**American Journal of Life Sciences**
2016; 4(6): 146-151
http://www.sciencepublishinggroup.com/j/ajls
doi: 10.11648/j.ajls.20160406.12
ISSN: 2328-5702 (Print); ISSN: 2328-5737 (Online)

# Gap Filling for a Human MHC Haplotype Sequence

**Yuanwei Zhang[1, 2], Tao Zhang[2], Zuhong Lu[1]**

[1]School of Biological Science and Medical Engineering, Southeast University, Nanjing, China

[2]BGI-Shenzhen, Shenzhen, China

## Email address:

zhangyuanwei@genomics.cn (Yuanwei Zhang), tao.zhang@genomics.cn (Tao Zhang), zhlu@seu.edu.cn (Zuhong Lu)

## To cite this article:

Yuanwei Zhang, Tao Zhang, Zuhong Lu. Gap Filling for a Human MHC Haplotype Sequence. *American Journal of Life Sciences.* Vol. 4, No. 6, 2016, pp. 146-151. doi: 10.11648/j.ajls.20160406.12

**Received**: November 15, 2016; **Accepted**: November 28, 2016; **Published**: December 1, 2016

**Abstract:** The major histocompatibility complex (MHC) is recognized as the most variable region in the human genome and has susceptibility to > 100 diseases. We constructed a complete MHC haplotype sequence of MCF cell line by gap filling based on whole genome sequencing (WGS) data. Gaps spanning ~ 1 Mb were filled and 31 genes were annotated in these gaps. This sequence could be used as reference to identify disease associations within this haplotype or similar haplotypes. The method for gap filling can be applied to other MHC haplotypes or other genomic region.

**Keywords:** Gap Filling, MHC, Haplotype, WGS

## 1. Introduction

The MHC in human is a gene-dense region on Chromosome 6p21.31. The MHC region spans ~4 Mb and covers >120 expressed genes [1]. Forty percent of the expressed loci encode proteins with functions related to immune defense [2]. These include the highly polymorphic human leukocyte antigen (HLA) membrane glycoproteins that present peptides for recognition by T lymphocytes. Many diseases especially autoimmune diseases were found strongly associated with MHC [3].

The high density, polymorphism, linkage disequilibrium (LD) and frequent non-Mendelian inheritance of gene loci in MHC region have made it challenging to identify variations that cause or contribute to disease phenotypes. To partly resolve these problems, the MHC haplotype project has constructed eight common and disease associated MHC haplotype sequences in European populations [4]. These sequences and their variants have helped to identify a second MHC susceptibility locus for multiple sclerosis [5]. Many diseases are associated with some particular MHC haplotypes [4, 6]. The complete sequences of these haplotypes could be used as references to identify variants responsible for the disease associations. However, six of the eight sequenced MHC haplotype still obtain gaps [4]. For example, the MCF haplotype, which has been reported to be associated with rheumatoid arthritis (RA) [7], has 22 gaps covering about 1Mb.

There have been a few tools available for gap filling of genome assembly like GapBlaster [8], FGAP [9], G4ALL [10], GapFiller [11] and GapCloser [12]. They use two approaches to fill gaps: paired end reads and assembled contigs from different software. GapBlaster aligns contigs obtained in the assembly of the genome to a draft of the genome, using BLAST or Mummer, and all identified alignments of contigs that extend through the gaps in the draft sequence are presented to the user for further evaluation via the GapBlaster graphical interface [8]. FGAP also aligns multiple contigs against a draft genome assembly to find sequences that overlap gaps [9]. G4ALL is a multiuser software that allows the visualization and curation of a group of contigs that are aligned locally to a reference genome, and can also be used to fill gaps [10]. GapFiller method seeks to find read pairs of which one member matches within a sequence region and the second member falls (partially) within the gap, and the latter reads are then used to close the gap through sequence (k-mer) overlap, and the process is iteratively repeated until no further gaps can be closed [11]. GapCloser also uses the information from paired end reads to extend the sequences of contigs between gaps [12]. GapBlaster and G4ALL are fit for bacterial genomes. FGAP, GapFiller and GapCloser can be used in human genomes, but the complexity of MHC makes it challenging to enclose all gaps by applying these methods. Also, the resource of eight

assembled MHC haplotypes is extremely useful for filling gaps. Thus, a particular gap-filling method fit for human MHC is needed.

Here we constructed a complete MHC haplotype sequence of MCF cell line by gap filling based on WGS data from the same cell line. Firstly, sequences from other haplotypes were filled into the gaps as patches. Then the first intermediate sequence with patches was aligned by WGS reads. Using various signals (reads depth, ratio of single end mapping to pair end mapping, frequency of soft clipping and SNP & Indel calling), errors of the sequence were identified and revised. The final complete MCF haplotype sequence was obtained by repeating the reads alignment and mistake revision until no errors were found.

# 2. Methods

### *2.1. Filling Gaps by Patches from Other Haplotypes*

Sequences as patches were extracted from the other seven sequenced haplotypes. Positions of patches were determined by alignment of adjacent sequences of gaps. For each gap, the type of nearest HLA gene was regarded as symbol of sequence similarity, and the most similar haplotype was selected for patch extraction (Table 1). All gaps represented by N in raw MCF haplotype sequence were replaced by these patches. Blat [13] was used for alignment. The HLA type is from IMGT/HLA database [14] using cell query tool.

***Table 1.*** *Gaps and selected haplotypes for patching.*

| Gap id | Gap position | Gap length | Selected haps | Patch position | Patch length |
|---|---|---|---|---|---|
| 0 | - | - | PGF | 1-218,807 | 218,807 |
| 1 | 135,142-138,043 | 2,902 | SSTO | 172,506-175,408 | 2,903 |
| 2 | 179,430-186,757 | 7,328 | SSTO | 216,790-224,108 | 7,319 |
| 3 | 229,134-282,732 | 53,599 | SSTO | 266,488-320,061 | 53,574 |
| 4 | 383,989-404,415 | 20,427 | SSTO | 421,327-441,751 | 20,425 |
| 5 | 491,591-560,421 | 68,831 | SSTO | 528,965-597,804 | 68,840 |
| 6 | 676,600-685,588 | 8,989 | SSTO | 713,959-722,946 | 8,988 |
| 7 | 994,251-1,009,296 | 15,046 | DBB | 994,599-1,009,643 | 15,045 |
| 8 | 1,177,760-1,292,220 | 114,461 | DBB | 1,178,123-1,203,693 | 25,571 |
| 9 | 1,562,953-1,726,699 | 163,747 | MANN | 1,529,610-1,692,936 | 163,327 |
| 10 | 1,815,770-1,838,106 | 22,337 | MANN | 1,781,984-1,804,320 | 22,337 |
| 11 | 2,091,200-2,189,158 | 97,959 | DBB | 2,002,976-2,101,344 | 98,369 |
| 12 | 2,241,825-2,251,998 | 10,174 | DBB | 2,153,984-2,164,154 | 10,171 |
| 13 | 2,300,851-2,305,647 | 4,797 | DBB | 2,213,024-2,217,819 | 4,796 |
| 14 | 2,712,452-2,811,903 | 99,452 | PGF | 2,854,875-2,954,326 | 99,452 |
| 15 | 2,947,577-2,958,999 | 11,423 | SSTO | 2,898,675-2,910,108 | 11,434 |
| 16 | 3,154,499-3,188,719 | 34,221 | DBB | 3,060,384-3,094,607 | 34,224 |
| 17 | 3,329,673-3,355,524 | 25,852 | SSTO | 3,282,538-3,308,390 | 25,853 |
| 18 | 3,602,942-3,659,279 | 56,338 | SSTO | 3,570,835-3,627,163 | 56,329 |
| 19 | 3,899,295-3,929,849 | 30,555 | SSTO | 3,867,405-4,024,432 | 157,028 |
| 20 | 4,149,580-4,149,626 | 47 | SSTO | 4,243,476-4,243,522 | 47 |
| 21 | 4,284,354-4,302,274 | 17,921 | PGF | 4,469,687-4,487,607 | 17,921 |
| 22 | 4,422,173-4,594,253 | 172,081 | PGF | 4,607,560-4,641,767 | 34,208 |
| 23 | - | - | PGF | 4,881,847-4,970,558 | 88,712 |

### *2.2. Revise Errors by WGS Data*

By adding patches, the first intermediate sequence (MCF1) was obtained. WGS data was mapped to hg19 whose MHC region (chr6: 28477797-33448354) was replaced by MCF1. BWA [15] was used for reads alignment with parameters: aln -o 1 -e 63 -i 15 -L -k 2 -l 31 -t 4 -q 10 -I. Reads depth, ratio of single end mapping reads to paired end mapping reads and frequency of soft clipping were used to identify big errors (Figure 1). Low reads depth implied insertion errors because artificial sequence could not be covered by reads, and it also implied deletion errors because reads spanning the breakpoint could hardly be mapped properly. High ratio of single end mapping reads to paired end mapping reads implied deletion errors since loss of reference sequence would make one read of a pair not aligned. High frequency of soft clipping indicated both insertion and deletion errors since the edge of errors would clip a read to a mapped part and an unmapped part. Big insertion errors were revised by simply deleting it. Big deletion errors were revised by adding patches from other haplotypes. Break points of big insertion and deletion errors were determined by the clipping positions of soft clipped reads. Complex errors which caused by small but dense differences were revised by replacement of consensus sequences. The consensus sequences were obtained by Samtools pileup [16]. Small errors were identified by variation calling pipeline of Samtools [17] and revised by replacement of called variations. Repeating the reads alignment and error revision until no more error was found.

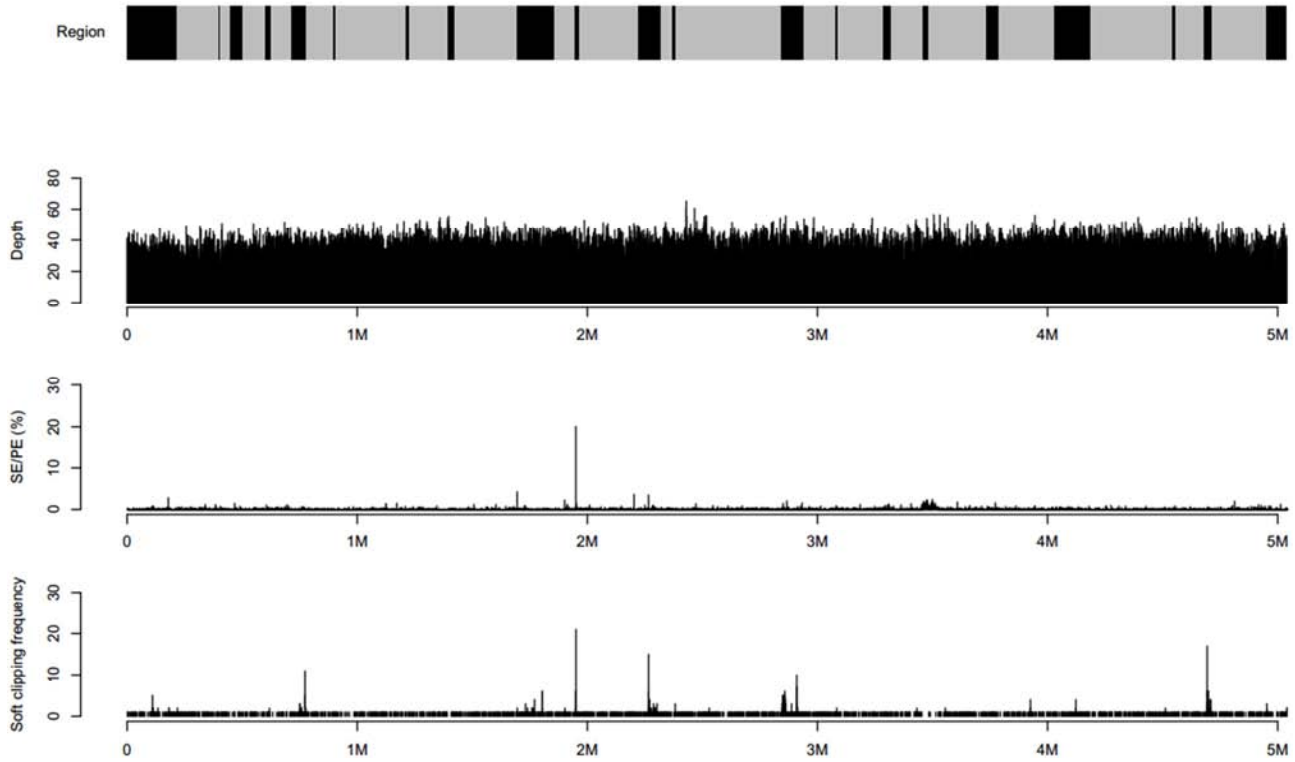


***Figure 1.*** *The first intermediate sequence (MCF1). The "Region" bar presents gap region (black) and raw MCF region (grey). SE: single end mapping reads. PE: paired end mapping reads. Depth and SE/PE were calculated in 1Kb window.*

### *2.3. Gene Annotation*

Gene information of the eight MHC haplotypes from refGene annotations (downloaded from UCSC Genome Browser Data [18]) were used to get sequences of transcripts from all eight MHC haplotypes. These sequences were mapped to our final complete MCF haplotype sequence (MCF6) by blast [19] to determine positons. For those transcripts whose sequences could not be mapped perfectly as a whole, exons were picked out for alignment to help positioning.

### *2.4. Whole Genome Sequencing of MCF Cell Line*

High molecular weight genomic DNA (gDNA) was extracted form MCF cell line. The gDNA was sheared into ~500bp fragments. Paired-end shotgun libraries were generated followed by the manufacturer's instruments (Illumina). DNA libraries were sequenced with Hiseq2000 (Illumina) platform to generate 2 x 91 bp paired-end reads with ~30X genome coverage on average.

## 3. Results

### *3.1. Revised Errors*

All changes during gap filling were listed in Figure 2. Firstly, 24 patches from other haplotypes were filled into gaps. Then one big deletion error was revised and small errors were revised to reduce noises. Then one big insertion error, two complex errors and four Indels were revised. After two run of small-error revision, final sequence was obtained. There are two big errors: one big deletion error (~2.4Kb) and one big insertion error (311bp). For the big deletion error, the reads depth was low, and there were many single end mapping reads and soft clipped reads around error position as expected (Figure 3). Alignment of this error region to other haplotypes by UCSC Blat indicated a deleted sequence with ~2.4 Kb in length when comparing to haplotype SSTO (Table 2). For the big insertion error, depth of reads covering the inserted sequence was extremely low, and there were two high-frequency soft clipping positions which marked the edge of the insertion (Figure 4).

***Table 2.*** *Results of UCSC Blat for the big deletion error in MCF1.*

| Query | Start | End | Query Size | Identity | Object | Start | End | Span |
|---|---|---|---|---|---|---|---|---|
| mcf1:1950000-1950999 | 1 | 1000 | 1000 | 100.00% | chr6_mann_hap4 | 1786745 | 1787744 | 1000 |
| mcf1:1950000-1950999 | 1 | 1000 | 1000 | 100.00% | chr6_dbb_hap3 | 1732349 | 1733348 | 1000 |
| mcf1:1950000-1950999 | 1 | 1000 | 1000 | 100.00% | chr6 | 30438667 | 30439666 | 1000 |
| mcf1:1950000-1950999 | 1 | 1000 | 1000 | 99.90% | chr6_qbl_hap6 | 1731586 | 1732585 | 1000 |
| mcf1:1950000-1950999 | 1 | 1000 | 1000 | 99.90% | chr6_cox_hap2 | 1950734 | 1951733 | 1000 |
| mcf1:1950000-1950999 | 1 | 1000 | 1000 | 98.10% | chr6_ssto_hap7 | 1768545 | 1771963 | 3419 |

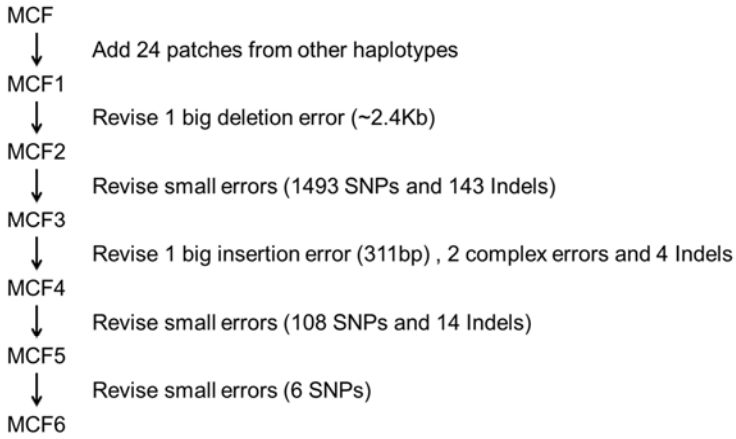


***Figure 2.*** *All changes during gap filling of MCF.*

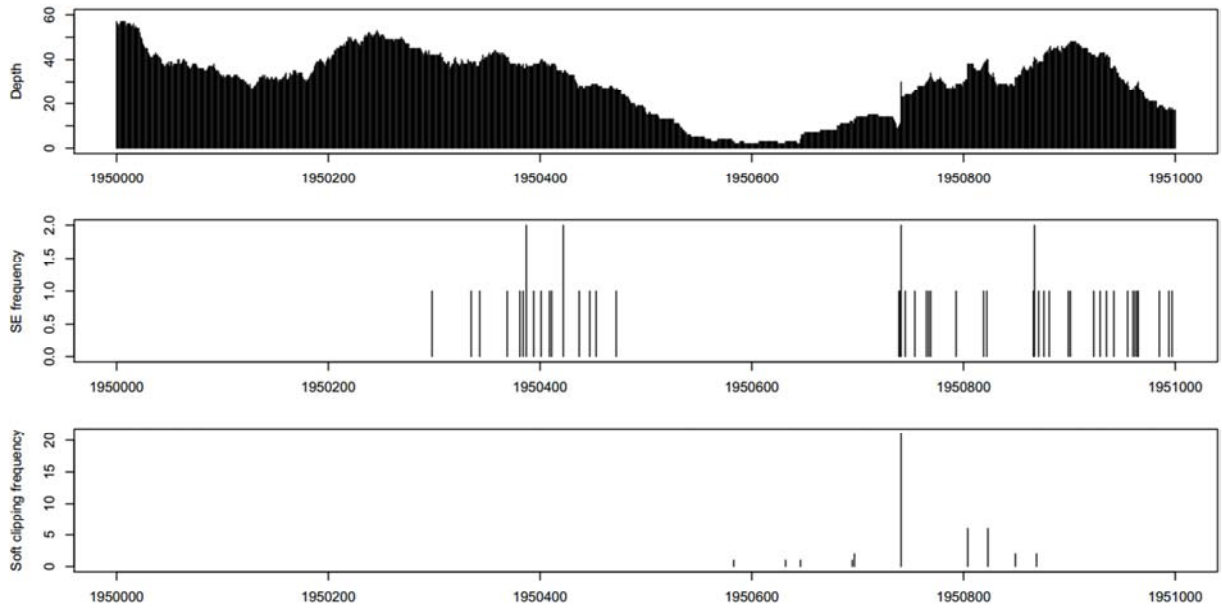


***Figure 3.*** *The big deletion error in MCF1. SE: single end mapping reads. Location of SE was according to leftmost position of single end mapping reads.*

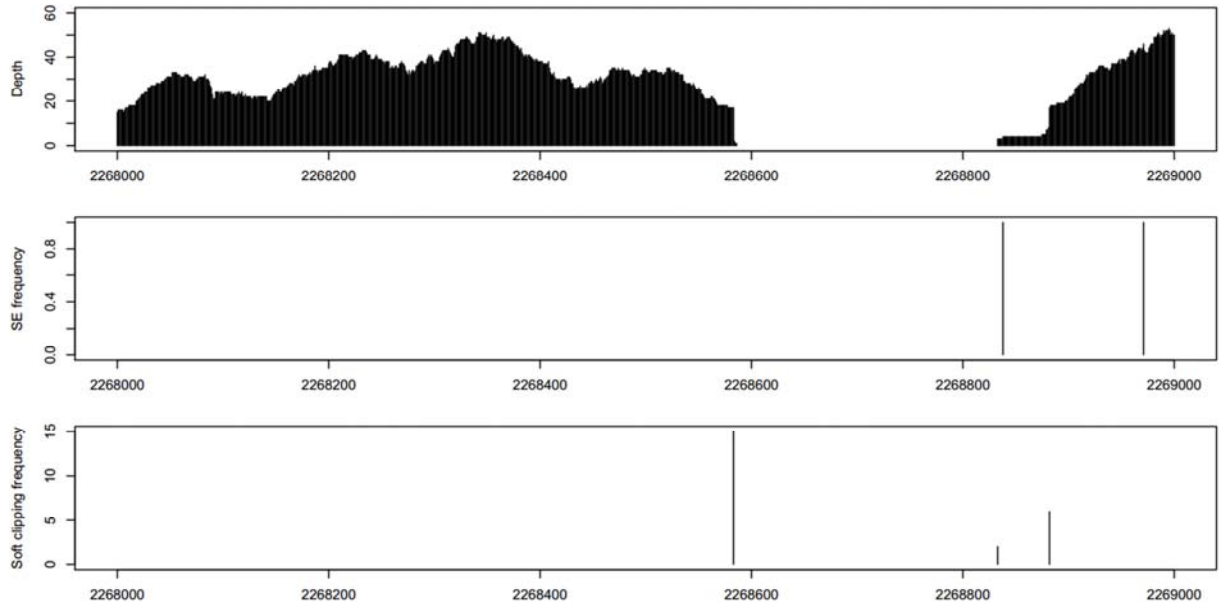


***Figure 4.*** *The big insertion error in MCF3. SE: single end mapping reads. Location of SE was according to leftmost position of single end mapping reads.*

### *3.2. Evaluation*

The final complete MCF haplotype sequence did not have any significant signals of error (Figure 5). Type of HLA-DRB1 was DRB1*04:01 in the MCF haplotype (recorded in IMGT/HLA database). Region containing HLA-DRB1 in the raw MCF haplotype sequence was a gap. After gap filling, coding sequence of DRB1 from the final complete MCF sequence 100% matched that of DRB1*04:01.

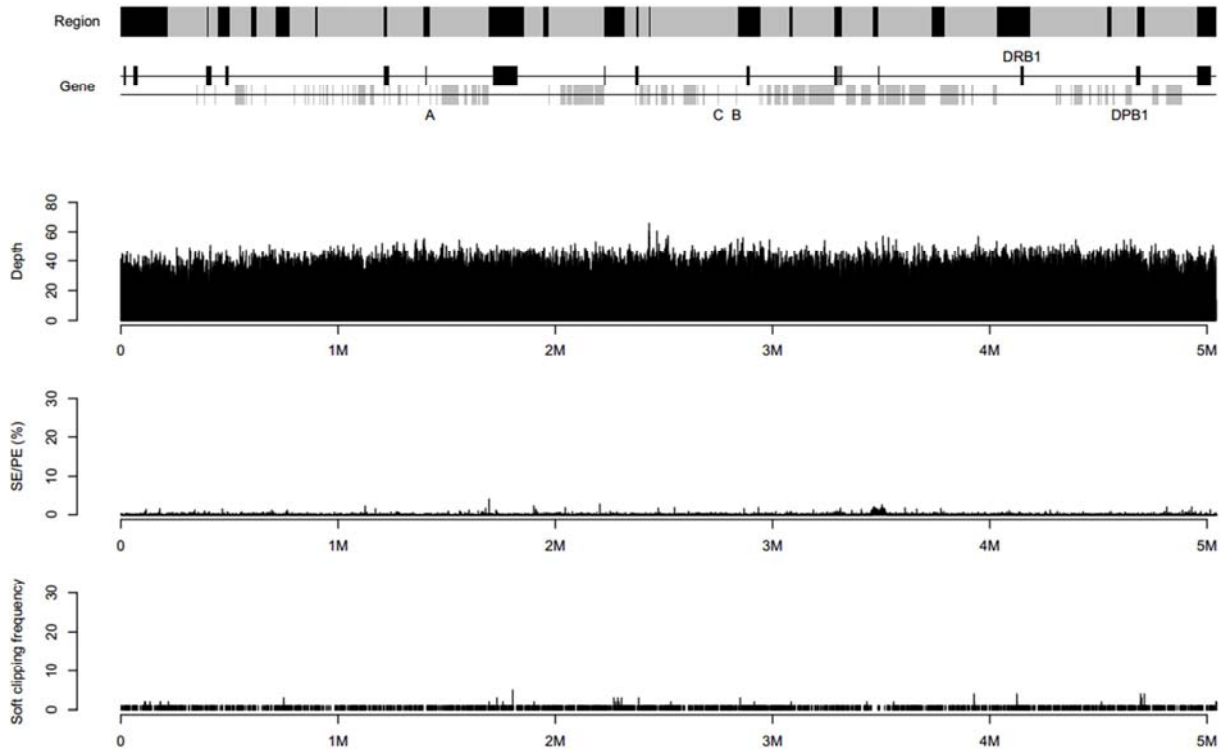


***Figure 5.*** *The complete MCF haplotype sequence (MCF6). The "Region" bar presents gap region (black) and raw MCF region (grey). The "Gene" bar has two parts: black in gap region and grey in raw MCF region. Depth and SE/PE were calculated in 1Kb window.*

### *3.3. Gene Annotation*

A total of 220 genes and 416 transcripts were annotated. There are 31 genes and 62 transcripts in gap region while 189 genes and 354 transcripts in the original raw MCF (Figure 5). Several important immune genes are in gap region, for example, MICA and HLA-DRB1.

## 4. Discussion

The eight MHC haplotypes were used for extraction of patch sequences and revision of a big deletion error. It is suitable for MCF cell line since they are all European. But for haplotypes from other ethnic groups, use of the existing eight MHC haplotypes would not be appropriate, since European haplotypes could miss structural variations. This problem could be solved by new assembly of haplotypes from different ethnic groups or using other tools like GapFiller [11] or GapCloser [12] for local assembly. In this method, short reads from Illumina was used, however, haplotypes with huge structural variations or repeat elements may not be solved by short reads. Shotgun sequencing from the third generation sequencing platform like single molecule real time (SMRT) sequencing by pacbio could be applied to resolve this limitation. For example, SMRT sequencing was successfully used for filling a gap composed of macrosatellite repeats in the facioscapulohumeral muscular dystrophy locus in human chromosome 4 [20]. FGAP can use SMRT sequencing reads for automated gap filling [9]. For now, our method needs some manual operations like using blat in the UCSC genome browser and identification of sequence errors from signals. Also, the whole process is divided into many small steps which need manual linking between two steps. It is necessary to develop algorithm for sequence-error classification and combine the steps as an automated or half-automated pipeline.

## 5. Conclusion

We have filled the gaps spanning 1Mb of a MHC haplotype named MCF which is associated with RA. This complete MCF haplotype sequence can be used as reference in disease studies for more precise reads mapping, more accurate variation detection and thus new discoveries of risk loci. The method uses WGS data and existing sequence of MHC haplotypes to fill gaps. It can be applied to any other draft MHC haplotype assemblies.

# References


[1] The, M. H. C. s. c., Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. Nature, 1999. 401 (6756): p. 921-923.

[2] Stewart, C. A., et al., Complete MHC haplotype sequencing for common disease gene mapping. Genome research, 2004. 14 (6): p. 1176-1187.

[3] Burton, P. R., et al., Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 2007. 447 (7145): p. 661-678.

[4] Horton, R., et al., Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. Immunogenetics, 2008. 60 (1): p. 1-18.

[5] Yeo, T. W., et al., A second major histocompatibility complex susceptibility locus for multiple sclerosis. Annals of neurology, 2007. 61 (3): p. 228-236.

[6] Miller, F. W., et al., Genome-wide association study identifies HLA 8.1 ancestral haplotype alleles as major genetic risk factors for myositis phenotypes. Genes and immunity, 2015. 16 (7): p. 470-480.

[7] Lagha, A., et al., HLA DRB1/DQB1 alleles and DRB1 - DQB1 haplotypes and the risk of rheumatoid arthritis in Tunisians: a population - based case–control study. HLA, 2016. 88 (3): p. 100-109.

[8] De Sá, P. H., et al., GapBlaster—A Graphical Gap Filler for Prokaryote Genomes. PloS one, 2016. 11 (5): p. e0155327.

[9] Piro, V. C., et al., FGAP: an automated gap closing tool. BMC research notes, 2014. 7 (1): p. 1.

[10] Ramos, R. T. J., et al., Graphical contig analyzer for all sequencing platforms (G4ALL): a new stand-alone tool for finishing and draft generation of bacterial genomes. Bioinformation, 2013. 9 (11): p. 599.

[11] Boetzer, M. and W. Pirovano, Toward almost closed genomes with GapFiller. Genome biology, 2012. 13 (6): p. 1.

[12] Luo, R., et al., SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience, 2012. 1 (1): p. 1.

[13] Kent, W. J., BLAT—the BLAST-like alignment tool. Genome research, 2002. 12 (4): p. 656-664.

[14] Robinson, J., et al., The IPD and IMGT/HLA database: allele variant databases. Nucleic acids research, 2014: p. gku1161.

[15] Li, H. and R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 2009. 25 (14): p. 1754-1760.

[16] Li, H., et al., The sequence alignment/map format and SAMtools. Bioinformatics, 2009. 25 (16): p. 2078-2079.

[17] Li, H., A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 2011. 27 (21): p. 2987-2993.

[18] Speir, M. L., et al., The UCSC Genome Browser database: 2016 update. Nucleic acids research, 2016. 44 (D1): p. D717-D725.

[19] Altschul, S. F., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 1997. 25(17): p. 3389-3402.

[20] Morioka, M. S., et al., Filling in the Gap of Human Chromosome 4: Single Molecule Real Time Sequencing of Macrosatellite Repeats in the Facioscapulohumeral Muscular Dystrophy Locus. PloS one, 2016. 11 (3): p. e0151963.