이학박사 학위논문

Pangenome Reference and Missing Genomic Regions : Human and *Mycobacterium tuberculosis*

참조 게놈의 손실된 유전체 발견 및 범유전체 참조게놈:

인간과 결핵균을 중심으로

2021 년 8 월

서울대학교 자연과학대학

생물정보학 전공

김 지 나

Pangenome Reference and Missing Genomic Regions : Human and *Mycobacterium tuberculosis*

참조 게놈의 손실된 유전체 발견 및 범유전체 참조게놈:
인간과 결핵균을 중심으로

지도교수 성 주 헌

이 논문을 이학 박사 학위논문으로 제출함

2021년 8월

서울대학교 자연과학대학
생물정보학 전공

김 지 나

김지나의 박사 학위논문을 인준함

2021년 8월

위 원 장 원 성 호

부위원장 조 성 일

위 원 성 주 헌

위 원 김 창 훈

위 원 한 규 동

Abstract

Pangenome Reference and Missing Genomic Regions : Human and *Mycobacterium tuberculosis*

Jina Kim
Interdisciplinary Program in Bioinformatics
The Graduate School of Natural Science
Seoul National University

DNA sequencing is the pivotal point of mordern biology. To accomplish cost-efficiency, the re-sequencing approaches based on reference genomes are use by the vast majority of sequencing platforms. Because reference genomes play an important role in mapping short reads and detecting several variants on next generation sequencing (NGS), there are reference genomes in several species. For example, in humans, GRCh (human reference genome of the Genome Reference Consortium) has been the reference genome since the Human Genome Project. H37Rv, the most studied strain, has been used as the reference genome in *Mycobacterium tuberculosis*. It was previously thought that determining individuals' genetic variants would require only a single global reference genome. However, there are some skepticism whether reference genomes are truly representative of all individuals in a given species. Many researchers have pointed out the diversity of structural variation among different ethnic or lineage groups and reported novel sequences that are not present in the reference genome but are present in at least a few individuals or strains. In the sequencing process, this could bring about missing or limited information through "unmapped reads" or incorrect variant calling so on. This study attempts to bridge the gap and identify missed genomic regions of the reference genome in human and *Mycobacterium tuberculosis*.

In human genome, this study used a highly contiguous ethnic genome assembly (AK1) to complement missing parts in the human reference genome (GRCh38), which consists of genomes from >50 individuals including those with African ancestry. To find the missing

regions on GRCh38, this study directly compared the reference genome (GRCh38) with the AK1 and using “unmapped” reads of fourteen individuals’ whole genome sequencing data (5 East Asian, 4 European, and 5 African ancestry).

The direct comparison between GRCh38 and AK1 was performed with chain file, which describes a pairwise alignment that allow gaps in both sequences. Another way of using unmapped reads were newly re-aligned to AK1. Each way discovered 3,333 unique genomic regions (size > 200 bp) of AK1 as compared to GRCh38 and 38 estimated missing regions (by ≥ 7 individuals’ unmapped reads) that did not exist in GRCh38. In using unmapped reads, the average 0.90% of the unmapped reads was newly re-aligned to AK1. Furthermore, the alignment rate for East Asian was 0.95%, which was higher than other ethnic groups.

For further research on the estimated missing regions, which were defined as unique AK1 genomic sequences aligned by seven or more individuals’ unmapped reads, this study analyzed the sequences with BLASTx to identify the suggested functional roles of the sequences and Repeat Masker to take a look into the repetitive characteristics of the AK1 regions.

In *Mycobacterium tuberculosis*, this study was performed using another method to complement the missing parts in the reference genome. New pan-genome sequences of *Mycobacterium tuberculosis’* reference genome (H37Rv) were constructed. To build alternative sequences on H37Rv, this study assembled sequences (gap size > 50 bp) of 176 complete genome assemblies and “unmapped” reads of 724 whole genome sequencing data (de novo assembly). 454 contigs were finalized as pan-genome sequences after quality control. To identify the effects of constructed pan-genome sequences, this study analyzed alignment and variant calling results as compared to using only H37Rv.

Finally, this study provides more understanding for reference genome and sequencing. Also, this study raises the need for further investigations on the missing regions of reference genomes in human and *Mycobacterium tuberculosis* and illuminates the possibility of bridging the gap in the reference with using genome data of *Mycobacterium tuberculosis* as a practical example.

Keywords: reference genome, human, *Mycobacterium tuberculosis*, missing information

Student number: 2015-30119

Contents

List of Tables

List of Figures

List of Supplementary materials

Chapter 1.

Introduction

1.1. Overview of sequencing technology

The method of determining the sequence of the four bases on DNA is known as DNA sequencing. Since 2003 when the human genome project completed, the technologies of genome sequencing have advanced and led to a reduced cost. For a long time, a vast number of researchers have worked to develop technologies that allow DNA and RNA sequencing. Sanger sequencing is the basis of DNA sequencing, which was developed by Fred Sanger and based on the detection of DNA fragments with two-dimensional fractionation(1). After the Sanger sequencing, Next-generation sequencing (NGS) , which is the deep, high-throughput, in-parallel sequencing technologies, had evolved over the few decades (2, 3). Unlike the Sanger method, the NGS technologies provide rapidly high-throughput from multiple samples and massively parallel analysis. They would be able to sequence billions of DNA nucleotides in parallel, lowering the need for the fragment-cloning approaches used in Sanger sequencing. The time needed to make the gigabase-sized sequences by NGS was reduced from many years to only a few days or hours, with an massive price reduction(4).

Over time, third-generation single molecular sequencing technologies have been introduced that could compensate for the shortcomings of next-generation sequencing; The third-generation sequencing can generate significantly longer reads than second generation sequencing (5). Such a benefit has important implications for the research of biology in general. However, at start, third generation sequencing had far higher error rates than previous technologies, which could make interpretation of downstream genome data difficult. In current, several companies such as Pacific Biosciences and Oxford Nanopore Technology have continually developed to alleviate sequencing error rates. PacBio invented the sequencing platform of single molecule real time sequencing (SMRT), which is built on zero-mode waveguides properties (6). Each nucleotide inserted by a DNA polymerase attached to the bottom of the zL well emits fluorescent light. The

sequencer captures light signal and determine the sequence of DNA. On the other hand, Oxford Nanopore's technology makes a DNA molecule move into a Nano-size pore structure and then calculates changes in the electrical field around the pore (7).

In recent, the techniques of various sequencing platforms have undergone big improvements of decreasing error rate and having more longer reads. For example, Pacbio HiFi(high-fidelity) reads are generated by multiple passes of the enzyme around a circularized template, which called the circular consensus sequencing (CCS). HiFi reads provide base-level resolution with >99.9% single-molecule read accuracy(8). Oxford Nanopore also launched R10. R10,a new design of nanopore, has a longer barrel and dual reader head to improve resolution of homopolymer and accuracy(9). Although interpretation of sequencing outcomes still presents many drawbacks(3, 10, 11), the increasing number and diversity of sequencing platforms are making progresses in advanced biological and medical research.

1.2. *De novo* assembly vs. Resequencing

1.2.1. *De novo* assembly

After production of sequencing data by sequencers, the data has undergone several steps. Because the sequencing platforms cannot read the whole genome sequence at once, the various steps are necessary for determining the organism's genome sequence. The process of determining the order of genome data can be divided into two, *de novo* assembly or resequencing(12).

De novo assembly is usually performed by assembling individual sequence reads into longer contiguous and correctly ordered sequences without reference sequence. This approach reconstructs the initial sequence of DNA from fragmented reads. In terms of

complexity and time requirements, the *de novo* assemblies requires more time and computation than resequencing (13). This is largely because of the fact that the assembly algorithm necessity comparing every read with every other read. In *de novo* assembly, the most difficult problem is that there is no exact answer. Because of no answer, the approach is time consuming for comparing the reads. However, if there is no problem of time-consuming and computation, it allows us to obtain various genomic data that is not stereotyped and to perform downstream analysis more accurately(12, 14).

1.2.2. Resequencing

Another method dealing with sequencing data is “Resequencing”. This method use aligning reads against a backbone sequence. By comparing the reference genome sequence, the method allows new variations and sequences of genes to be found (15, 16). This method requires already representative genetic information (It is called reference sequence) of living organisms that are significantly closer to study. This is because it is necessary to align genomic sequences against the reference sequence. The reference sequences should be available within the same species. Aligning short reads against a reference sequences and finding SNPs is generally done in resequencing method. It is much easier to discover what makes the new genome different from the reference genome than to construct the new genome (17, 18). Each new organism's sequencing reads are mapped to the most similar part of the reference genome and placed there, which is called “Alignment step”. Following the read alignment and quality control stage, the variant discovery steps are carried out, which

include identifying variations between the mapped reads and the reference genome of the species.

1.2.3. Sequence alignment

In the field of bioinformatics, sequence alignment is a dynamic research area when dealing with sequencing data. It also plays an important role since it helps many tasks such as phylogenetic analysis, functional prediction, and structure prediction of biological many molecules (DNA, RNA, and Protein). In bioinformatics, sequence alignment is commoly the first stage in determining an unknown sequence. Aligning the unknown sequence with existing sequences from database helps to predict the functional and structural role of unknown sequences (19).

The optimal alignment is maximizing the number of identical or related residues that are matched (20). During alignment, the process of rearrangement may be done by inserting several gaps or spaces in the alignment of sequences. The gaps or spaces implies the potential loss or gain of a residue. Those present insertions or deletions (INDEL), translocations, and inversion in the genome sequences.

In sequence alignment, there are two types, pairwise sequence alignment and multiple sequence alignment according to the number of sequences. While the pairwise sequence alignment considers two sequences, multiple sequence alignment considers multiple sequences. The pairwise sequence alignment ,in particular, has three methods; dot-matrix methods, dynamic programming, and word methods (20).

To begin, the dot-matrix approach is quite simple, though this approach takes a long time to analyze large data. The approach can help to visually identify some sequence characteristics, such as INDELs, repeats, or inverted repeats, from a dot-matrix plot.

For example, there are two sequences A and B. The sequence A is written on the top of the dot matrix and sequence B written vertically on the left side of the matrix. If characters of sequence A and B are same, put the dot where the character of sequence A and sequence B match. This approach clearly presents the concordance between the two sequences that are closely related. But, the main drawback of this approach does not provide optimum alignment.

Second, the dynamic programming approach is used to obtain the best alignments. The dynamic programming methods are classified into local alignment and global alignment. The Smith-Waterman method is used for local alignment, while the Needleman-Wunch method is used for global alignment. Needleman-Wunsch (global alignment) uses one diagonal value, a second for match or miss match, and a third for gap penalty, while Smith-waterman (local alignment) uses four values, including zero. When comparing between the sequences with different lengths, local alignment is generally performed because global alignment is done over the entire length of the sequences (21).

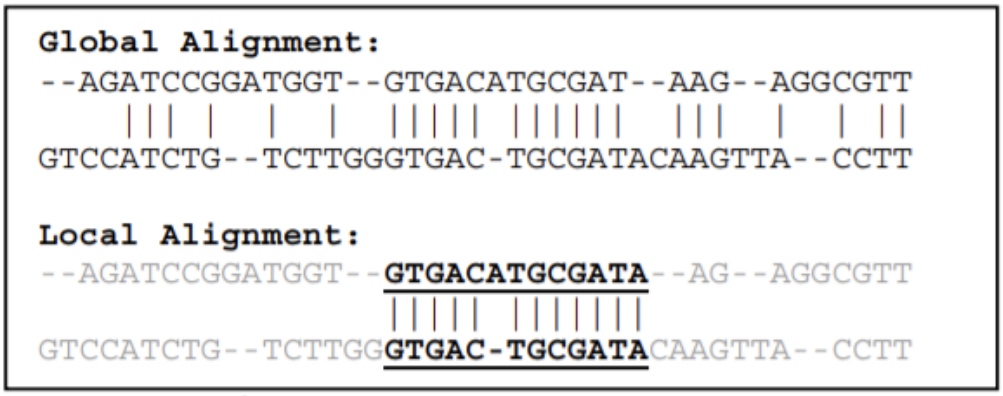


Figure 1. The illustration of comparison between global alignment and local alignment – adopted from Scott E.Coull et al.(2003)

Finally, the word (or K-tuple) approach is a heuristic method that is performed in large databases or alignment. A k-tuple is a k string, which is also known as a k-mer. The K tuple approach is used in the BLAST family(20). The user specifies the length of k word when using BLAST to scan a database, and this method is quite fast.

1.3. The usage of the reference genome in sequencing data analysis

1.3.1 Reference genome

- Human

Currently, the vast majority of sequencing technologies involves mapping massive short reads to reference genome, the GRCh38/39 human genome assembly. After the Human Genome Project, which published the initial draft of the human reference genome in 2001, the reference genomes have updated until GRCh38/GRCh39 (22, 23). The genome was made up of sequences from about 20 volunteers, who were anonymous. The draft of human reference genome was constructed as a mosaic of these sequenced individuals, and had about 150,000 gaps in the sequence size of a 2.69Gb (22). Since 2001, the reference genome has undergone several significant updates. The GRCh38, now in its current form, has been further enriched by adding genomes from more than 50 individuals including people of African ancestry (24), and the genome has only 349 gaps in the sequence size of 2.95Gb. The filling in gaps, replacing rare alleles with the common variants, and adding alternative sequences representing the genomic information of diverse population have been performed in the process of updating reference genome. However, the reference genome's fundamental genetic information has remained the same as in the draft version included genetic background of a small number of anonymous individuals. A study performed comparison between human reference genome of anonymous individuals and Neanderthal genome with the original BAC information and determined the ancestry of each donor with population-specific SNPs (25). The results in this study discovered that nearly two-thirds of the reference genome sequence consisted of RPCI-11, and genome of the sample was almost composed of genomic components of African and European(Figure 2).

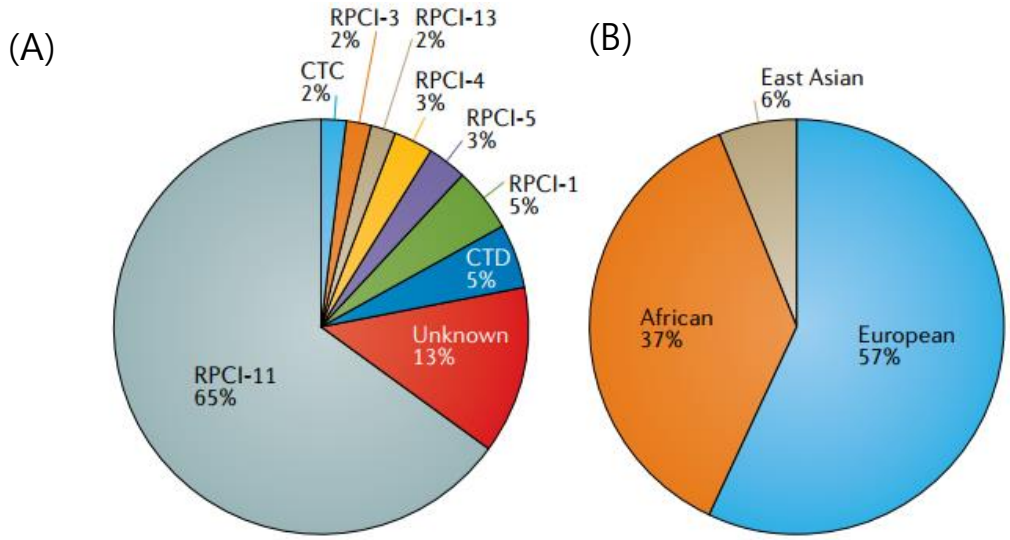


Figure 2 (A) BAC clones in human reference genome (% of total BACs) (B) Inferred ancestry make-up of BAC clones in human reference genome – adopted from Rachel M. Sherman and Steven L. Salzberg (2020)

- *Mycobacterium tuberculosis* (hereafter *'M. tuberculosis'*)

In *M. tuberculosis*, H37Rv was the first strain sequenced by whole genome sequencing and is reference genome of *M. tuberculosis*. H37Rv was named after H37, the initial strain obtained from a patient of the Trudeau Institute in 1905. The genome of the strain was sequenced, and published as a main research in the tuberculosis field (26). Since the research, H37Rv as a reference strain has been used widely in *M. tuberculosis* research. In genome-based studies, the genome of H37Rv has been importantly used in many studies on the identification of drug resistance variants(27), *M. tuberculosis* phylogeny (28), and molecular epidemiology (29). The reference genome of *M. tuberculosis* is one of the most finely curated of any bacterial species, and is linked with a number of resources such as transcriptomic, proteomic, and functional data(30). In addition, a research paper reported

that the selection of reference genome in *M. tuberculosis* has minimal effects on several phylogenetic and epidemiological results with showing alignments of 162 whole genome sequencing data to 7 reference genomes(31).

1.3.2. The shortcomings of reference genome

Although reference genome facilitates to help handling sequencing data, all reference genomes are unlikely to be perfect. In human, most scientists use human reference genome for approximately all human genetics studies. However, it is necessary to know that it may not represent all variant combinations that exist in any individual despite a mosaic genome of many individuals. Recently, the researchers find it considerable that their findings discovered various structural variations among ethnic groups (32, 33) and has raised concerns about whether any portions of the DNA sequences are missed by the recent resequencing methods(34, 35). Researchers have started to recognize the many constraints that a single reference genome imposes upon genomic analysis of various population as the number of sequencing data has increased dramatically(36-39). Actually, when using current reference genome (GRCh38 or GRCh39), the lack of reference genome could result in a large number of unmapped reads in alignment step(40, 41) (Figure 3).

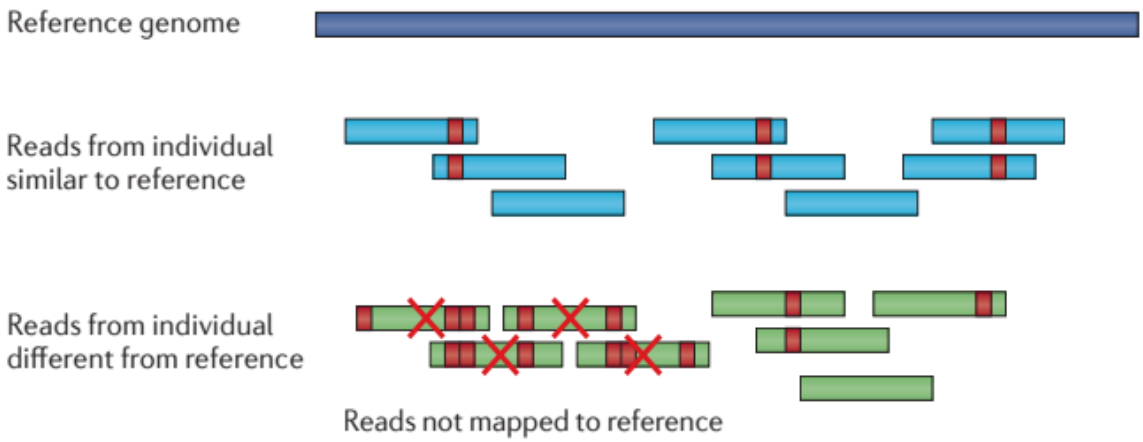


Figure 3. The occurrence reason of unmapped read using reference genome – adopted from Vitor Sousa and Jody Hey (2013)

Also, in *M. tuberculosis*, shortcomings have been discovered on the usage of H37Rv reference genome when comparing with genomes of clinical isolates of *M. tuberculosis*. For instance, a study provided information that H37Rv did not cause caseous necrosis in mice or production of multinucleate giant cells in comparison to other clinical isolates of *M. tuberculosis* (42). Because of the observed differences between strains in other researches, some researchers have suggested the usage of H37Rv as the reference could be inaccurate in pathology (43, 44). The H37Rv strain's various characteristics including genetic components could restrict its robustness in studies involving *M. tuberculosis* pathogenicity.

1.3.3. The efforts to bridge the gap on reference genomes

As we mentioned before, there is the fact that the differences in what genes exist often have effects on pathogenicity, drug resistance and other phenotypes. Due to this problem, the reference genome should be representative with as much information as possible not

to be biased the analysis of new sequencing data. Thus, some studies have performed several efforts because they thought one reference genome is insufficient (45).

One of the attempts is the concept of a pan-genome. This idea was first presented in 2005 (46). It described a pan-genome as a “core genome including genes present in all strains and a dispensable genome consists of genes absent from one or more strains and genes that are unique to each strain”. Another attempt is inclusion of alternative sequences (47, 48). This concept has been the most commonly used in the development of human reference. By including various human sequences in a genetic analysis, human reference genome has been improved. Despite the fact that their inclusion does not capture all human variations, several hundred of these alternative sequences are already present in the human reference genome.

These methods and attempts have still been on going to bridge the gap on reference genome since the studies dealing with sequencing data could miss various results. The recent scientists take efforts to improve reference genome in many species, from human to *M. tuberculosis*. The efforts will increase power of research to connect variants and diseases to human diversity.

1.4. Objectives

In this study, we aimed to study putative missing regions on the reference genome of humans. We also aimed to study *M. tuberculosis* and illustrate the *M.tuberculosis* genome as a practical example for supplementing reference genomes. In order to fulfill the aim, for the human genome, we first performed new methods for finding missing regions and verified the found missing regions through experiments. Second, we investigated several characteristics of the putative missing regions by predicting functions and searching for repeated sequences. Third, for *M.tuberculosis*, we identified missing genomic sequences

on the reference genome of *M. tuberculosis* with the same method used for finding missing regions in the human genome. Furthermore, we constructed pan-genome sequences with the found missing genomic sequences to show an illustration of supplementing reference genomes, and investigated the possibility of the constructed pan-genome sequences on the reference genome of *M. tuberculosis*.

1.5. Outline of the thesis

This thesis is organized as follows:

Chapter 1 introduces this study and the general background of NGS and reference genomes. This chapter presents that the characteristics of current sequencing technologies and using reference genome. Chapter 2 contains identification of putative missing genomic regions in human reference genome with using highly contiguous genome assembly, AK1. Chapter 3 describes the repetitive features and the estimated functional role of the missing genomic regions presented in the previous chapter. Chapter 4 deals with the process of constructing pan-genome sequences and its effects to complement current reference genome of *M. tuberculosis*. Finally, Chapter 5 presents the summary and conclusion.

This chapter was published in *Genes*
as a partial fulfillment of Jina Kim's Ph.D. program.

Chapter 2.

Finding Missing Regions with Human Reference Genome

2.1. Introduction

Since the human genome project was launched, large-scale genomic analysis has become increasingly common. Because resequencing depends on a reference to determine the genomic variations of individuals, the common idea has been that a single reference genome was satisfactory. However, some studies identify the substantial diversity of structural variation among ethnic groups (49, 50). This point has led to questions about whether some portions of human genomic information are missed by the current resequencing methods (34, 35). As various efforts started to solve the problems, several studies collected missing information and identified ethnic specific alternative sequences such as African (51), Danish (34), and Chinese (52). Under efforts to find the missing regions and to discover as many human SNPs(51) and structural variants as possible, some researches have used the *de novo* assembly of “unmapped” reads, which fail to align to the reference, from the RNA (35) and DNA sequencing data (37, 52, 53) or other studies discovered missing regions with long read sequences comparing to GRCh38 and identified the possibility on the usage of the long sequences as a reference alternative patch by finding new structural variants (54) and alternate alleles (55).

In this chapter 2, we first performed to compare the two human genome assemblies, GRCh38 and AK1 (56), with high contiguity, and described the differences between the two assemblies. Also, we used the “unmapped” reads from whole genome sequencing data to further specify the putative missing parts by re-aligning “unmapped” reads to AK1. In the chapter 3, after searching missing regions, we searched for the putative functions of the missed genomic regions and investigated characteristics on repetitive patterns of the missing regions. Therefore, by exploring the common missing regions in two chapters, we

addressed the necessity and potential of “pan reference” for bridging the gap of one reference genome.

2.2. Materials and Methods

2.2.1. Genome assembly data and making chain file between genome assemblies

There are several human reference genomes such as Hg19, GRCh37, and GRCh38 so on. Because the location of genes on each reference genomes varies slightly, liftover step, which converts genome position from one genome assembly to another genome assembly, is necessary (Figure 4). The chain file is also important in the liftover step because this file explains a pairwise alignment that allows gaps in both sequences. By the LASTZ program(57), we generated a chain file between two genome assemblies, the AK 1(https:// www.ncbi. nlm.nih .gov/assembly/ GCA_001750385.2/) and GRCh38 patch 12 including ALT sequences (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38/). When we made a chain file, we first used the written parameters (-gapped -gap = 600,150,-hspthresh = 4500,-seed = 12of19 –notransition -ydrop= 15000) in AK1 article (56). To enhance the reliability of our LASTZ results, we consider the parameters of set1 to use when several sets of parameters were performed and each result of the several sets was compared(Table S2). To generate the chain file after LASTZ, we did the chaining and netting process by UCSC Kent utilities (https://github.com/ ENCODEDCC/kentUtils).

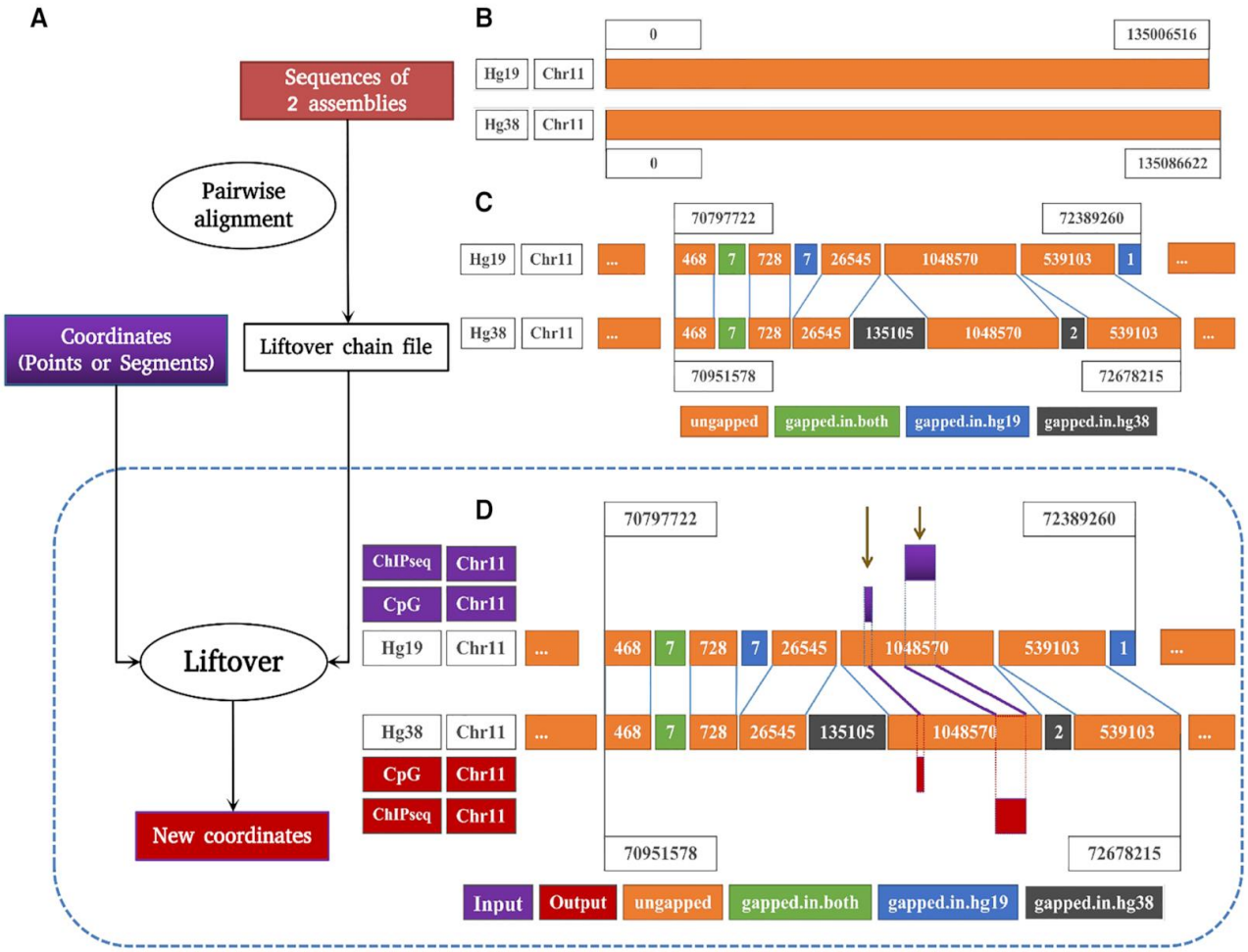


Figure 4. The illustration of the liftover (A) overall steps of the liftover process. (B) An example of discordance between two reference assemblies hg19 and hg38. (C) Ungapped, gapped-in-hg19, gapped-in-both, gapped-in-hg38 regions and the principle of conversion between reference genomes. (D) Results of liftover CpGs and ChIP-Seq data on the ungapped region. - adopted from Phuc-Loi Luu et al. (2020)

2.2.2. Comparison between the reference genome (GRCh38) and the AK1 genome with chain files.

After we made a chain file presenting both ungapped and gapped regions, we divided a total of 2,832 scaffolds of AK1 into three groups based on the match percents of alignment on the

chain file (Figure 6).

Group 1: The scaffolds have of ≥ 99% matches with GRCh38 genome (n=945, ~2.70 Gbp in total).

Group 2: The scaffolds have partial (0% < X <99%) matches (n=467, ~165Mbp in total).

Group 3: The scaffolds have no synteny with GRCh38 (n=1,420 ~41 Mbp).

2.2.3. Sample data

Downloading the whole genome sequencing data (bam files) aligned to the GRCh38 full analysis set with HLA sequences (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project), we extracted unmapped reads from the data. All data was made by Illumina HiSeq platforms. The PCR-free procedures were performed in all data sequencing process. We used only deeply sequenced (depth >50X) data of 3 ethnic groups. The data was mapped to GRCh38 with BWA-MEM(version bwakit-0.7.12.) (58), and underwent quality control procedures including sorting, marking duplicates, and realigning INDEL. The written quality control processes in 1000 Genome ftp server (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project) were performed by SAMtools (version 1.2), BioBamBam (version 0.0.191) (59), GATK-3.3-0 (60) and CRAMtools.3.0. The descriptions of the 14 finally selected data are presented in Table 1. The methods and data of this study were approved by the IRB(Institutional Review Board) of Seoul National University (IRB No. E1912/002-009).

2.2.4. The processing of unmapped reads extracted from sample files

We first investigated the characteristics and qualities of mapped/unmapped reads of the 14 multiethnic samples' BAM files by FastQC (61) and RepeatMasker(62). After extracting the

unmapped reads from the BAM files, we re-mapped the unmapped reads against the new genome assembly, AK1, by BWA-MEM (Figure 7). The re-mapped bam files were processed by sorting and getting rid of duplicates with SAMtools (version 1.3) and Picard Tools(version 2.0.1).

In this analysis, we only focused on reads of primary alignments and discarded secondary alignments, which occurs when a read could align reasonably well to more than one place.

To deal with only primary alignments, we performed the “Samtools view -F 256 input.bam” command (-F option is “Do not output alignments with any number in the FLAG field” and 256 is the flag of secondary alignments) on extracting unmapped reads, and removed reads of secondary alignments. To identify the putative microbial sequences from unmapped reads, we also used GATK-pathSeq(63). In the analysis, we calculated the depth/breadth (15) by BEDTools(version 2.25.0) (64) and Samtools (version 1.3). we described coverage and counted depth by genomic positions of output data from BEDTools with R (version 3.4.3).

For further study on the genomic regions located on Group1 scaffolds that are estimated that missing globally, which is defined as common missing regions in seven or more individuals, we investigated the locations of the missing regions in the GRCh38 genome using a chain file (“lifting” AK1 over GRCh38). Also, to investigate regions as missing identified in previous studies, BLASTn search was used and outcomes of BLASTn were filtered with e-value < 10^{-10}, identity >= 70%, and coverage >= 70%

2.2.5. Visualization

The UCSC genome browser(65) and Integrative Genomics Viewer (IGV) (66) were adopted to visualize the merged 14 BAM files. With the tools, it is easy that study could visualize specific regions and speculate features on genomic regions near the putative missing

regions.

2.3. Results

2.3.1. Discovery of missing information with systematic comparison between GRCh38 p.12 and AK1

For discovery of missing information, we used the chain file to perform systematic comparison between the AK1 and GRCh38. Firstly, to enhance the reliability of our LASTZ results, we consider the first used parameters as set1 and several sets of parameters were performed to compare results. Consequently, the sets of parameters didn't make difference each chain files (Table S2).

Based on the ungapped and gapped regions in the chain file, a total of 53.4 Mbp (~1.8%) of the AK1 genome lacks homology with GRCh38 (Table 2). Categorizing GRCh38 genome sequences with sequence types (chromosome; fix; random; unknown chromosome), we calculated matching size between the AK1 scaffolds and GRCh38 chromosomes by sequence type of GRCh38 genome. The Group 1 and 2 scaffolds of AK1 matched simultaneously with multiple chromosomes of GRCh38 amounted to ~22.2 Mbp(~0.76%). The N50 of the third group was 34.6 Kb, although the N50 of AK1 genome data from NCBI was 44.85 Mb. This shows the almost scaffolds have small size among scaffolds of AK1 genome in group 3 (Table 2). Also, among gapped regions, there were 3,333 regions, which of difference between two assemblies was larger than 200 bp.

2.3.2. Profile of the "Unmapped Reads"

In addition to an exact comparison between AK1 and GRCh38, we performed realignment of

“unmapped reads” to find putative missing regions. We chose 14 individuals’ high-depth (>50X) whole genome sequencing data downloaded from the 1000G database. The samples were Caucasians (4 individuals), Asians (5 individuals), and Africans (5 individuals). On average, ~2.6 M out of 54.6M total reads per individual’s bam file failed to align with GRCh38 and its alternative sequences. This value was amounted to ~4.7% of the WGS data. The whole genome sequencing data of Africans had the lowest alignment rate, and that of Caucasians had the highest mapping rate to GRCh38 (Table 1). Scrutinizing characteristics of unmapped reads, the most part of the unmapped reads (~59%) were the “unpaired reads” (Figure 8). This is because there are the differences in sequencing quality between read1 and read2. Considering the quality and components related to sequencing data, technical underlying features to the sequencing platform rather than the flaw of the reference genome are likely to have brought about many unmapped reads. Among the unmapped reads, the quality of the re-mapped reads to AK1 with MAQ >=10 was about 7. This was higher than average quality of general unmapped reads. In the Figure 9 and Table 3, the distribution of the qualities of reads shows a similar distribution of each reads mapped to AK1 and GRCh38, although the qualities of the reads recovered by the AK1 genome are a little lower than those of the reads mapped to GRCh38. This implies that the read quality seems to be low on average but high-quality reads among the unmapped reads are usually re-mapped to AK1. Besides general low quality of unmapped reads, repetitive sequences of unmapped reads were ten times more low-complexity and >2 times more simple repeats and satellites than general reads (Table 4).

2.3.3. Discovery of missing information with “unmapped reads” by realignment to AK1

With mapping quality >10, reads newly mapped to AK1 were average 72 K of the ~2.6 M reads per individual, and there was a very small proportion of reads of microbial origins. The

remapping rates from realignment to AK1 were somewhat low (0.92% or 0.49%) and did not show significant differences between each population (Table 1). The regions with remapped reads amounted to ~0.2% (5.3Mb) of the AK1 genome. The remapped reads to the scaffolds were classified by three groups as described in Figure 10. The majority of realigned reads was harbored on the Group1 scaffolds. However, unmapped reads of the Group3 scaffolds were harbored more broadly than those of other groups (Table 5). Many unmapped reads were also most densely mapped on regions in Group1's scaffold having high homology with chromosomes 19 and 21 (Figure S1). Meanwhile, according to results of remapping "unmapped reads" to AK1, we narrowed down 110 regions where shared by ≥2 individuals with read depth >10X for each and 38 regions where shared by ≥7 individuals with read depth >10X for each. This study considers those regions, which were not on GRCh38, as the estimated missing regions. By performing BLASTn searches with the mammalian genome database, we identified Sixty-four of the 110 recovered genomic regions showed on previous studies (9, 22, 23) (Table S3).

2.3.4. Verification of presence on missing regions by comparing with GRCh38 and experimenting PCR

For further investigation on the regions that are shared by ≥7 individuals (depth >10X for each) and missing globally, the 31 regions located on Group1 scaffolds were searched the locations in the GRCh38 genome to study genomic features near the estimated missing regions, and discovered that the regions have repetitive elements. Because Group 1 scaffolds (≥ 99% homology with GRCh38) could be annotated by comparing with GRCh38, 31 of the 38 common missing regions belong to Group 1 scaffolds could be identified with GRCh 38 annotation when investigating flanked missing regions. The inserted regions were mostly flanked by some repeat elements like *Alu* or LINE elements (Figure 11).

After identification on the regions using chain file, it was necessary to verify the presence of the regions with experiments. To confirm the regions, Kim et al (2020) extracted +- 2KB of adjacent sequences of the 31 regions on Group 1 scaffolds to verify the existence of the regions. The study performed PCR experiments with the DNA of AK1, four Europeans and a chimpanzee. The results showed that 20 out of 31 putative insertions on AK1 genome and 9 regions on chimpanzee genome were corroborated by the experiments. Kim et al (2020) also found that the missing regions on European genomes were polymorphic (Figure 5). For example, European genomes had either homozygous or heterozygous for insertions/ deletions on the regions. As a result, the study revealed that putative missing regions as insertions exist and that each missing part has heterogeneous genomic structure according to ethnicity.

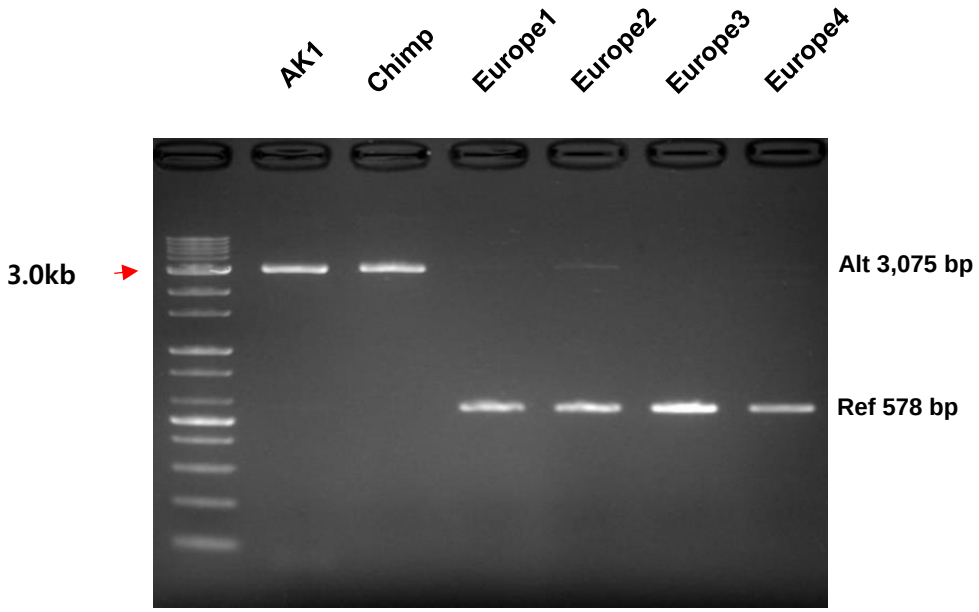


Figure 5. The verification on the presence of missing regions by PCR - adopted from Kim et al.(2020)

2.4. Discussion

As so far, this study found candidate missing regions on GRCh38 using two methods; the comparisons between the reference genome and a precise genome assembly and re-alignment of unmapped reads.

Comparing AK1 with reference genome, this study showed the difference of ~1.8%. offered that genomic similarity between individuals are lower than “99.9% sharing” which was primarily derived from human genome variation projects) and are far higher than similarity derived from the study of African ancestry assembly with unmapped reads (53). This difference may be either conservative or inflated; Considering that GRCh38 consists of genomes of >50 individuals, it may be conservative. On the other hand, considering that scaffolds of Group 3, which might not have been completely identified on GRCh38, have a high proportion of repetitive sequences such as satellite, it may be overestimated. Also, it might be difficult to explain the exact genomic differences due to the different technical sequencing platforms for having performed *de novo* assembly of each GRCh38 and AK1. It is unlikely that the level of quality of the two genome assemblies had largely affected the difference of a 1.8% because both AK1 and GRCh38 were assembled by the factors and strategy of *de novo* assembly technology applied depending on the best possible technology at that time despite differences of the technology.

Besides the way of systematic comparison between two genomes, realignment of unmapped reads was another way to find the estimated missing regions on GRCh38. From the unmapped reads, only a tiny portion of the “missing information” was recovered (<0.2% of AK1 sequences). When two methods of this study were used to identify missing information, the differences of results might be attributable to the high proportion of repetitive sequences in unique AK1 regions and the underlying limitations of the sequencing platform (e.g., extremely large numbers of repetitive sequences among the unmapped reads).

In addition to only finding missing regions *in silico*, Kim et al. (2020), which is mother research of this study, verified the presence of missing information on reference genome. The study also revealed each inserted missing region has heterogeneous genomic structure according to ethnicity. This provided that the regions were incompletely deleted in recent European people.

In conclusion, this study shows the first attempts of mapping unmapped reads against new genome assembly to discover missing information instead of *de novo* assembly with unmapped reads. The attempts suggest the possibility on recovery and usage of unmapped reads which fail to align with human reference genome. This implies that one reference is not perfect and several ways are necessary to complement reference genome.

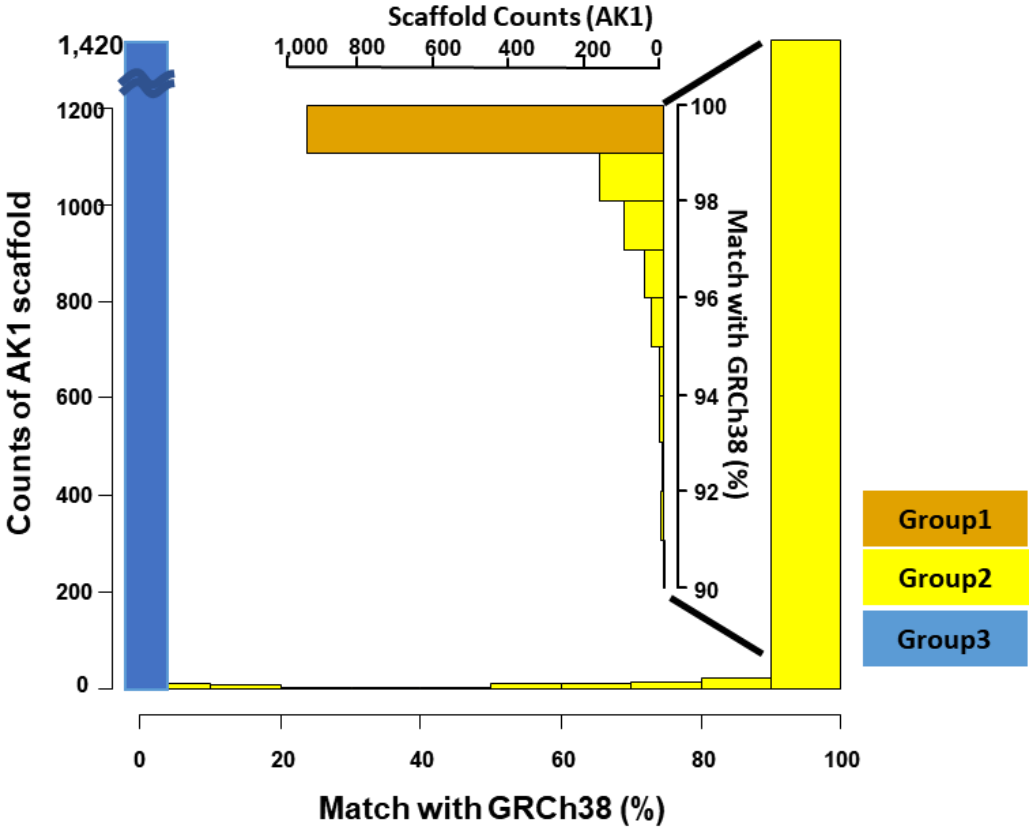


Figure 6. The degree of match divided AK1 scaffolds into three distinct patterns of synteny by chain file. The x axis (and vertical pop-up axis for group 1) represents the percent of matches between AK1 scaffold and GRCh38.p12 chromosomes, and the y axis represents the count of scaffolds.

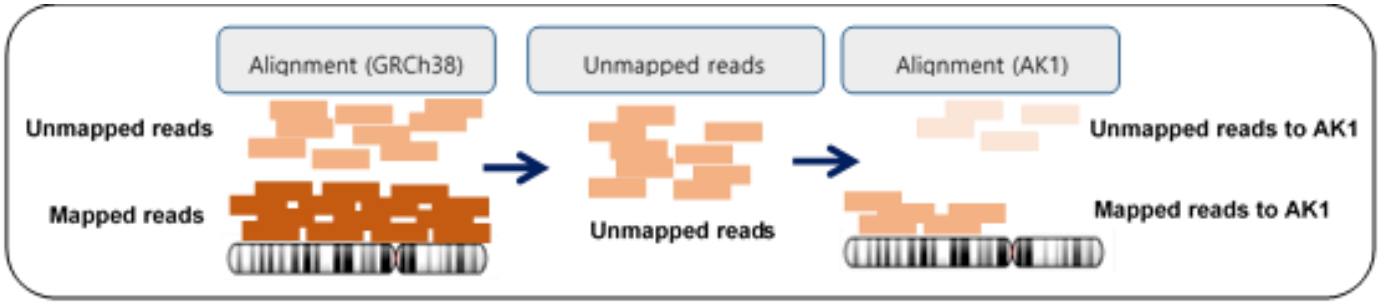


Figure 7. The steps of remapping unmapped reads from GRCh38 reference. After extracting the unmapped reads from the 14 BAM files from 1000 genome database, we used BWA-MEM for re-mapping the unmapped reads of bam files to the AK1 genome.

Table 1. The summary on counts of unmapped reads by samples

Sample ID	Ancestry	Population	Total number of unmapped reads (K)		Unpaired reads, counts (K) (%)	Mapped on AK1, read counts (K) mapping rate (%)*				Suggestive microbial origin, read count
						Overall		Mapping quality >10		
HG02922	AFR	Esan	59,751	Average 42,613	36,871 (61.7)	205 (0.9)	Mean % 0.90	110 (0.5)	Mean % 0.46	318
HG03052		Mende	34,958		21,174 (60.6)	127 (0.9)		67 (0.5)		401
NA19625		African-American SW	48,718		34,396 (70.6)	121 (0.8)		63 (0.4)		353
HG01879		African-Caribbean	35,674		198,064 (55.5)	165 (1.0)		78 (0.5)		1,191
NA19017		Luhya	33,965		20,442 (60.2)	96 (0.7)		56 (0.4)		2,188
HG00419		South. Han Chinese	34,935		22,398 (64.1)	131 (1.0)		66 (0.5)		527
NA18525		Han Chinese	15,620		8,759 (56.1)	51 (0.7)		34 (0.5)		517

HG01595	EAS	Kinh Vietnamese	59,355	**Average 36,474**	31,507 (53.1)	265 (1.0)	**Mean % 0.95**	140 (0.5)	**Mean % 0.51**	3,405
NA18939		Japanese	27,950		15,520 (55.5)	127 (1.0)		66 (0.5)		522
HG00759		Dai Chinese	44,510		21,418 (48.1)	234 (1.0)		117 (0.5)		512
NA20502	EUR	Tuscan	26,343		19,640 (74.6)	57 (0.9)		33 (0.5)		1,557
HG00096		British	29,915	**Average**	16,773 (56.1)	108 (0.8)	**Mean %**	64 (0.5)	**Mean %**	1,878
HG01500		Spanish	31,331	**26,711**	15,726 (50.2)	164 (1.1)	**0.88**	76 (0.5)	**0.49**	2,423
HG00268		Finnish	19,255		12,139 (63.0)	58 (0.8)		36 (0.5)		289
Total Average (Mean±sd)				35,877 ± 13,193	21,184 ± 8,091 (59.0%)		137± 65 (**0.92%)**		71 ±31 (**0.49%)**	1,149 ± 988

Suggestive microbial origin was analyzed by GATK-pathSeq. African-American SW, African-American Southwest

* Mapping rate $= \frac{No.\ of\ reads\ re{-}aligned\ to\ AK1}{(total\ unmapped\ reads\ -unpaired\ read)}$

Table 2. Statistics of the three groups of AK1 scaffolds according to the above matching patterns. Fix, the patches represent changes (error corrections or assembly improvements) to GRCh38 genome.; Random, the unlocalized contigs of GRCh38.

	All	group1	group2	group3
Number of Scaffolds	2,832	945	467	1420
Total scaffold size (Scaffold N50)	2,904 Mb (44.85 Mb)	2,697 Mb (45.09 Mb)	165 Mb (13.74Mb)	41 Mb (34.60Kb)
Size matched with GRCh38.p12 (%)	2,851 Mb (98.2)	2,691 Mb (99.8)	160 Mb (96.2)	0
by Sequence types Chromosomes (or alternative)	2,839 Mb	2,681 Mb	158 Mb	0
Fix	8,047 Kb	7,831 Kb	216 Kb	0
Random	2,783 Kb	1,906 Kb	878 Kb	0
Unknown chromosomes	1,005 Kb	648 Kb	358 Kb	0
Scaffolds matched multiple chromosomes of GRCh38.p12	487	343	144	0
Total size of scaffolds contributed from multiple chromosomes*	22,2 Mb	21,1 Mb	1.1 Mb	0

* Size of sum of minor contributing chromosomes

Table 3. Average mapping quality and depth of mapped reads on GRCh38 and AK1

	Mapped reads on GRCh38	Remapped reads on putative missing regions of AK1
Average mapping quality	**30.8**	**37.6**
Average depth		
>5 Reads count by position	**56.23**	**42.32**
>10 Reads count by position	**56.39**	**56.54**

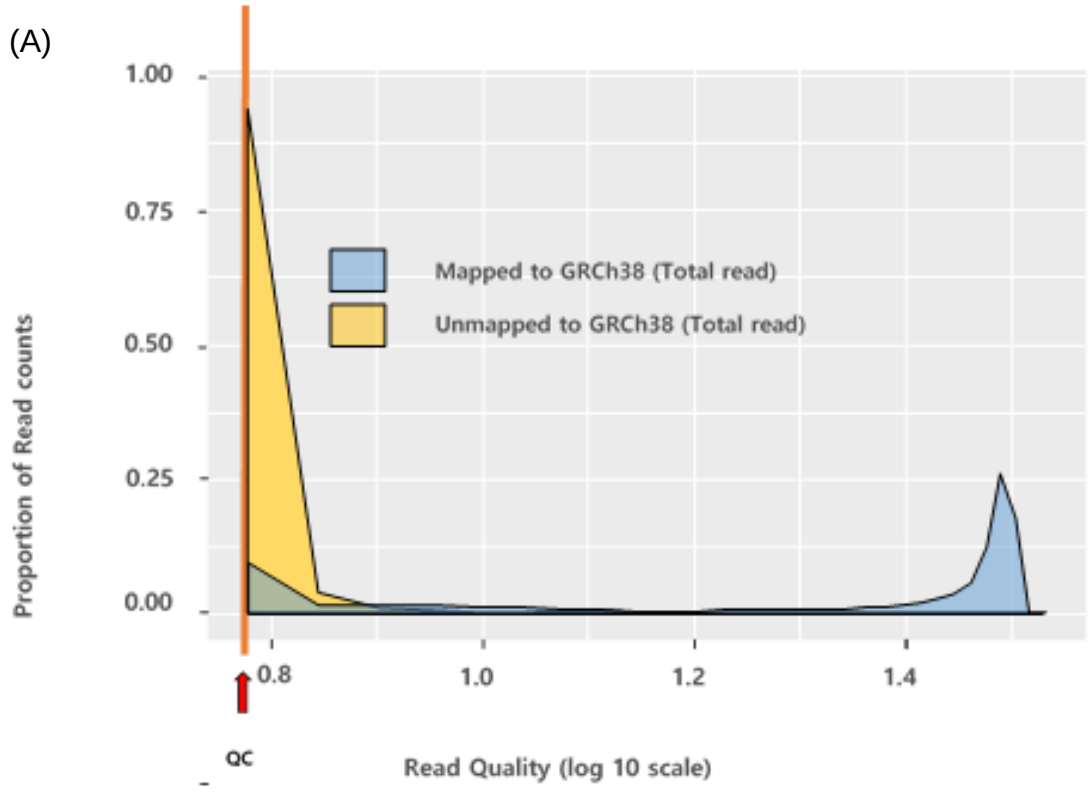


Figure 8. The trend of the read count proportions by read quality of read counts on mapped and unmapped reads on GRCh38. The proportion of read counts on mapped and unmapped groups has different trends by read quality. (A) The comparison of read count proportion by read quality between total mapped reads and total unmapped reads on GRCh38.

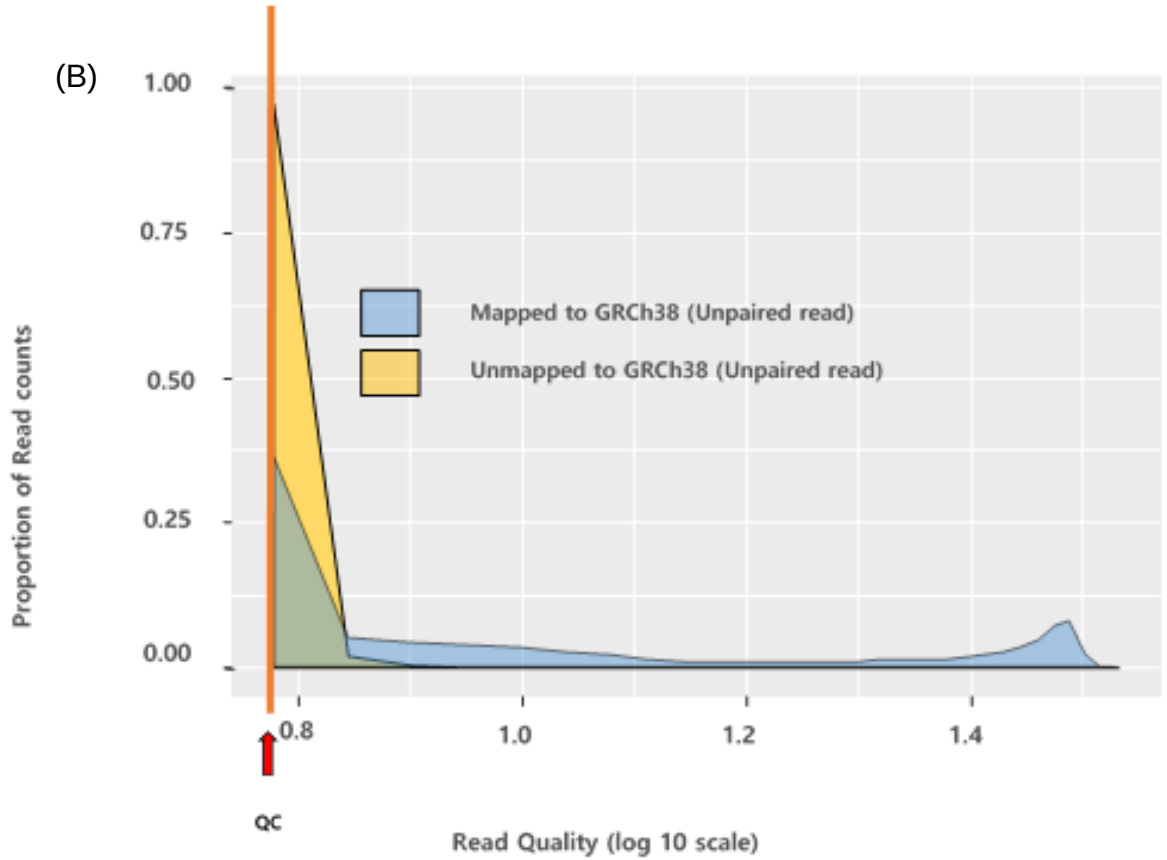


Figure 8. The trend of the read count proportions by read quality of read counts on mapped and unmapped reads on GRCh38. The proportion of read counts on mapped and unmapped groups have different trend by read quality. (B) The comparison of read count proportion by read quality between unpaired-mapped reads and unpaired-unmapped reads on GRCh38

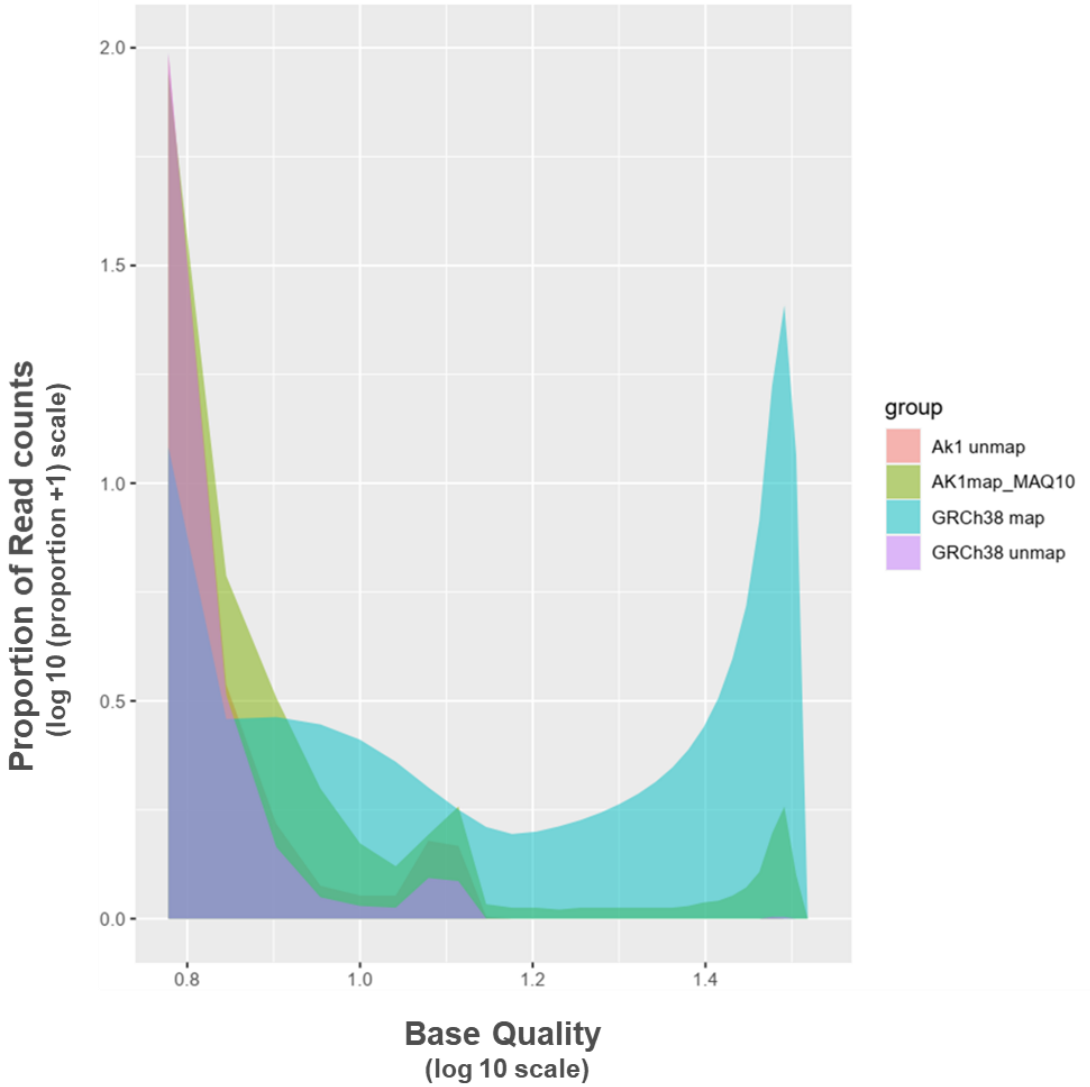


Figure 9. The trend of the read count proportions by read quality of the mapped reads and unmapped reads on each two genome assemblies, GR Ch38 and AK1. This graph compares the read count proportion by read quality of the total mapped reads and unmapped reads on AK1 and GRCh38, respectively. The proportion of read counts on mapped and unmapped groups follows various patterns depending on read quality. The light green and light blue indicate the mapped reads to AK1 and GRCh38, respectively. The light red and light purple indicate the unmapped reads to AK1 and GRCh38, respectively.

Table 4. The distribution of repetitive sequences on reference genome (GRCh38) and sequencing reads from 14 samples by Repeat Masker.

		GRCh38 reference genome (hs38d1+hla sequence)	Unmapped Reads of 14 samples Mean% (SD)
SINE	**All**	11.75	1.92 (1.44)
	ALUs	9.73	1.82 (1.38)
	MIRs	2	0.11 (0.06)
LINE	**All**	18.31	4.43 (1.82)
	LINE1	15.49	4.34 (1.76)
	LINE2	2.52	0.08 (0.05)
	L3/CR1	0.23	0.01 (0.01)
LTR	**All**	7.84	1.70 (0.79)
	ERVL	1.59	0.21 (0.11)
	ERVL-MaLRs	3.18	0.63 (0.31)
	ERV-class I	2.67	0.72 (0.32)
	ERV-class II	0.3	0.14 (0.05)
DNA	**All**	2.82	0.42 (0.21)
	hAT-Charlie	1.19	0.13 (0.08)
	TcMar-Tigger	1.07	0.24 (0.10)
Unclassified		0.2	0.27 (0.17)
Small RNA		0.14	0.1 (0.05)
Satellites		2.15	4.98 (4.99)
Simple repeats		1.25	3.43 (3.03)
Low complexity		0.21	1.98 (1.68)

SINE = Short interspersed elements

MIR = Mammalian-wide interspersed repeats.

LINE = Long interspersed elements

LTR = Long terminal repeat

ERVL = Endogenous retrovirus-L

ERVL-MaLRs = Endogenous retrovirus-L-Mammalian apparent LTR Retrotransposons

ERV = Endogenous retroviruses

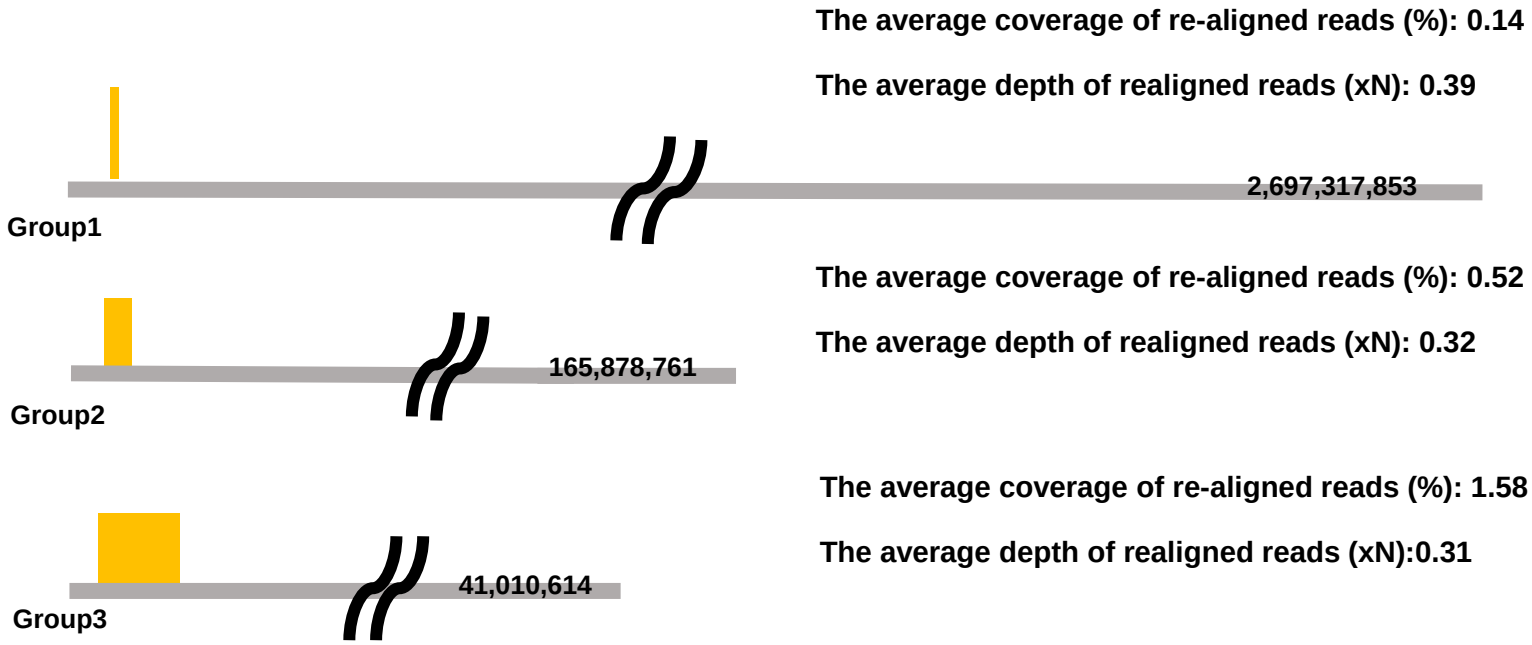


Figure 10. The overall descriptions of "unmapped reads" realigning with AK1. The breadth of coverage and average depth of coverage by position by groups of AK1. The width and the height of the yellow box shows each the breadth of coverage and the average depth of coverage by groups. The average coverage of re-aligned reads (%) = (Breadth of coverage / Total size of synteny)*100, The average depth of realigned reads (xN) = (sum of average depth at position /breadth of coverage)

Table 5. The summary of "unmapped reads" realigning with AK1. The breadth of coverage, average depth of coverage by position, and read counts per individual by groups of AK1.

AK1 scaffold Group	Total size of Scaffolds (Kbp)	Breadth of re-alignment* (Kbp, coverage* %)	Average depth of realigned reads (xN)	Average read count per individual
Group1	2,697,318	3,795 (0.14)	0.39	58,340
Group2	165,879	860 (0.52)	0.32	7,499
Group3	41,011	646 (1.58)	0.31	5,931
Total	2,904,208	5,303 (0.18)	0.34	71,771

*The average coverage of re-aligned reads (mapping quality >10) (%)= (Breadth of coverage / Total size of scaffold)*100,
The average depth of realigned reads (xN) = (sum of average depth at position /breadth of coverage),
Average read count = Sum read counts / a number of samples

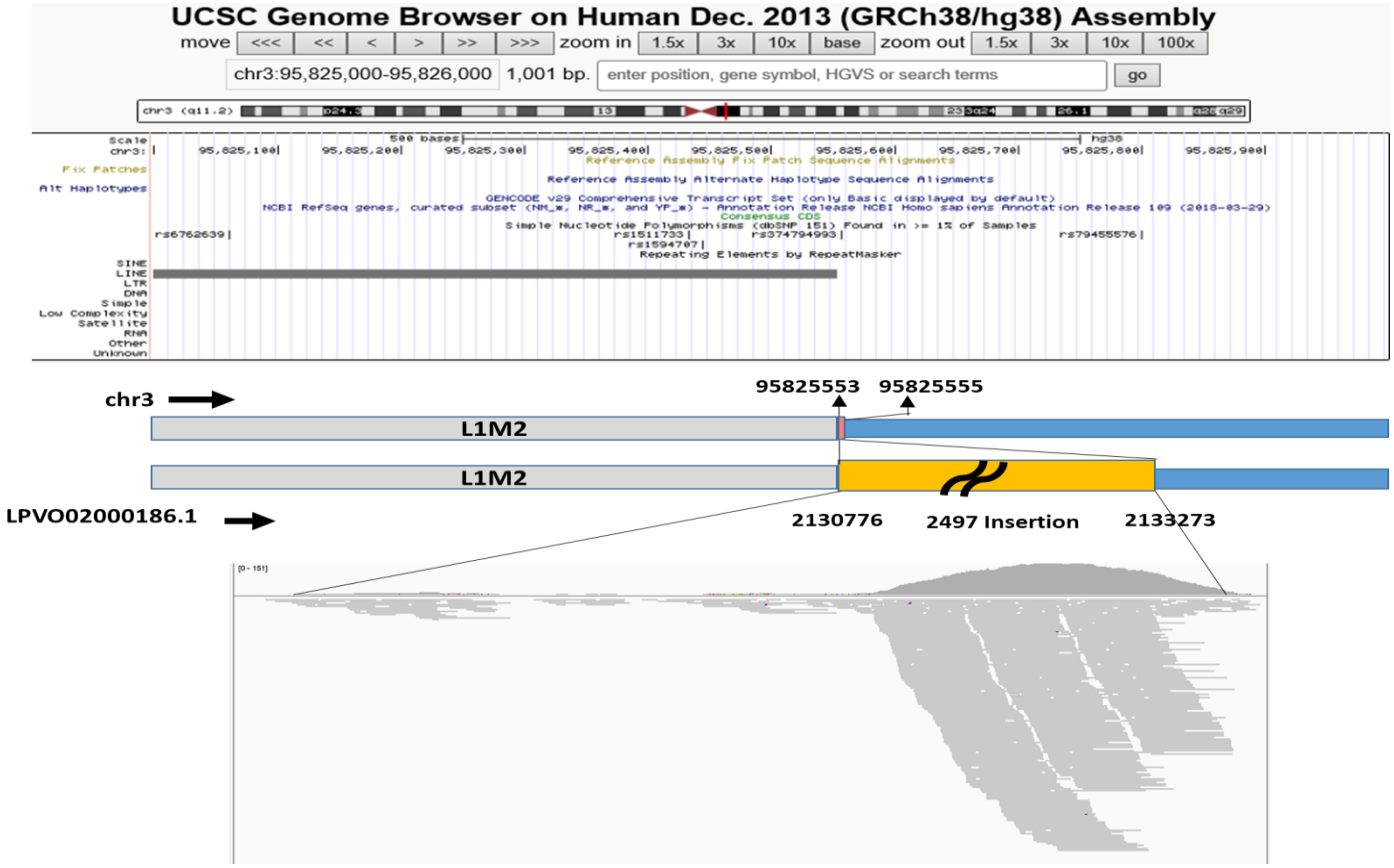
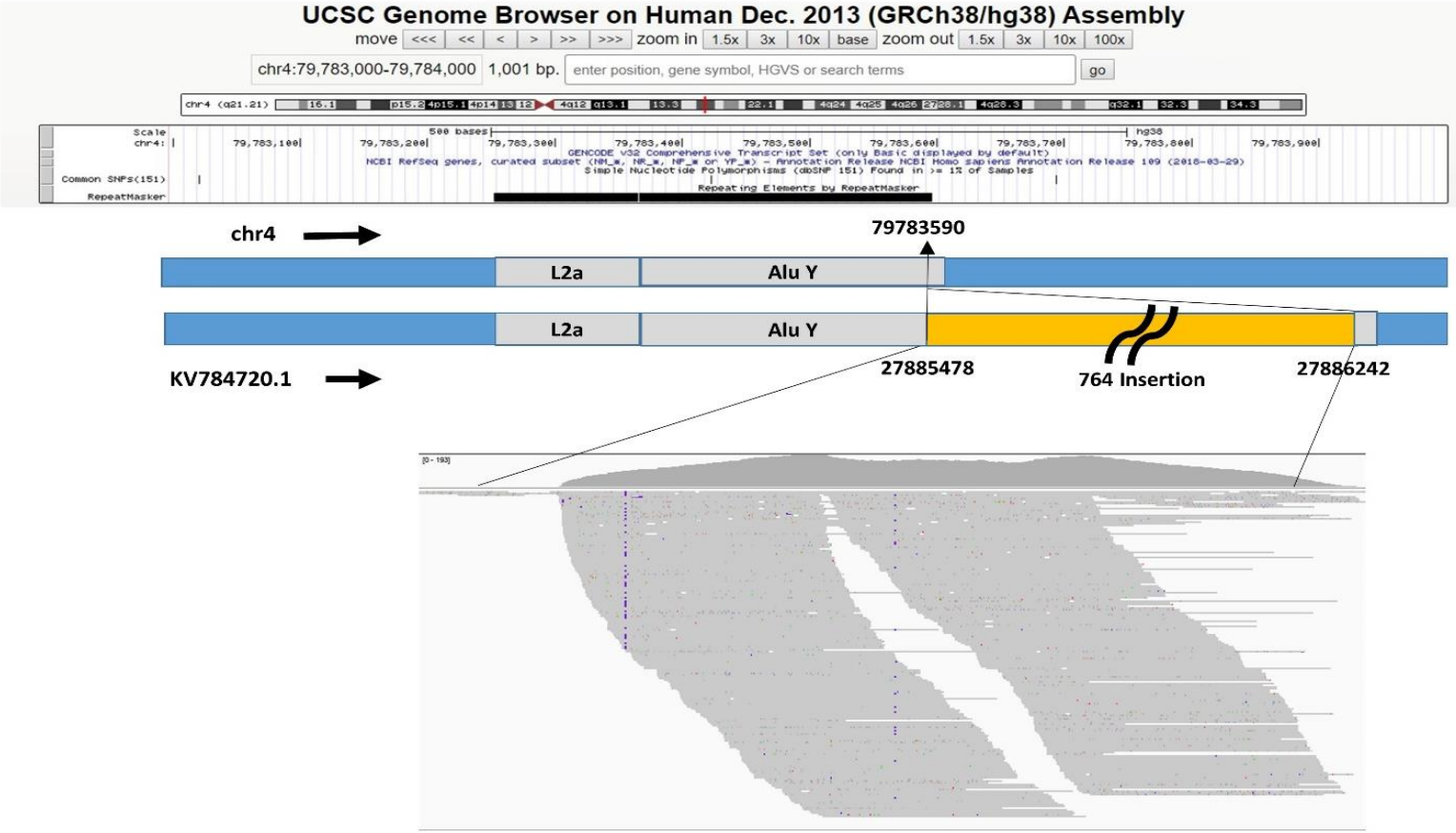
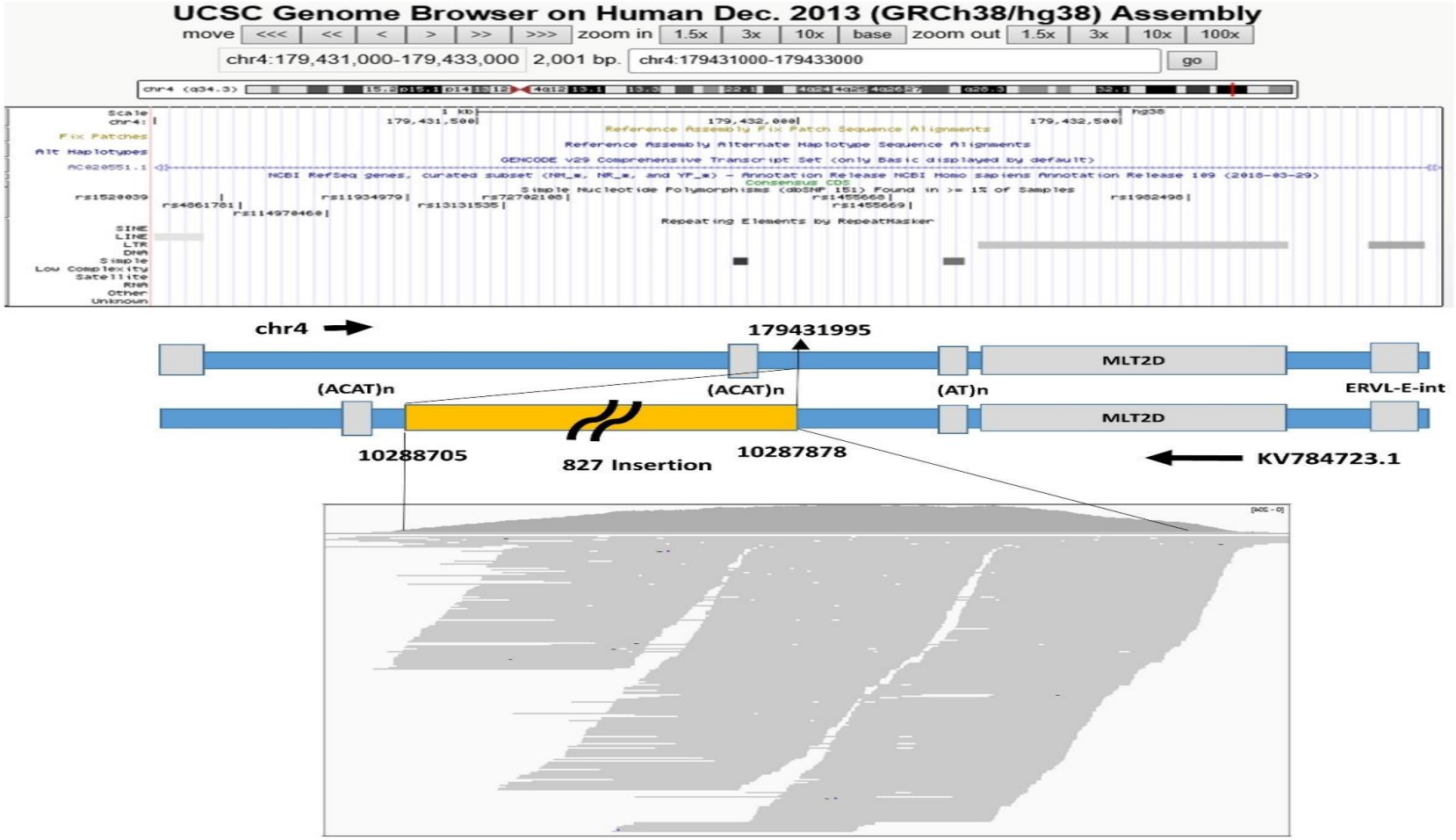


Figure 11. The examples on globally missing regions of GRCh38 investigated with UCSC Genome browser. The 38regions (>=10X,>=7indiv) were found in the inserted sequences(yellow block). (A) The G1-26 region(Insertion into chr3:95,825,553-95,825,555) was near L1M2. The yellow block is the estimated insertion against GRCh38 on the chain file; The grey blocks are repetitive sequences. The purple block is the sequence only on GRCh38 genome.



(B) The region was inserted in *Alu*Y (chr4:79,781,761-79,785,451(G1-5)).



(C) The region was near the repetitive sequences (chr4:179,430,209-179,433,860(G1-7))

This chapter was published in *Genes*
as a partial fulfillment of Jina Kim's Ph.D. program.

Chapter 3.

Characterization of the Common Missing Genomic Regions

3.1. Materials and Methods

3.1.1. Sample data

Genomic data for this study is downloaded which mentioned in the previous chapter. Please refer to Materials and Methods section of Chapter 2 in details.

3.1.2. *In silico* functional search on candidate missing regions - BLAST (Basic Local Alignment Search Tool) search

In this chapter, missing genomic regions identified on previous chapter were further investigated with *in silico* functional analysis. The *insilico* functional study was performed by BLAST search. Generally, BLAST (Basic Local Alignment Search Tool) can be used for comparing nucleotide or protein sequences from sequence public databases and calculating the statistical significance of matches. BLAST search also has various function; BLASTn, tBLAST, BLASTx and so on. In this study, BLASTn and BLASTx were used to discover functional roles and gene families of sequences. This study performed translated BLASTx search to further investigate the estimated missing regions on previous chapter. Especially, this study searched the recovered regions; the regions identified as unique parts to the AK1 (>200bp), the regions where remapped unmapped reads (with a depth >10) of two or more individuals (common missing regions), and the regions where remapped unmapped reads (with a depth >10) of seven or more individuals (globally common missing regions). The searches were performed against the nr database and with default options during BLAST search. After searches, the results of BLASTx were filtered with e-value < 10^{-10}, identity>= 70%, and alignment length >= 50bp.

3.1.3. Identifications of transposable elements for studying the characteristics on missing regions by Repeat Masker

The defined 110 regions and 38 regions were not on GRCh38 but on AK1. Also, the defined regions as well as the third group scaffolds are unique to AK1. Because the genomic regions were not on GRCh38 and newly discovered, the features the genomic regions were not known. For further description, this study analyzed repetitive sequence pattern on the regions with using the RepeatMasker (62).

3.2. Results

3.2.1. Finding estimated functions of missing genomic regions

First, this study found 3,333 regions whose difference between GRCh38 and AK1 was larger than 200bp with using a chain file. When the 3,333 regions were searched through a translated BLAST(67) within mammals to identify protein-coding functions, a 1390 (e-value $<10^{-10}$, identity **≥** 70%, and alignment length **≥** 50bp) of 3,333 regions were predicted to have putative protein-coding elements. Second, with read depth >10X for each, 110 regions where shared by ≥2 individuals and 38 regions where shared by ≥7 individuals were not on GRCh38 and simultaneously considered as the putative missing regions. Remarkably, through the translated BLAST search with NCBI's nr database, 25 of the 110 regions had putative mammalian protein-coding functions. The list of the 25 regions showing putative protein-coding functions is described in Table 6 and the list was filtered with e-value $<10^{-10}$, identity **≥**70%, and alignment length **≥**50bp. In this list, there were a GPALPP motifs-containing protein 1, alternative protein DYZ1L14 and so on. Specially, P150 is a protein that is largely absent or greatly reduced in ovarian cancer, and MYB Isoform 6 is a transcription

factor associated with some human diseases. Also, when BLAST results of the 38 regions (≥7 individuals with read depth >10X for each) was filtered with same filtering criteria, one of the 38 regions searched to have homology with *zinc finger protein 454 isoform 2* (Table 7).

3.2.2. Characteristics of candidate missing genomic regions on the repetitive sequences

When this study use comparison between two assemblies to identify putative missing regions, the group 3 scaffolds had no synteny with GRCh 38, which means unique to AK1. The putative missing regions were analyzed by the RepeatMasker(62) to characterize missing regions. Compared with the proportion of repeat sequences on GRCh38, satellite repeat sequences (> 87%) outnumber a higher proportion of simple repeats (Table 8). We found that the characteristics of missing regions on the AK1 genome is frequently repeated sequences. The percentage of SINE and LINE regions was slightly higher, while the value of low complex regions and simple repeats account for approximately 12% (Table 9).

3.2.3. Identifying the occurrence mechanism of insertions related with missing genomic regions

When many re-alignments of “unmapped reads” were observed, they found on putative insertions of AK1 compared with the GRCh38 genome (Figure 12). To further examine characteristics of on the regions, Kim et al (2020) used BioEdit(68). The results of the study showed the putative breakpoints on the regions and revealed that the most common occurrence mechanism of insertions is nonhomologous end-joining (NHEJ, n=26) with microhomology, followed by nonallelic homologous recombination (NAHR, n=3).

3.3. Discussion

According to this chapter, the characteristics of the common missing parts were revealed. The most of the candidate “globally and commonly missing” regions, which were found with unmapped reads of various populations, might be absent in the GRCh38. On behalf of, the insertions were discovered on the genomes of other populations. This discovery is in line with a previous finding (54).

This chapter also used the BLAST search to identify the functions of the putative missing regions. According to the results, it is revealed that sequences of missing regions might have some functions. However, it was exploratory on the functional search of missing regions, and the functional search was narrow to the coding sequences. Therefore, further functional validation using experiments would be required to confirm the functional missing regions.

In addition to the function of missing parts, the proportion of repetitive sequences on common missing parts is greater and the flanked sequences of the regions had SINE, LINE or repetitive sequences on previous chapter 2 as we mentioned. This feature confirms again that transposable elements would play an important role in population diversity and result in the different structural variation of various population (69).

Kim et al (2020) found the identified insertions were caused by different mechanisms although NHEJ with microhomology was discovered on occurrence mechanisms of many insertions. This means that occurrence mechanisms of common missing regions might also be different although the missing regions were shared in several population genomes. It is necessary to more study the ethnically specific structures and occurrence mechanism with heterogeneity in the common missing regions.

In conclusion, the chapter 2 and 3 showed the approach based on usage of precise ethnic genomes for obtaining missing genomic information. As various precise ethnic genomes frequently appear in the future, they will profoundly provide understandings of complete

genome functions and a precise evolutionary history of humans. Precise ethnic genomes would also help the discovery of other missing information with closing the scientific gap between genomes of various populations

Table 6. The putative proteins of translated BLAST search on the 25 of 110 regions (>=10 reads >=2 sample by position)

Group	Scaffold	Start	End	Putative protein	Species	E-value	Identity(%)
Group1	LPVO02000231.1	510488	533621	chloride channel Kb, isoform CRA_c	*Homo sapiens*	3.60E-36	100
	KV784731.1	15610446	15612082	ubiquitin-conjugating enzyme E2Q-like protein 1	*Pan troglodytes*	3.78E-27	100
	KV784727.1	1910039	1911015	ceramide glucosyltransferase isoform X2	*Homo sapiens*	1.71E-25	92.59
	KV784734.1	23227641	23232362	putative p150	*Homo sapiens*	0	92.07
	KV784772.1	6472482	6482727	putative p150	*Homo sapiens*	9.09E-29	91.55
	KV784725.1	1338204	1338992	hCG2038537, partial	*Homo sapiens*	1.47E-24	89.47
	KV784805.1	52976540	52978099	hypothetical protein EGK_08355, partial	*Macaca mulatta*	1.21E-22	89.09
	KV784803.1	21187527	21189031	putative uncharacterized protein encoded by LINC00269, partial	*Theropithecus gelada*	3.84E-27	87.50
	KV784762.1	941858	944985	hypothetical protein EGK_14950, partial	*Macaca mulatta*	1.10E-27	86.00
	KV784719.1	93468677	93473936	EBPL isoform 3	*Pan troglodytes*	4.70E-26	85.94
	LPVO02000621.1	1215578	1217611	hypothetical protein EGK_08355, partial	*Macaca mulatta*	1.61E-20	85.71
	KV784803.1	15594781	15596171	hypothetical protein EGM_09555, partial	*Macaca fascicularis*	1.10E-19	85.71
	KV784797.1	27752714	27755221	hypothetical protein EGM_17106, partial	*Macaca fascicularis*	2.24E-17	80.77
	KV784723.1	8349058	8350242	GPALPP motifs-containing protein 1 isoform X2	*Piliocolobus tephrosceles*	6.01E-28	80.28
	KV784719.1	93450257	93457305	hypothetical protein EGK_08749, partial	*Macaca mulatta*	1.50E-22	80.00

	LPVO02000621.1	1219247	1220501	hypothetical protein EGK_19543, partial	*Macaca mulatta*	1.43E-27	78.08
	KV784804.1	4078651	4079471	hCG1993336	*Homo sapiens*	2.83E-28	77.78
	KV784811.1	3732844	3735258	hypothetical protein EGK_18118, partial	*Macaca mulatta*	2.03E-29	77.38
	KV784726.1	9049712	9056232	hCG1997218	*Homo sapiens*	2.15E-38	74.23
	KV784734.1	77101351	77102330	PREDICTED: uncharacterized protein LOC103229289	*Chlorocebus sabaeus*	1.68E-21	72.73
Group2	LPVO02000185.1	2178029	2179621	hypothetical protein EGM_17921, partial	*Macaca fascicularis*	1.12E-35	86.84
	KV784763.1	490557	494476	PREDICTED: protein GVQW1-like, partial	*Callithrix jacchus*	1.97E-38	83.95
	KV784740.1	23347990	23348752	MYB isoform 6	*Pan troglodytes*	2.85E-15	72.88
Group3	LPVO02001414.1	24394	28402	hypothetical protein I79_026279	*Cricetulus griseus*	4.90E-30	100.00
	LPVO02001464.1	36803	40809	alternative protein DYZ1L14	*Homo sapiens*	4.87E-18	88.46

Table 7. The results of translated BLAST search of 38 globally missing regions (>=10 reads >=7 sample by position)

Group	AK1 scaffold	Start position	End position	ID	Putative protein	Species	Score	E-value	Align length	Identity (%)
Group1	KV784719.1	30209977	30210924							
	KV784719.1	79001655	79002640							
	KV784719.1	93452303	93455222	gi|1020158921|ref|NP_001310237.1|	*zinc finger protein 454 isoform 2*	*Homo sapiens*	265	5.61E-24	79	74.6835
	KV784719.1	93470705	93471918							
	KV784720.1	27885647	27886104							
	KV784723.1	8349171	8349628							
	KV784723.1	10288012	10288493							
	KV784723.1	34400763	34401227							
	KV784731.1	15610509	15611959							
	KV784736.1	6179476	6184176							
	KV784736.1	18433040	18435697							
	KV784738.1	33432222	33432240							
	KV784747.1	1225842	1227344							
	KV784754.1	50234036	50235663							
	KV784761.1	2374854	2374857							
	KV784762.1	646396	646455							
	KV784762.1	942159	943260							

	KV784774.1	387226	387651	
	KV784797.1	27753978	27754392	
	KV784800.1	13617523	13617941	
	KV784803.1	15594978	15595455	
	KV784803.1	21188206	21188829	
	KV784804.1	4078861	4078900	
	KV784806.1	65330325	65332270	
	KV784811.1	3734091	3735143	
	LPVO02000186.1	2132760	2132810	
	LPVO02000191.1	8716140	8716258	
	LPVO02000230.1	3020537	3020573	
	LPVO02000423.1	11658530	11658908	
	LPVO02000423.1	13811264	13811292	
	LPVO02000621.1	1217413	1217481	
Group2	KV784740.1	23348347	23348356	
	KV784740.1	49263566	49263958	
	KV784763.1	493662	494031	
	LPVO02000309.1	281355	281370	
	LPVO02001070.1	97325	97709	
Group3	LPVO02002730.1	7950	8114	
	LPVO02002730.1	12869	12969	

Table 8. The distribution of non-repetitive and repetitive sequences between GRCh38 genomes and AK1 Group3 scaffolds by Repeat Masker

		GRCh38 reference genome (hs38d1+hla sequence)		AK1 Group3 scaffolds	
		Size (Kbp)	Proportion (%)	Size (Kbp)	Proportion (%)
Non-repetitive sequence		1,780,160	55.33	1,041	2.54
SINE	All	377,884	11.75	100	0.24
	ALUs	313,121	9.73	100	0.24
	MIRs	64,186	1.99	-	-
LINE	All	589,075	18.31	805	1.96
	LINE1	498,467	15.49	805	1.96
	LINE2	81,160	2.52	-	-
	L3/CR1	7,437	0.23	0.05	0.00
LTR	All	252,386	7.84	10	0.02
	ERVL	51,149	1.59	-	-
	ERVL-MaLRs	102,207	3.18	-	-
	ERV-class I	85,823	2.67	2	0.00
	ERV-class II	9,664	0.30	9	0.02
DNA	All	90,580	2.82	0.1	0.00
	hAT-Charlie	38,281	1.19	-	-
	TcMar-Tigger	34,299	1.07	0.1	0.00
Unclassified		6,351	0.20	39	0.10
Small RNA		4,482	0.14	11	0.03
Satellite	All	69,287	2.15	35,998	87.78
	centromere	62,756	1.95	30,824	75.16
	others	6,531	0.20	5,174	12.62
Simple repeats		40,357	1.25	3,003	7.32
Low complexity		6,786	0.21	4	0.01
Total		3,217,347	100.0	41,011	100.0

SINE = Short interspersed elements

MIR = Mammalian-wide interspersed repeats.

LINE = Long interspersed nuclear elements

LTR = Long terminal repeat

ERVL = Endogenous retrovirus-L

ERVL-MaLRs = Endogenous retrovirus-L-Mammalian apparent LTR Retrotransposons

ERV = Endogenous retroviruses

Table 9. The proportion of repetitive sequences and transposable elements on 110 regions (>=10X, >=2indiv) and 38 regions (>=10X,>=7indiv) by Repeat Masker.

		110 regions (more ten reads are mapped in more than two samples	38 regions (more ten reads are mapped in more than seven samples)
		Mean % (S.D.)	Mean % (S.D.)
SINE	**All**	8.01(9.85)	2.54 (5.41)
	ALUs	6.41 (12.25)	0.27 (1.63)
	MIRs	1.6 (6.65)	2.27 (7.37)
LINE	**All**	7.34 (13.35)	3.64 (13.80)
	LINE1	5.13 (15.50)	3.64 (13.80)
	LINE2	2.21 (10.77)	0
	L3/CR1	0	0
LTR	**All**	2.47(4.79)	0.56 (2.50)
	ERVL	0.88 (5.86)	0
	ERVL-MaLRs	0.98 (4.35)	0.56 (2.50)
	ERV-class I	0.60 (3.93)	0
	ERV-class II	0	0
DNA	**All**	0.14 (0.70)	0
	hAT-Charlie	0.14 (0.70)	0
	TcMar-Tigger	0	0
Unclassified		0.48 (5.01)	0
Small RNA		0.05 (0.51)	0
Satellite		8.94 (26.92)	7.85 (26.82)
Simple repeats		17.62 (33.73)	10.82 (29.95)
Low complexity		11.80 (31.59)	0.52 (2.00)

SINE = Short interspersed elements
MIR = Mammalian-wide interspersed repeats.
LINE = Long interspersed elements
LTR = Long terminal repeat
ERVL = Endogenous retrovirus-L
ERVL-MaLRs = Endogenous retrovirus-L-Mammalian apparent LTR Retrotransposons
ERV = Endogenous retroviruses

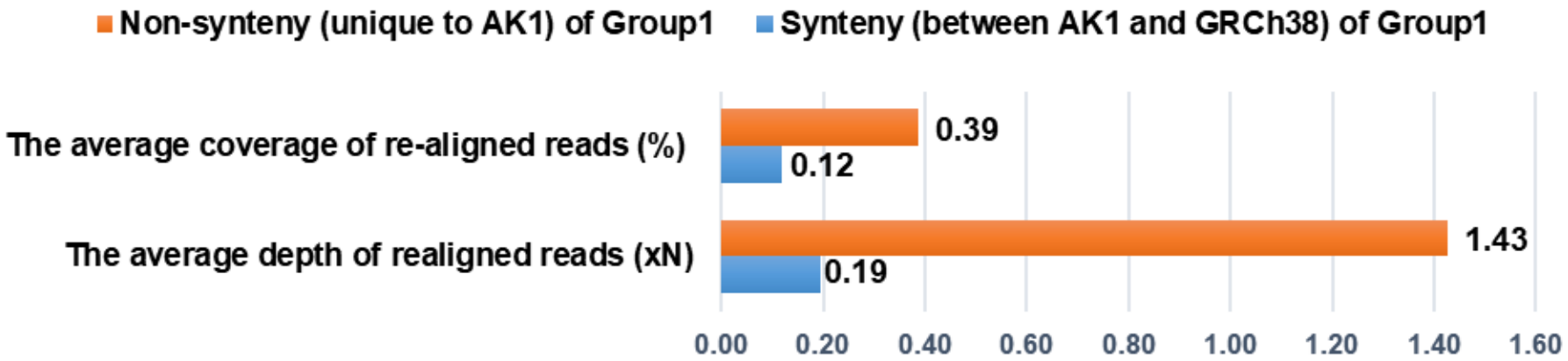


Figure 12. The average depth and coverage of the remapped reads on Group1 AK1 scaffolds by synteny status from 14 individuals. The average coverage of re-aligned reads (%)= (Breadth of coverage / Total size of synteny)*100, The average depth of realigned reads (xN) = (sum of average depth at position /breadth of coverage)

Chapter 4.

Construction of a Pan-tuberculosis Reference

4.1. Introduction

Short-read aligning to one reference genome is a generally used method in bacterial genomic studies related with variant discovery and building complete genomes of isolates (70-72). Notwithstanding, there are reasons for questioning that this way might result in biases by the reference used for mapping (73, 74, 75)). The genetic differences between the reference and sample sequencing data arise most of these errors, and they can have effect on sequencing data analyses (76-78). In recent study, evaluating the effect of reference choice on the analysis of sequence data from five bacteria (*Klebsiella pneumoniae, Legionella pneumophila, Neisseria gonorrhoeae, Pseudomonas aeruginosa* and *Serratia marcescens*), the choice of different reference genomes proved to affect almost all the parameters such as on mapping statistics and variant calling in the various species (45).

Meanwhile, the reference of *M. tuberculosis* is H37Rv, which was the most used in laboratories and the first strain of *M. tuberculosis* to perform whole-genome sequencing(26). Actually, the effects of *M. tuberculosis'* reference selection on epidemiological inferences have been also described (31). However, using whole genome sequencing to identify mutations associated with drug resistance (79), some studies have found small but noteworthy variations, which often affect *M. tuberculosis* pathogenicity, in gene content from comparisons between various genomes of clinical isolates of *M. tuberculosis* (30, 80-82).

For the reasons, researchers have tried to complement reference genome with using several ways (83, 84). One of attempts is using outbreak-specific (29, 85) or lineage-specific (86) genomes as a reference genome to dwindle alignment errors. Recently, a study constructed a Beijing lineage reference genome using assembled genome of TCDC11 with high quality and all-known base, and revealed that it contains several genes not found in the standard reference H37Rv (87).

Despite the attempts including lineage specific or outbreak-specific, there are still lack of investigation on hidden parts between overall pan-lineages of *M. tuberculosis* genomes. This is why it brings up the possibility of missing important variants and necessity of alternative sequences complementing *M. tuberculosis'* reference.

In this study, we first present identification of missing regions by comparing *M.tuberculosis'* various complete genomes with H37Rv reference genome. We used chain files between complete genomes of the pan-lineages as well as the Beijing specific strain and H37Rv. Secondly, we attempt to construct alternative sequences based on putative missing information. Constructing sequences, we performed *de novo* assembly with gaps from complete genomes and contigs from unmapped reads. Finally, we conducted a comparative genomics analysis between assembled genomic sequences and H37Rv, and investigated the effects and utilities of constructed alternative sequences with outcomes of variant detection.

4.2. Materials and Methods

4.2.1. Sample data

- Complete genome data from public DB

Until August 2020, 198 complete genomes of *M. tuberculosis* were downloaded from NCBI. Among 198 complete genomes, 22 complete genomes were excluded *M.tuberculosis* complex species or *M. bovis* so on(Table S4). Finally, total 176 complete genomes were compared with H37Rv.

- Whole genome data of *M. tuberculosis*

Whole genome sequencing data of total 724 strains, 348 strains of Korea and 376 strains of Japan, consists of the GReAT project and the Korean Institute of Tuberculosis (KIT) data.

After the WGS data was aligned to H37Rv, unmapped reads were extracted from aligned bam files with Samtools.

4.2.2. The identification of differences between complete genome data by using chain files.

To identify differences of each complete genomes compared H37Rv, 176 chain files were made between 176 *M. tuberculosis* and H37Rv as previous chapter 2. In making chain files, the commonly used blat (88) program was used instead of LASTZ, which was mentioned in part2. Using 176 Chain files, one to one genome analysis was performed and made statistics of missed regions. The gaps by each lineage were identified via the statistics, and gaps were collected with >= 50bp. The sequences were used of constructing H37Rv alternative sequences.

4.2.3. The *de novo* assembly of unmapped reads from whole genome data

After the unmapped reads from mapping WGS data were extracted, *De novo* assembly was performed with extracted unmapped reads. Using MEGAHIT assembler, contigs were assembled. MEGAHIT is mainly used in large and complex metagenomics NGS reads, and it adopted succinct *de Bruijn* graph to achieve low memory assembly (89).

4.2.4. Building pangenome reference by hybrid *de novo* assembly

To merge gaps (>=50bp) from complete genomes and contigs of unmapped reads' *de novo* assembly, hybrid *de novo* assembly was performed. Because hybrid *de novo* assembly could combine and assemble sequences of various size from two ways, this study used hybrid *de novo* assembly. This consists of two step's *de novo* assembly, and outcomes from

this process were used as alternative sequence of H37Rv. After assembly, contigs were undergone quality control of three steps; similarity with H37Rv and other assembled contigs, reasonable portion of GC contents, and simple repeats.

First, using pairwise alignment with BLASTn search, the assembled contigs were compared with H37Rv to exclude contigs of high similarity(Identity >=80%, coverage>=80%) from alternative sequences set. For second quality control, it was used that GC contents of H37Rv is 65.6%. GC contents of assembled contigs had widely various values, and contigs with extreme values of GC contents compared to 65.6% were removed(<40% or 80%=<). The third step was performed with Repeat Masker to check a portion of simple repeat (%). Thereby, a contig having many simple repeats was excluded. Finally, after quality control, contigs were annotated with Prokka (90), which is a software tool to annotate bacterial, archaeal and viral genomes quickly, and BLASTx.

4.2.5. Identification of effects on alignments and variant call results with alternative sequences

After variant call processes of 724 samples with Bowtie-GATK variant call pipeline, vcf files were filtered with GQ(genotype)<15 and RD(read depth) <10. The outcomes of alignments and variant calls with alternative sequences were compared to those of variant call with H37Rv. Samtools, vcftools, and RTG tools were used when comparing outcomes of using H37Rv with using H37Rv+ALTv2. Calculating statistics of vcfs was also performed with vcf-stats. The results of data were visualized with R-4.0.5.

4.3. Results

4.3.1 *In silico* analysis on candidate genomic gaps of 176 scaffolds based on H37Rv

We first compared 176 scaffolds and H37Rv with BLAT to create 176 chain files. With a difference of more than 50 bp between each scaffold and H37Rv, a total of 4,390 sequences were extracted from 89 complete genomes through 176 chain files. The average size of the sequences is 517.87 bp, and the largest is 13469 bp (Figure 14). To present difference of each lineage, we grouped the extracted sequences of complete genome, listing sequences by lineages. As a result, there are relatively large sequences found in lineage 2, followed by lineage 1, lineage 4 and lineage 3 (Figure 15). many sequences showing large differences in lineage 2 group on average suggest a far-relevant relationship between lineage 4, including H37Rv, and lineage 2(Figure 16).

4.3.2. *De novo* assembly of unmapped reads from whole genome sequencing data of TB

After alignment to H37Rv with whole genome sequencing data of 724 samples received from GReAT project and KIT, unmapped reads were extracted from 724 bam files. The unmapped reads have an average of 7161 reads per samples and average length of unmapped reads is 176.54bp (Table 10). Performing *de novo* assembly at different k-mers with Megahit, 7997 contigs were constructed at the optimal k-mer at 111. The genome assembly statistics for the contigs are summarized in Table 11. The largest sequence of the contigs has 220,364 bp and the shortest contig has 202 bp.

4.3.3. Merging gaps from complete genomes and contigs of unmapped reads using hybrid *de novo* assembly

By hybrid d*e novo* assembly, the new contigs merged with different gaps of complete genomes and unmapped reads at optimal k-mer at 141 have a variety of lengths. As a result, a total of 600 contigs were assembled and their average size is 4550 bp (The largest contig: 103,214 bp, the smallest contig: 270 bp), and the N50 value of the contigs is 17,003 (Table 11).

Before using these contigs as an alternative sequence, the quality control process had undergone with three criteria. When the first filtering criterion was to remove contigs having many simple repeats, this study resulted in the removal of one contig, which occupied 41.62% of simple repeat on the contig(Table 12). When, secondly, filtered contigs with GC contents, this study eliminated 78 contigs whose GC content was <40% or 80%=< in the extreme range compared to 65.6%, general GC contents of *M. tuberculosis* (Figure 17). Finally, comparing to H38Rv, the identity of 600 contigs was used as filter to avoid overlapping sequences. 31 contigs were excluded on both identity and coverage >=80%, and having > 3000 bp identical size(Figure 18). Also, 42 contigs were removed, which had identity and coverage >=80% comparing with each other contigs of hybrid *de novo* assembly or had more reads of secondary alignment than those primary alignment after alignment. As a result, after quality control, a total of 146 contigs including contigs that overlap with each criterion were removed and 454 sequences were used for variant calls with the *M. tuberculosis* reference. The final alternative sequences had about 0.9% repetitive sequences(Table 13). This showed that it was lower than repetitive proportion of alternative sequences before quality control(about 1.4%). By annotation with only prokka, the final alternative sequences consist of 923 genes, which were excluded annotated genes of hypothetical protein in a total annotated 2575 genes (Figure 19).

4.3.4. The effects on alignment and variant call results with final alternative sequences

After final contigs had been defined, this study aligned and called whole genome sequencing data with only H37Rv or H37Rv+ALT.v2 sequences.

In alignment step, this study investigated how much missing information can be salvaged. Comparing bam files, the mapping rate increased from 99.37 % to 99.79 %, which was equivalent to the value of an average 10,000 reads by samples. The number of unmapped reads, which were considered as missing information, were declined to one-third, comparing the results aligned to only H37Rv (Figure 20).

When this study compared variant call results of H37Rv+ALT.v2 to the results of H37Rv, the variant call results between H37Rv and H37Rv+ALT.v2 do not differ significantly in terms of read depth, mapping quality, and genotype quality as you can show on the Figure S2. In addition, when it comes to substitution, the counts by each type of substitution were shown on Figure 21. In allele frequencies of variant call results, the allele frequencies of variants called on the same position between two vcfs were not much different but same. Dividing into rare and common variants, almost variants on H37Rv genome sequence and alternative sequences were rare. Although two variants call results were not much different in several aspect, the variants call results with Pan-Reference(H37Rv+ALT.v2) have discovered new variants or genes. Among newly found 503 SNVs, there are about 88% SNPs and 12% INDELs (Table 14, Figure 22). By contigs, the k141-146 contig had the largest number of variants, 103 SNPs and 60 INDELs(Table 14). As the k141-146 contig was searched with BLASTn to further investigate the contig, it was searched as the genomic sequence producing PE-PGRS family protein. It is that PE-PGRS genes are multiple tandem repetitive sequences, and encodes related proteins including exceptionally many glycines and alanines.

In addition to SNVs, 50 gene groups were newly found. The 326 products and variants of 50 gene groups were annotated and searched by Prokka and BLASTx. The Table S7 has shown all lists and Table 15 presents the list of selected variants. Scrutinizing gene groups of the results in Table 15, *msrP_3, pimC,* and *tuf_3* were annotated by Prokka and the gene groups that are not in the Prokka DB were annotated by BLASTx. Methionine sulfoxide reductase, the product of *msrP_3*, converts methionine sulfoxide to methionine and shield bacteria from reactive oxygen intermediates (ROI) and reactive nitrogen intermediates (RNI). In case of pimC, the gene encodes GDP-mannose-dependent alpha-(1-6)-phosphatidylinositol dimannoside mannosyltransferase, which catalyzes the addition of a mannose residue from GDP-D-mannose to the position 6 of the alpha-1,6-linked mannose residue to produce triacyl phosphatidylinositol trimannoside. This enzyme is involved in phosphatidylinositol metabolism. The product of *tuf_3* was EF-Tu, which is a multifunctional protein in various pathogenic bacteria. Although the important roles of this gene are yet to be revealed, EF-Tu is associated with antibiotic-mediated inhibition of translation (91).

4.4. Discussion

According to the research, this study generated alternative sequences complementing H37Rv with complete genome sequences from public database and whole genome sequencing data of 724 samples. Currently, in human, the usage of alternative sequences has compensated the lack of reference genome. Some studies have developed the definition of “pan genomes” (92), which tried to be included at most the genomic content and variation of a species, instead of a single reference genome. However, although *M. tuberculosis* has relatively large genome in bacterial species large, the reference genome of *M. tuberculosis* comprised of genomic sequence of only one strain. This increases the need of supplementing reference genome. This study also led to attempts to complement reference

genome by constructing alternative sequence of *M. tuberculosis* from missing information. At first, we found missing information with using 198 complete genomes of public data and unmapped reads of 724 samples. In complete genomes, there were various gaps of complete genomes comparing to H37Rv. Particularly, the comparison with respect to lineage shows that the average gap sizes comparing to H37Rv were larger in other lineages than in lineage 4. This might imply that some problems arise when the H37Rv of lineage 4 is used as a reference genome. In addition to complete genomes, one of reasons on the unmapped reads, which were fail to align reference genome, might be incompleteness of reference genome, although the reasons are various. Actually, alternative sequences merged with unmapped reads make mapping rates increase and average 10,000 unmapped reads reduced. This might be another evidence of a lack of reference genome mentioned in previous studies. Meanwhile, in this study, genome sequences when merging sequences did not include sequences on several lineages. This might be a possible scarcity due to lack of information on specific lineages.

Although the results of two variant calls against different genome sets were not much different on overall statistics, constructing and use of alternative sequences provided the possibility on detection of new critical variants or genes. Actually, it revealed the sequences have 2575 genes including gene groups of hypothetical protein after annotating the sequences. This implies that it might be a more variety of discovering new variant calls or genes when several sequencing data of *M. tuberculosis* have undergone aligning to H37Rv+ALT.v2.

In the results of this study, new variants or genes were identified as this study expected. For instance, the product of pimC on several detected genes plays role in the phosphatidylinositol metabolism by transferring a mannose residue. The metabolism is associated with the cell wall of mycobacteria, which is important for survival and pathogenicity in mycobacteria. This suggest that the identified pimC gene, which involves

cell wall and product, could be a target for the new antibiotics. Furthermore, although the pimC gene is present in *M. tuberculosis* CDC1551, it is absent from other mycobacterial genomes such as H37Rv (93).

In sum, to supplement reference genome of *M. tuberculosis*, this study tried to merge various missing information as “Pangenome” by new methods and successfully constructed alternative sequences. The utility of alternative sequences with reference sequence is critical for obtaining missing information. It is shown that the diversity of genomic analysis was enriched with using alternative sequences in *M. tuberculosis* research. Furthermore, when other studies use the constructed alternative sequences including various genes, it is possible that new variants and genes, which could not be identified in this study, might be newly found.

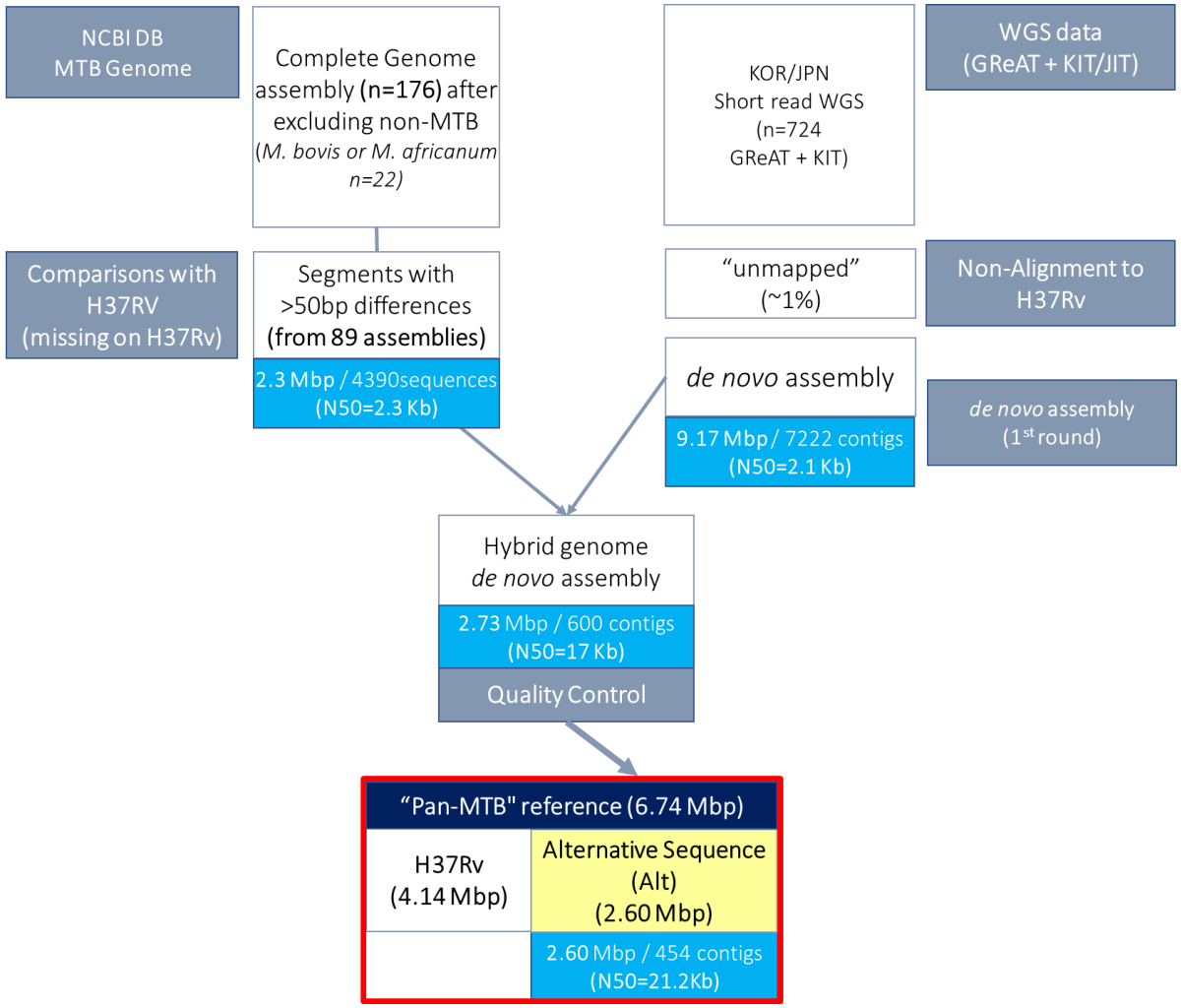
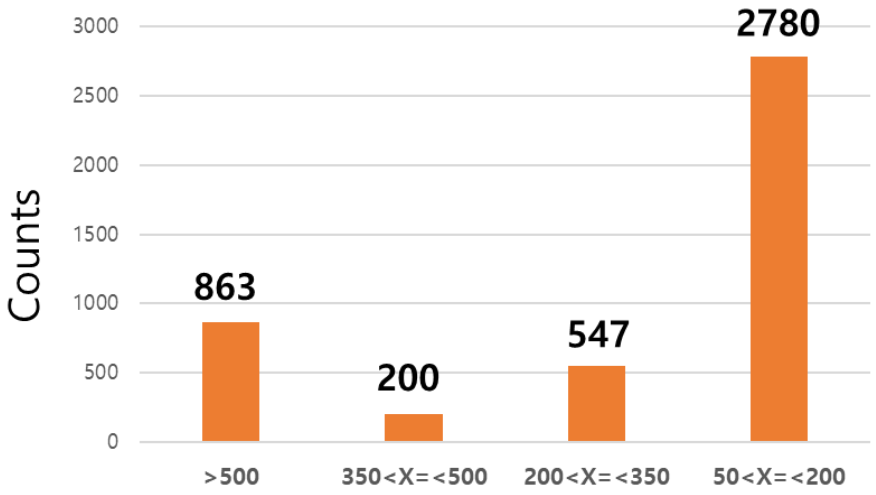


Figure 13. The workflow of constructing Pan-tuberculosis genome.



Total N	**4,390**
The sum of length of total sequences	2,273,460 bp
Average size of sequences	517.87 bp
The median size of sequences	131.00 bp
The value of N50	2,257 bp
The minimum size of sequences	51 bp
The maximum size of sequences	13,469 bp

Figure 14. The statistics of > 50bp different sequences of complete genomes comparing to H37Rv

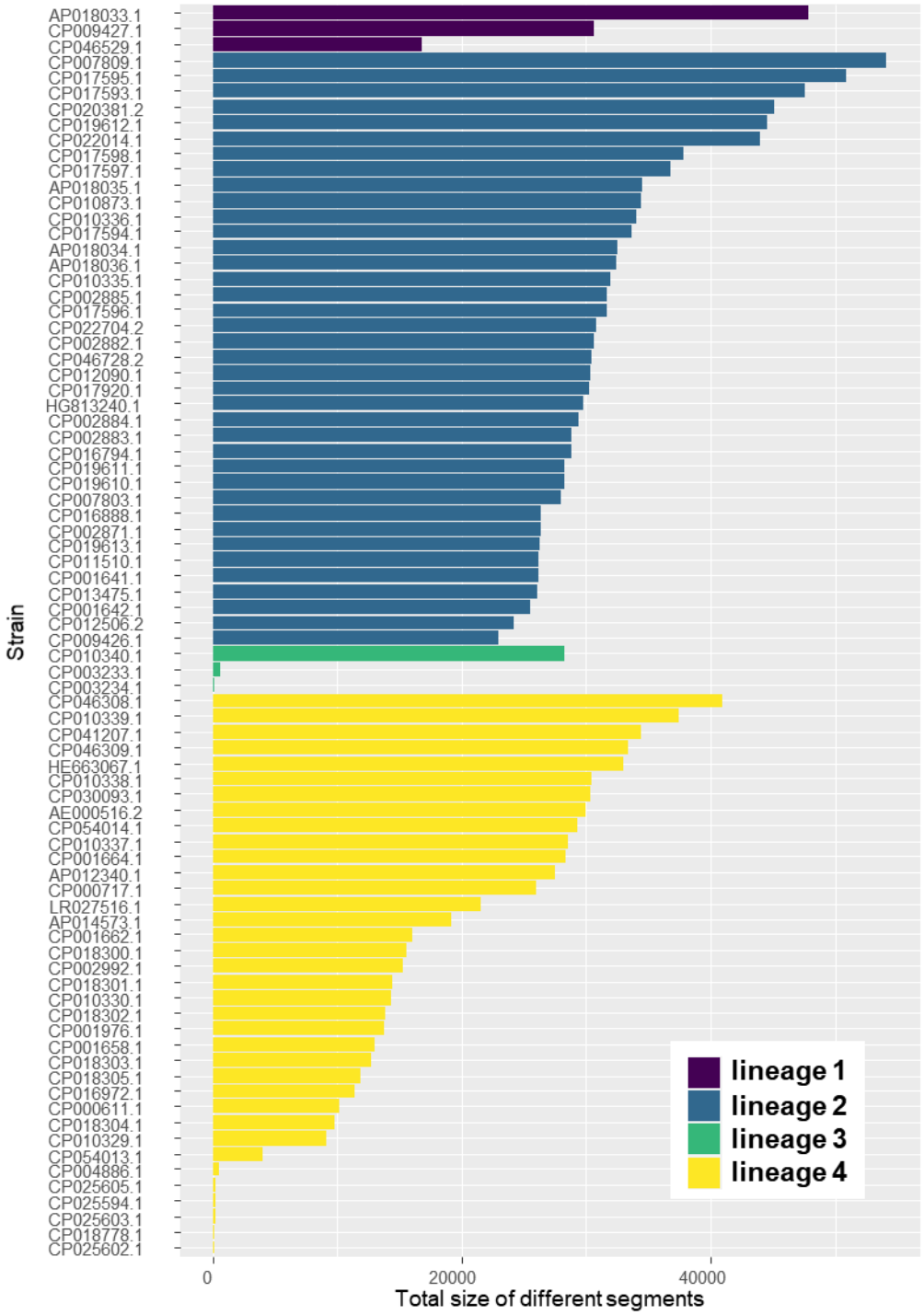


Figure 15. The distribution of sum of > 50bp different sequence sizes comparing to H37Rv by strains

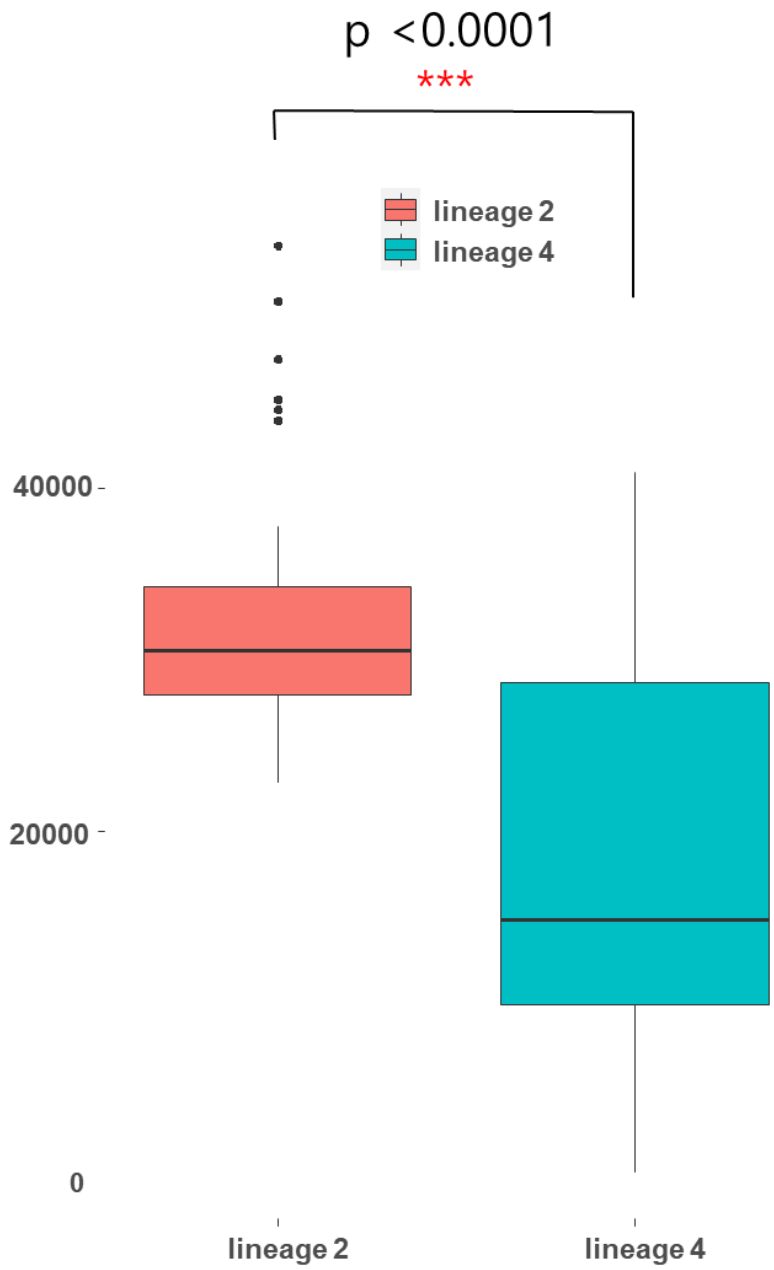


Figure 16. The comparison of sum of > 50bp different sequences between lineage 2 and 4.

Table 10. The summary statistics of whole genome sequencing data from KIT and GReAT consortium.

KOR, the strains collected in Korea; JPN, the strains collected in Japan

	KOR	JPN	TOTAL
The number of samples	348	376	724
The number of total reads	5,610,735	4,758,198	10,368,933
Average reads	8,061	6,327	7,161
Total length (bp)	891,541,830	891,541,830	1,822,854,011
Average length (bp)	167.67	184.74	176.54
Minimum read length (bp)	25	25	25
Maximum read length (bp)	351	350	351

Table 11. The statistics of results on 1st *de novo* assembly of unmapped reads, 2nd *de novo* assembly (1st contigs and gapped sequences from complete genomes), and final pan-genome sequences

	1st assembly (724 sample)	2nd assembly (1st contigs and gapped sequences of complete genomes)	Final alternative sequence (after quality controls)
The number of contigs	7222	600	454
Total Size (bp)	9,165,331	2,730,013	2,599,898
Mean size of total contigs (bp)	1,146	4,550	5,727
The size of longest sequence (bp)	220,364	103,200	103,200
The size of shortest sequence (bp)	202	270	270
The N50 of contigs (bp)	2098	17,003	21,194
GC Content (%)	57.27	61.39	61.68

Table 12. The statistics on simple repeats by contigs

	The number of simple repeats	Length (bp)	Percentage of sequence (%)
k141_10	4	206	41.62
k141_578	1	139	31.52
k141_568	2	123	27.89
k141_278	2	96	25.07
k141_8	1	156	25
k141_118	1	246	22.99
k141_541	1	87	21.64
k141_584	1	51	16.5
k141_468	1	55	15.76
k141_562	1	154	15.37
k141_62	1	56	13.53
k141_282	1	106	13.4
k141_335	2	128	12.45
k141_393	4	173	12.05
k141_258	6	363	11
k141_1	2	56	10.79
k141_34	1	33	10.48
k141_483	1	50	10.35
k141_124	1	57	9.5

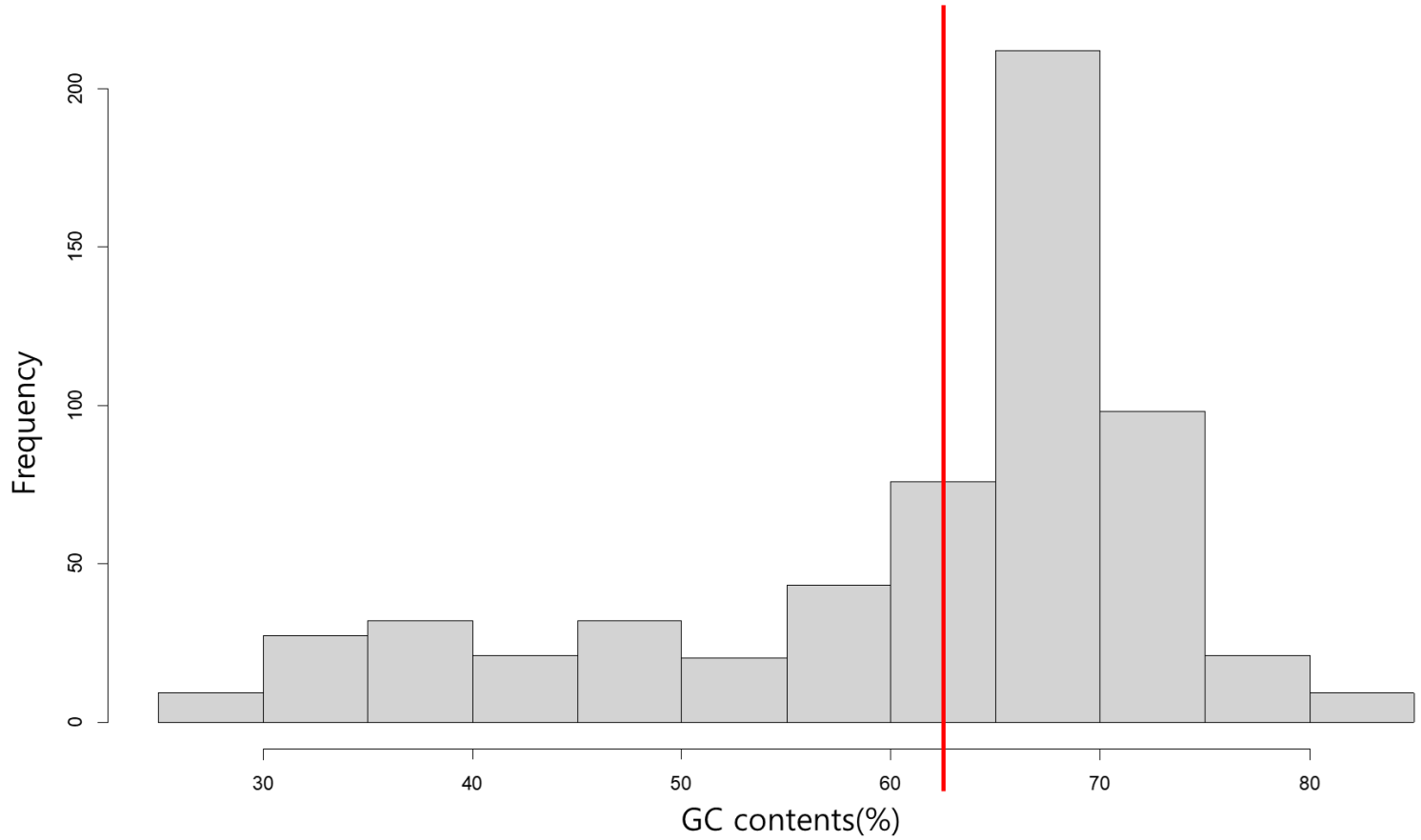


Figure 17. The frequency of contigs by GC contents(%) The red line indicates the standard GC contents(%)of *M. tuberculosis*; 65.6%

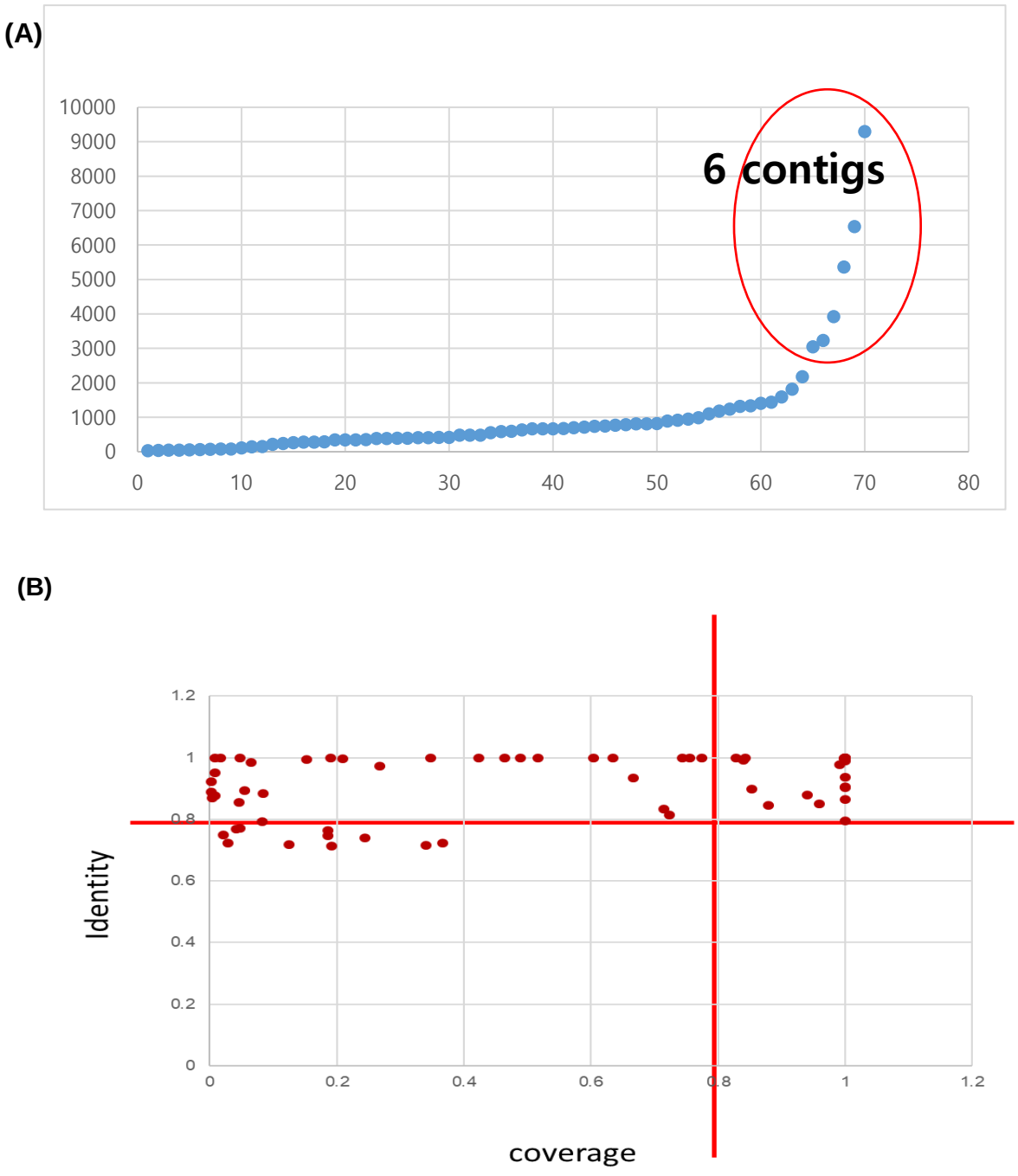
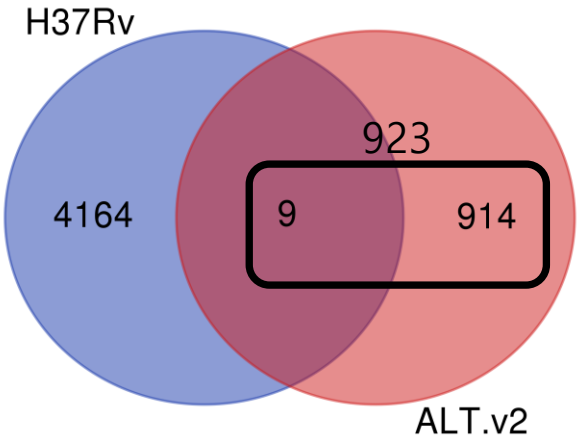


Figure 18. The results of pairwise alignment between alternative sequences and H37Rv to filter the similar sequences with H37Rv (A) The size of similar sequences to H37Rv on each contigs. The red circle indicates the criteria of filtering **(B) The identity and coverage of contigs having similar sequence with H37Rv.** The red lines indicate the criteria of filtering**.**

Table 13. The statistics of repetitive sequences on final pan-genome sequences.

		Contigs of hybrid denovo assembly		
		number of elements	Size (bp)	Proportion (%)
Non-repetitive sequence		-	2,576,974	99.118
SINE	All	2	311	0.011
	ALUs	2	311	0.011
	MIRs	-		-
LINE	All	4	1,650	0.060
	LINE1	2	1,468	0.054
	LINE2	-	-	-
	L3/CR1	1	72	-
LTR	All	-	-	-
DNA	hAT-Charlie	-	-	-
	TcMar-Tigger	-	-	-
Unclassified		-	-	-
Small RNA		9	4,816	0.176
Satellite		-	-	-
Simple repeats		307	15,624	0.572
Low complexity		12	523	0.019
Total			2,599,898	100.00

(A)

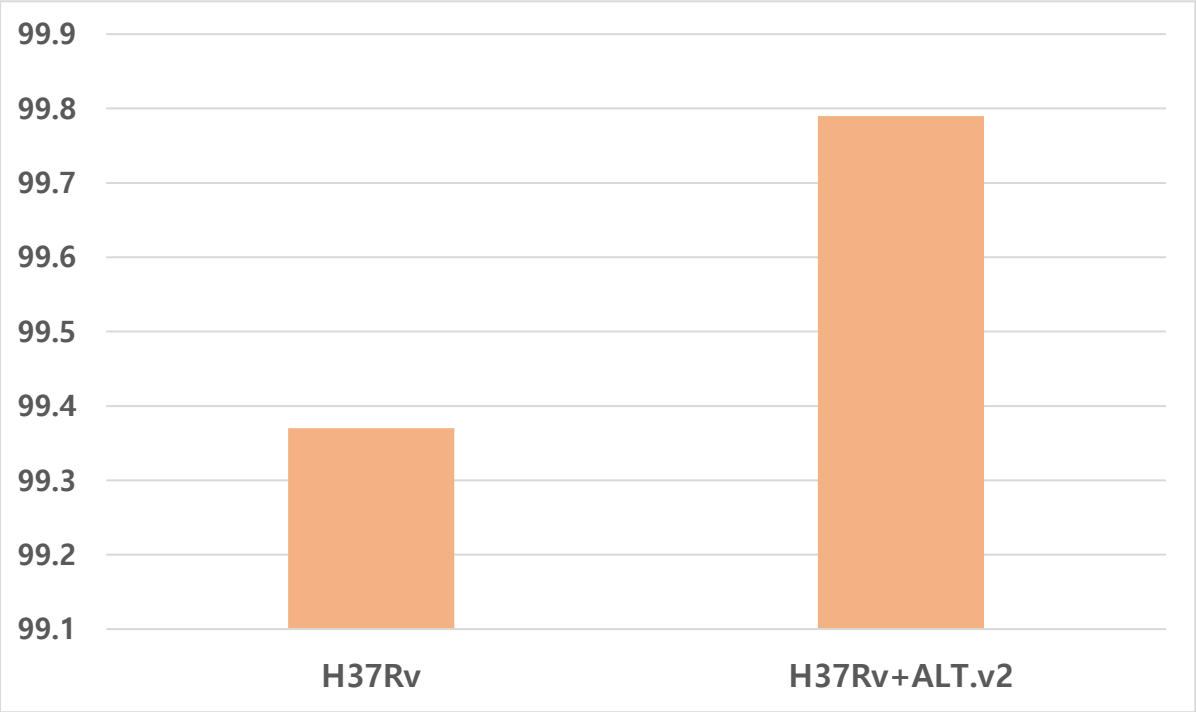


(B)

Number of genes	CDS		tRNA
	Annotated	Hypothetical protein	
2575	923 (including one Transfer-Messenger RNA)	1635	17

Figure 19. The investigations on annotated pan-genome sequences (A) The Venn diagram of annotated genes between pan-genome sequences and H37Rv (B) The number and type of genes annotated on alternative sequences

(A)



(B)

	H37Rv	Pan-Reference (H37Rv+ALT.v2)
Average total reads of 724 samples	2,196,151±1,151,997	
Average Mapped reads	2,181,829±1,143,554	2,191,581±1,149,541
Average Unmapped reads	14,322±24,969	4,570±10,897
Mapping rate (%)	99.37±0.98	99.79±0.47

Figure 20. The statistics of alignment with using pan-genome sequences comparing to H37Rv (A) The alignment rates with H37Rv and H37Rv+ALT.v2 (B) The statistics on the number of reads (mapped or unmapped) with H37Rv and Pan-Reference (H37Rv+ALT.v2)

Substitution	ALT.v2
C>T	60
G>A	79
A>G	35
T>C	31
G>C	22
C>G	30
G>T	18
C>A	17
T>G	20
A>C	15
A>T	4
T>A	2
Total	333
Transition	205
Transversion	128

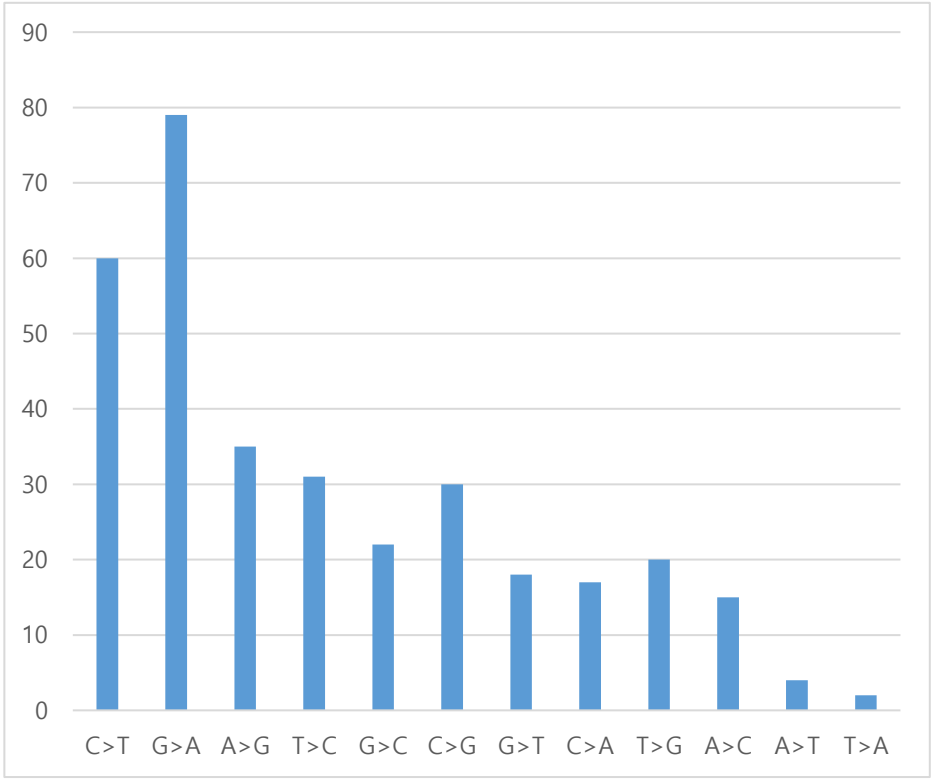


Figure 21. The counts of substitutional variants on pan-genome sequences.

Table 14. The number of SNVs (rare SNP, common SNP, INDEL) of pan-genome sequences.

	Total variants	SNP			INDEL
		Total SNPs	rare	common	
Alternative sequences	503	387	303	84	116
K141-146	163	103	88	15	60
K141-375	81	66	51	15	15
K141-363	67	54	48	6	13
K141-258	58	45	30	15	13
K141-143	45	34	34	0	11
K141-345	24	24	16	8	0
K141-565	23	23	14	9	0
K141-179	18	18	11	7	0
K141-284	13	10	6	4	3
K141-527	6	5	3	2	1
K141-561	5	5	2	3	0

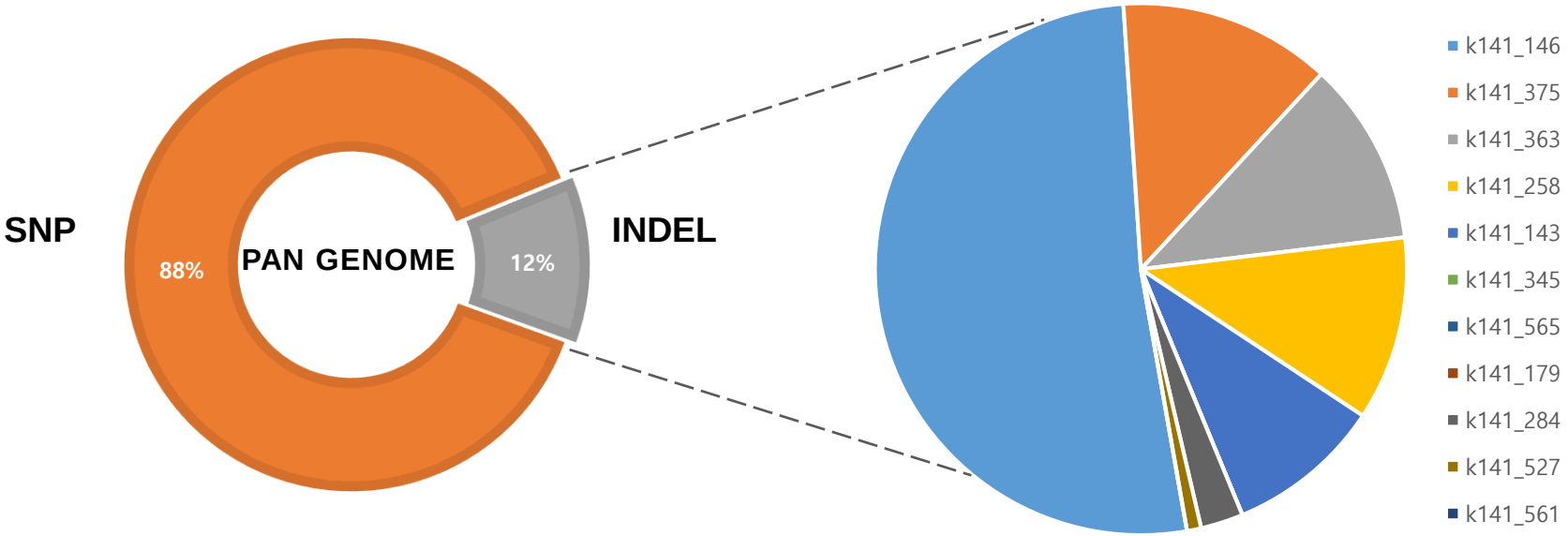


Figure 22. The composition of variants called with using pan-genome and the distribution of INDEL by contigs.

Table 15. The summary of selected variants called with using Pan-genome reference

Contig	Position	Average read depth	REF	ALT	N	The frequency of REF allele	The frequency of ALT allele	Gene region	Gene product
345	3,792	17.4	G	C	119	0.950	0.050	*msrP_3* (k141_345:3561-4664)	Protein-methionine-sulfoxide reductase catalytic subunit MsrP
345	3,900	18.1	G	A	116	0.991	0.009	*msrP_3* (k141_345:3561-4664)	Protein-methionine-sulfoxide reductase catalytic subunit MsrP
345	4,632	16.8	G	A	117	0.991	0.009	*msrP_3* (k141_345:3561-4664)	Protein-methionine-sulfoxide reductase catalytic subunit MsrP
345	5,196	17.4	T	C	123	0.992	0.008	*pimC* (k141_345:4890-6035)	GDP-mannose-dependent alpha-(1-6)-phosphatidylinositol dimannoside mannosyltransferase
345	5,875	16.6	C	T	120	0.950	0.050	*pimC* (k141_345:4890-6035)	GDP-mannose-dependent alpha-(1-6)-phosphatidylinositol dimannoside mannosyltransferase
375	1,078	91.3	C	T	610	0.993	0.007	*tuf_3* (k141_375:952-1221)	Elongation factor Tu
375	1,090	92.2	A	G	610	0.998	0.002	*tuf_3* (k141_375:952-1221)	Elongation factor Tu

Chapter 5.

Summary and Conclusions

5.1. General Discussion

In this study, we provided another contribution that strengthens the need of bridging the gap in reference genomes. This study also attempted to find missing information and supplement representative reference genomes in humans and *M. tuberculosis*.

For human reference genomes, we identified the missing regions in human reference genome in two methods, using AK1 as a high-quality genome map. By trying two new approaches to find missing regions against reference genomes, this research may be meaningful among many other attempts in human genome reference. Using precise ethnic genome in our two methods helped to acquire missing genomic information. Recently, the precise ethnic genome has been increased worldwide. The Human Pangenome Reference Consortium (https://humanpangenome.org/) is an example. Like our results, such precise ethnic genomes will also play an important role in finding other missing information and redress the research gap between populations by supplementing the shortage of reference genome.

Also, the functions of the found missing regions were predicted by our *in silico* analysis. In the analysis, functional genes on missing regions were searched and associated with human diseases. If the missing regions include disease-related variants, the variant is not detected. When using our methods to complement the reference genome, new drug targets might be detected by finding new variants and genes. For instance, in 2017, the Icelandic human-sequencing project found a 766-bp insertion with high allele frequency, and the insertion was correlated with low risk of myocardial infarction (37).

Aside from predicting the function of missing regions, we investigated occurrence mechanisms of common missing regions by simultaneously verifying the presence of common missing regions. We found that most of the common missing regions in this research were completely or incompletely deleted in European genomes and caused by

nonhomologous end-joining(NHEJ) with microhomology. It may signify that our results show a more complex admixture history of modern human populations. Furthermore, it is known that NHEJ plays a significant part in the production of double minutes, which are small fragments of extrachromosomal DNA that have been detected in a large number of human tumors. For this reason, a research in 2015 reported that NHEJ may be targeted for the treatment of certain type colon cancer(94). Considering this role of NHEJ and our results, we need to further investigate not only the heterogeneous genomic structures and occurrence mechanism of the common missing regions by populations but also the link between the extrachromosomal DNA occurred by NHEJ, which could be one of the causes of cancer, and the missing regions.

For the *M. tuberculosis* reference genome*,* this study tried to complement the reference genome of one strain by constructing pan-genome sequences because it consisted of only one strain unlike human reference genomes. The endeavor has reduced the number of unmapped reads that could actually be missing and allowed more information to be discovered. In addition, like our previous human reference research, this study identified the newly discovered genes, which are involved in critical pathways and could become targets for the development of new drugs. In Helicobacter pylori(95) and Escherichia coli(96, 97),which have representative reference genome, as well as *M. tuberculosis,* collecting and merging genomes of various strains for searching for missing information of reference genome have performed and led to find novel pathogenic variants. Therefore, there will be continuous efforts on supplementing reference genomes because certain missing DNA sequences exist in populations or lineages around the world that are not also found in the reference genome of various species as well.

5.2. Summary and Conclusions

Rapid improvement in next-generation sequencing technology have made us obtain many genetic information, but there is missing genetic information on several steps when carrying out common sequencing technologies. In terms of reference genomes, researchers have known for several years that considering more genomes than just a single representative genome is necessary to identify various genes and variants in several species.

In chapter 2, we used two methods for identifying putative missed genomic regions in human reference genomes by using highly contiguous genome assembly. Comparing one genome to the reference genome directly uncovered the 3,333 regions (>200bp size of gap), and alignments by unmapped reads from bam files helped find 110 regions. We have identified common missing regions that were estimated with the data of several populations. Through experiments, we verified the presence of common missing regions on other genomes and discovered that the common missing regions had been deleted on European genomes.

In chapter 3, we predicted the function of common missing regions through the translated BLAST search with NCBI's nr database. Also, we investigated features and occurrence mechanisms of common missing regions. 1,390 regions (e-value $<10^{-10}$, identity **≥** 70%, and alignment length **≥** 50bp) of 3,333 regions found by directly comparing between two genomes were estimated to have putative protein coding elements. 25 of the 110 regions (e-value $<10^{-10}$, identity **≥**70%, and alignment length **≥**50bp) had putative protein coding functions of mammalian. Among the detected proteins, it has been known that P150 is a protein that is largely absent or greatly reduced in ovarian cancer, and MYB Isform 6 is a transcription factor associated with some human diseases. Also, most of the common missing regions of the AK1 genome had high proportions of repeated sequences compared

with the proportion of repetitive sequences on GRCh38. In occurrence mechanism of missing regions, this study showed that common missing regions were mostly occurred by NHEJ(n=26) with microhomology.

In chapter 4, we described the process of constructing pan-genome sequences and their effects to complement the current reference genome of *M. tuberculosis*. This study constructed pangenome sequences to complement the reference genome of one strain with a variety of complete genomic sequence from public database and unmapped reads from whole genome sequencing data. A total of 4,390 sequences with a difference of more than 50 bp between each complete genome and H37Rv were generated from 89 complete genomes via 176 chain files. After the 1st *de novo* assembly of unmapped reads extracted from 724 bam files, the 2nd *de novo* assembly was performed with outputs of the 1st *de novo* assembly and sequences extracted from complete genomes. As a result, a total of 600 contigs were assembled and 454 contigs were used for variant calls with the *M. tuberculosis* reference after quality control. In the effects on alignment stage, the mapping rate increased from 99.37 % to 99.79 %, which was equivalent to the value of an average 10,000 reads by samples. In the effects on variant call results of H37Rv+pan-genome sequences (ALT.v2), the variant call results between H37Rv and H37Rv+pan-genome sequences (ALT.v2) did not differ significantly in terms of read depth, mapping quality, and genotype quality. However, the 326 variants of 50 gene groups were newly found using H37Rv+pan-genome sequences(ALT.v2). In particular, the product of pimC, which was newly found and is absent from H37Rv, is involved with the formation of cell walls and could be a possible target for new antibiotics.

In conclusion, this study found missing regions of reference genome on human and *M. tuberculosis* and aimed to complement lack of reference genomes. For the human reference genome, the number of high contiguous genome assembly of diverse ethnic genomes was known to be increasing in current studies as the ability to assemble genomes is improving (ex. T2T Consortium; the first telomere-to-telomere assembly of a human X chromosome (98)). As with the current stream of studies, our results showed that the usage of a high contiguous ethnic genome might be useful for obtaining missing genomic information. This suggests that it may help find new missing genomic information and close the known scientific gap between the genomes of various populations.

Also, the fact that missing regions had been deleted in European genomes, had repetitive sequences, and were caused by one dominant mechanism may provide understanding on the evolutionary history of humans or reasons for the occurring genomic differences between populations.

Finally, we could find variants that are absent from a single reference genome involved in beneficial or harmful traits of humans and *M. tuberculosis*. In this way, it is shown that the newly discovered genes are involved in critical pathways and may be new targets on the development of drugs.

References

1. Sanger F. BG, Barrell B. A two-dimensional fractionation procedure for radioactive nucleotides. J Mol Biol. 1965.
2. Stein LD. The case for cloud computing in genome informatics. Genome Biology 2010;11:207.
3. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9):418-26.
4. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333-51.
5. Gupta PK. Single-molecule DNA sequencing technologies for future genomics research. Trends Biotechnol. 2008;26(11):602-11.
6. John Eid AF, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex deWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, Stephen Turner. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323
7. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 2009;4(4):265-70.
8. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37(10):1155-62.
9. Karst SM, Ziels RM, Kirkegaard RH, Sorensen EA, McDonald D, Zhu Q, et al. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. Nat Methods. 2021;18(2):165-9.
10. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21(1):30.

11. Lappalainen T, Scott AJ, Brandt M, Hall IM. Genomic Analysis in the Age of Human Genome Sequencing. Cell. 2019;177(1):70-84.
12. Chaisson MJ, Wilson RK, Eichler EE. Genetic variation and the de novo assembly of human genomes. Nat Rev Genet. 2015;16(11):627-40.
13. Compeau PE, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. 2011;29(11):987-91.
14. Liao X, Li M, Zou Y, Wu F-X, Yi P, Wang J. Current challenges and solutions of de novo assembly. Quantitative Biology. 2019;7(2):90-109.
15. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121-32.
16. Raphael BJ. Chapter 6: Structural Variation and Medical Genomics. PLOS Computational Biology. 2012 8(12): e1002821. https://doi.org/10.1371/journal.pcbi.1002821.
17. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
18. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-20.
19. Thompson JD, Linard B, Lecompte O, Poch O. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PLoS One. 2011;6(3):e18093.
20. Mount D. Bioinformatics: Sequence and Genome Analysis, Second Edition. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2004.
21. Lakshmi NJ, Gavarraju, P., Jeevana, J. K., Karteeka, P. A literature survey on multiple sequence alignment algorithms. . Int J Adv Res Comput Sci Softw 2016;Eng, 6:280-8.
22. Consortium IHGS. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921
23. Venter JCea. The sequence of the human genome. Science 2001; 291:1304–51.
24. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27(5):849-64.

25. Richard E. Green JK, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, , Markus Hsi-Yang Fritz NFH, Eric Y. Durand, Anna-Sapfo Malaspinas,Jeffrey D. Jensen, Tomas Marques-Bonet, Can Alkan, Kay Prüfer, Matthias Meyer,Hernán A. Burbano, Jeffrey M. Good, Rigo Schultz, Ayinuer Aximu-Petri, Anne Butthof, Barbara Höber, Barbara Höffner, Madlen Siegemund, Antje Weihmann, Chad Nusbaum,Eric S. Lander, Carsten Russ, Nathaniel Novod, Jason Affourtit, Michael Egholm,Christine Verna, Pavao Rudan, Dejana Brajkovic, Željko Kucan, Ivan Gušic,Vladimir B. Doronichev, Liubov V. Golovanova, Carles Lalueza-Fox, Marco de la Rasilla,Javier Fortea,Antonio Rosas, Ralf W. Schmitz,Philip L. F. Johnson, Evan E. Eichler, Daniel Falush, Ewan Birney,James C. Mullikin, Montgomery Slatkin, Rasmus Nielsen, Janet Kelso, Michael Lachmann, David Reich, Svante Pääbo. A Draft Sequence of the Neandertal Genome. Science. 2010;328:710-22.

26. S. T. Cole RB, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry III,, F. Tekaia KB, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles,, N. Hamlin SH, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail,, M.-A. Rajandream JR, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, Barrell BG. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998;396.

27. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. The Lancet Infectious Diseases. 2015;15(10):1193-202.

28. Mestre O, Luo T, Dos Vultos T, Kremer K, Murray A, Namouchi A, et al. Phylogeny of Mycobacterium tuberculosis Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. PLoS One. 2011;6(1):e16020.

29. Roetzer A, Diel R, Kohl TA, Ruckert C, Nubel U, Blom J, et al. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. PLoS Med. 2013;10(2):e1001387.

30. O'Toole RF, Gautam SS. Limitations of the Mycobacterium tuberculosis reference genome H37Rv in the detection of virulence-related loci. Genomics. 2017;109(5-6):471-4.

31. Lee RS, Behr MA. Does Choice Matter? Reference-Based Alignment for Molecular Epidemiology of Tuberculosis. J Clin Microbiol. 2016;54(7):1891-5.

32. Auton A, Abecasis, G., Altshuler, D. et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
33. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity, population stratification, and selection of human copy-number variation. Science (New York, NY). 2015;349(6253):aab3761-aab.
34. Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, et al. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. Nature. 2017;548(7665):87-91.
35. Genovese G, Handsaker RE, Li H, Kenny EE, McCarroll SA. Mapping the human reference genome's missing sequence by three-way admixture in Latino genomes. Am J Hum Genet. 2013;93(3):411-21.
36. Wong KHY, Levy-Sakin M, Kwok PY. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. Nat Commun. 2018;9(1):3040.
37. Kehr B, Helgadottir A, Melsted P, Jonsson H, Helgason H, Jonasdottir A, et al. Diversity in non-repetitive human sequences not found in the reference genome. Nat Genet. 2017;49(4):588-93.
38. Telenti A, Pierce LC, Biggs WH, di Iulio J, Wong EH, Fabani MM, et al. Deep sequencing of 10,000 human genomes. Proc Natl Acad Sci U S A. 2016;113(42):11901-6.
39. Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, et al. Long-read sequencing and de novo assembly of a Chinese genome. Nat Commun. 2016;7:12065.
40. Faber-Hammond JJ, Brown KH. Anchored pseudo-de novo assembly of human genomes identifies extensive sequence variation from unmapped sequence reads. Hum Genet. 2016;135(7):727-40.
41. Sousa V, Hey J. Understanding the origin of species with genome-scale data: modelling gene flow. Nat Rev Genet. 2013;14(6):404-14.
42. Al Shammari B, Shiomi T, Tezera L, Bielecka MK, Workman V, Sathyamoorthy T, et al. The Extracellular Matrix Regulates Granuloma Necrosis in Tuberculosis. J Infect Dis. 2015;212(3):463-73.
43. De Groote MA, Gruppo V, Woolhiser LK, Orme IM, Gilliland JC, Lenaerts AJ. Importance of confirming data on the in vivo efficacy of novel antibacterial drug regimens against various strains of Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2012;56(2):731-8.
44. Marquina-Castillo B, Garcia-Garcia L, Ponce-de-Leon A, Jimenez-Corona ME, Bobadilla-Del Valle M, Cano-Arellano B, et al. Virulence, immunopathology and

transmissibility of selected strains of Mycobacterium tuberculosis in a murine model. Immunology. 2009;128(1):123-33.
45. Valiente-Mullor C, Beamud B, Ansari I, Frances-Cuesta C, Garcia-Gonzalez N, Mejia L, et al. One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. PLoS Comput Biol. 2021;17(1):e1008678.
46. Hervé Tettelin VM, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, Jonathan Crabtree, Amanda L. Jones, A. Scott Durkin, Robert T. DeBoy, Tanja M. Davidsen, Marirosa Mora, Maria Scarselli, Immaculada Margarit y Ros, Jeremy D. Peterson, Christopher R. Hauser, Jaideep P. Sundaram, William C. Nelson, Ramana Madupu, Lauren M. Brinkac, Robert J. Dodson, Mary J. Rosovitz, Steven A. Sullivan, Sean C. Daugherty, Daniel H. Haft, Jeremy Selengut, Michelle L. Gwinn, Liwei Zhou, Nikhat Zafar, Hoda Khouri, Diana Radune, George Dimitrov, Kisha Watkins, Kevin J. B. O'Connor, Shannon Smith, Teresa R. Utterback, Owen White, Craig E. Rubens, Guido Grandi, Lawrence C. Madoff, Dennis L. Kasper, John L. Telford, Michael R. Wessels, Rino Rappuoli, and Claire M. Fraser. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial ''pan-genome''. PNAS. 2005;102.
47. Sherman RM, Salzberg SL. Pan-genomics in the human genome era. Nat Rev Genet. 2020;21(4):243-54.
48. Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, Chin CS, et al. Extending reference assembly models. Genome Biol. 2015;16:13.
49. The1000GenomeProjectConsortium. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
50. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity, population stratification, and selection of human copy-number variation. Science. 2015;349(6253):aab3761.
51. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet. 2019;51(1):30-5.
52. Duan Z, Qiao Y, Lu J, Lu H, Zhang W, Yan F, et al. HUPAN: a pan-genome analysis pipeline for human genomes. Genome Biol. 2019;20(1):149.
53. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet. 2018;51:30–5.

54. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al. Characterizing the Major Structural Variant Alleles of the Human Genome. Cell. 2019;176(3):663-75 e19.
55. Li R, Tian X, Yang P, Fan Y, Li M, Zheng H, et al. Recovery of non-reference sequences missing from the human reference genome. BMC Genomics. 2019;20(1):746.
56. Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, et al. De novo assembly and phasing of a Korean human genome. Nature. 2016;538(7624):243-7.
57. Harris RS. Improved Pairwise Alignment of Genomic DNA. PhD thesis, Penn State Univ. 2007.
58. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
59. Tischler G LS. biobambam: tools for read pair collation based algorithms on BAM files. Source Code Biol Med. 2014;9:13.
60. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.
61. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.
62. Smit A, Hubley, R.,Green,P. . RepeatMasker Open-4.0. http://www.repeatmasker.org/. 2015.
63. Walker MA, Pedamallu CS, Ojesina AI, Bullman S, Sharpe T, Whelan CW, et al. GATK PathSeq: A customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. Bioinformatics. 2018;34(24):4287–9.
64. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.
65. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu Y, et al. The UCSC genome browser database. Nucleic acids research. 2003;31(1):51-4.
66. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24-6.
67. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
68. TA H. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser. 1999;41:95-98.

69. Rishishwar L, Tellez Villa CE, Jordan IK. Transposable element polymorphisms recapitulate human evolution. Mob DNA. 2015;6:21.
70. Ellington MJ, Heinz E, Wailan AM, Dorman MJ, de Goffau M, Cain AK, et al. Contrasting patterns of longitudinal population dynamics and antimicrobial resistance mechanisms in two priority bacterial pathogens over 7 years in a single center. Genome Biol. 2019;20(1):184.
71. McAdam PR, Templeton KE, Edwards GF, Holden MT, Feil EJ, Aanensen DM, et al. Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant Staphylococcus aureus. Proc Natl Acad Sci U S A. 2012;109(23):9107-12.
72. Mentasti M, Cassier P, David S, Ginevra C, Gomez-Valero L, Underwood A, et al. Rapid detection and evolutionary analysis of Legionella pneumophila serogroup 1 sequence type 47. Clin Microbiol Infect. 2017;23(4):264 e1- e9.
73. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011;8(1):61-5.
74. Takishita K, Inagaki Y. Eukaryotic origin of glyceraldehyde-3-phosphate dehydrogenase genes in Clostridium thermocellum and Clostridium cellulolyticum genomes and putative fates of the exogenous gene in the subsequent genome evolution. Gene. 2009;441(1-2):22-7.
75. Pightling AW, Petronella N, Pagotto F. Choice of reference sequence and assembler for alignment of Listeria monocytogenes short-read sequence data greatly influences rates of error in SNP analyses. PLoS One. 2014;9(8):e104579.
76. Pightling AW, Petronella N, Pagotto F. Choice of reference-guided sequence assembler and SNP caller for analysis of Listeria monocytogenes short-read sequence data greatly influences rates of error. BMC Res Notes. 2015;8:748.
77. Usongo V, Berry C, Yousfi K, Doualla-Bell F, Labbe G, Johnson R, et al. Impact of the choice of reference genome on the ability of the core genome SNV methodology to distinguish strains of Salmonella enterica serovar Heidelberg. PLoS One. 2018;13(2):e0192233.
78. Bush SJ, Foster D, Eyre DW, Clark EL, De Maio N, Shaw LP, et al. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. Gigascience. 2020;9(2).

79. Consortium CR, the GP, Allix-Beguec C, Arandjelovic I, Bi L, Beckert P, et al. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. N Engl J Med. 2018;379(15):1403-15.
80. Periwal V, Patowary A, Vellarikkal SK, Gupta A, Singh M, Mittal A, et al. Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of Mycobacterium tuberculosis pangenome. PLoS One. 2015;10(4):e0122979.
81. Gao Q, Kripke KE, Saldanha AJ, Yan W, Holmes S, Small PM. Gene expression diversity among Mycobacterium tuberculosis clinical isolates. Microbiology (Reading). 2005;151(Pt 1):5-14.
82. Midori Kato-Maeda JTR, Thomas R. Gingeras, Hugh Salamon, Jorg Drenkow, Nat Smittipat, and Peter M. Small. Comparing Genomes within the Species Mycobacterium tuberculosis. Genome Research. 2001;11:547–54.
83. Elghraoui A, Modlin SJ, Valafar F. SMRT genome assembly corrects reference errors, resolving the genetic basis of virulence in Mycobacterium tuberculosis. BMC Genomics. 2017;18(1):302.
84. Okumura K, Kato M, Kirikae T, Kayano M, Miyoshi-Akiyama T. Construction of a virtual Mycobacterium tuberculosis consensus genome and its application to data from a next generation sequencer. BMC Genomics. 2015;16:218.
85. Lillebaek T, Andersen AB, Rasmussen EM, Kamper-Jorgensen Z, Pedersen MK, Bjorn-Mortensen K, et al. Mycobacterium tuberculosis outbreak strain of Danish origin spreading at worrying rates among greenland-born persons in Denmark and Greenland. J Clin Microbiol. 2013;51(12):4040-4.
86. Bainomugisa A, Duarte T, Lavu E, Pandey S, Coulter C, Marais BJ, et al. A complete high-quality MinION nanopore assembly of an extensively drug-resistant Mycobacterium tuberculosis Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. Microb Genom. 2018;4(7).
87. Wang WF, Lu MJ, Cheng TR, Tang YC, Teng YC, Hwa TY, et al. Genomic Analysis of Mycobacterium tuberculosis Isolates and Construction of a Beijing Lineage Reference Genome. Genome Biol Evol. 2020;12(2):3890-905.
88. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12(4):656-64.
89. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674-6.

90. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9.
91. Harvey KL, Jarocki VM, Charles IG, Djordjevic SP. The Diverse Functional Roles of Elongation Factor Tu (EF-Tu) in Microbial Pathogenesis. Front Microbiol. 2019;10:2351.
92. Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. Genome Res. 2017;27(5):665-76.
93. Kremer L, Gurcha, S. S., Bifani, P., Hitchen, P. G., Baulard, A.,Morris, H. R., Dell, A., Brennan, P. J. and Besra, G. S. Characterization of a putative alpha-mannosyltransferase involved in phosphatidylinositol trimannoside biosynthesis in Mycobacterium tuberculosis. Biochemical Journal. 2002b;363.
94. Meng X, Qi X, Guo H, Cai M, Li C, Zhu J, et al. Novel role for non-homologous end joining in the formation of double minutes in methotrexate-resistant colon cancer cells. J Med Genet. 2015;52(2):135-44.
95. Ali A, Naz A, Soares SC, Bakhtiar M, Tiwari S, Hassan SS, et al. Pan-genome analysis of human gastric pathogen H. pylori: comparative genomics and pathogenomics approaches to identify regions associated with pathogenicity and prediction of potential core therapeutic targets. Biomed Res Int. 2015;2015:139580.
96. Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM, et al. Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains. Genome Res. 2015;25(1):119-28.
97. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, et al. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol. 2008;190(20):6881-93.
98. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature. 2020;585(7823):79-84.

Supplementary materials

Box 1. Commands of creating chain files.

```
#CHAIN file

#OLD: non-ref New: H37Rv

#faToTwoBit NC_000962.3.fa NC_000962.3.2bit

#twoBitInfo NC_000962.3.2bit NC_000962.3.chrom.sizes

##faToTwoBit GCA_000008585.1_ASM858v1_genomic.fna.gz GCA_000008585.1_ASM858v1_genomic.2bit

##twoBitInfo GCA_000008585.1_ASM858v1_genomic.2bit
GCA_000008585.1_ASM858v1_genomic.chrom.sizes

for i in *.fna.gz; do faToTwoBit $i "$i".2bit; done

for i in *.2bit; do twoBitInfo $i "$i".chrom.sizes; done

##blat /BiO3/TB/refs/complete_genome/fasta/GCA_000008585.1_ASM858v1_genomic.fna.gz.2bit
/BiO3/TB/refs/H37Rv/NC_000962.3.fa /BiO3/TB/refs/chain/psl/GCA_000008585.1_ASM858v1_genomic.psl
-tileSize=11 -minScore=30 -minIdentity=90

axtChain -linearGap=medium -psl GCA_000008585.1_ASM858v1_genomic.psl
/BiO3/TB/refs/complete_genome/fasta/GCA_000008585.1_ASM858v1_genomic.fna.gz.2bit
/BiO3/TB/refs/H37Rv/NC_000962.3.2bit ../net/GCA_000008585.1_ASM858v1_genomic.chain

chainNet GCA_000008585.1_ASM858v1_genomic.chain
/BiO3/TB/refs/complete_genome/fasta/GCA_000008585.1_ASM858v1_genomic.fna.gz.2bit.chrom.sizes
/BiO3/TB/refs/H37Rv/NC_000962.3.chrom.sizes GCA_000008585.1_ASM858v1_genomic.net /dev/null

netChainSubset GCA_000008585.1_ASM858v1_genomic.net
GCA_000008585.1_ASM858v1_genomic.chain ../over/GCA_000008585.1_ASM858v1_genomic-Ref.chain

chainSwap GCA_000008585.1_ASM858v1_genomic-Ref.chain Ref-
GCA_000008585.1_ASM858v1_genomic.chainfile
```

Supplementary Table S1. LASTZ parameters we performed in our article.

Parameters	Meaning
--gapped	Perform gapped extension of HSPs (high scoring segment pairs), after first reducing them to anchor points.
--gap = 600,150	Set the score penalties for opening and extending a gap.
--hspthresh	Set the HSP score threshold for the x-drop extension method
--seed = 12of19	Seeds require a 19-bp word with matches in 12 specific positions
--ydrop	Set the threshold for terminating gapped extension; this restricts the endpoints of each local alignment by limiting the local region around each anchor in which extension is performed.
--notransition	Don't allow any match positions in seeds to be satisfied by transitions

Group	Scaffold	Set1	Set2	Set3	Set4	Set5
		--gapped --gap = 600,150 --hspthresh = 4500 --seed = 12of19 --notransition --ydrop= 15000	--gapped --gap = 600,150 --hspthresh=4000 --seed = 12of19 --notransition --ydrop= 15000	--gapped --gap = 600,150 --hspthresh =5000 --seed = 12of19 --notransition --ydrop= 15000	--gapped --gap = 600,150 --hspthresh =5000 --seed = 12of19 --notransition --ydrop= 10000	--gapped --gap = 600,150 --hspthresh = 5000 --seed = 12of19 --notransition --ydrop= 20000
1	KV784802.1	99.65	99.40	99.41	99.40	99.40
2	KV784739.1	97.32	97.30	97.30	97.30	97.30
3	LPVO02003073.1	0	0	0	0	0

Supplementary Table S2. The match % between scaffolds and GRCh38 applied with different parameter sets

Supplementary Table S3. The 110 regions not on GRCh38 reference of Group 1, 2, and 3 including the regions with more ten reads of more than two samples and the 64 similar sequences of 110 on BLASTn search

Group	Scaffold	Start	End	ID	Title	Species	E-value	Identity (%)	Align length	Covering	Ref
group1	KV784719.1	30209877	30211073	gi|1149048289|gb|KY503317.1|	Homo sapiens clone 099F breakpoint junction genomic sequence	*Homo sapiens*	0	99.9	1196	100.0	Kehr et al. 2017
	KV784719.1	79001435	79003312								
	KV784719.1	93450257	93457305								
	KV784719.1	93468677	93473936								
	KV784720.1	27885478	27886242	gi|1444895772|gb|MH534279.1|	Homo sapiens chr4:79783587-79783591 non-reference unique insertion sequence	*Homo sapiens*	0	99.9	765	100.1	Wong et al. 2018
	KV784723.1	8349058	8350242	gi|1444895845|gb|MH534352.1|	Homo sapiens chr4:181368160-181368169 non-reference unique insertion sequence	*Homo sapiens*	0	99.9	1184	100.0	Wong et al. 2018
	KV784723.1	10287878	10288705	gi|1149050568|gb|KY505596.1|	Homo sapiens clone 2091 breakpoint junction genomic sequence	*Homo sapiens*	0	100.0	827	100.0	Kehr et al. 2017
	KV784723.1	29721947	29722858	gi|1444895820|gb|MH534327.1|	Homo sapiens chr4:160015467-160015475 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	911	100.0	Wong et al. 2018
	KV784723.1	34400553	34401454	gi|1444895813|gb|MH534320.1|	Homo sapiens chr4:155349306-155349248 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	843	93.6	Wong et al. 2018
	KV784724.1	45194420	45194971	gi|1444895632|gb|MH534139.1|	Homo sapiens chr3:45133491-45133740 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	551	100.0	Wong et al. 2018
	KV784725.1	1338204	1338992	gi|1444895063|gb|MH533570.1|	Homo sapiens chr15:33894781-33894683 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	649	82.4	Wong et al. 2018
	KV784726.1	9049712	9056232	gi|1444896040|gb|MH534547.1|	Homo sapiens chr7:9051188-9051328 non-reference unique insertion sequence	*Homo sapiens*	0	99.6	6537	100.3	Wong et al. 2018
	KV784726.1	31660959	31661535	gi|1444896055|gb|MH534562.1|	Homo sapiens chr7:31657166-31657167 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	576	100.0	Wong et al. 2018
	KV784727.1	1910039	1911015	gi|1444895415|gb|MH533922.1|	Homo sapiens chr2:128082412-128082368 non-reference unique insertion sequence	*Homo sapiens*	0	99.6	934	95.7	Wong et al. 2018

KV784727.1	2800890	2801225	gi|1444895413|gb|MH533920.1|	Homo sapiens chr2:127192753-127192757 non-reference unique insertion sequence	*Homo sapiens*	6.15E-170	100.0	335	100.0	Wong et al. 2018
KV784727.1	18558006	18558391								
KV784727.1	18559929	18560255								
KV784728.1	32362110	32362747								
KV784730.1	25575121	25576012								
KV784730.1	35420749	35420904								
KV784731.1	15610446	15612082	gi|1395189198|gb|AC277922.1|	Homo sapiens chromosome 5 clone CH17-423E10, complete sequence	*Homo sapiens*	0	99.9	1636	100.0	
KV784734.1	23227641	23232362	gi|1444894887|gb|MH533394.1|	Homo sapiens chr12:61061267-61061266 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	4720	100.0	Wong et al. 2018
KV784734.1	77101351	77102330	gi|1444894916|gb|MH533423.1|	Homo sapiens chr12:114828447-114828445 non-reference unique insertion sequence	*Homo sapiens*	0	99.6	977	99.8	Wong et al. 2018
KV784736.1	6179367	6184327	gi|671686054|emb|AL512455.8|	Human DNA sequence from clone RP11-380E6 on chromosome 6, complete sequence	*Homo sapiens*	0	100.0	4960	100.0	
KV784736.1	18432910	18435802	gi|157385269|gb|AC206742.4|	Homo sapiens FOSMID clone ABC9-41243500D16 from chromosome 6, complete sequence	*Homo sapiens*	0	100.0	2892	100.0	
KV784737.1	881299	882209								
KV784738.1	33431793	33434575								
KV784741.1	24128831	24129687								
KV784741.1	81737080	81738171								
KV784742.1	40854245	40854927	gi|1149051568|gb|KY506596.1|	Homo sapiens clone 3091 breakpoint junction genomic sequence	*Homo sapiens*	0	99.9	682	100.0	Kehr et al. 2017
KV784747.1	1225640	1227675	gi|1444895954|gb|MH534461.1|	Homo sapiens chr6:28176186-28176185 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	2034	100.0	Wong et al. 2018
KV784747.1	17495881	17496490	gi|1353793181|gb|CP027091.1|	Bos mutus isolate yakQH1 chromosome 23	*Bos mutus*	7.01E-65	70.1	591	97.0	
KV784754.1	50233885	50235842	gi|1444896213|gb|MH534720.1|	Homo sapiens chr8:136026912-136026908 non-reference unique insertion sequence	*Homo sapiens*	0	99.7	1953	99.8	Wong et al. 2018

KV784754.1	57453598	57455146								
KV784756.1	22164261	22164870	gi|1444895665|gb|MH534172.1|	Homo sapiens chr3:103593531-103593528 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	606	99.5	Wong et al. 2018
KV784757.1	31404118	31407278								
KV784760.1	7628417	7628679								
KV784760.1	12907992	12909035								
KV784761.1	2374584	2375127	gi|1149048279|gb|KY503307.1|	Homo sapiens clone 089F breakpoint junction genomic sequence	*Homo sapiens*	0	99.6	543	100.0	Kehr et al. 2017
KV784762.1	644258	644871								
KV784762.1	646383	646607								
KV784762.1	941858	944985	gi|1142969548|gb|KY429348.1|	Homo sapiens clone CHM1_19_1162229_1162230 genomic sequence	*Homo sapiens*	0	99.0	3112	99.5	Kehr et al. 2017
KV784768.1	3651009	3659687								
KV784772.1	6472482	6482727	gi|164607370|gb|AC193179.2|	Pan troglodytes BAC clone CH251-322A3 from chromosome x, complete sequence	*Pan troglodytes*	0	99.0	9093	88.8	
KV784772.1	15341131	15341705	gi|1444896333|gb|MH534840.1|	Homo sapiens chrX:99134437-99134468 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	574	100.0	Wong et al. 2018
KV784773.1	40559	40646								
KV784774.1	385408	388147								
KV784780.1	1170968	1171114								
KV784797.1	27752714	27755221	gi|1444894579|gb|MH533086.1|	Homo sapiens chr1:93876121-93876139 non-reference unique insertion sequence	*Homo sapiens*	0	99.7	2508	100.0	Wong et al. 2018
KV784800.1	9712941	9714910								
KV784800.1	13617320	13618083	gi|1149049429|gb|KY504457.1|	Homo sapiens clone 0952 breakpoint junction genomic sequence	*Homo sapiens*	0	100.0	753	98.7	Kehr et al. 2017
KV784802.1	1344895	1349218								
KV784803.1	15594781	15596171	gi|1444895039|gb|MH533546.1|	Homo sapiens chr14:88711382-88711377 non-reference unique insertion sequence	*Homo sapiens*	0	99.9	1386	99.7	Wong et al. 2018

KV784803.1	21187527	21189031	gi|1444895031|gb|MH533538.1|	Homo sapiens chr14:83120833-83120831 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	1502	99.9	Wong et al. 2018
KV784803.1	82104835	82105368	gi|1444894995|gb|MH533502.1|	Homo sapiens chr14:22306949-22306946 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	530	99.4	Wong et al. 2018
KV784804.1	4078651	4079471	gi|1444895204|gb|MH533711.1|	Homo sapiens chr17:40523190-40523190 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	820	100.0	Wong et al. 2018
KV784805.1	20610925	20611477	gi|1444896179|gb|MH534686.1|	Homo sapiens chr8:65040539-65040543 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	552	100.0	Wong et al. 2018
KV784805.1	32075963	32076918								
KV784805.1	52976540	52978099	gi|1444896160|gb|MH534667.1|	Homo sapiens chr8:30740468-30740442 non-reference unique insertion sequence	*Homo sapiens*	0	98.8	1539	98.7	Wong et al. 2018
KV784805.1	56280501	56281043	gi|1444896155|gb|MH534662.1|	Homo sapiens chr8:27438465-27438463 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	540	99.6	Wong et al. 2018
KV784806.1	2835010	2835679	gi|1149050196|gb|KY505224.1|	Homo sapiens clone 1719 breakpoint junction genomic sequence	*Homo sapiens*	0	100.0	669	100.0	Kehr et al. 2017
KV784806.1	31982729	31985575								
KV784806.1	65330230	65332390	gi|272991732|gb|GU267905.1|	Homo sapiens contig freeze2_3024 genomic sequence	*Homo sapiens*	0	100.0	1914	88.6	
KV784806.1	77428763	77428904								
KV784806.1	82987348	82989812								
KV784811.1	3732844	3735258	gi|1444896080|gb|MH534587.1|	Homo sapiens chr7:68761514-68761512 non-reference unique insertion sequence	*Homo sapiens*	0	99.8	2413	100.0	Wong et al. 2018
LPVO02000023.1	26736264	26738235								
LPVO02000045.1	5682802	5683381	gi|1444894780|gb|MH533287.1|	Homo sapiens chr11:42959928-42959928 non-reference unique insertion sequence	*Homo sapiens*	0	99.1	579	100.0	Wong et al. 2018
LPVO02000140.1	366103	367210	gi|1149049876|gb|KY504904.1|	Homo sapiens clone 1399 breakpoint junction genomic sequence	*Homo sapiens*	0	99.6	988	89.3	Kehr et al. 2017
LPVO02000140.1	1171059	1172590	gi|157694627|gb|AC192208.3|	Pan troglodytes BAC clone CH251-8K13 from chromosome 14, complete sequence	*Pan troglodytes*	9.66E-163	70.3	1555	101.6	

	LPVO02000186.1	2130776	2133273	gi|1444895660|gb|MH534167.1|	Homo sapiens chr3:95825553-95825556 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	2497	100.0	Wong et al. 2018
	LPVO02000190.1	4768875	4770445								
	LPVO02000190.1	6164420	6164742								
	LPVO02000191.1	8716004	8716719								
	LPVO02000230.1	3020481	3020803								
	LPVO02000231.1	510488	533621								
	LPVO02000257.1	880951	885107								
	LPVO02000351.1	4447475	4449128								
	LPVO02000423.1	11658344	11659150	gi|1149048356|gb|KY503384.1|	Homo sapiens clone 162F breakpoint junction genomic sequence	*Homo sapiens*	0	98.5	752	93.3	Kehr et al. 2017
	LPVO02000423.1	13810934	13811513	gi|1444894825|gb|MH533332.1|	Homo sapiens chr11:104078567-104078909 non-reference unique insertion sequence	*Homo sapiens*	0	100.0	578	99.8	Wong et al. 2018
	LPVO02000492.1	16244000	16246658								
	LPVO02000493.1	1700956	1701530	gi|1444894812|gb|MH533319.1|	Homo sapiens chr11:87455440-87455447 non-reference unique insertion sequence	*Homo sapiens*	0	99.8	574	100.0	Wong et al. 2018
	LPVO02000493.1	6515210	6515927	gi|1444894805|gb|MH533312.1|	Homo sapiens chr11:82643721-82643728 non-reference unique insertion sequence	*Homo sapiens*	0	99.2	593	82.7	Wong et al. 2018
	LPVO02000621.1	1215578	1217611	gi|1142969735|gb|KY429535.1|	Homo sapiens clone CHM1_X_2238509_2238510 genomic sequence	*Homo sapiens*	0	98.4	1997	98.2	Fan et al. 2017
	LPVO02000621.1	1219247	1220501	gi|1142969735|gb|KY429535.1|	Homo sapiens clone CHM1_X_2238509_2238510 genomic sequence	*Homo sapiens*	0	94.1	1293	103.1	Fan et al. 2017
	LPVO02000630.1	917544	920306								
group2	KV784740.1	20117326	20118654								

	KV784740.1	23347990	23348752
	KV784740.1	49263424	49264109
	KV784763.1	490557	494476
	LPVO02000148.1	1228733	1229240
	LPVO02000185.1	2178029	2179621
	LPVO02000309.1	280240	281953
	LPVO02000309.1	282973	283267
	LPVO02000618.1	2277807	2277891
	LPVO02000673.1	660	1022
	LPVO02000673.1	2130	2541
	LPVO02000674.1	885	923
group3	LPVO02001414.1	24394	28402
	LPVO02001464.1	36803	40809
	LPVO02001464.1	39636	43644
	LPVO02001567.1	1888	5889
	LPVO02001985.1	11729	16962
	LPVO02002168.1	14010	19178

LPVO02002189.1	0	5126	
LPVO02002189.1	15811	20606	
LPVO02002189.1	23010	27010	
LPVO02002730.1	3065	7091	
LPVO02002730.1	5918	10156	
LPVO02002730.1	10110	15002	

Supplementary Table S4. The list of excluded species

The list of excluded species	Count
Mycobacterium tuberculosis variant bovis AF2122/97 (high GC Gram+)	1
Mycobacterium tuberculosis variant bovis (high GC Gram+)	8
Mycobacterium tuberculosis variant bovis BCG str. Tokyo 172 (high GC Gram+)	2
Mycobacterium tuberculosis variant africanum (high GC Gram+)	1
Mycobacterium tuberculosis variant microti (high GC Gram+)	1
Mycobacterium tuberculosis variant bovis BCG (high GC Gram+)	3
Mycobacterium tuberculosis variant bovis BCG str. Korea 1168P (high GC Gram+)	1
Mycobacterium tuberculosis variant africanum GM041182 (high GC Gram+)	1
Mycobacterium tuberculosis variant bovis BCG str. Mexico (high GC Gram+)	1
Mycobacterium tuberculosis variant bovis BCG str. Moreau RDJ (high GC Gram+)	1
Mycobacterium tuberculosis variant bovis BCG str. ATCC 35743 (high GC Gram+)	1
Mycobacterium tuberculosis variant bovis BCG str. Pasteur 1173P2 (high GC Gram+)	1
Total	22

Supplementary Table S5. The list of annotated genes on ALT sequences.

Total	The number of Gene	Gene name
ALT.v2 and H37Rv	9	fucA plsC deoD sodA narH aroD gyrB hisA garA
Only ALT.v2	914	thpR, ald_2, cca_2, pimB_3, parA_2, purK_2, cut3_4, smc_3, ldh2, caeA_2, glgE1, glgM_2, mviN_2, pfkA_2, yjiB_1, sigL_3, mhpB, echA8_4, metP, comR, sucC_2, recD2, mmpS4_3, nreC_1, dasA, gtaB_2, hcaB_5, devS_2, ytrA_2, manA_2, wbbD, alsS, lpqB_2, aam_2, dus_2, btuD_6, camK, infC_2, caeB_3, dgt_2, lipB_2, ccpA, orn_2, sasA_2, ftsE_2, metXA_2, ppa_2, aidA, yhdN_2, nlhH_5, ppsB_2, paaG_6, glgE_2, ricR_2, gltB_2, selB, sapB, hldD, fdnH, panB_2, paaG_7, hpf_2, fgd_4, bdcA, msrP_1, baiE_2, iolT, yxlF, exoA, tilS_2, uvrD1_2, cpnT_2, rpsO_2, proB_2, aroK_2, nagF, mrp_2, trpE_2, moaC2_2, ppgK_2, mmpS5_2, dcsG_2, trpA_2, aes_7, rip1_2, mmpS4_4, ogt_2, rph_2, feoB, murG_2, nudL_2, golD, ephG_2, lipY_6, glpQ1_2, yciC, rplC_2, ktrA, albA, gltC, nhaK, kdpB_2, ppsD_2, gltX1, Hgd, etfB_2, metQ, aroC_2, cysQ_2, yjiB_2, mhpA, sasA_1, ctpG_2, yfiR, fadA_5, hmgA, mutY_2, aroH, ribY, yidE_2, yqeN, suhB_2, ylbL_2, rne_2, nanK_2, ftsW_2, sepF_2, hmp_5, parA_1, pstC2_2, ngcG, espG3_2, metY_2, glpC, ugpQ_2, rsmI_2, xylE_1, yidD, moeZ_3, yegS_2, yteP, mdh_2, iolC, dasC_4, glnQ_1, helY_2, bsdB, ponA1_2, pimB_2, ndkA_2, yggS, frc_1, aroB_2, dcd_3, ybaK, btuD_7, scrK, nucS_3, pat_2, ord_4, cmaA1_2, ectD, dnaE1_2, trmL_2, moaA1_2, betI_4, neo_2, mmpS4_5, trkA_4, lprN_5, yedK_2, paaG_4, dadD, thyA_2, fldA_2, tmpC, trpGD_2, deoC_2, folE_2, deoA_2, yfeW_2, echA8_5, sseB_2, rhaS, asd_2, pheA_2, cut3_3, msrP_2, aspS_3, proB_3, acuR, PE3_3, pncB1_3, lysP, pyrF_2, ifcA, ytrA_1, mftF_2, yrdA_2, typA_2, rlmB_3, yknY, nreC_2, hsaD_2, ubiD, htpG_2, eda, apt_2, strE, metF, lipY_5, fhs, mobA_2, hemW_2, dinG_2, ackA_2, sstT, gatA_2, maiA_1, btuD_4, amyS, xerC_4, thlA, ptlE, hisS_2, pckG_2, opuCA, mshA_3, ptsJ, rsfS_2, nagC_3, thiB, pepO_2, dasC_3, etfA_3, dnaN_3, pimC, pds, smc_2, ecm, pfkA2, xylE_2, folE_3, smtB_2, btuD_3, hup_2, dhaM, raaS_2, esxG_2, cycA_2, xseA_3, lldD_2, rsmH_2, pknD_4, infB_2, selD, pheT_2, echA8_6, malP, ald_4, moaE2_3, rpmF_2, pup_2, esxJ_3, bacA_2, rpmA_2, msiK_3, murJ, ybiT_2, glpB,

		hslR, uvrC_2, ribX, fdhA, efp_2, tgs2_2, ksgA_2, ydiO, yajL, tesB_3, truB_2, rip2_2, ychF_2, aceE_2, prmB, gdh_3, crt_3, bla, ruvB_2, cobIJ_2, ahcY_2, fdhD_2, yehY, gph_2, rpsJ_2, pks5_2, oppF, accD5_4, rpmB_3, mmpL4_2, dinG_3, ddlB, gsiD, embR_2, ftsY_2, tadA_4, pfp, kdpD_2, hpt_2, mog_2, prmC_3, manC1, ftsQ_2, recG_2, clpX_2, hsaD_3, adhA_2, tcrX_4, gsiC, caiD_7, dhaK, frr_2, clsA_2, nadE_3, bbsG_2, espI_4, egtE_2, pntA, nixA_2, adhC2, cdd_2, priA_4, mbtH_2, btuE, csd_2, desA3_2, rplY_2, rhlG, malL_3, ispH2_2, recF_2, styD_2, tesB_4, pepPI, nusB_2, ilvB1_2, fdnG-3, dapL, aplIIM, fadD3_4, epsH, rutE, tgt, fbpB_2, mtcA1_2, czcD, murD_2, purB_2, cinA_2, btuD_5, ruvA_2, yehZ_2, metE_2, recR_2, aroQ_2, xerC_5, pgsA_2, macB, pdhD, narG_2, lipY_4, eis_2, dasB_2, frc_3, coaBC_2, whiB1_2, socA, gdh_2, bfrB_2, mtrB_2, auaH_3, fabR_1, ydfJ, truB_3, malX, fdr_2, pafA_2, potA, narX_3, mce2R_2, pknD_5, rpf2, ald_3, gatC_2, lprN_6, menE_7, mtrA_2, mqo_2, rnc_2, nagC_2, mcrC, fadR, ktrB, nadD_2, leuD_2, def_2, yhdN_1, modC_2, pepN_2, hmuU, yngG, lnrN, nimT, thrB_2, styD_3, dut_2, mscL_2, cbh, apaH, appC_2, mak_2, impA_2, nagC_1, yicL, menH_5, fmt_2, opuCB, yccF_2, dxr_2, gabD2_2, ilvC_2, lolD_1, rpmI_2, msrP_4, lcfB_6, folD_2, bacC_3, lnt_3, recF_3, ligA_2, prmC_2, pgi_2, trpS2, cobB_3, dnaB_2, degU, ispE_2, mntB_2, clsA_1, lacA, ilvK, cydD_2, ksdD_2, narK_4, modA_2, mmpS5_3, exoI, idnD, yeaD, folA_2, glyA1_2, truA_2, rpsF_2, nitA, zitB, glfT2_2, esxN_2, yfdE, tspO, nagR_2, clcB_2, cobB_4, ribF_2, ndhB_3, menH_4, thiL_2, COQ3_4, menA_2, chuR, folP1_2, sdrM, dnaK_4, hupR1, mntB_4, glnE_4, galE, glgX_2, car_2, virS_3, melE, pdtaS_2, mrpD, fabG_2, ask_2, polC, xseA_2, birA_2, xpkA, ycdF, nadE_2, cobT_2, carC_1, mraZ_2, cobQ_2, mngR, menE_5, ravA_2, rlmB_2, pno, glyQS_2, gpr, btuD_1, mmpS4_6, purR_1, PPE4_4, proP_1, fabG_3, tagU_2, fldA_1, apeB_2, lpqI_2, pstS2_2, psaA, ung_2, nusA_2, mlaE_2, rplU_2, dasC_2, crcB_3, dpgD, PPE3_2, gpgP_2, tyrS_2, ypdF_2, pyrG_2, secF_2, rarD, ndhB_2, gyrA_3, ghrA, metG_2, acdA_3, desA2_2, mmgC_6, menG_2, hdpA_2, folK_2, carC_2, hsaC_2, caiA_6, fadD3_3, accA1_3, lolD_2, glgB_2, pstA1_2, speA, rep_4, cobB_5, pyrK, fabG_5, lcfB_5, xseB_3, ybjI, pepN_3, hutI_3, sigL_2, poxB, cobK_2, hflX_2, yheS_2, aes_4, sdcS, btuD_2, leuB_3, fni, argF_2, mntB_1, copA, topA_3, kstD_2, pnp_2, COQ5_3, sugA_2, msrP_3, yfkM, fgd_3, fosA, yefM, mnmA_2, ksdD_1, deaD_3, mmpS5_4, cobD_2, fadJ_2, xylA, caiD_6, priA_3, xerC_6, paaG_8, ruvC_2, feaB, smpB_2, nohA, dnaE2_2, menE_6, fadD, xylB_3, rbfA_2, crcB_2, lcpB_2, MTAP, coaD_2, gph_3, nagB, rlmN_2, glmS_2, mraY_2, ephA_4, mutA_2, mntR_2, bcd, cmtR_2, murE_2, secD_2, proS_2, mfd_2, fra, dinB_2, pstS3_2, ppsE_2, choD_2, pheS_2, pheT_3, kimA_2, sucD_2, gatB_2, rbsA, yidE_1,

		mlaE_3, napA_2, mdtK, dtd3, ord_3, carA2, melD_1, xerD_2, srrA, kstR_2, glmM_2, lexA_2, mpa_2, mpdB, hprA, proA_2, lldD_3, clpP2_2, pknD_3, zraS, fabG_6, trxC_3, pstB1_2, nfrA1, nlhH_4, rep_3, mmgC_7, deaD_4, vapC5_3, metK_2, relA_2, etfA_2, ileS_2, ssrA, trkA_3, thrS_2, rafA, dnaN_2, tag_2, ftsZ_2, gloC, rbsK/rbiA_2, mbtK_2, menA_3, bdhA, menH_3, acdA_4, COQ5_4, maiA_2, mmpS4_7, gdh_4, oppA_2, metG_3, apeX, tsf_2, cydB_2, hsaA_3, rimP_2, greA_2, dhmA, crcB_5, argH_2, murC_3, nirQ, drrA_3, fusA_2, malQ_2, pbpB_2, selA, fadK_3, mrpC, yidC_2, ydaP_1, trpGD_1, ethA_4, glnQ_2, glnE_2, upp_2, ylmA_2, clpP1_2, codB, tuf_3, pntE, gsiA_2, yheS_3, desR, rpoD, valS_2, yegS_3, ectC, dop_2, auaH_2, fadA_6, leuB_2, fabR_2, aes_5, pdxS_2, nqo1, menH_2, melD_2, dnaG_2, cobL_2, ssuD_1, todF, deaD_5, ltaE, fkbP, dcd_2, glf_2, rpoZ_2, cytR_2, btuD_8, ddl_2, dinB_3, uppP_2, trmH_2, glgC_2, ilvD_2, prs_2, ectB, mbtG_2, alsT_2, lutA, whiB_2, pyrH_2, hisF_2, yoeB, eno_2, ddn_3, lnrL, paaG_5, fgd_2, ydaP_2, nucS_2, ftsQ_3, glnE_3, dps2, rplD_2, proP_2, aspS_2, crgA_2, xecD, trkA_5, ble, desA3_3, ligC_2, btuR_2, sufU, pimA_2, fabG2_3, cynR, dnaA_2, serS_2, obg_2, xseB_2, aspA, murF_2, mmpL4_3, por, fadK_2, crcB_4, mcm, kshA_2, araQ_3, gntR, phrA, hsaA_4, nicB, lipA_2, vgb, cobM_2, metC_2, crtB_2, gloB_5, glnA, yflN, murC_2, codA, malL_2, disA_2, paaG_3, subB, moaA2_2, rbpA_2, guaA_2, mntB_3, xylB_2, trxB_2, nudC_2, spk1, mbtI_2, ilvH_2, eccE5_2, aes_6, map_2, ppx2_2, devR_2, sucP, ycsE, dltA_2, uvrB_2, calA, nhaP, treS_2, hin, fadK_1, pth_2, papA5_2, melC, dasB_1, clpS_3, proC_2, msrP_5, nat_2, uxuA, prfC, mftE_2, murAA, ftsX_2, prfB_2, nimR_2, ppsD_3, dnaA_3, rplI_2, lysN, lspA_2, tfdA, alsT_1, purR_2, fabG_4, frc_2, clpS_2, ilvA_2, rpe_2, ftsH_3, btuD_9, lacZ, topA_2, pncB1_2, pntB_2, rplT_2, ssuD_2, cynS, gyrA_2, pmt_2, yagU, bcp_2, ydaD, fliY, yedI_2, rpsR1_3, pdxT_2, tuf_2, cytR_1, hemH_2, paiA, ptsI, ppsC_2, uctC_3, pncA_2, yiaX1, puuC, soxC, tagU_3, arcC1, fgd_5, metH_2, osmC, mmpL5_2, cnbA, narB, ettA_3, trpB_2, PPE4_3, lpd, ptrB

Supplementary Table S6. The coverage of mapped reads on contigs by the number of samples.

	The number of mapped contigs	The total size of mapped contigs	The total size of mapped sequences on contigs	The mapping coverage*
sample >=1	454	2,599,898	2,598,912	99.96
sample >=2	310	2,422,889	214,649	8.86
sample >=72	31	77,148	43,321	56.15

The mapping coverage $= \frac{\text{"The total size of mapped sequences on contigs"}}{\text{"The total size of mapped contigs"}} * 100$

Supplementary Table S7. The list and summary statistics of annotated 326 variants.

Contig	Position	Average read depth	REF	ALT	N	The frequency of REF allele	The frequency of ALT allele	Gene region	Gene product	Annotation
k141_345	3792	17.4	G	C	119	0.94958	0.0504202	msrP_3 (k141_345:3561-4664)	Protein-methionine-sulfoxide reductase catalytic subunit MsrP	PROKKA
k141_345	3900	18.1	G	A	116	0.991379	0.00862069	msrP_3 (k141_345:3561-4664)	Protein-methionine-sulfoxide reductase catalytic subunit MsrP	PROKKA
k141_345	4632	16.8	G	A	117	0.991453	0.00854701	msrP_3 (k141_345:3561-4664)	Protein-methionine-sulfoxide reductase catalytic subunit MsrP	PROKKA
k141_345	5196	17.4	T	C	123	0.99187	0.00813008	pimC (k141_345:4890-6035)	GDP-mannose-dependent alpha-(1-6)-phosphatidylinositol dimannoside mannosyltransferase	PROKKA
k141_345	5875	16.6	C	T	120	0.95	0.05	pimC (k141_345:4890-6035)	GDP-mannose-dependent alpha-(1-6)-phosphatidylinositol dimannoside mannosyltransferase	PROKKA
k141_375	1078	91.3	C	T	610	0.993443	0.00655738	tuf_3 (k141_375:952-1221)	Elongation factor Tu	PROKKA
k141_375	1090	92.2	A	G	610	0.998361	0.00163934	tuf_3 (k141_375:952-1221)	Elongation factor Tu	PROKKA
k141_345	712	17.6	G	T	119	0.991597	0.00840336	mmpL5_2 (k141_345:534-3360)	Siderophore exporter MmpL5	PROKKA
k141_345	743	17.8	G	A	120	0.983333	0.0166667	mmpL5_2 (k141_345:534-3360)	Siderophore exporter MmpL5	PROKKA
k141_345	804	17.5	A	G	118	0.855932	0.144068	mmpL5_2 (k141_345:534-3360)	Siderophore exporter MmpL5	PROKKA
k141_345	838	17.7	G	A	120	0.991667	0.00833333	mmpL5_2 (k141_345:534-3360)	Siderophore exporter MmpL5	PROKKA
k141_345	933	18.0	C	A	126	0.992063	0.00793651	mmpL5_2 (k141_345:534-3360)	Siderophore exporter MmpL5	PROKKA
k141_345	1015	17.6	T	C	125	0.992	0.008	mmpL5_2 (k141_345:534-3360)	Siderophore exporter MmpL5	PROKKA
k141_345	1660	17.4	G	A	117	0.982906	0.017094	mmpL5_2 (k141_345:534-3360)	Siderophore exporter MmpL5	PROKKA
k141_345	1748	16.8	C	A	119	0.857143	0.142857	mmpL5_2 (k141_345:534-3360)	Siderophore exporter MmpL5	PROKKA
k141_345	2650	16.9	C	T	115	0.991304	0.00869565	mmpL5_2 (k141_345:534-3360)	Siderophore exporter MmpL5	PROKKA

k141_345	3088	16.9	C	T	116	0.982759	0.0172414	mmpL5_2 (k141_345:534-3360)	Siderophore exporter MmpL5	PROKKA
k141_345	3096	16.9	A	C	115	0.947826	0.0521739	mmpL5_2 (k141_345:534-3360)	Siderophore exporter MmpL5	PROKKA
k141_143	2865	108.5	T	C	669	0.994021	0.00597907	moaA1_2 (k141_143:2612-3748)	GTP 3',8-cyclase 1	PROKKA
k141_143	2989	109.4	C	T	667	0.997001	0.0029985	moaA1_2 (k141_143:2612-3748)	GTP 3',8-cyclase 1	PROKKA
k141_143	3026	109.9	C	T	668	0.998503	0.00149701	moaA1_2 (k141_143:2612-3748)	GTP 3',8-cyclase 1	PROKKA
k141_143	3050	109.9	T	C	668	0.997006	0.00299401	moaA1_2 (k141_143:2612-3748)	GTP 3',8-cyclase 1	PROKKA
k141_143	3085	109.8	T	C	668	0.998503	0.00149701	moaA1_2 (k141_143:2612-3748)	GTP 3',8-cyclase 1	PROKKA
k141_143	3175	108.2	G	A	665	0.998496	0.00150376	moaA1_2 (k141_143:2612-3748)	GTP 3',8-cyclase 1	PROKKA
k141_143	3255	110.0	T	G	665	0.998496	0.00150376	moaA1_2 (k141_143:2612-3748)	GTP 3',8-cyclase 1	PROKKA
k141_143	3256	109.9	C	A	665	0.996992	0.00300752	moaA1_2 (k141_143:2612-3748)	GTP 3',8-cyclase 1	PROKKA
k141_143	3504	125.5	C	G	670	0.998507	0.00149254	moaA1_2 (k141_143:2612-3748)	GTP 3',8-cyclase 1	PROKKA
k141_143	3698	120.0	T	C	673	0.998514	0.00148588	moaA1_2 (k141_143:2612-3748)	GTP 3',8-cyclase 1	PROKKA
k141_143	1070	80.9	C	T	492	0.995935	0.00406504	embR_2 (k141_143:894-2039)	Transcriptional regulatory protein EmbR	PROKKA
k141_143	1432	82.9	C	T	502	0.998008	0.00199203	embR_2 (k141_143:894-2039)	Transcriptional regulatory protein EmbR	PROKKA
k141_143	1494	84.7	G	A	497	0.997988	0.00201207	embR_2 (k141_143:894-2039)	Transcriptional regulatory protein EmbR	PROKKA
k141_143	1738	82.0	G	A	556	0.994604	0.00539568	embR_2 (k141_143:894-2039)	Transcriptional regulatory protein EmbR	PROKKA
k141_143	1797	76.9	C	T	490	0.997959	0.00204082	embR_2 (k141_143:894-2039)	Transcriptional regulatory protein EmbR	PROKKA
k141_143	1915	108.5	T	C	666	0.998498	0.0015015	embR_2 (k141_143:894-2039)	Transcriptional regulatory protein EmbR	PROKKA
k141_143	1978	108.2	C	T	668	0.997006	0.00299401	embR_2 (k141_143:894-2039)	Transcriptional regulatory protein EmbR	PROKKA
k141_143	1981	107.5	A	T	668	0.998503	0.00149701	embR_2 (k141_143:894-2039)	Transcriptional regulatory protein EmbR	PROKKA
k141_527	484	57.3	C	G	692	0.998555	0.00144509	NKLMHFJF_05921 (k141_527:173-499)	hypothetical protein	PROKKA
k141_258	521	90.1	T	C	646	0.580495	0.419505	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	545	88.7	C	T	648	0.998457	0.00154321	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	586	86.9	G	A	646	0.987616	0.0123839	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	588	87.1	G	T	646	0.998452	0.00154799	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	637	86.5	C	G	647	0.998454	0.0015456	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA

k141_258	680	85.2	G	A	647	0.998454	0.0015456	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	707	83.9	G	A	644	0.998447	0.0015528	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	793	69.7	C	T	582	0.987973	0.0120275	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	804	88.3	G	A	600	0.56	0.44	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	813	81.3	C	T	622	0.734727	0.265273	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	818	82.4	T	A	631	0.825674	0.174326	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	819	83.0	G	A	633	0.840442	0.159558	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	824	84.8	T	C	692	0.998555	0.00144509	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	829	85.2	G	A	648	0.929012	0.0709877	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	839	86.3	C	T	662	0.942598	0.0574018	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	853	47.5	C	G,*	663	0.948718	0.0377074	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	859	49.5	G	A,*	663	0.950226	0.0361991	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	861	50.5	A	C,*	663	0.951735	0.0346908	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	864	51.7	A	C,*	661	0.951589	0.0347958	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	869	53.4	G	C,*	664	0.951807	0.0346386	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	877	57.1	G	A,*	664	0.951807	0.0346386	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	883	59.0	C	G	665	0.95188	0.0481203	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	888	60.9	G	C	664	0.953313	0.0466867	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	1019	99.9	G	A	634	0.998423	0.00157729	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	1043	100.6	G	A	634	0.998423	0.00157729	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	1155	100.8	C	A	635	0.974803	0.0251969	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	1326	97.0	C	T	641	0.99844	0.00156006	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	1650	99.4	G	A	643	0.990669	0.00933126	NKLMHFJF_04534 (k141_258:28-2007)	putative PPE family protein PPE40	PROKKA
k141_258	2276	104.7	C	T	647	0.972179	0.0278207	NKLMHFJF_04535 (K141_258:2255-3142)	putative PPE family protein PPE42	PROKKA
k141_258	2312	100.1	G	A	645	0.99845	0.00155039	NKLMHFJF_04535 (K141_258:2255-3142)	putative PPE family protein PPE42	PROKKA
k141_258	2351	99.7	G	C	644	0.947205	0.052795	NKLMHFJF_04535 (K141_258:2255-3142)	putative PPE family protein PPE42	PROKKA
k141_258	2500	100.5	C	T	651	0.998464	0.0015361	NKLMHFJF_04535 (K141_258:2255-3142)	putative PPE family protein PPE42	PROKKA
k141_258	2527	101.8	G	A	652	0.998466	0.00153374	NKLMHFJF_04535 (K141_258:2255-3142)	putative PPE family protein PPE42	PROKKA
k141_258	2552	102.5	C	T	653	0.992343	0.00765697	NKLMHFJF_04535 (K141_258:2255-3142)	putative PPE family protein PPE42	PROKKA
k141_258	2652	105.8	C	T	649	0.996918	0.00308166	NKLMHFJF_04535 (K141_258:2255-3142)	putative PPE family protein PPE42	PROKKA
k141_258	2730	98.3	C	T	644	0.998447	0.0015528	NKLMHFJF_04535 (K141_258:2255-3142)	putative PPE family protein PPE42	PROKKA
k141_258	2824	87.7	C	A	641	0.99688	0.00312012	NKLMHFJF_04535 (K141_258:2255-3142)	putative PPE family protein PPE42	PROKKA
k141_258	2850	86.7	C	T	640	0.996875	0.003125	NKLMHFJF_04535 (K141_258:2255-3142)	putative PPE family protein PPE42	PROKKA
k141_258	3018	39.9	T	G	631	0	1	NKLMHFJF_04535 (K141_258:2255-3142)	putative PPE family protein PPE42	PROKKA

k141_258	3019	39.5	C	T	634	0.996845	0.00315457	NKLMHFJF_04535 (K141_258:2255-3142)	putative PPE family protein PPE42	PROKKA
k141_284	230	100.2	A	G	620	0.998387	0.0016129	NKLMHFJF_04590 (k141_284: 62-829)	putative PPE family protein PPE51	PROKKA
k141_284	490	74.7	G	A	462	0.997835	0.0021645	NKLMHFJF_04590 (k141_284: 62-829)	putative PPE family protein PPE51	PROKKA
k141_146	188	90.4	C	T,*	720	0.993056	0.00277778	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	193	90.0	G	A,*	720	0.994444	0.00138889	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	348	76.1	C	G,*	720	0.997222	0.00138889	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	353	73.5	C	G,*	719	0.962448	0.00139082	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	356	73.4	C	G,*	719	0.962448	0.00278164	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	363	74.9	C	T,*	720	0.993056	0.00555556	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	366	74.7	C	G,*	721	0.997226	0.00138696	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	369	74.2	C	A,*	720	0.997222	0.00138889	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	384	71.4	T	A,*	720	0.997222	0.00138889	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	387	70.8	G	C,*	720	0.997222	0.00138889	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	389	70.1	A	G,*	720	0.997222	0.00138889	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	408	67.1	T	C,*	715	0.997203	0.0013986	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	442	66.8	G	A,*	715	0.997203	0.0013986	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	445	66.8	G	A,*	715	0.997203	0.0013986	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	446	66.5	G	C,*	716	0.997207	0.00139665	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	447	66.4	C	T,*	716	0.997207	0.00139665	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	456	65.2	C	T,*	717	0.997211	0.0013947	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	505	60.3	C	A,*	704	0.984375	0.00142045	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	507	60.2	C	G,*	709	0.984485	0.00141044	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	510	60.1	C	G,*	710	0.984507	0.00140845	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	516	57.5	C	G,*	709	0.984485	0.00141044	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	526	56.0	G	A,*	705	0.97305	0.012766	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	538	55.8	C	T,G,*	702	0.982906	0.0014245	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	540	55.4	C	T,*	709	0.963329	0.022567	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	574	50.1	G	A,*	693	0.98557	0.001443	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	598	50.4	G	A,*	679	0.992636	0.00294551	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	660	44.7	C	A,*	646	0.995356	0.00154799	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	674	48.5	G	T,*	639	0.953052	0.00312989	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	680	51.8	C	G,*	640	0.9875	0.003125	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	708	49.9	C	T,*	654	0.984709	0.00611621	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	758	50.2	A	T,*	651	0.99232	0.0015361	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA

k141_146	901	56.0	C	T,*	702	0.994302	0.002849	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	904	56.0	G	T,*	701	0.99572	0.00142653	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	905	56.0	C	T,*	701	0.99572	0.00142653	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	919	54.5	C	T,*	703	0.992888	0.00284495	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	921	54.8	C	G,*	706	0.941926	0.0552408	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	976	51.2	C	G	708	0.998588	0.00141243	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	977	51.0	C	G	708	0.998588	0.00141243	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	978	51.1	C	G	711	0.998594	0.00140647	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1001	48.4	C	T	711	0.936709	0.0632911	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1007	48.3	G	T	712	0.936798	0.0632022	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1017	47.3	A	G	706	0.998584	0.00141643	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1020	47.9	C	T	707	0.998586	0.00141443	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1054	55.1	C	T	697	0.998565	0.00143472	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1134	80.2	C	T	722	0.998615	0.00138504	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1180	87.3	G	T	722	0.98615	0.0138504	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1201	90.3	C	T	722	0.981994	0.0180055	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1214	92.0	G	T	722	0.99723	0.00277008	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1215	91.7	G	C	722	0.99723	0.00277008	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1217	91.4	G	C	722	0.99723	0.00277008	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1222	92.1	G	A	722	0.99723	0.00277008	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1224	92.1	C	T	722	0.99723	0.00277008	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1251	96.4	C	T	722	0.965374	0.034626	NKLMHFJF_04197 (k141_146:158-1390)	hypothetical protein	PROKKA
k141_146	1986	93.7	G	C,*	721	0.997226	0.00138696	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2004	91.5	G	A,*	720	0.997222	0.00138889	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2007	91.2	A	G,*	720	0.997222	0.00138889	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2008	91.5	T	C,*	721	0.997226	0.00138696	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2126	71.6	T	C	721	0.995839	0.00416089	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2127	71.7	C	G	721	0.995839	0.00416089	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2131	71.7	C	T	718	0.998607	0.00139276	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2134	71.0	G	A	718	0.998607	0.00139276	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2135	70.2	A	C	718	0.998607	0.00139276	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2136	70.2	G	C	718	0.998607	0.00139276	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2140	70.2	C	G	719	0.998609	0.00139082	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2144	68.6	A	C	719	0.998609	0.00139082	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA

k141_146	2156	67.0	T	G	719	0.998609	0.00139082	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2212	75.3	C	G,*	719	0.997218	0.00139082	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2223	75.8	G	T	719	0.998609	0.00139082	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2369	80.3	C	A,*	718	0.981894	0.0139276	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2398	59.1	T	A,*	717	0.981869	0.013947	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2486	53.5	C	A,*	717	0.973501	0.0223152	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2552	52.7	C	T,*	693	0.994228	0.002886	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2566	56.5	C	T,*	692	0.995665	0.00144509	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2609	67.5	C	T,*	707	0.97454	0.0226308	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_146	2725	64.2	G	T	710	0.998592	0.00140845	NKLMHFJF_04198 (k141_146:1869-2759)	hypothetical protein	PROKKA
k141_143	578	71.0	G	A	483	0.995859	0.00414079	NKLMHFJF_04184 (k141_143:362-799)	hypothetical protein	PROKKA
k141_143	655	70.9	G	A	483	0.99793	0.00207039	NKLMHFJF_04184 (k141_143:362-799)	hypothetical protein	PROKKA
k141_143	690	71.2	G	A	481	0.995842	0.004158	NKLMHFJF_04184 (k141_143:362-799)	hypothetical protein	PROKKA
k141_143	709	71.4	C	G	487	0.99384	0.00616016	NKLMHFJF_04184 (k141_143:362-799)	hypothetical protein	PROKKA
k141_565	318	96.4	G	A	694	0.0461095	0.95389	NKLMHFJF_06619 (k141_565:121-465)	hypothetical protein	PROKKA
k141_565	419	101.0	T	G	668	0.998503	0.00149701	NKLMHFJF_06619 (k141_565:121-465)	hypothetical protein	PROKKA
k141_363	3610	122.0	A	G	695	0.998561	0.00143885	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_363	3631	122.4	A	G	695	0.981295	0.018705	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_363	3661	120.9	A	G	695	0.998561	0.00143885	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_363	3687	120.8	C	G	694	0.998559	0.00144092	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_363	3731	117.6	C	T	693	0.998557	0.001443	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_363	3800	120.2	T	C	692	0.998555	0.00144509	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_363	3831	120.8	C	T	691	0.998553	0.00144718	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_363	3877	120.7	C	T	692	0.943642	0.0563584	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_363	3905	119.7	A	C	693	0.998557	0.001443	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_363	3913	120.1	C	G	693	0.998557	0.001443	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_363	3935	119.3	A	G	693	0.998557	0.001443	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_363	4157	122.7	T	C	696	0.998563	0.00143678	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_363	4222	125.0	G	A	695	0.98705	0.0129496	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_363	4308	124.8	T	G	696	0.998563	0.00143678	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_363	4788	137.6	A	G	683	0.998536	0.00146413	NKLMHFJF_04962 (k141_363:3490-4386)	hypothetical protein	PROKKA
k141_375	700	95.7	C	T	614	0.998371	0.00162866	NKLMHFJF_05031 (k141_375:695-874)	hypothetical protein	PROKKA
k141_375	726	94.8	T	C	613	0.998369	0.00163132	NKLMHFJF_05031 (k141_375:695-874)	hypothetical protein	PROKKA
k141_375	729	95.1	C	G	613	0.998369	0.00163132	NKLMHFJF_05031 (k141_375:695-874)	hypothetical protein	PROKKA

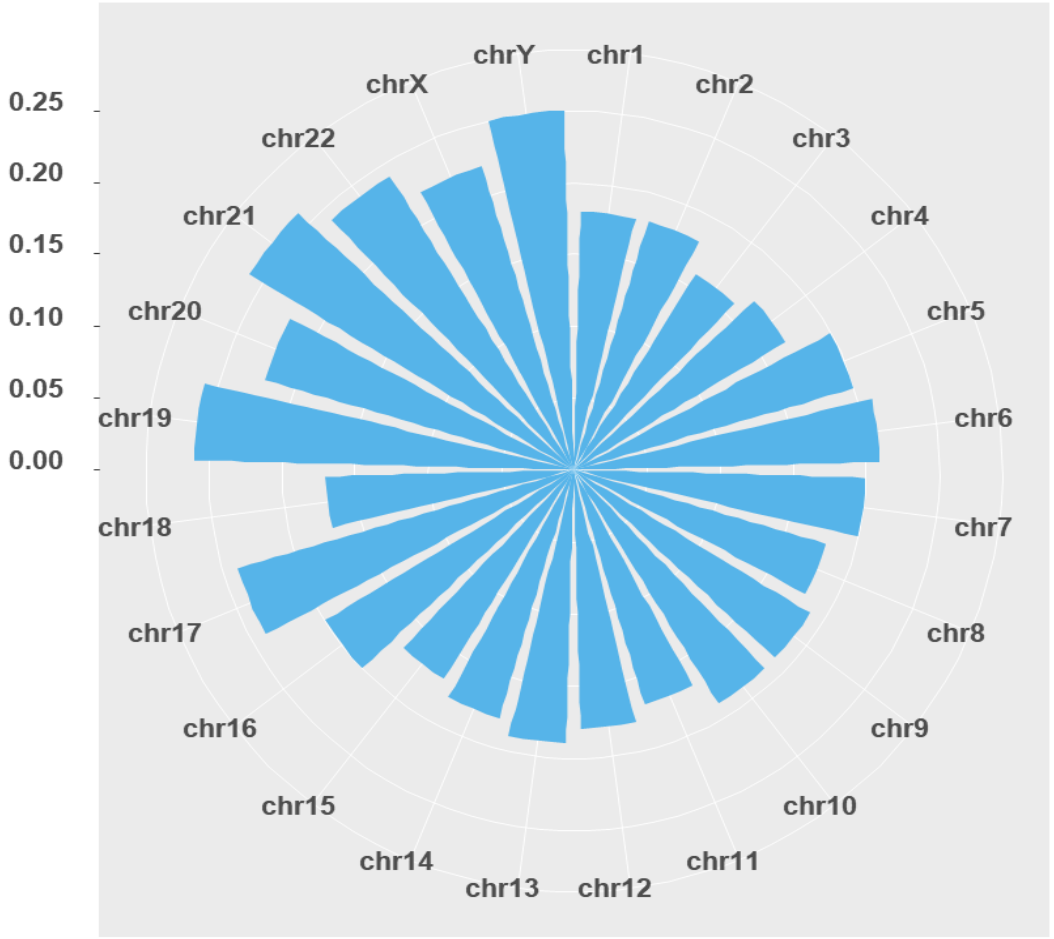
k141_375	732	95.2	T	C	613	0.998369	0.00163132	NKLMHFJF_05031 (k141_375:695-874)	hypothetical protein	PROKKA
k141_375	744	94.8	A	G	613	0.998369	0.00163132	NKLMHFJF_05031 (k141_375:695-874)	hypothetical protein	PROKKA
k141_375	750	94.8	A	C	612	0.998366	0.00163399	NKLMHFJF_05031 (k141_375:695-874)	hypothetical protein	PROKKA
k141_375	753	94.7	C	T	612	0.998366	0.00163399	NKLMHFJF_05031 (k141_375:695-874)	hypothetical protein	PROKKA
k141_375	759	95.1	G	C	612	0.998366	0.00163399	NKLMHFJF_05031 (k141_375:695-874)	hypothetical protein	PROKKA
k141_375	1388	86.6	G	A	610	0.998361	0.00163934	NKLMHFJF_05033 (k141_375:1274-2020)	hypothetical protein	PROKKA
k141_375	1419	87.9	G	A	610	0.996721	0.00327869	NKLMHFJF_05033 (k141_375:1274-2020)	hypothetical protein	PROKKA
k141_375	1421	87.7	A	C	610	0.996721	0.00327869	NKLMHFJF_05033 (k141_375:1274-2020)	hypothetical protein	PROKKA
k141_375	1495	90.0	C	T	611	0.973813	0.0261866	NKLMHFJF_05033 (k141_375:1274-2020)	hypothetical protein	PROKKA
k141_375	1707	88.5	T	A	611	0.998363	0.00163666	NKLMHFJF_05033 (k141_375:1274-2020)	hypothetical protein	PROKKA
k141_375	1744	88.3	G	A	610	0.990164	0.00983607	NKLMHFJF_05033 (k141_375:1274-2020)	hypothetical protein	PROKKA
k141_375	1803	92.2	G	A	616	0.996753	0.00324675	NKLMHFJF_05033 (k141_375:1274-2020)	hypothetical protein	PROKKA
k141_375	1806	92.4	G	C	616	0.998377	0.00162338	NKLMHFJF_05033 (k141_375:1274-2020)	hypothetical protein	PROKKA
k141_375	1851	94.1	C	T	615	0.996748	0.00325203	NKLMHFJF_05033 (k141_375:1274-2020)	hypothetical protein	PROKKA
k141_375	1898	95.3	G	A	617	0.998379	0.00162075	NKLMHFJF_05033 (k141_375:1274-2020)	hypothetical protein	PROKKA
k141_375	1916	95.0	C	T	617	0.991896	0.00810373	NKLMHFJF_05033 (k141_375:1274-2020)	hypothetical protein	PROKKA
k141_375	1925	95.6	G	A	615	0.969106	0.0308943	NKLMHFJF_05033 (k141_375:1274-2020)	hypothetical protein	PROKKA
k141_375	1973	95.9	T	G	615	0.996748	0.00325203	NKLMHFJF_05033 (k141_375:1274-2020)	hypothetical protein	PROKKA
k141_375	2022	96.0	A	G	614	0.995114	0.00488599	NKLMHFJF_05034 (k141_375:2017-2766)	hypothetical protein	PROKKA
k141_375	2038	96.7	C	G	613	0.998369	0.00163132	NKLMHFJF_05034 (k141_375:2017-1766)	hypothetical protein	PROKKA
k141_375	2039	96.3	A	G	613	0.998369	0.00163132	NKLMHFJF_05034 (k141_375:2017-1766)	hypothetical protein	PROKKA
k141_375	2143	93.6	A	C	609	0.998358	0.00164204	NKLMHFJF_05034 (k141_375:2017-1766)	hypothetical protein	PROKKA
k141_375	2154	94.7	G	A	612	0.998366	0.00163399	NKLMHFJF_05034 (k141_375:2017-1766)	hypothetical protein	PROKKA
k141_375	2237	93.3	G	A	611	0.998363	0.00163666	NKLMHFJF_05034 (k141_375:2017-1766)	hypothetical protein	PROKKA
k141_375	2292	91.4	G	A	611	0.97054	0.0294599	NKLMHFJF_05034 (k141_375:2017-1766)	hypothetical protein	PROKKA
k141_375	2344	94.5	G	A	611	0.99509	0.00490998	NKLMHFJF_05034 (k141_375:2017-1766)	hypothetical protein	PROKKA
k141_375	2451	95.6	C	T	614	0.995114	0.00488599	NKLMHFJF_05034 (k141_375:2017-1766)	hypothetical protein	PROKKA
k141_375	2453	95.8	C	T	614	0.998371	0.00162866	NKLMHFJF_05034 (k141_375:2017-1766)	hypothetical protein	PROKKA
k141_375	2480	96.1	T	G	614	0.998371	0.00162866	NKLMHFJF_05034 (k141_375:2017-1766)	hypothetical protein	PROKKA
k141_375	2599	95.2	T	C	616	0.998377	0.00162338	NKLMHFJF_05034 (k141_375:2017-1766)	hypothetical protein	PROKKA
k141_375	2723	91.7	G	A	613	0.880914	0.119086	NKLMHFJF_05034 (k141_375:2017-1766)	hypothetical protein	PROKKA
k141_375	4088	92.5	G	T	615	0.998374	0.00162602	NKLMHFJF_05036 (k141_375:3998-4192)	hypothetical protein	PROKKA
k141_375	4105	94.3	G	T	614	0.998371	0.00162866	NKLMHFJF_05036 (k141_375:3998-4192)	hypothetical protein	PROKKA
k141_375	4132	96.9	C	A	614	0.998371	0.00162866	NKLMHFJF_05036 (k141_375:3998-4192)	hypothetical protein	PROKKA

k141_375	4145	97.2	G	A	616	0.998377	0.00162338	NKLMHFJF_05036 (k141_375:3998-4192)	hypothetical protein	PROKKA
k141_146	1270	96.6	G	A	722	0.998615	0.00138504	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1300	92.9	A	G	722	0.984765	0.0152355	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1306	93.0	A	G	722	0.98615	0.0138504	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1307	93.0	T	G	722	0.98615	0.0138504	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1309	93.3	C	G	722	0.98615	0.0138504	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1342	97.5	G	C	722	0.174515	0.825485	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1376	100.6	G	A	722	0.99723	0.00277008	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1377	100.1	T	C	722	0.99723	0.00277008	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1408	100.0	A	G	722	0.951524	0.0484765	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1435	101.9	C	G	722	0.950139	0.0498615	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1475	103.6	G	A	723	0.998617	0.00138313	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1490	102.9	A	G	722	0.99446	0.00554017	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1492	102.7	C	T	721	0.932039	0.0679612	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1495	103.1	C	T	721	0.918169	0.0818308	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1613	98.1	C	G	719	0.998609	0.00139082	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1618	94.9	G	T	719	0.998609	0.00139082	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1627	95.2	C	G	719	0.998609	0.00139082	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1661	94.4	T	G	719	0.997218	0.00278164	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1684	95.3	G	A	720	0.998611	0.00138889	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1704	95.8	G	A	719	0.991655	0.00834492	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_146	1775	94.4	A	C	720	0.998611	0.00138889	k141_146: 1261-1878	PE-PGRS family protein	BLAST+
k141_258	1650	99.4	G	A	643	0.990669	0.00933126	K141_258:1495-2004	ppe family protein	BLAST+
k141_258	2044	104.5	C	A	641	0.99844	0.00156006	K141_258:1495-2005	ppe family protein	BLAST+
k141_363	156	112.7	G	C	698	0.998567	0.00143266	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	251	111.2	T	C	698	0.998567	0.00143266	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	332	111.9	T	C	698	0.977077	0.0229226	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	534	110.1	C	T	696	0.998563	0.00143678	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	872	114.5	C	T	696	0.971264	0.0287356	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	933	113.4	C	G	698	0.889685	0.110315	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	937	113.5	G	A	699	0.998569	0.00143062	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	1044	114.7	C	G	700	0.998571	0.00142857	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	1057	114.4	G	C	699	0.998569	0.00143062	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	1077	114.3	G	T	698	0.998567	0.00143266	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+

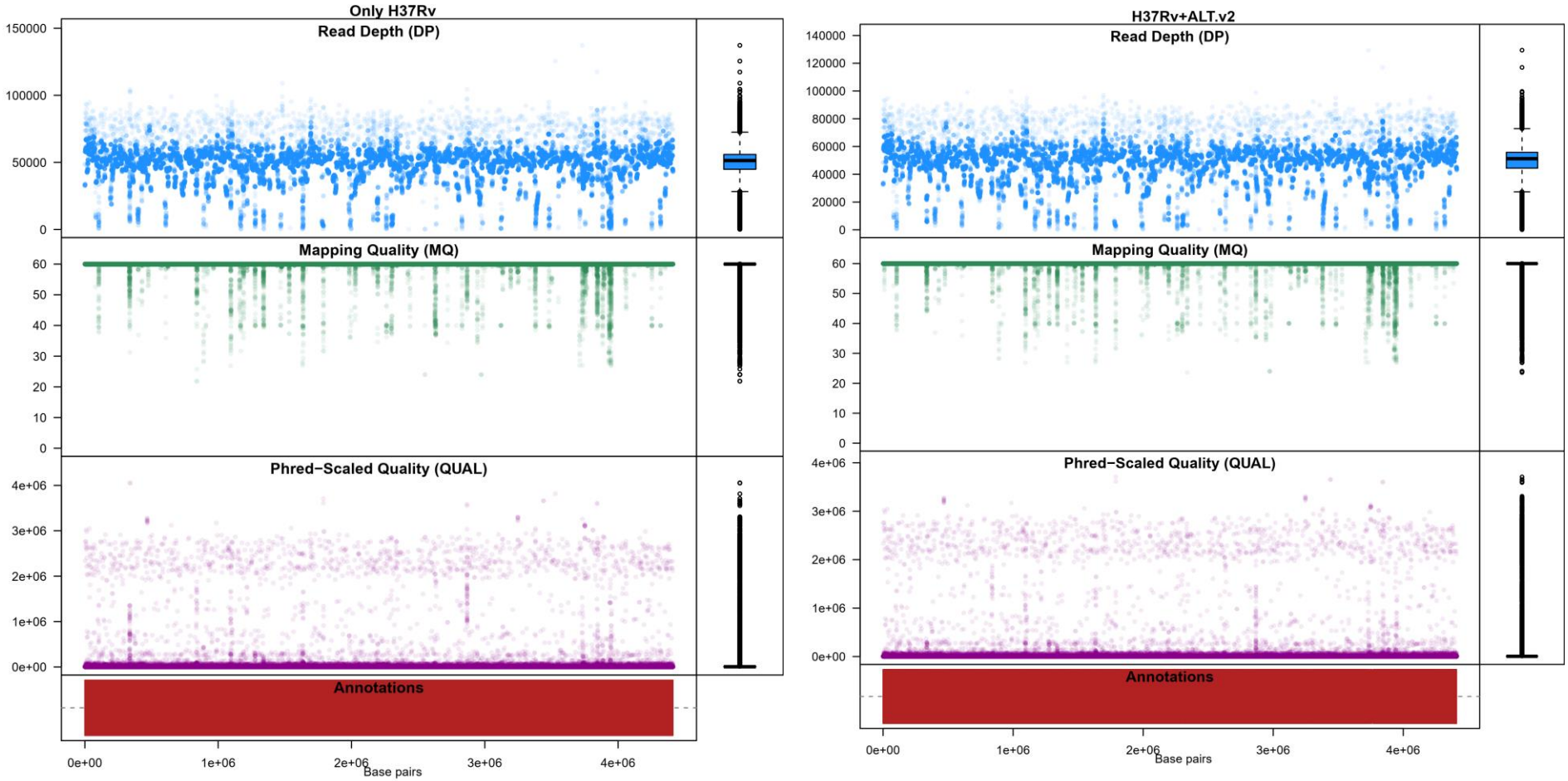
k141_363	1101	114.6	G	A	697	0.997131	0.00286944	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	1181	110.9	T	G	697	0.998565	0.00143472	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	1210	109.3	T	G	697	0.992826	0.0071736	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	1248	109.5	G	A	697	0.998565	0.00143472	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	1389	108.1	G	A	695	0.997122	0.0028777	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	1679	118.5	A	G	696	0.998563	0.00143678	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	1830	119.9	C	T	693	0.998557	0.001443	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	1970	123.2	G	A	693	0.998557	0.001443	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	2184	112.6	C	A	692	0.998555	0.00144509	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	2289	113.2	C	G	694	0.998559	0.00144092	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	2311	112.9	G	A	694	0.994236	0.00576369	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	2456	114.4	A	G	693	0.998557	0.001443	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	2496	115.7	A	T	694	0.998559	0.00144092	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	2612	115.0	G	T	695	0.998561	0.00143885	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	2688	112.8	C	T	695	0.998561	0.00143885	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	2749	111.4	T	G	696	0.971264	0.0287356	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	2768	110.5	T	G	696	0.998563	0.00143678	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	2802	107.9	T	C	696	0.998563	0.00143678	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	2875	111.5	G	A	695	0.998561	0.00143885	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	3109	128.0	T	C	693	0.998557	0.001443	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	3191	128.2	C	A	693	0.997114	0.002886	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	3230	128.3	C	T	693	0.98557	0.01443	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	3237	127.8	C	T	693	0.995671	0.004329	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_363	3272	125.5	T	G	693	0.998557	0.001443	k141_363:1-3303	DEAD/DEAH box helicase family protein	BLAST+
k141_565	658	104.8	G	A	672	0.998512	0.0014881	k141_565:472-1815	HAMP domain containing protein, partial	BLAST+
k141_565	769	104.1	G	C	672	0.997024	0.00297619	k141_565:472-1816	HAMP domain containing protein, partial	BLAST+
k141_565	781	103.5	G	C	672	0.998512	0.0014881	k141_565:472-1817	HAMP domain containing protein, partial	BLAST+
k141_565	936	106.1	T	C	671	0.997019	0.00298063	k141_565:472-1818	HAMP domain containing protein, partial	BLAST+
k141_565	977	107.2	A	G	671	0.997019	0.00298063	k141_565:472-1819	HAMP domain containing protein, partial	BLAST+
k141_565	1081	110.5	G	A	669	0.998505	0.00149477	k141_565:472-1820	HAMP domain containing protein, partial	BLAST+
k141_565	1092	111.6	G	A	669	0.998505	0.00149477	k141_565:472-1821	HAMP domain containing protein, partial	BLAST+
k141_565	1093	111.8	G	C	669	0.998505	0.00149477	k141_565:472-1822	HAMP domain containing protein, partial	BLAST+
k141_565	1095	110.6	A	C	669	0.998505	0.00149477	k141_565:472-1823	HAMP domain containing protein, partial	BLAST+
k141_565	1101	110.4	C	G	669	0.998505	0.00149477	k141_565:472-1824	HAMP domain containing protein, partial	BLAST+

k141_565	1104	109.9	A	G	670	0.998507	0.00149254	k141_565:472-1825	HAMP domain containing protein, partial	BLAST+
k141_565	1106	109.4	T	C	670	0.998507	0.00149254	k141_565:472-1826	HAMP domain containing protein, partial	BLAST+
k141_565	1114	108.8	A	C	670	0.998507	0.00149254	k141_565:472-1827	HAMP domain containing protein, partial	BLAST+
k141_565	1116	108.7	C	G	670	0.998507	0.00149254	k141_565:472-1828	HAMP domain containing protein, partial	BLAST+
k141_565	1155	106.1	G	T	671	0.99851	0.00149031	k141_565:472-1829	HAMP domain containing protein, partial	BLAST+
k141_565	1164	106.3	C	G	671	0.997019	0.00298063	k141_565:472-1830	HAMP domain containing protein, partial	BLAST+
k141_565	1241	105.0	T	C	672	0.99256	0.00744048	k141_565:472-1831	HAMP domain containing protein, partial	BLAST+
k141_565	1281	99.4	G	A	654	0.998471	0.00152905	k141_565:472-1832	HAMP domain containing protein, partial	BLAST+
k141_565	1288	98.1	A	T	654	0.937309	0.0626911	k141_565:472-1833	HAMP domain containing protein, partial	BLAST+
k141_565	1293	98.4	G	T	652	0.937117	0.0628834	k141_565:472-1834	HAMP domain containing protein, partial	BLAST+
k141_565	1351	98.9	G	A	653	0.998469	0.00153139	k141_565:472-1835	HAMP domain containing protein, partial	BLAST+
k141_179	142	71.0	G	C	659	0.995448	0.00455235	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	148	73.3	G	T	659	0.998483	0.00151745	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	252	93.9	G	C	659	0.892261	0.107739	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	287	93.9	G	A	658	0.99848	0.00151976	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	312	94.5	G	C	660	0.998485	0.00151515	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	361	97.1	G	A	659	0.998483	0.00151745	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	439	98.1	G	A	661	0.998487	0.00151286	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	523	92.8	G	A	663	0.998492	0.0015083	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	673	83.6	G	A	662	0.998489	0.00151057	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	690	84.8	T	C	662	0.998489	0.00151057	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	836	76.5	C	G	660	0.872727	0.127273	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	837	76.2	G	T	660	0.872727	0.127273	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	841	75.5	G	C	662	0.996979	0.00302115	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	845	74.6	G	A	662	0.996979	0.00302115	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	862	69.9	A	G	661	0.996974	0.00302572	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	877	64.9	G	A	661	0.990923	0.00907716	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	906	52.8	C	G	661	0.981846	0.0181543	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_179	917	47.3	A	G	661	0.995461	0.00453858	k141_179:73-951	pentapeptide repeat-containing protein	BLAST+
k141_561	765	86.2	A	G	558	0.0913978	0.908602	k141_561:306-926	Universal stress protein family	BLAST+
k141_561	844	76.4	T	C	534	0.996255	0.00374532	k141_561:306-926	Universal stress protein family	BLAST+
k141_561	923	62.2	A	G	533	0.998124	0.00187617	k141_561:306-926	Universal stress protein family	BLAST+
k141_527	29	108.4	T	C	693	0.997114	0.002886	k141_527:1-465	alpha-mannosidase	BLAST+

k141_527	145	106.0	G	A	693	0.998557	0.001443	k141_527:1-465	alpha-mannosidase	BLAST+
k141_527	265	94.3	G	A	693	0.998557	0.001443	k141_527:1-465	alpha-mannosidase	BLAST+
k141_143	2596	112.7	T	C	667	0.995502	0.00449775	k141_143: 2561-3745	molybdopterin cofactor biosynthesis protein A	BLAST+
k141_375	3334	93.4	A	G	615	0.998374	0.00162602	k141_375:3123-4034	PE family protein	BLAST+
k141_375	3529	88.0	A	G	613	0.998369	0.00163132	k141_375:3123-4034	PE family protein	BLAST+
k141_375	3540	88.3	A	G	613	0.998369	0.00163132	k141_375:3123-4034	PE family protein	BLAST+
k141_375	3558	87.3	T	C	612	0.998366	0.00163399	k141_375:3123-4034	PE family protein	BLAST+
k141_375	3751	88.4	C	T	612	0.993464	0.00653595	k141_375:3123-4034	PE family protein	BLAST+
k141_375	3798	88.2	C	A	612	0.998366	0.00163399	k141_375:3123-4034	PE family protein	BLAST+
k141_375	3877	83.2	C	A	613	0.998369	0.00163132	k141_375:3123-4034	PE family protein	BLAST+
k141_375	3907	84.5	G	A	614	0.996743	0.00325733	k141_375:3123-4034	PE family protein	BLAST+
k141_375	3936	83.4	G	A	614	0.998371	0.00162866	k141_375:3123-4034	PE family protein	BLAST+
k141_375	3992	85.1	T	G	613	0.983687	0.0163132	k141_375:3123-4034	PE family protein	BLAST+



Supplementary Figure S1. The average depth (xN) of coverage by chromosomes in Group 1



Supplementary Figure S2. The read depth, mapping quality, and genotype quality of vcf files. The left graph shows the results with only H37Rv. The right graph shows the results with Pan-Reference

Abstract in Korean

참조 게놈의 손실된 유전체 발견 및 범유전체 참조게놈: 인간과 결핵균을 중심으로

서울대학교 자연과학대학

생물정보학 전공

김 지 나

DNA 시퀀싱 기술은 현대 생물학의 중추적인 부분이다. 비용 효율성을 달성하기 위해 대부분의 시퀀싱 플랫폼에서는 참조 게놈에 기반한 리시퀀싱 접근 방식을 사용한다. 참조 게놈은 차세대 시퀀싱(NGS)에서 짧은 리드들을 매핑하고 변이들을 발견하는데 중요한 역할을 하기 때문에 여러 종들에서 참조 게놈들이 존재하고 있다. 예를 들어, 인간에서 GRCh(Genome Reference Consortium의 인간 참조 게놈)는 인간 게놈 프로젝트 이후부터 참조 게놈으로 사용되어져 왔고, 또한 결핵에서는 가장 많이 연구된 변종인 H37Rv이 참조 게놈으로 사용되어 왔다. 이전에는 개인의 유전적 변이들을 결정하는 데 하나의 참조 게놈만이 필요할 것으로 생각되었다. 그러나 참조 게놈이 특정 종의 모든 개인을 대표하는 것인지에 대해서는 여전히 회의적인 시각들이 있다. 많은 연구자들이 다른 인종 또는 혈통 집단들 간의 유전체간의 구조적 변화의 다양성을 지적하면서, 참조 게놈에는 없지만 적어도 소수의 개인들 또는 혈통들에 존재하는 새로운 유전체 서열들을 보고했다. 실제로, 시퀀싱 과정에서 “매핑되지 않은 리드“들이나 잘못된 변이 호출 등을

통해 누락되거나 제한된 정보들이 발생할수 있다. 따라서, 이 연구는 인간 및 미코박테리아 결핵균에서 기준 게놈의 누락된 유전체 영역을 확인하고 그 격차를 해소하는 시도를 하였다.

인간 유전체에서 이 연구는 아프리카 조상을 포함한 50명 이상의 개인 게놈으로 구성된 인간기준 게놈(GRCh38)에서 빠진 부분을 보완하기 위해, 고도로 연속된 게놈 조립체인 AK1을 사용했다. GRCh38에서 누락된 지역을 찾기 위해 기준 게놈(GRCh38)을 AK1과 직접 비교하는 방법과 14명의 전장 유전체 데이터(동아시아 5명, 유럽 4명, 아프리카 5명)에서 “매핑되지 않은 리드”들을 다시 AK1에 붙여보는 방법을 사용하였다.

먼저, GRCh38과 AK1 간의 직접 비교는 두 시퀀스에서 간격을 허용하는 쌍방향 정렬을 설명하는 체인 파일을 사용하였고, 매핑되지 않은 읽기를 사용하는 또 다른 방법은 AK1에 다시 정렬하였는데, 각 방법은 GRCh38에 존재하지 않았던 3,333개의 고유 게놈 영역(사이즈> 200bp)과 38개의 추정 결측 영역(7명 이상의 데이터의 매핑되지 않은 리드들이 붙은 영역)을 각각 발견했다. 또한, 매핑되지 않은 리드들을 사용할 때 여러 인종들의 데이터에서 매핑되지 않은 리드들의 평균 0.90%가 AK1에 새로 정렬되었고, 동아시아 인종의 매핑되지 않은 리드들의 정렬율은 0.95%로 다른 민족에 비해 높다는 것을 확인할수 있었다.

7명이상의 전장 유전체 데이터의 매핑되지 않은 리드들이 정렬된 AK1만의 유전자 서열이자 GRCh38에서는 결측되어 있을것이라 추정되는 영역에 대한 추가 연구를 위해, 본 연구는 BLASTx와 함께 서열을 분석하여 서열의 기능적 역할을 확인해보았고, Repeat Masker를 통해 누락된 것으로 보이는 유전체 영역에 대한 반복서열을 조사하였다.

미코박테리움 결핵균에서는 참조 게놈에서 누락된 부분을 보완하기 위해 다른 방법을 사용하여 이 연구를 수행하였다. 이 연구에서는 결핵균 참조 게놈(H37Rv)의 새로운 범유전자 서열을 구성하였는데, H37Rv에서 대체 서열을 구축하기 위해

176개의 전체 게놈 어셈블리로부터 추출한 시퀀스들(갭 사이즈> 50bp)과 724개의 전장 유전체 데이터에서 추출한 “매핑되지 않은” 리드들을 데노보 어셈블리를 하였다. 그 결과, 454개의 contigs들이 범유전체 시퀀스들로 최종 확정되었다. 본 연구에서는 구성된 범 유전체 시퀀스의 효과를 확인하기 위해 H37Rv만을 사용하는 것과 비교하여 정렬과 변이 호출 결과들을 분석하였다.

결론적으로, 이 연구는 본 연구는 인간 및 미코박테리아 결핵균의 참조 게놈과 염기서열들에 대한 더 많은 이해를 제공한다. 또한, 참조 게놈들에서 누락된 부위에 대한 추가 조사의 필요성을 제기하고, 특히 미코박테리아 결핵균의 유전체 데이터를 실제 사례로 활용하여 참조 게놈에서의 차이를 해소할 수 있는 가능성을 보여주고 있다.

주요어: 참조게놈, 인간, 결핵균, 누락된 정보

학번: 2015-30119