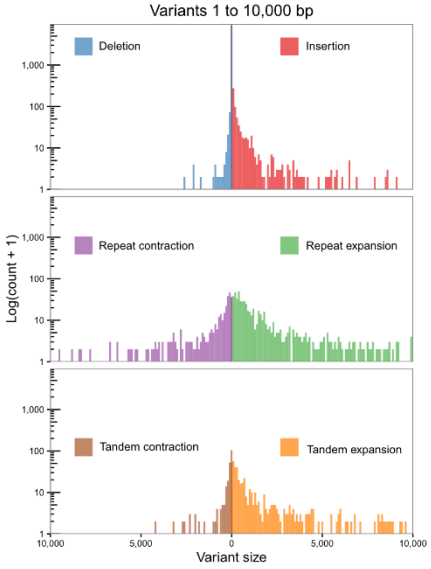


Figure 11. Structural variations (SVs) between the CB4856 and N2 genomes and their effects on chromosomal contents. (A) SVs between the N2 genome and the short-read-based CB4856 genome, previously reported (left), and between the N2 genome and the long read-based CB4856 genome (right). Repeat expansion, tandem expansion, and insertion SVs are more often detected when using long read-based genome than when using the previous short read-based genome. (B) Tracks representing density at 100-kb intervals; from outside to inside: 1, genomic positions (in Mb) of the six chromosomes based on the N2 genome; 2, density of local recombination rate in CB4856/N2 introgression lines; 3–9, types of SVs identified using Assemblytics: 3, size of SVs; 4, density of repeat-contraction SVs; 5, density of repeat-expansion SVs; 6, density of tandem-contraction SVs; 7, density of tandem-expansion SVs; 8, density of deletion SVs; 9, density of insertion SVs. (C) Tracks representing density at 100-kb intervals; from outside to inside: 1, genomic positions (in Mb) of the six chromosomes based on the N2 genome;

2–4, density of SVs estimated by SnpEff: 2, high-impact SVs; 3, low-impact SVs; 4, modifier SVs. (D) Annotation of SVs. SVs effects were categorized using SnpEff based on their position in the annotated N2 genome. “N2-specific genes” indicates the number of the genes that are completely deleted in CB4856. ‘Genic’ indicates the number of genes whose function is predicted to be affected by the SVs. ‘Intergenic’ indicates the number of SVs in the intergenic region. ‘Upstream’ indicates the number of SVs located within 5 kb upstream of a gene. ‘Downstream’ indicates the number of SVs located within 5 kb downstream from a gene.

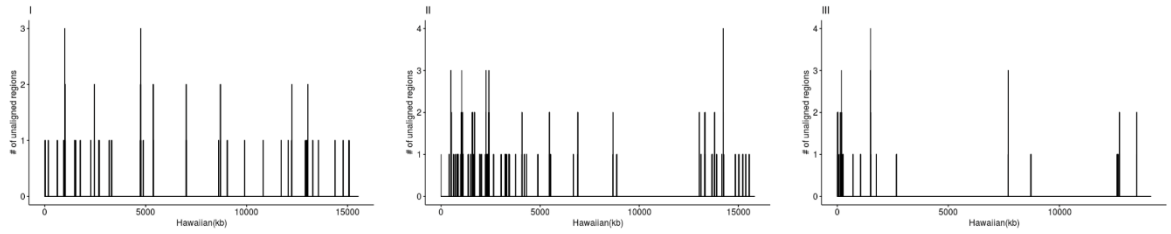
A



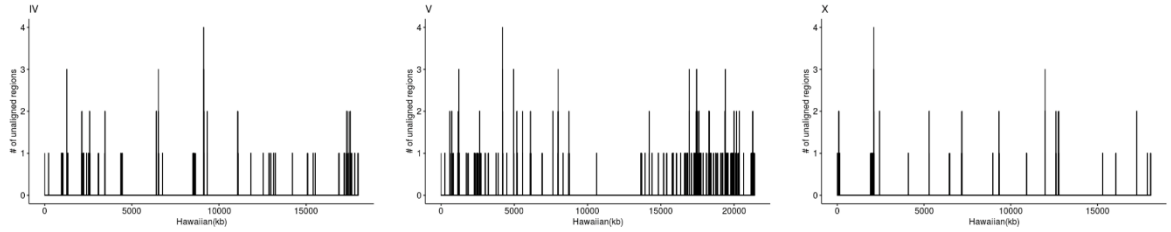
B

Variant type	Total base pairs
Insertion (Ins)	508,763
Deletion (Del)	58,165
Ins-Del	**450,598**
Tandem_expansion (Texp)	557,898
Tandem_contraction (Tcont)	47,733
Texp-Tcont	**510,165**
Repeat_expansion (Rexp)	999,426
Repeat_contraction (Rcont)	319,997
Rexp-Rcont	**679,429**
Unaligned base (Kim et al., 2019)	**4,019,041**
Unaligned base (Thompson et al., 2015)	**607,896**
Total change	**5,051,337**

C D E



F G H



I

Method	Subset	Subset	Intersection	Intersection
Offset	`+/- 0.5kb	`+/- 1kb	`+/- 0.5kb	`+/- 1kb
I	82.38%	84.63%	89.55%	90.16%
II	85.19%	87.59%	92.22%	92.41%
III	81.36%	85.04%	89.50%	90.81%
IV	82.55%	85.10%	90.59%	90.78%
V	85.77%	87.85%	91.16%	91.71%

Figure 12. Direct comparison of the Kim genome and the Thompson genome. (A) SVs in the Kim et al., 2019 genome with the Thompson et al., 2015 genome as a reference (B) Length of SVs in the Kim et al., 2019 genome. (C-H) Distribution of unaligned bases. (I) Comparing SVs found in the Thompson et al., 2015 genome and the Kim et al., 2019 genome. 'Subset' case means an SV from one genome are completely included within an SV of the other genome, and the 'intersection' case means that the two SVs partially overlap each other.

A

	N2	CB4856
Genes	46742	46238 (98.9%)
Protein coding genes	20039	19355 (96.6%)
Transferred protein coding genes		19355
Not transferred protein coding genes	684	
N2 specific genes	619	
Partially transferred protein coding genes	345	
Gene prediction		781

B

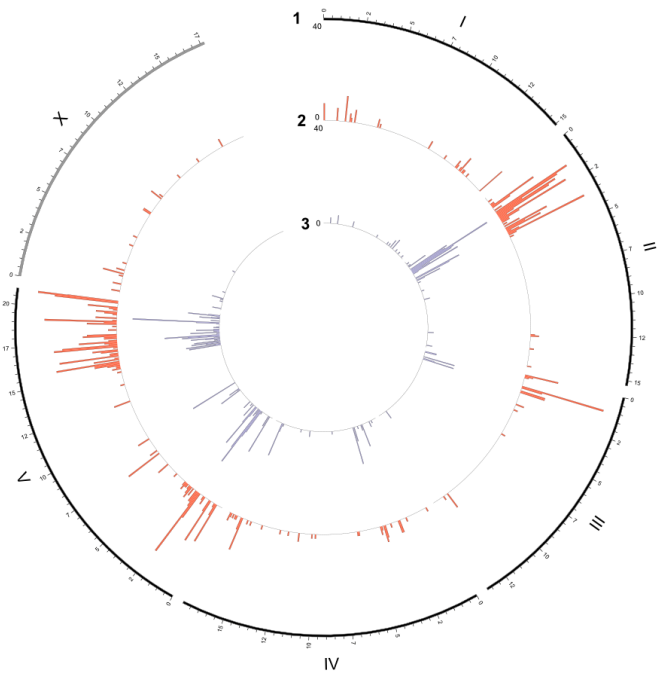


Figure 13. CB4856 gene annotation. (A) Annotation report of CB4856. Transferred protein coding genes: number of transferred genes from N2 by RATT; not transferred protein coding genes: number of not transferred genes from N2 by RATT; N2-specific genes: number of genes that were confirmed to be absent in the Kim et al., 2019 genome by blast; partially transferred protein coding genes: number of genes whose exons are partially deleted; gene prediction: number of de novo annotated genes by Maker in the 'not transferred region' by RATT. (B) Tracks representing density at 100 kb intervals; from outside to inside: 1, genomic positions (in Mb) of the six chromosomes based on the N2 genome; 2, density of CB4856-specific genes; 3, density of N2-specific genes.

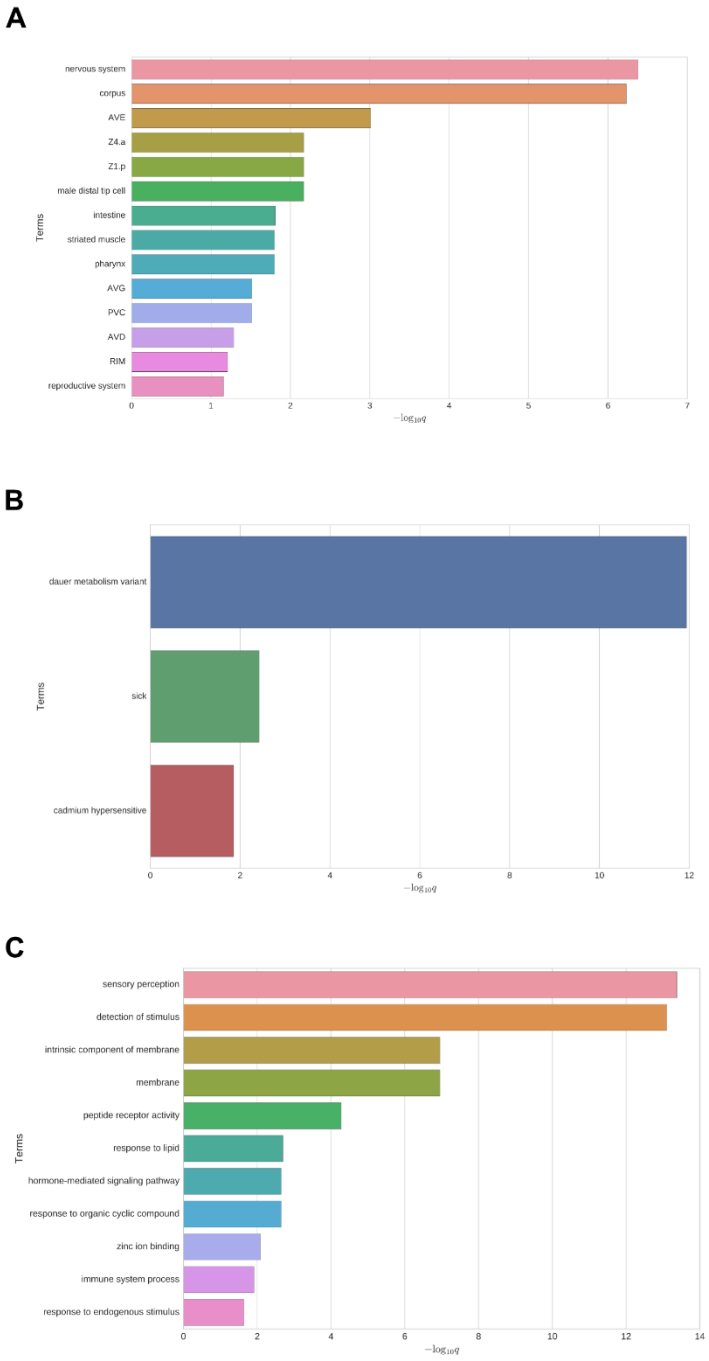
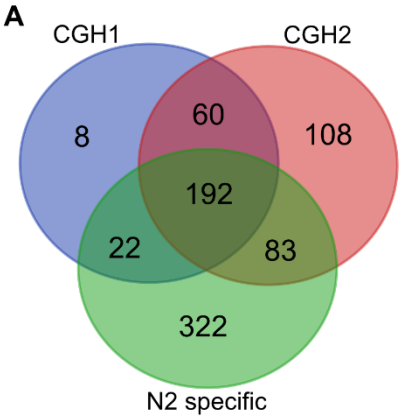


Figure 14. N2 Genes affected by SVs at CB4856. (A) Tissue Enrichment Analysis of high-impact SV genes. (B) Phenotype Enrichment Analysis of high-impact SV genes. (C) Gene Ontology Enrichment Analysis of high-impact SV genes.

A



B

	CGH1-specific	CGH2-specific	Intersection (CGH1 & CGH2)
High impact gene	5	64	49
Pseudogene	1	26	9
No_clue	2	18	2
Total	8	108	60

Figure 15. Validation of N2-specific genes. (A) N2-specific genes (completely deleted genes) compared with two CGH data. CGH1: Maydan et al., 2007, CGH2: Maydan et al., 2010. (B) Analysis of CGH specific genes. 67% of the genes that were reported as deleted by CGH analyses but were not confirmed as N2-specific genes are actually high impact genes in CB4856, and 20% of them were pseudogenes.

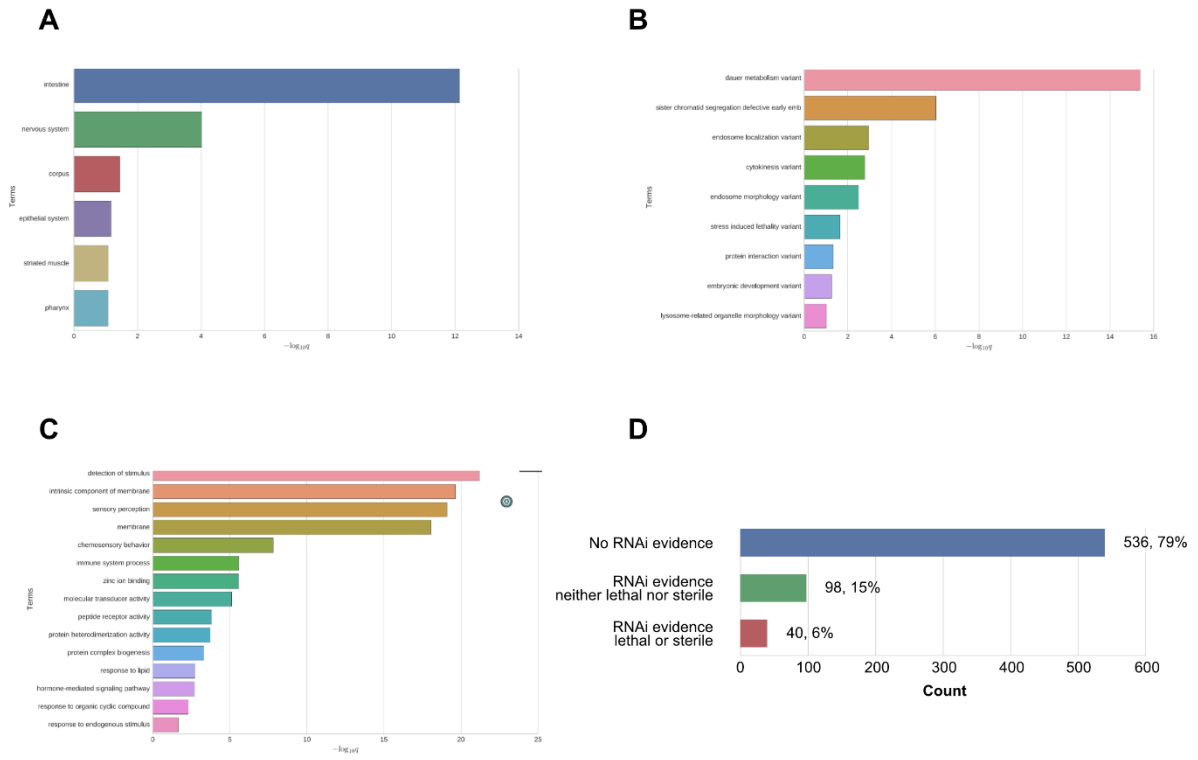


Figure 16. Gene ontology analysis of N2-specific genes. (A-D) Enrichment analysis of tissue, phenotype and GO in N2-specific genes. (A) Tissue Enrichment Analysis of high-impact SV genes. (B) Phenotype Enrichment Analysis of high-impact SV genes. (C) Gene Ontology Enrichment Analysis of high-impact SV genes. (D) RNAi evidence of lethal and sterile phenotype in N2-specific genes.

A

Location	N2	CB4856	Location	N2	CB4856
Chr IL	50,235	50,305	Chr IVL	5,937	4,396
Chr IR	64,274	24,112	Chr IVR	16,102	16,613
Chr IIL	3,823	3,803	Chr VL	3,356	3,354
Chr IIR	23,137	13,803	Chr VR	200,000	200,000
Chr IIIL	176,124	176,099	Chr XL	29,526	36,985
Chr IIIR	1,406	8,309	Chr XR	4,113	4,111

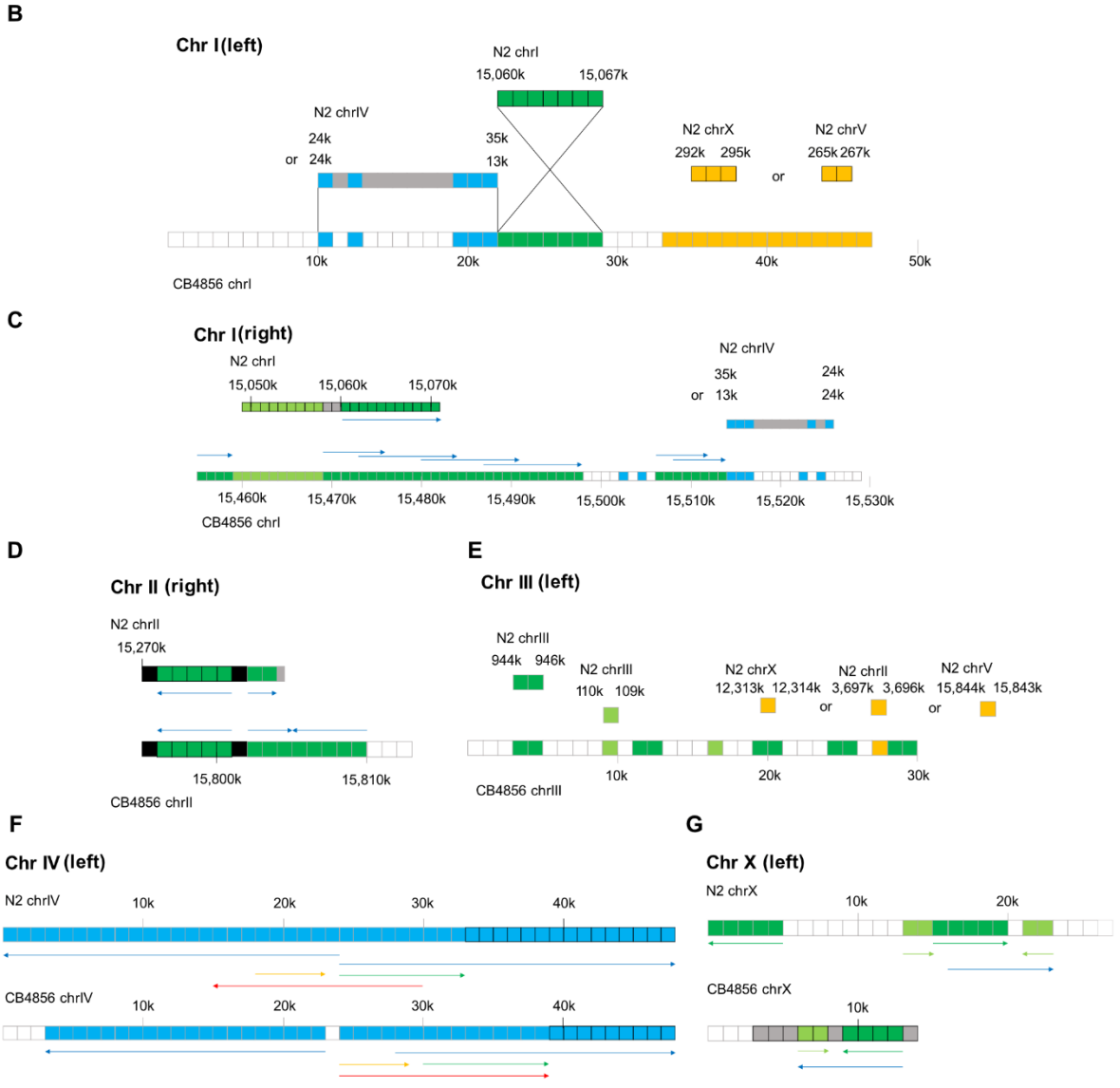


Figure 17. CB4856 subtelomeres. (A) The number of unaligned bases of each subtelomere (200 kb) between N2 and CB4856 chromosomes obtained using nucmer. (B-G) Schematic representation of BLAST results of unaligned bases in each CB4856 subtelomere of Chr IL (B), Chr IR (C), Chr IIR (D), Chr IIIL (E), Chr IVL (F), and Chr XL (G) to the N2 genome. White blocks represent regions without any homology found using BLAST, and other colored blocks represent aligned regions with homology detected longer than 600 bp. Each block represents 1 kb of genome length and alignments with less than 600 bp in length are not included. Arrows indicate directions.

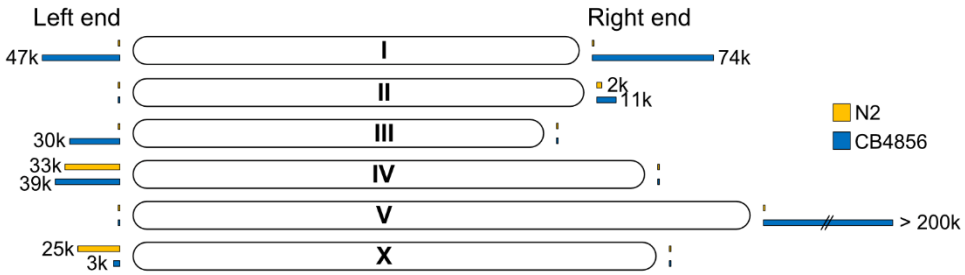


Figure 18. Schematic representation of subtelomere differences between the N2 and CB4856 chromosomes. Yellow bars and blue bars at the end of chromosomes indicate the ratio of unaligned bases of subtelomeres in N2 and CB4856 genome, respectively.

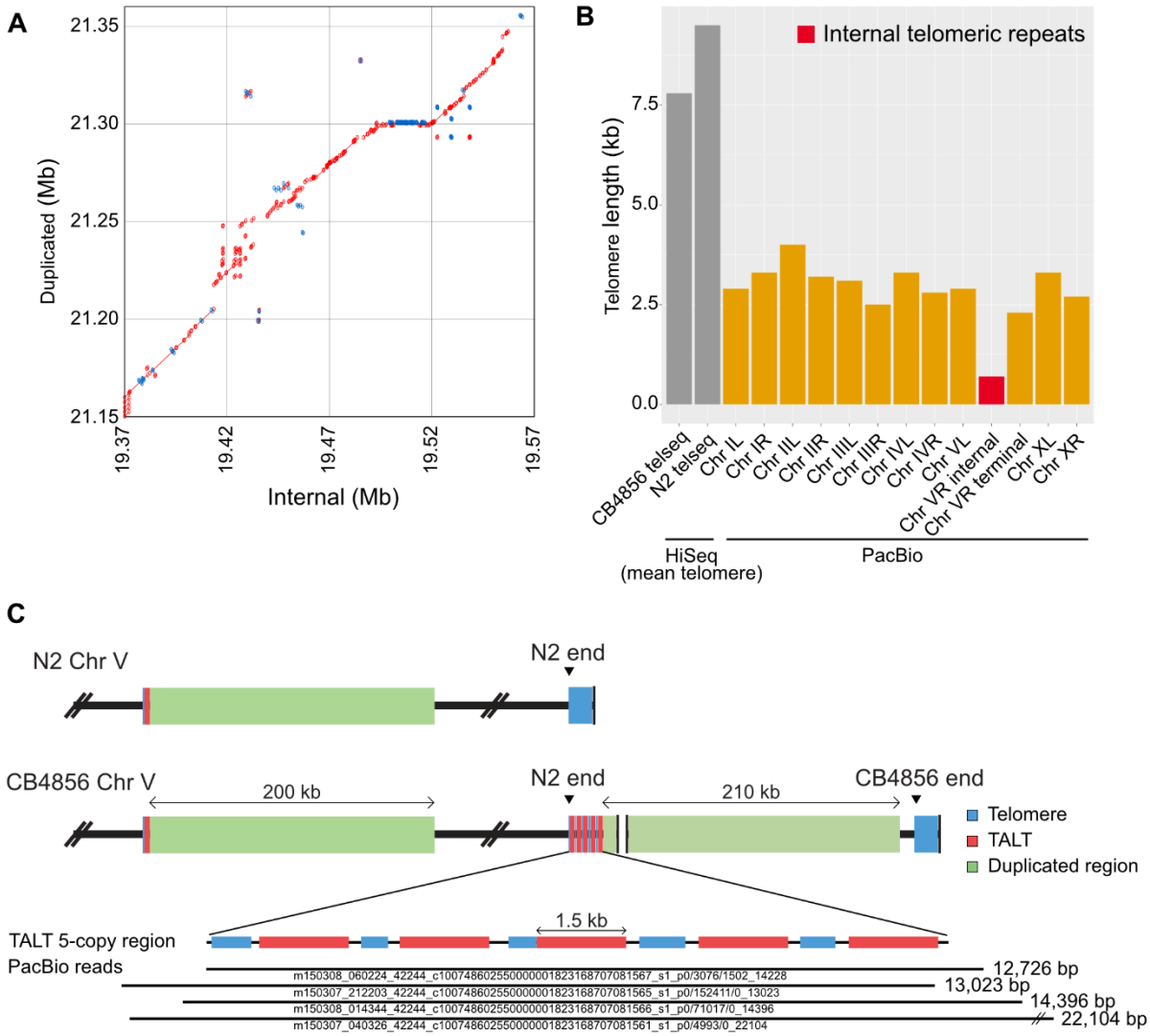


Figure 19. New subtelomere formation in CB4856 Chr VR using an alternative lengthening of telomeres (ALT) mechanism. (A) Dot plot representing alignment between internal segment (V: 19,377,978–19,606,221) and duplicated segment (V: 21,171,521–21,389,866) of CB4856 Chr VR; 63% of the two regions are aligned, and 91% of the aligned bases are identical. Red: forward strand matches; blue: reverse strand matches. (B) Telomere length of all chromosomes deduced from the long-read CB4856 genome. 'HiSeq' data are mean telomere lengths normalized by the telseq software (Ding et al. 2014). The red bar represents the end of N2 (Chr VR internal) in Chr VR of CB4856. Only small portions of the N2 telomere remain in CB4856, followed by a new subtelomere. 'Chr V terminal' is from the real end of Chr VR. (C) Schematic representation of Chr V subtelomere in CB4856. Five copies of template for ALT (TALT) (red) are connected to the duplicated segment from the internal segment close to the internal TALT (V: 19,366,148–19,367,611). The bottom shows PacBio raw reads on the tandemly repeated TALT region. Four raw reads almost fully cover this region.

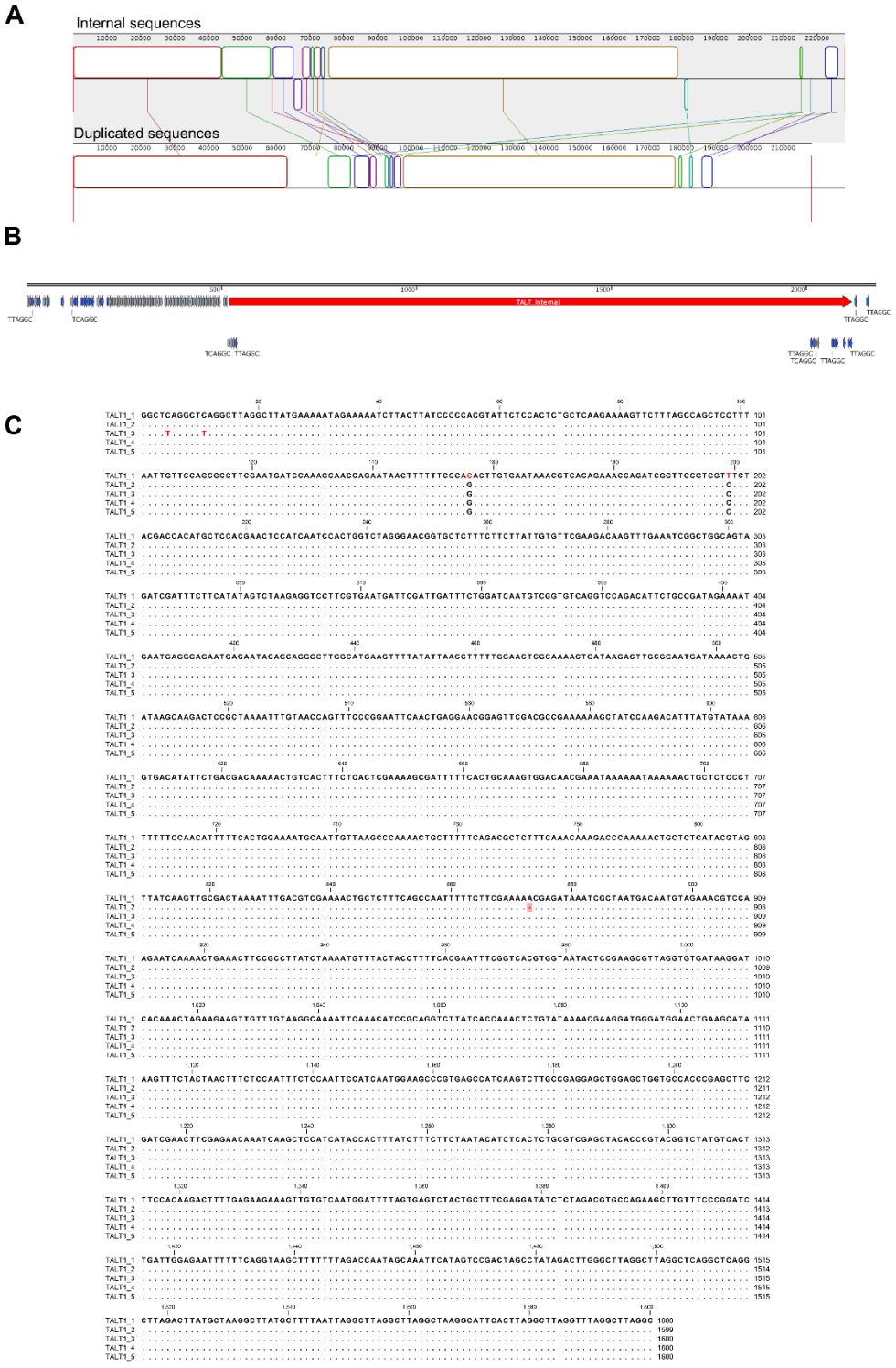


Figure 20. Alignment between internal and duplicated segments and TALT structures. (A) Schematic alignment between two sequences, obtained using Mauve. (B) Schematic depicting internal TALT structure. Gray arrows: TTAGGC repeats; blue arrows: TCAGGC repeats; red bar: TALT sequence. (C) SNPs between tandem-copied TALT sequences.

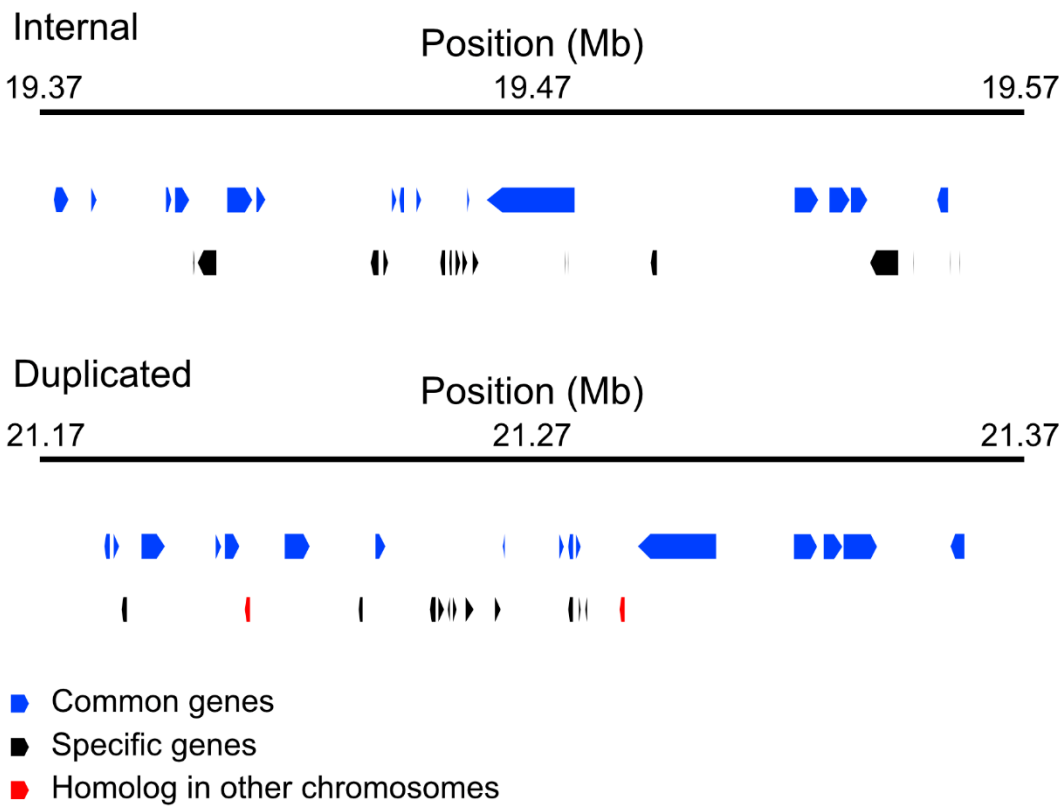


Figure 21. New subtelomere and new genes. Internal genes were duplicated to Chr VR subtelomere. The figure shows a putative gene model of the Chr VR subtelomere. Upper panel: internal gene model; lower panel: subtelomeric gene model.

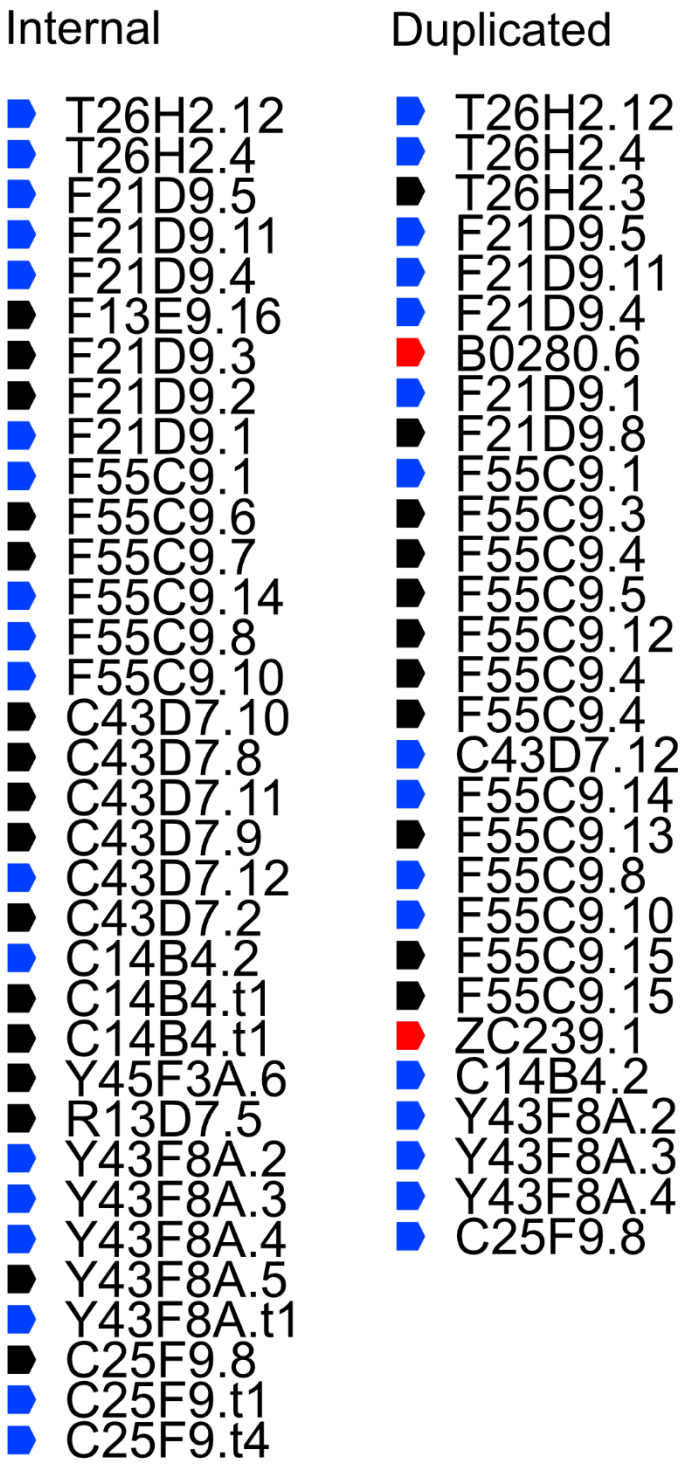


Figure 22. Gene lists of the internal and subtelomeric region. Blue genes are common between the internal and subtelomeric region; black genes are specific to one region; red genes have a homologue in another chromosome.

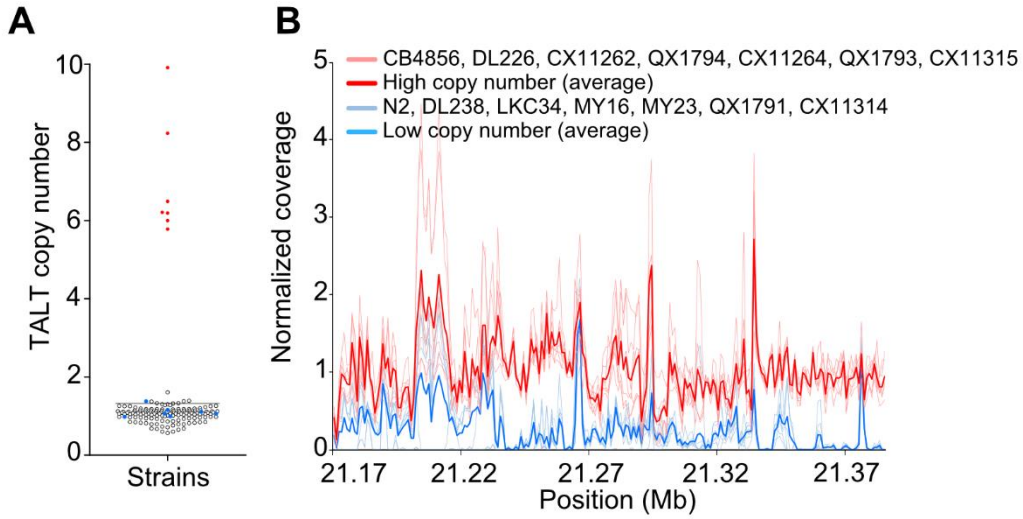


Figure 23. New subtelomere formation in wild isolates. (A) TALT copy numbers among wild isolates (Table 6). (B) Normalized coverage mapped on the duplicated segment of wild isolates with high TALT copy number (red) strains and low TALT copy number (blue) strains.

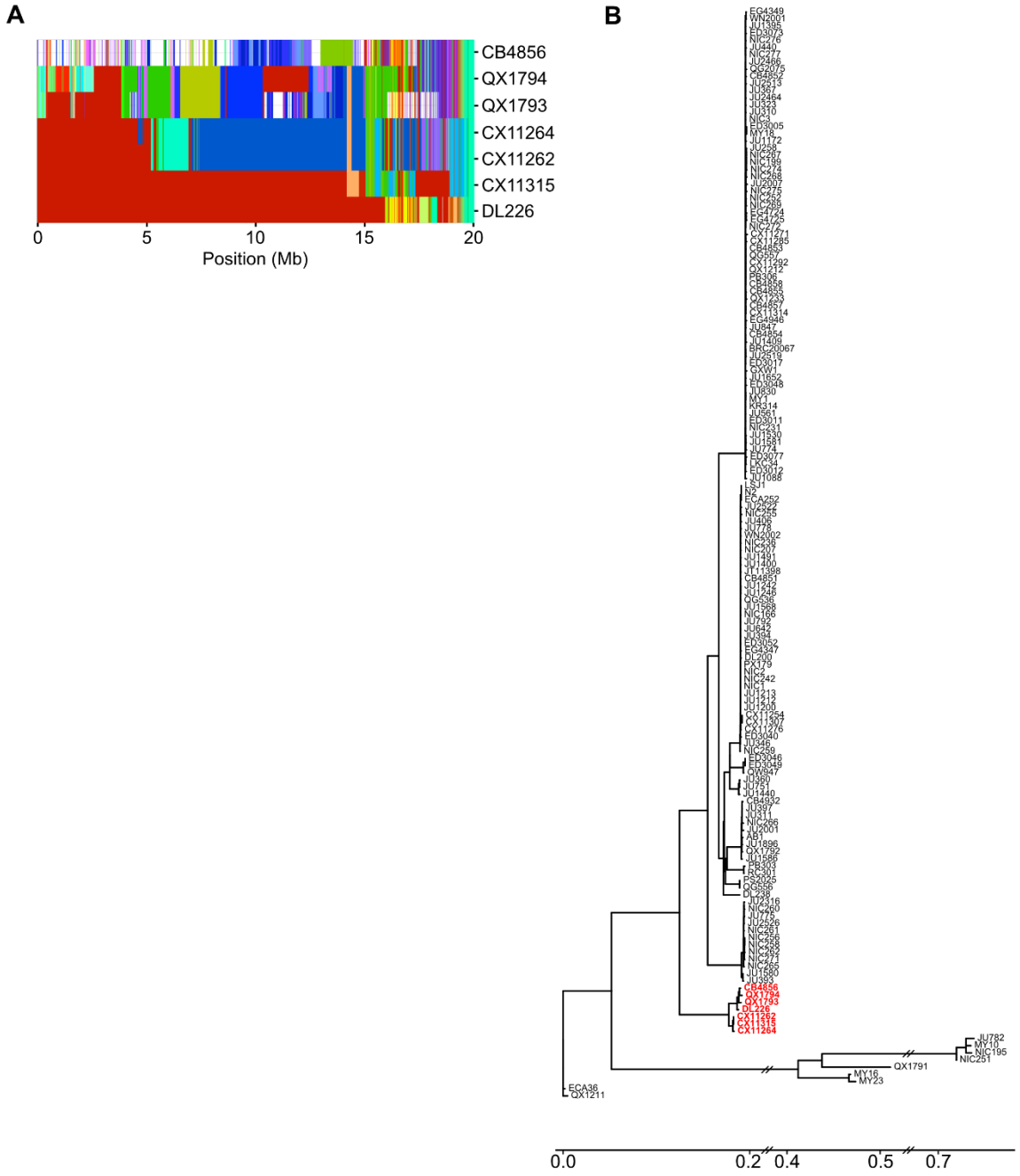


Figure 24. Haplotype and phylogenic tree of wild isolates. (A) Haplotype blocks on Chr V of seven strains that have high TALT copy numbers. (B) Phylogenic tree of reference N2 and 151 wild strains whose genomes have been fully sequenced. Strains marked with red color contain several copies of TALT.

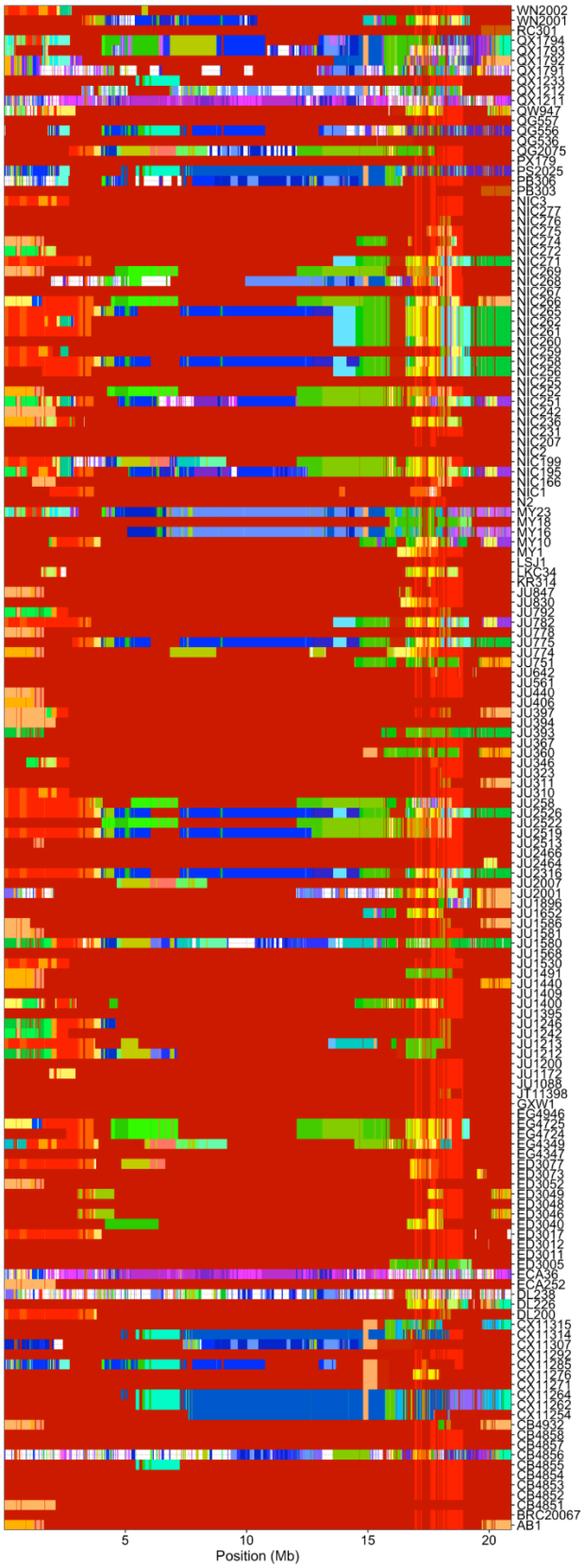


Figure 25. Haplotype block of reference N2 and 151 wild strains

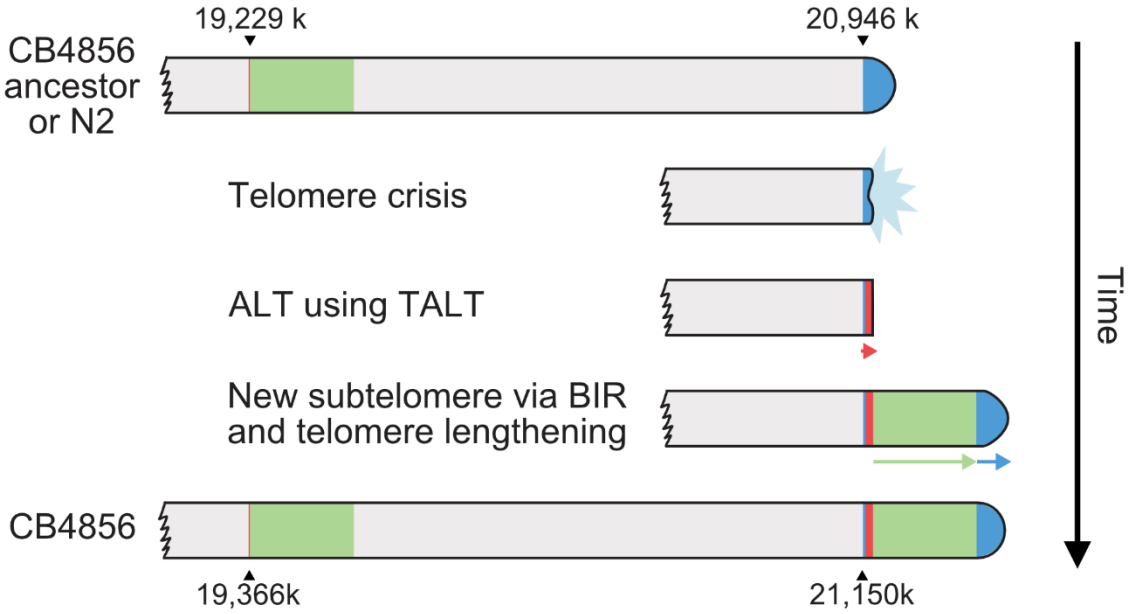


Figure 26. A model of Chr VR subtelomere formation in CB4856. The CB4856 ancestor underwent telomere crisis, and two sequential telomere-damage repair events, one using ALT and the other using BIR, formed new subtelomeres. Finally, the duplicated block end was repaired by telomerase, ending with at least 3-kb-long telomeric repeats.

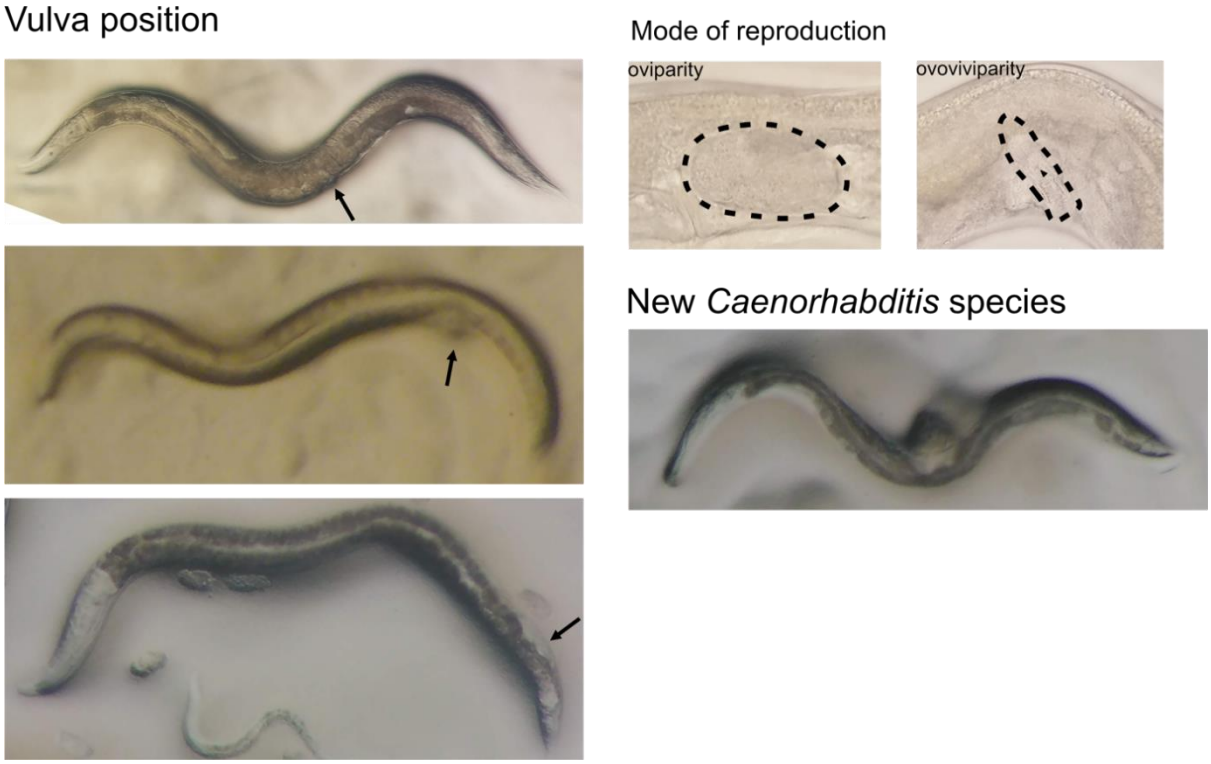


Figure 27. Phenotypic variation in Korean nematodes.

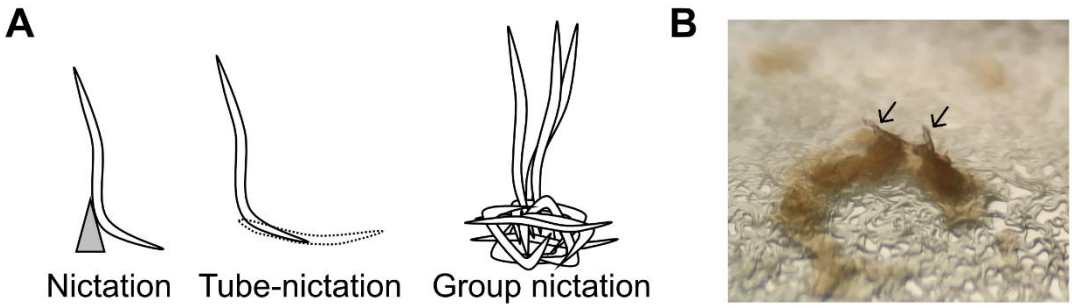


Figure 28. Repertoire of nictation behaviors. (A) *C. elegans* worms do not show nictation behavior on the smooth NGM plate and require any physical support. Most *Auanema* species worms show tube-nictation, and *Auanema* sp. APS14 species do not only tube-nictation, but also group nictation. (B) Typical image of group nictation in the *Auanema* sp. APS14 species. Arrows indicate the top parts of each groups.

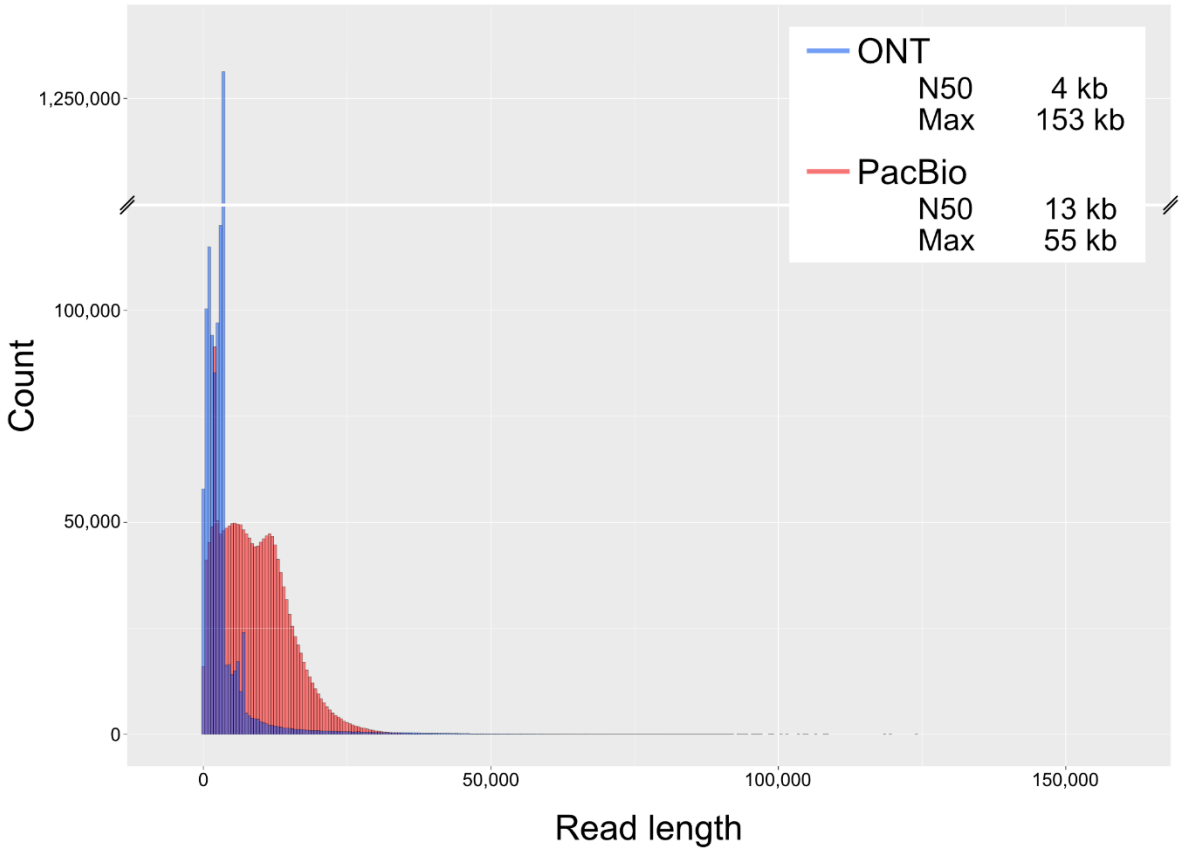


Figure 29. Raw read length distribution for ONT and PacBio. Blue bars and read bars represent ONT MinION and PacBio RSII platforms, respectively. ONT has more biased distribution to shorter reads, especially enriched near the 3.6 kb.

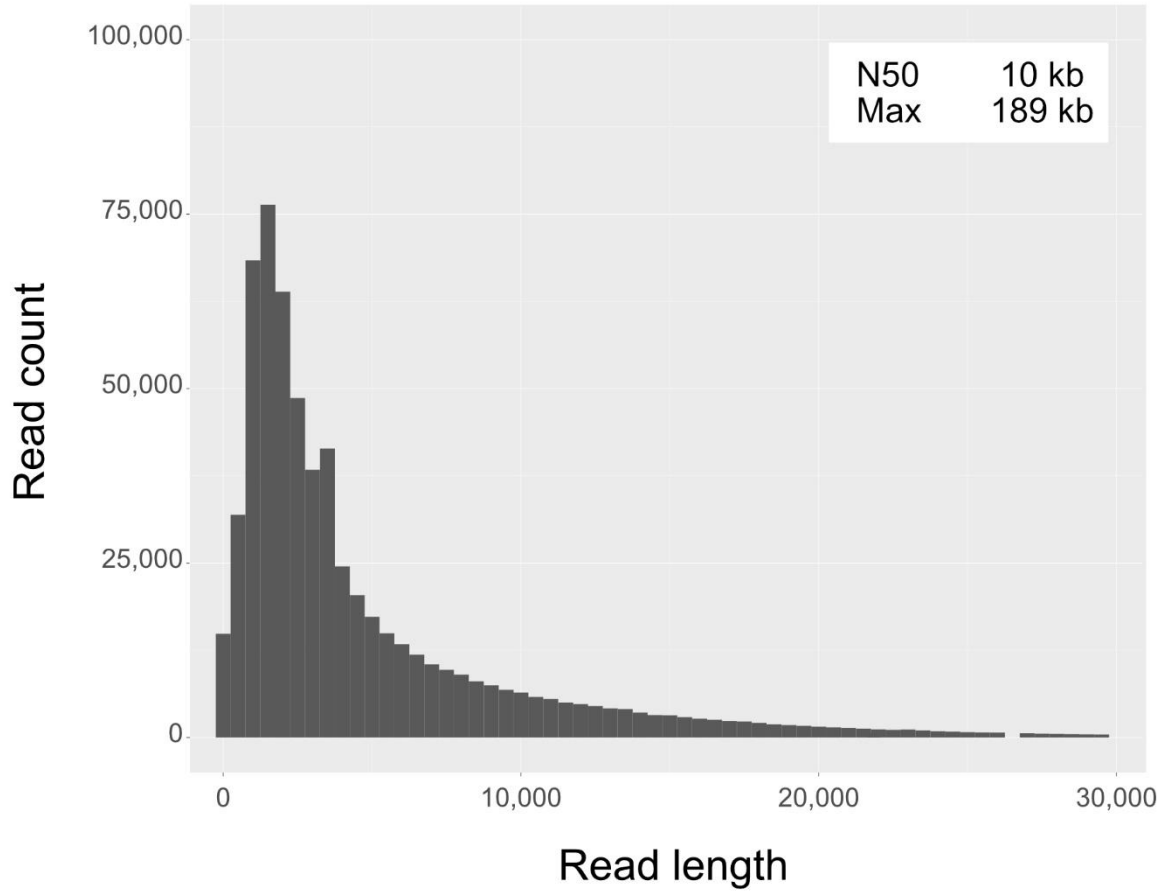


Figure 30. Raw read length distribution for the *Rhabditella axei* sequencing result using ONT.

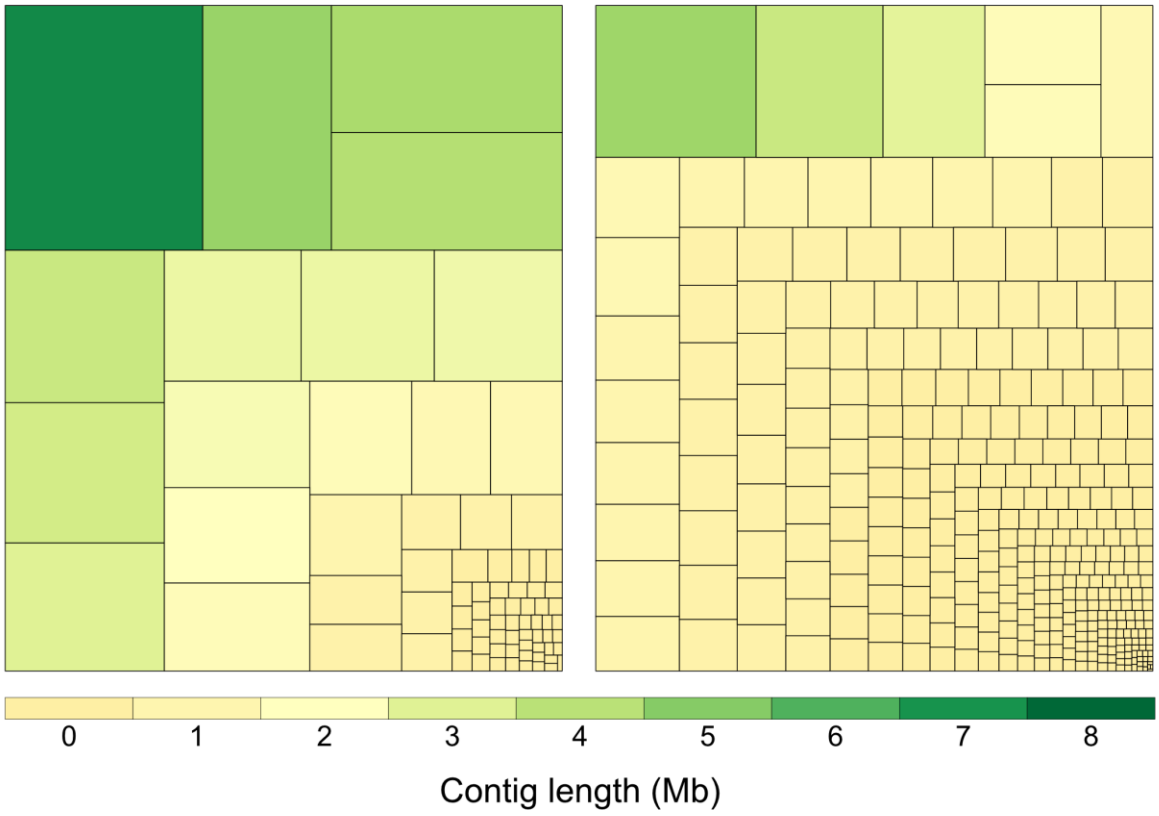


Figure 31. Treemaps for *A. freiburgensis* APS7 (left) and *Auanema* sp. APS14 (right) genome assemblies. The total area represents the total assembly size, and each rectangular area represents the length of each contigs.

Table 1. Long-read sequencing-based genome assemblies of CB4856

	Canu	**Canu +polishing**	**Canu +polishing +tiling**	**Canu +polishing +tiling +Fosmid**
Polishing	N/A	Quiver x2 + Pilon x2	Quiver x2 + Pilon x2	Quiver x2 + Pilon x2
Bacterial contigs removal	No	Yes	Yes	Yes
Removed contigs N50 (bp)	N/A	15,531	21,629	N/A
Number of contigs or scaffolds	137	128	76	26
Number of bases (bp)	104,001,098	103,898,092	102,856,938	102,862,938
N50 (bp)	2,786,743	2,786,967	2,786,967	6,622,535
Maximum length (bp)	9,649,103	9,650,681	9,650,681	19,875,540
Minimum length (bp)	4,093	4,093	22,460	25,081

Table 2. Comparisons between pairs of N2/Thompson, and N2/Kim genomes

	N2 vs. Thompson et al., 2015		N2 vs. Kim et al., 2019	
	N2	Thompson et al	N2	Kim et al.
Aligned bases (bp)	96,233,595 (95.96%)	95,534,154 (97.19%)	96,278,605 (96.00%)	97,205,531 (94.45%)
Unaligned bases (bp)	4,052,806 (4.04%)	2,757,262 (2.81%)	4,007,796 (4.00%)	5,709,254 (5.55%)
Identity between alignments (%)	99.54	99.54	99.39	99.39
Number of SNPs	170,250		176,543	
Number of single nucleotide indels	222,323		256,747	
Number of SVs with > 50 bp	2,965		3,349	
Number of mapped corrected reads[1]	316,299 (99.30%)		317,669 (99.73%)	
Average mapping ratio of each reads[2]	93.67%		98.16%	
Number of unqualified reads[3]	5,500		3,111	

[1] Total number of corrected reads were 318,534, in total 3,711,901,354 bp
[2] Average number of mapped bases of each reads divided by their lengths
[3] Number of reads that have MAPQ<254

Table 3. Phenotypic diversification in the genus *Auanema*

	Arsenic resistance	**Gender**	**Mode of reproduction**	**Phoretic behavior**
C. elegans	No[1]	♀♂	Oviparity	Nictation
Auanema sp.	Yes[1]	♀♀♂[1]	Ovoviviparity[1]	Tube-nictation[2]
A. rhodensis	Yes[1]	♀♀♂[2]	Oviparity[1]	Tube-nictation[2]
A. freiburgensis	Yes[1]	♀♀♂[2]	Oviparity[1]	Tube-nictation[2]
Auanema sp. APS14 (this study)	?	♀♀♂[3]	Oviparity[3]	Tube-nictation[3] Group nictation[3]

[1] Shih et al. 2019
[2] Kanzaki et al. 2017
[3] This study

Table 4. Comparisons between the draft genomes of *A. freiburgensis* and *Auanema* sp. APS14

	A. freiburgensis	***Auanema* sp. APS14**
Polishing	Quiver x1 + Pilon x4	Nanopolish x1
Number of contigs	75	327
Number of bases (bp)	55,262,204	69,089,623
N50 (bp)	3,036,121	555,637
Maximum length (bp)	7,206,631	4,537,356
Minimum length (bp)	11,256	1,899

References

Alföldi J, Lindblad-Toh K. 2013. Comparative genomics as a tool to understand evolution and disease. *Genome research* **23**: 1063-1068.

Andersen EC, Bloom JS, Gerke JP, Kruglyak L. 2014. A variant in the neuropeptide receptor npr-1 is a major determinant of Caenorhabditis elegans growth and physiology. *PLoS Genet* **10**: e1004156.

Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, Bloom JS, Félix M-A, Kruglyak L. 2012. Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. *Nature genetics* **44**: 285.

Angeles-Albores D, RY NL, Chan J, Sternberg PW. 2016. Tissue enrichment analysis for C. elegans genomics. *BMC Bioinformatics* **17**: 366.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**: 11.

Barry J, Ginger ML, Burton P, McCulloch R. 2003. Why are parasite contingency genes often associated with telomeres? *International journal for parasitology* **33**: 29-45.

Brenner S. 1974. The genetics of *Caenorhabditis elegans*. *Genetics* **77**: 71-94.

Blackburn EH. 1991. Structure and function of telomeres. *Nature* **350**: 569-573.

Blasco MA, Lee H-W, Hande MP, Samper E, Lansdorp PM, DePinho RA, Greider CW. 1997. Telomere shortening and tumor formation by mouse cells lacking telomerase RNA. *Cell* **91**: 25-34.

Bosco G, Haber JE. 1998. Chromosome break-induced DNA replication leads to nonreciprocal translocations and telomere capture. *Genetics* **150**: 1037-1047.

Brown CA, Murray AW, Verstrepen KJ. 2010. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Curr Biol* **20**: 895-903.

Bryan TM, Englezou A, Dalla-Pozza L, Dunham MA, Reddel RR. 1997. Evidence for an alternative mechanism for maintaining telomere length in human tumors and tumor-derived cell lines. *Nat Med* **3**: 1271-1274.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC bioinformatics* **10**: 421.

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**: 188-196.

Ceccaldi R, Rondinelli B, D'Andrea AD. 2016. Repair pathway choices and consequences at the double-strand

break. *Trends in cell biology* **26**: 52-64.

Cesare AJ, Reddel RR. 2010. Alternative lengthening of telomeres: models, mechanisms and implications. *Nat Rev Genet* **11**: 319-330.

Cho Nam W, Dilley Robert L, Lampson Michael A, Greenberg Roger A. 2014. Interchromosomal Homology Searches Drive Directional ALT Telomere Movement and Synapsis. *Cell* **159**: 108-121.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**: 80-92.

The C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*: 2012-2018.

Cook DE, Andersen EC. 2017. VCF-kit: assorted utilities for the variant call format. *Bioinformatics* **33**: 1581-1582.

Cook DE, Zdraljevic S, Roberts JP, Andersen EC. 2016a. CeNDR, the Caenorhabditis elegans natural diversity resource. *Nucleic acids research* **45**: D650-D657.

Cook DE, Zdraljevic S, Tanny RE, Seo B, Riccardi DD, Noble LM, Rockman MV, Alkema MJ, Braendle C, Kammenga JE. 2016b. The genetic basis of natural variation in Caenorhabditis elegans telomere length. *Genetics*: genetics. 116.191148.

Costantino L, Sotiriou SK, Rantala JK, Magin S, Mladenov E, Helleday T, Haber JE, Iliakis G, Kallioniemi OP, Halazonetis TD. 2014. Break-induced replication repair of damaged forks induces genomic duplications in human cells. *Science* **343**: 88-91.

Cutter AD, Choi JY. 2010. Natural selection shapes nucleotide polymorphism across the genome of the nematode Caenorhabditis briggsae. *Genome research* **20**: 1103-1111.

Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics* **14**: 262.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS one* **5**: e11147.

de Bono M, Bargmann CI. 1998. Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in C. elegans. *Cell* **94**: 679-689.

Dilley RL, Verma P, Cho NW, Winters HD, Wondisford AR, Greenberg RA. 2016. Break-induced telomere synthesis underlies alternative telomere maintenance. *Nature* **539**: 54-58.

Ding Z, Mangino M, Aviv A, Consortium UK, Spector T, Durbin R. 2014. Estimating telomere length from whole genome sequence data. *Nucleic acids research* **42**: e75-e75.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.

Fatt HV, Dougherty EC. 1963. Genetic Control of Differential Heat Tolerance in Two Strains of the Nematode *Caenorhabditis elegans*. *Science* **141**: 266.

Félix M-A. 2004. Alternative morphs and plasticity of vulval development in a rhabditid nematode species. *Development genes and evolution* **214**: 55-63.

Félix M-A. 2006. Oscheius tipulae. In *WormBook: The Online Review of C elegans Biology [Internet]*. WormBook.

Garavís M, González C, Villasante A. 2013. On the origin of the eukaryotic chromosome: the role of noncanonical DNA structures in telomere evolution. *Genome Biol Evol* **5**: 1142-1150.

Glover T, Stein C. 1987. Induction of sister chromatid exchanges at common fragile sites. *American journal of human genetics* **41**: 882.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**: 1494-1512.

Harley CB, Futcher AB, Greider CW. 1990. Telomeres shorten during ageing of human fibroblasts. *Nature* **345**: 458.

Hunt VL, Tsai IJ, Coghlan A, Reid AJ, Holroyd N, Foth BJ, Tracey A, Cotton JA, Stanley EJ, Beasley HJNg. 2016. The genomic basis of parasitism in the Strongyloides clade of nematodes. **48**: 299.

Johnson TE, Wood WB. 1982. Genetic analysis of life-span in Caenorhabditis elegans. *Proc Natl Acad Sci U S A* **79**: 6603-6607.

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M. 2003. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421**: 231.

Kammenga JE, Doroszuk A, Riksen JA, Hazendonk E, Spiridon L, Petrescu A-J, Tijsterman M, Plasterk RH, Bakker J. 2007. A Caenorhabditis elegans wild type defies the temperature–size rule owing to a single nucleotide polymorphism in tra-3. *PLoS genetics* **3**: e34.

Kanzaki N, Kiontke K, Tanaka R, Hirooka Y, Schwarz A, Müller-Reichert T, Chaudhuri J, Pires-daSilva AJSr. 2017. Description of two three-gendered nematode species in the new genus Auanema (Rhabditina) that are models for reproductive mode evolution. **7**: 11135.

Kern AD, Hahn MW. 2018. The neutral theory in light of natural selection. *Molecular biology and evolution* **35**: 1366-1371.

Kim C, Sung S, Lee J. 2016. Internal genomic regions mobilized for telomere maintenance in C. elegans. In *Worm*, Vol 5, p. e1146856. Taylor & Francis.

Kim J, Lee, D., & Lee, J. 2017. A quantitative trait locus for nictation behavior on chromosome V. *microPublication Biology* doi:https://doi.org/10.17912/W23D39.

Koch R, van Luenen HG, van der Horst M, Thijssen KL, Plasterk RH. 2000. Single nucleotide polymorphisms in wild isolates of Caenorhabditis elegans. *Genome Res* **10**: 1690-1696.

Koepfli K-P, Paten B, Scientists GKCo, O'Brien SJ. 2015. The Genome 10K Project: a way forward. *Annu Rev Anim Biosci* **3**: 57-111.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*: gr. 215087.215116.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**: 59.

Kramara J, Osia B, Malkova A. 2018. Break-induced replication: the where, the why, and the how. *Trends in Genetics*.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**: 1639-1645.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome biology* **5**: R12.

Lee D, Yang H, Kim J, Brady S, Zdraljevic S, Zamanian M, Kim H, Paik Y-K, Kruglyak L, Andersen EC et al. 2017. The genetic basis of natural variation in a phoretic behavior. *Nat Commun* **8**: 273.

Lee D, Zdraljevic S, Cook DE, Frézal L, Hsu J-C, Sterken MG, Riksen JAG, Wang J, Kammenga JE, Braendle C et al. 2019. Selection and gene flow shape niche-associated copy-number variation of pheromone receptor genes. *bioRxiv* doi:10.1101/580803: 580803.

Lee RYN, Howe KL, Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Davis P, Gao S, Grove C et al. 2018. WormBase 2017: molting into a new stage. *Nucleic Acids Res* **46**: D869-d874.

Leister D. 2004. Tandem and segmental gene duplication and recombination in the evolution of plant disease

resistance genes. *Trends Genet* **20**: 116-122.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987-2993.

Li H. 2013a. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997*.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.

Lundblad V, Blackburn EH. 1993. An alternative pathway for yeast telomere maintenance rescues *est1*− senescence. *Cell* **73**: 347-360.

Lydeard JR, Jain S, Yamaguchi M, Haber JE. 2007. Break-induced replication and telomerase-independent telomere maintenance require Pol32. *Nature* **448**: 820-823.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: a fast and versatile genome alignment system. *PLoS computational biology* **14**: e1005944.

Mason JM, Randall TA, Capkova Frydrychova R. 2016. Telomerase lost? *Chromosoma* **125**: 65-73.

Mason JMO, McEachern MJ. 2018. Mild Telomere Dysfunction as a Force for Altering the Adaptive Potential of Subtelomeric Genes. *Genetics* **208**: 537-548.

Maupas E. 1901. Modes et formes de reproduction des nematodes. *Archives de Zoologie expérimentale et générale* **8**: 463-624.

Maydan JS, Flibotte S, Edgley ML, Lau J, Selzer RR, Richmond TA, Pofahl NJ, Thomas JH, Moerman DG. 2007. Efficient high-resolution deletion discovery in Caenorhabditis elegans by array comparative genomic hybridization. *Genome research* **17**: 337-347.

Maydan JS, Lorch A, Edgley ML, Flibotte S, Moerman DG. 2010. Copy number variation in the genomes of twelve natural isolates of Caenorhabditis elegans. *BMC genomics* **11**: 62.

McEachern MJ, Haber JE. 2006. Break-induced replication and recombinational telomere elongation in yeast. *Annu Rev Biochem* **75**: 111-135.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**: 1297-1303.

McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**: e61217.

Meier B, Clejan I, Liu Y, Lowden M, Gartner A, Hodgkin J, Ahmed S. 2006. trt-1 is the Caenorhabditis elegans catalytic subunit of telomerase. *PLoS genetics* **2**: e18.

Nakamura TM, Cooper JP, Cech TR. 1998. Two modes of survival of fission yeast without telomerase. *Science* **282**: 493-496.

Nattestad M, Schatz MC. 2016. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**: 3021-3023.

Neumann AA, Watson CM, Noble JR, Pickett HA, Tam PP, Reddel RR. 2013. Alternative lengthening of telomeres in normal mammalian somatic cells. *Genes & development* **27**: 18-23.

Nigon V, Dougherty EC. 1950. A DWARF MUTATION IN A NEMATODE: A Morphological Mutant of *Rhabditis briggsae*, a free-living soil nematode. *Journal of Heredity* **41**: 103-109.

Nigon VM, Felix MA. 2017. History of research on C. elegans and other free-living nematodes as model organisms. *WormBook : the online review of C elegans biology* **2017**: 1-84.

O'sullivan RJ, Karlseder J. 2010. Telomeres: protecting chromosomes against genome instability. *Nature reviews Molecular cell biology* **11**: 171.

Otto TD, Dillon GP, Degrave WS, Berriman M. 2011. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Research* **39**: e57-e57.

Palopoli MF, Rockman MV, TinMaung A, Ramsay C, Curwen S, Aduna A, Laurita J, Kruglyak L. 2008. Molecular basis of the copulatory plug polymorphism in Caenorhabditis elegans. *Nature* **454**: 1019.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289-290.

Pedersen BS, Quinlan AR. 2017. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**: 867-868.

Pich U, Schubert I. 1998. Terminal heterochromatin and alternative telometric sequences in Allium cepa. *Chromosome Res* **6**: 315-322.

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D et al. 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* doi:10.1101/201178: 201178.

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD,

Levy-Moonshine A, Roazen D et al. 2018. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* doi:10.1101/201178: 201178.

Prabh N, Roeseler W, Witte H, Eberhardt G, Sommer RJ, Rödelsperger C. 2018. Deep taxon sampling reveals the evolutionary dynamics of novel gene families in Pristionchus nematodes. *Genome research* **28**: 1664-1674.

Rödelsperger C, Meyer JM, Prabh N, Lanz C, Bemm F, Sommer RJ. 2017. Single-molecule sequencing reveals the chromosome-scale genomic architecture of the nematode model organism Pristionchus pacificus. *Cell reports* **21**: 834-844.

Rockman MV, Kruglyak L. 2009. Recombinational landscape and population genomics of Caenorhabditis elegans. *PLoS Genet* **5**: e1000419.

Rudd MK. 2014. Human and primate subtelomeres. In *Subtelomeres*, pp. 153-164. Springer.

Schluter D. 2001. Ecology and the origin of species. *Trends Ecol Evol* **16**: 372-380.

Schulenburg H, Müller S. 2004. Natural variation in the response of Caenorhabditis elegans towards Bacillus thuringiensis. *Parasitology* **128**: 433-443.

Seidel HS, Ailion M, Li J, van Oudenaarden A, Rockman MV, Kruglyak L. 2011. A novel sperm-delivered toxin causes late-stage embryo lethality and transmission ratio distortion in C. elegans. *PLoS Biol* **9**: e1001115.

Seidel HS, Rockman MV, Kruglyak L. 2008. Widespread genetic incompatibility in C. elegans maintained by balancing selection. *Science* **319**: 589-594.

Seo B, Kim C, Hills M, Sung S, Kim H, Kim E, Lim DS, Oh H-S, Choi RMJ, Chun J et al. 2015. Telomere maintenance through recruitment of internal genomic regions. *Nat Commun* **6**: 8189.

Shih P-Y, Lee JS, Shinya R, Kanzaki N, Pires-daSilva A, Badroos JM, Goetz E, Sapir A, Sternberg PWJCB. 2019. Newly Identified Nematodes from Mono Lake Exhibit Extreme Arsenic Resistance. **29**: 3339-3344. e3334.

Sfeir A, Kosiyatrakul ST, Hockemeyer D, MacRae SL, Karlseder J, Schildkraut CL, de Lange T. 2009. Mammalian telomeres resemble fragile sites and require TRF1 for efficient replication. *cell* **138**: 90-103.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210-3212.

Slos D, Sudhaus W, Stevens L, Bert W, Blaxter M. 2017. Caenorhabditis monodelphis sp. n.: defining the stem morphology and genomics of the genus Caenorhabditis. *BMC Zoology* **2**: 4.

Smit A, Hubley R, Green P. 2016. RepeatMasker Open-4.0. 2015. *Google Scholar*.

Sommer RJ. 2006. Pristionchus pacificus. In *WormBook: The Online Review of C elegans Biology [Internet]*. WormBook.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**: W435-439.

Stanley E, Coghlan A, Berriman M. 2018. A MAKER pipeline for prediction of protein-coding genes in parasitic worm genomes.

Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A et al. 2003. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biol* **1**: E45.

Stevens L, Félix MA, Beltran T, Braendle C, Caurcel C, others. 2018. Comparative genomics of ten new Caenorhabditis species. *bioRxiv*.

Tange OJTUM. 2011. Gnu parallel-the command-line power tool. *The USENIX Magazine* **36**: 42-47.

Thompson OA, Snoek LB, Nijveen H, Sterken MG, Volkers RJM, Brenchley R, Van't Hof A, Bevers RPJ, Cossins AR, Yanai I et al. 2015. Remarkably Divergent Regions Punctuate the Genome Assembly of the Caenorhabditis elegans Hawaiian Strain CB4856. *Genetics* **200**: 975-989.

Tijsterman M, Okihara KL, Thijssen K, Plasterk RH. 2002. PPW-1, a PAZ/PIWI protein required for efficient germline RNAi, is defective in a natural isolate of C. elegans. *Current Biology* **12**: 1535-1540.

Tyson JR, O'Neil NJ, Jain M, Olsen HE, Hieter P, Snutch TP. 2018. MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. *Genome Res* **28**: 266-274.

Vannier J-B, Pavicic-Kaltenbrunner V, Petalcorin MI, Ding H, Boulton SJ. 2012. RTEL1 dismantles T loops and counteracts telomeric G4-DNA to maintain telomere integrity. *Cell* **149**: 795-806.

Viney ME, Lok JBJWtoroCeB. 2007. Strongyloides spp. 1-15.

Weissgerber TL, Milic NM, Winham SJ, Garovic VD. 2015. Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. *PLOS Biology* **13**: e1002128.

Werner MS, Sieriebriennikov B, Prabh N, Loschko T, Lanz C, Sommer RJ. 2018. Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome research* **28**: 1675-1687.

Wicks SR, Yeh RT, Gish WR, Waterston RH, Plasterk RH. 2001. Rapid gene mapping in Caenorhabditis elegans using a high density polymorphism map. *Nat Genet* **28**: 160-164.

Wu C-I, Ting C-T. 2004. Genes and speciation. *Nat Rev Genet* **5**: 114-122.

Yin D, Schwarz EM, Thomas CG, Felde RL, Korf IF, Cutter AD, Schartner CM, Ralston EJ, Meyer BJ, Haag ES. 2018. Rapid genome shrinkage in a self-fertile nematode reveals sperm competition proteins. *Science* **359**: 55-61.

Zalzman M, Falco G, Sharova LV, Nishiyama A, Thomas M, Lee S-L, Stagg CA, Hoang HG, Yang H-T, Indig FE. 2010. Zscan4 regulates telomere elongation and genomic stability in ES cells. *Nature* **464**: 858.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol* **18**: 292-298.

국문 초록

비교유전체학을 이용한 선충의 서브텔로미어 진화와 표현형 변이 연구

김준
서울대학교 생명과학부

긴 길이 염기서열분석 기법은 서로 다른 종을 이용한 비교유전체학 연구는 물론, 한 종 내에서 서로 다른 계통의 유전체를 비교하는 일도 빠르게 발전시키고 있다. 한 종 내에서 얼마나 크고 많은 유전체 변이가 축적될 수 있는지 확인하고자, 본 연구에서는 *C. elegans* 야생 계통 중 표준 계통 N2와 유전적으로 가장 멀다고 알려진 CB4856 계통의 유전체를 N2의 표준 유전체와 비교하였다. CB4856 유전체는 Pacific Biosciences (PacBio) 사의 RSII 기법을 활용해 염기서열 분석을 진행하였고(80×, N50 리드 길이 11.8 kb), 이후 유전체 이어붙이기 과정을 거쳐 염색체에 가까운 수준(76 contigs, N50 contig 2.8 Mb)으로 완성할 수 있었다. 두 유전체를 비교한 결과 2,694개 유전자에서 구조 변이를 확인할 수 있었고 그 중 상당수는 염색체 바깥쪽에 몰려있었다. 염색체 끝에 인접한 서브텔로미어(subtelomere) 지역은 가장 구조 변이가 심각한 지역으로, 그 중에는 새롭게 서브텔로미어가 생겨난 곳도 있었다. 5번 염색체 오른쪽의 서브텔로미어 구조는 CB4856 계통의 조상에서 텔로미어(telomere) 손상이 일어났고, 텔로머레이즈(telomerase) 유전자가 분명 존재했음에도 그 대신 대안적 텔로미어 연장(Alternative Lengthening of telomeres)을 통해 손상이 회복됐으며, 이후 절단 유도 복제(break-induced replication)이 일어나면서 새롭게 서브텔로미어가 형성됐다는 것을 암시하고 있다. 본 연구는 구조 변이와 새로운 서브텔로미어를 포함한 상당한 유전체 변화가 한 종 내에서도 유지될 수 있고, 이러한 변화가 종 내의 유전 다양성을 높일 수 있다는 것을 보여준다. 다음으로, 예쁜꼬마선충의 근연종이면서도 성별(암수한몸, 암컷, 수컷)과 행동(튜브 닉테이션)에서 확연한 차이를 보이는 *Auanema freiburgensis*와 *Auanema* sp. APS14 두 종의 유전체 초안 또한 본 연구에서 분석됐다. *A. freiburgensis*와 *Auanema* sp. APS14의 유전체는 각각 PacBio RSII (270×, N50 리드 길이 12.5 kb)와 Oxford Nanopore Technologies (ONT) 사의 MinION (113×, N50 리드 길이 3.6 kb)을 통해 염기서열이 분석됐으며, 유전체 이어붙이기 결과 예쁜꼬마선충(~100 Mb)에 비해 유전체 크기 또한 상당히 작다는 것(각각 55 Mb와 69 Mb) 또한 확인되었다. 이 두 유전체는 어떻게 유전체 내에 생긴 변화가 새로운 형질의 진화에 영향을 줄 수 있었을지 이해하는 데에 기여할 수 있을 것으로 내다본다.

주요어: 예쁜꼬마선충, 선충, 긴 길이 염기서열분석법, 유전체 이어붙이기, 텔로미어, 서브텔로미어, 대안적 텔로미어 연장(ALT)
학번: 2015-20422

감사의 인사

학위과정 동안 정말 많은 분들이 도움을 주셨습니다. 이준호 선생님은 지난 5년 동안 기어코 사람 하나 만들겠다고 물심양면 지원을 아끼지 않으셨습니다. 지도교수를 scientific mother/father라고들 하던데, 제가 사람 구실 할 수 있도록 이만큼 도와주신 분은 제 어머니 빼고는 정말이지 이준호 선생님이 유일했습니다. 워낙 훌륭하신 분이셔서 길 가는 사람들도 이준호 선생님 사람 좋은 건 다 알고 있습니다만, 같이 일하면서 더욱 존경하게 됐습니다. 저는 선생님처럼 좋은 지도교수가 될 수 없을 것 같다는 생각은 학위 과정 내내 했을 정도로, 연구와 관련된 것뿐만 아니라 연구 외에 다양한 사안에 대해서도 가르침을 주시고 항상 챙겨주시며 제자들에게 좋은 스승이란 어떤 존재인지 몸소 보여주시곤 하셨습니다. 교내 장학금을 알아보거나 외부 펠로십 지원하는 일에도 항상 열과 성을 다하셔서 연구비나 인건비 걱정도 없이 학위를 무사히 마칠 수 있었습니다. 항상 감사할 따름입니다.

연구실 동료들이 없었다면 이만큼 일하기는 어려웠을 것 같습니다. 특히 김천아 박사는 CB4856 프로젝트를 함께 하며 너무나 많은 도움을 줬는데, 그가 없었다면 이 연구를 마무리 짓는 것은 불가능했을 겁니다. 딱 두 번, 진지하게 학위를 그만 두고 다른 일을 알아봐야겠다는 생각을 했을 때도 김천아 박사가 힘을 주고 붙잡았습니다. 그때 나갔으면 제 삶이 어떻게 흘러가고 있을지 상상은 잘 안 됩니다만, 본인이 말린 만큼 앞으로 인생 안 풀리면 알아서 도와줄 거라 믿습니다. 연구를 하고 연구자로 성장하는 데 있어 정말 많은 도움을 받았다는 점, 이 자리를 빌어 다시 한 번 감사 드립니다. Erik 연구실에서는 CB4856 프로젝트에서 wild isolate 분석을 도와줬습니다. CB4856은 CGC에서 받았습니다.

한국에서 다양한 선충을 채집하고 분석하는 일은 김원주, 이보연, 임성희, 임지선 네 분과 함께 하고 있습니다. 혼자서는 결코 할 수 없을 많은 일을 나눠주셔서 더 빠르게 결과를 얻을 수 있었습니다. 한국을 돌아다니며 선충을 채집하는 일이 여행처럼 즐거웠던 건 모두 같이 일하는 분들이 정말 좋은 분이셨기 때문입니다. 특히 지선 씨는 저랑 다른 일도 함께 하며 겹치는 일이 많아 욕도 많이 하고 싸운 적도 더러 있는데, 연구실 사람들이 모두 알고 있듯 그건 다 제 탓이고 제 잘못입니다. 지선 씨랑 원주 씨랑 같이 일하는 건 정말 좋았고, 앞으로 어떤 일을 하든 잘 되기만 바라며 앞으로도 충성충성 하겠습니다. 선충을 채집하고 모으는 데에는 Andre Pires da Silva 와 Natsumi Kanzaki, 그리고 김영재 선생님과 이복남 선생님도 도움을 주셨습니다. 참고로 김영재 씨는 제 어머니시고, 맨바닥에서 닉테이션 하는 벌레를 주워다 주셨습니다. 이복남 씨는 원주 씨 어머님이십니다.

김혜숙 선생님은 연구실 행정 처리를 너무 잘해 주셨고, 김성경 선생님은 벌레 키울 media를 워낙 잘 만들어 주셨습니다. 두 분 덕분에 연구를 한결 수월하게 할 수 있었습니다.

연구실 구성원 모든 분들께 다시 한 번 진심으로 감사 드립니다.

장혜식 선생님은 ONT를 세팅하는 데 많은 도움을 주셨습니다. 선생님께서 도와주지 않으셨다면 ONT를 도입하는 건 시도도 못했을 겁니다. 남궁석 선생님도 장혜식 선생님과 함께 informatics 공부를 해야 한다고 등 떠밀어 주셔서 정말 감사했습니다. 초반에 컴퓨터 공부를 할

때에는 Linux에 대해서 아무것도 몰랐는데, 친구인 김준성 정일채 등이 많이 도와줬습니다. 무엇보다 Google의 도움을 많이 받았습니다.

연구비는 삼성재단과 포스코재단에서 지원을 받았습니다.

마지막으로 (여자)아이들, 특히 서수진 선생님께 감사 드립니다. 논문 한창 쓸 때는 몇 달 동안 내내 쉬는 날 없이, 잠자는 시간 빼고는 연구실에 틀어박혀서 코딩하고 벌레 잡는 일만 했습니다. 너무 지쳐 있던 때라 힘을 내기는 커녕 아침에 눈 뜨는 일조차 버거웠던 시기였는데, 그럴 때마다 너무나도 멋있는 (여자)아이들 공연 영상을 찾아봤습니다. 영상을 보면서 나도 열심히 살아야지, 그래야 돈 벌어서 갖다 바치지, 라는 생각을 하며 힘을 냈습니다. 만약 (여자)아이들이 없었다면, 그래도 논문은 끝낼 수는 있었겠지만, 실제로 마무리하는 데 들었던 시간보다 적어도 두 배는 더 오래 걸렸을 것 같습니다. 앞으로도 꾸준히 활동하시면서 돈 많이 벌고 행복하시길 정말 간절히 바랍니다.