DEVELOPMENT OF A NOVEL, REFERENCE-FREE TOOL FOR THE COMPREHENSIVE EVALUATION OF GENOME ASSEMBLY QUALITY AND ITS APPLICATION TO ESTABLISH A REFERENCE ASSEMBLY FOR CHINESE HAMSTER OVARY (CHO) CELLS

by

Madolyn L. MacDonald

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Bioinformatics and Systems Biology

Summer 2019

DEVELOPMENT OF A NOVEL, REFERENCE-FREE TOOL FOR THE COMPREHENSIVE EVALUATION OF GENOME ASSEMBLY QUALITY AND ITS APPLICATION TO ESTABLISH A REFERENCE ASSEMBLY FOR CHINESE HAMSTER OVARY (CHO) CELLS

by

Madolyn L. MacDonald

Approved: _____
Cathy H. Wu, Ph.D.
Chair of Bioinformatics & Computational Biology

Approved: _____
Levi T. Thompson, Ph.D.
Dean of the College of Engineering

Approved: _____
Douglas Doren, Ph.D.
Interim Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Kelvin H. Lee, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Shawn Polson, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Hagit Shatkay, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Millicent Sullivan, Ph.D.
Member of dissertation committee

ACKNOWLEDGEMENTS

I would like to thank my advisor, Kelvin Lee, for supporting and guiding me through graduate school and this project. I am very appreciative of the freedom I was given to choose the direction of my project regarding assembly quality once the CH PacBio genome assemblies were completed. In addition, he helped me improve both my written and oral communication skills throughout my time at UD. I am also particularly grateful for the support and project feedback given by my committee members, Dr. Shawn Polson, Dr. Hagit Shatkay, and Dr. Millicent Sullivan. I am especially thankful for the feature selection and machine learning advice provided by Dr. Shatkay and the bioinformatics help provided by Dr. Polson. I would also like to specifically thank my co-authors for the publications that chapters 2-4 are based on as well as Karol Miaskiewicz for all of his quick solutions to any problems I had while working on the BioMix computing cluster.

My sincerest thanks can be extended to my labmates for providing a welcoming, encouraging, and collaborative environment. They were always willing to provide general project advice and fresh insight on papers and presentations. I especially appreciate the kindness of Jongyoun Baik and Ben Kremkow who helped me settle into the lab when I first joined, and of Nate Hamaker who provided tremendously helpful feedback on my computational pipeline and manuscripts.

My heartfelt thanks to my family and friends especially to my mom, Ruth, for continually supporting me and joining me on much needed coffee breaks. Also, many thanks to my husband, Scott, for his endless optimism, motivation, and good sense of humor and to my fellow UD bioinformatics buddies; Parth, Irem, and Mengxi.

The project was made possible through funding from IGERT and NSF. In addition, support from the UD Center for Bioinformatics and Computational Biology Core

facility and use of the BioMix compute cluster was made possible through funding from Delaware INBRE (NIH GM103446) and the Delaware Biotechnology Institute.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Whole genome assemblies are regularly becoming available for more organisms due to the reduced time and costs of DNA sequencing. Multiple assemblies may be created for the same species with one being selected as the reference genome to guide wet-lab and bioinformatics studies. To select the most complete, continuous, and accurate assembly for an organism of interest, improved methods for quality assessment of assemblies is necessary. Currently, most methods to evaluate genome assembly quality focus on completeness or continuity only. If accuracy is assessed, a high quality reference genome for the organism of interest is often required for a direct sequence comparison.

Here, we emphasize the need for assembly quality assessment by using as a case study the creation of multiple genome assemblies for the Chinese hamster (CH) and Chinese hamster ovary (CHO) cells, the preferred platform for therapeutic protein production. The highest quality assembly, CH PICR, was created from combining multiple assemblies where the primary, base assembly was developed from long-read sequencing data. CH PICR was selected through manual quality assessment, annotated, and made available on the NCBI RefSeq database as the new reference genome.

We then describe the development of a novel tool, EvalDNA (Evaluation of *De Novo* Assemblies) to facilitate the evaluation of mammalian genome assembly quality and the selection of the reference genome. EvalDNA overcomes the requirement of an additional genome assembly by using a machine-learning model to integrate a variety of quality metrics into a single, comprehensive quality score. The provided model can explain approximately 86% of the variation in reference-based quality scores in the test data, consisting of different draft chromosome assemblies with real/simulated errors. EvalDNA also distinguishes itself from current assembly evaluation tools because

EvalDNA quality scores generated by the same model are comparable across different organisms.

EvalDNA was used to evaluate the novel assemblies of the CH genome. The resulting scores showed that CH PICR was of the highest quality, agreeing with the manual quality evaluation. This observation confirms EvalDNA's ability to score assemblies from organisms not used in the training data. EvalDNA's ability to compare assemblies from different assemblers and organisms is also examined.

Finally, we demonstrate the benefits of having an improved CH reference genome assembly in CHO cell genetic engineering. Successful gene knock-downs and knock-outs in CHO cells can prevent the expression of difficult-to-remove host cell proteins (HCPs). HCPs, if not removed, can cause problems in the stability, safety, and efficacy of the biotherapeutic protein being produced. Here, the CH PICR reference genome was used to identify new knockout targets with similar predicted functions and characteristics as several difficult-to-remove HCPs.

Chapter 1

BACKGROUND AND SIGNIFICANCE

A genome is the entire genetic information of an organism, including genes and non-coding DNA. An organism's genome can be sequenced using high-throughput DNA sequencing technology. Due to limitations of the current sequencing technology, the genome needs to be fragmented before it can be sequenced. Thus, DNA sequencing produces many short pieces of DNA, called sequencing reads, that need to be put together to form a whole genome assembly. Because of errors in sequencing reads and errors that can occur in the assembly process, there has been a significant push to design more accurate sequencing technology and assembly methods. Now, genome assemblies are regularly becoming available for more organisms with a greater than three-fold increase in NCBI's RefSeq assembly database since August of 2015 [1] (Table 1.1).

Table 1.1: Assemblies in the NCBI RefSeq Databases in August 2015 [1] and in February 2019 (counts taken on February 7th, 2019). The 'All' taxonomic group contains viruses and viroids, invertebrates, and protists in addition to the groups listed here. The total assemblies and species counts for Feb. 2019 were determined from the 'assembly_summary_refseq.txt' file located at ftp://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/.

Taxonomic Group	**NCBI RefSeq Aug. 2015**	**NCBI RefSeq Feb. 2019**
Archaea	414	810
Bacteria	34,514	143,385
Fungi	167	283
Plants	62	94
Mammals	94	127
All	40,390 (for 12,964 species)	153,355 (for 53,048 species)

1.1 DNA Sequencing and Whole Genome Assembly

DNA sequencing, the process in which nucleotide sequences of DNA fragments are determined, is the first step towards generating a genome assembly. The increased availability of whole genome assemblies is the direct result of the reduced monetary and time costs of DNA sequencing. The cost to create a high-quality 'draft' assembly for the human genome in 2006 was approximately $14 million. This cost was drastically decreased to $4,000 by 2015, and was as low as $1,000–$1,500 in 2016 [2].

Today, two of the most commonly used sequencing methods are Illumina's sequencing-by-synthesis method and Pacific Biosciences' (PacBio) Single-Molecule Real-Time sequencing method. In Illumina sequencing, the DNA to be sequenced is fragmented into pieces often 200-600 base pairs (bp) long. These pieces are hybridized to flow cells using adapters and then amplified to create clusters of template strands. Next, fluorescently-tagged nucleotide bases with attached reversible terminators are added to the flow cell. The base that pairs with the next base on the template strand is incorporated, while the terminator assures that only one base is added each round. At the end of a round, the terminator is cleaved allowing the next nucleotide to be added. Each base emits a specific flourescent signal when incorporated that is identified by the sequencing machine as an A,C,T, or G (Figure 1.1).

Illumina sequencing outperforms many other types of sequencing in cost, throughput, and read accuracy (∼0.1% error rate) [3]. However, Illumina sequencing produces short reads that are typically only 100-300 bp long. To mitigate the limitations of short reads, a method called paired-end sequencing is frequently used, where each DNA fragment results in two reads. One read is sequenced from the beginning of the fragment and the other read is sequenced from the end. This technique facilitates the assembly process by providing information about the distance between reads. Another method to overcome the short-read limitation of Illumina sequencing is mate-pair sequencing where the ends of a longer DNA strand, around 1 to 10 kbp, are sequenced. Mate-pair sequencing information helps to connect contigs during the assembly process (Figure 1.2). Although the sequence between the mate-pair reads is unknown, the distance

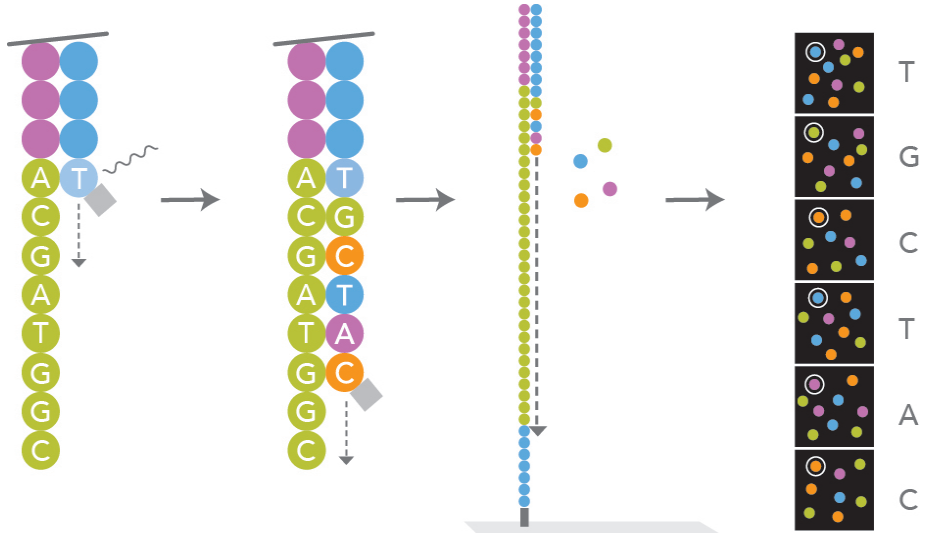

Figure 1.1: Illumina's sequencing-by-synthesis method. The first step shows a fluorescently-tagged thymidine (T) nucleotide, with a reversible terminator, being added to the template DNA strand. More nucleotides get incorporated in subsequent rounds, emitting specific flourescent signals for A,C,T, or G. Image courtesy of Illumina, Inc.

is known and can be used to link two contigs. The sequence created by linking and ordering two or more contigs is called a scaffold.

PacBio sequencing is another prevalent sequencing method. In PacBio's Single-Molecule Real-Time (SMRT) sequencing (Figure 1.3), the DNA to be sequenced is sheared often into 5-35 kilobase (kbp) pieces [4] and each DNA fragment is ligated to a chip called a SMRT cell. The DNA strand then diffuses into a zero-mode waveguide (ZMW) unit where a single polymerase for replication is located. Next, the four types of nucleotides each with a unique fluorophore attached are added to the SMRT cell. As each base is added to the DNA strand by the polymerase, there is a detectable pulse of light that indicates which base was incorporated [3].

PacBio sequencing, in contrast to Illumina, produces long reads with a relatively high error rate. Although the PacBio error rate (∼13%) [3] is much higher than that of Illumina, the errors are randomly distributed and thus, can be corrected by increasing sequencing coverage. Assemblers that use PacBio reads only often recommend having

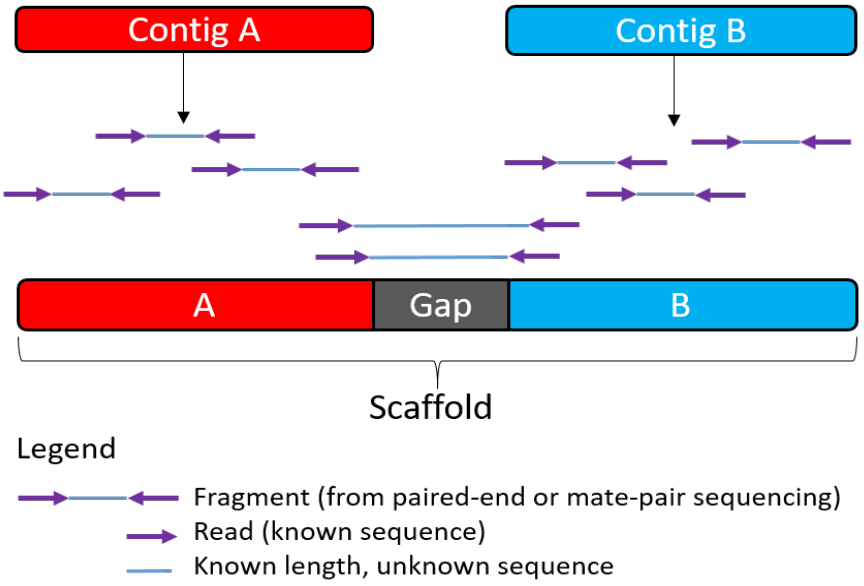

Figure 1.2: Contigs can be joined and ordered into a scaffold over gaps when there is enough evidence from paired-end and mate-pair sequencing.

at least 50x sequencing coverage (for the so-called P5C3 chemistry) of the genome to produce high quality assemblies [6, 7, 8]. Error-correction is highly recommended before assembly to avoid the generation of misassemblies [9]. Once error-corrected, PacBio's longer reads often lead to better assemblies as they are able to resolve repetitive regions and lead to fewer gap regions [8].

The differences in read length and sequencing error rate between these two sequencing methods impact the subsequent genome assembly process. Longer and more accurate reads can improve assembler output, especially when building an assembly *de novo* i.e. without an existing reference genome [10]. Because the benefits of each sequencing method, the highly accurate reads of Illumina and the long reads of PacBio, are complementary to one another, methods to integrate Illumina and PacBio sequencing data have been developed. For instance, errors in PacBio reads can be resolved prior to assembly by using highly accurate Illumina reads [9, 11]. In addition, merging assemblies created from the different sequencing methods has been shown to

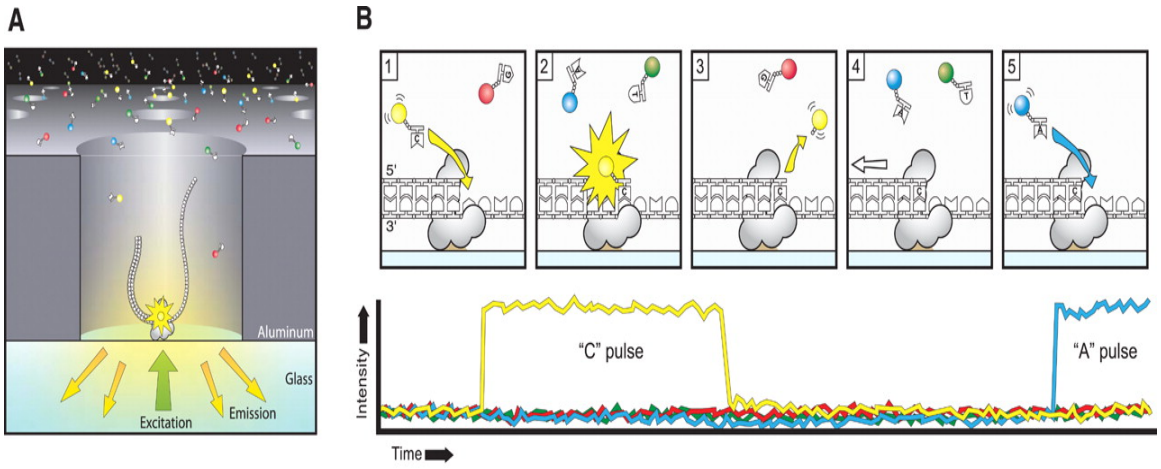

Figure 1.3: PacBio's Single-Molecule Real-Time (SMRT) sequencing method. A) A single molecule of DNA polymerase is attached at the bottom of a ZMW. The nano-scale size of the ZMW enables the detection of a single nucleotide as it is added to the DNA strand. B) A cytosine (C) is incorporated into the DNA strand, followed by an adenine (A), with an example time trace of fluorescence intensity. Image [5] is used with permission from The American Association for the Advancement of Science (see Appendix D).

further improve the quality of the final assembly [12]. The Metassembler tool iteratively updates a starting assembly based on pair-wise alignments to other assemblies [12]. Conflicts between assemblies are resolved by selecting the local sequence with the best compression–expansion (CE) statistic [13].

1.2 Reference Genome Assemblies

Due to advances in both DNA sequencing methods and genome assembly algorithms, multiple genome assemblies may be created for the same species. Before the advent of PacBio sequencing, most assemblies were created using Illumina sequencing reads only. More recently, reads from PacBio sequencing and Illumina sequencing are often combined through a variety of algorithms to address the short read length limitation of Illumina and the high error rate of PacBio [9, 14, 15]. However, the standard procedure is to select one genome assembly for a species as the reference genome. This

reference genome will become an important tool to study and/or manipulate the genetics of an organism. Therefore, the selected reference assembly should be the one that is the closest to the organism's true genome, i.e. the one of highest quality.

Reference genome assemblies are frequently used to guide wet lab experiments. For instance, to measure the expression of coding and noncoding RNAs in a genome using DNA microarrays, probes need to be designed to be complementary to the sequences of the areas of interest. In addition, most genetic engineering techniques, such as gene knockdowns and knockouts, require the sequence of the genome to be known. While the target gene's sequence could be determined through targeted sequencing, the sequence of the whole genome would still be necessary for certain genetic engineering methods to mitigate possible off-target effects. For example, the CRISPR-Cas9 genetic engineering system [16, 17, 18] cleaves based on a specific nucleotide sequence, which results in the deletion or addition of nucleotides knocking out the target gene. However, if this target sequence occurs multiple places in the genome, breaks and mutations would be incorporated multiple times which could cause unintended genes and/or regulatory regions to be disrupted.

Reference assemblies are also used in a significant portion of bioinformatics work. Genome annotations rely on the correct reference sequence for both *ab initio* prediction of genes and the alignment of mRNA data. Comparative genomic studies, such as finding homologs or syntenic gene regions, benefit greatly from having available reference genomes of the organisms being compared [19, 20, 21]. Additionally, RNA-sequencing data analysis regularly involves the mapping of sequencing reads to a reference genome to quantify gene expression and identify transcript variants [22].

1.3 Reference Assembly for Chinese Hamster Ovary (CHO) Cells

1.3.1 CHO cells

Chinese hamster ovary (CHO) cells are the preferred platform for the production of biotherapeutic monoclonal antibodies (mAbs). CHO cells were isolated from a female Chinese hamster in 1957 by Dr. Theodore Puck, who was able to establish

the cells in culture [23]. The first recombinant therapeutic produced from a mammalian cell host, tissue plasminogen activator, was produced from CHO cells and was approved for clinical use in 1987 [24]. Now, the majority of recombinant biotherapeutic proteins are produced from CHO cells. Between 2014 and 2018, 57 of the 68 (85%) approved monoclonal antibodies (mAbs) were produced by CHO cell systems [25]. These mAbs are used to treat a variety of diseases, with the majority targeting cancer and autoimmune disorders.

CHO cells are used for biotherapeutic production largely because of their high growth rate, ease of genetic manipulation, resistance to viral infection, and ability to form glycosylation patterns similar to those found in humans. They are also easily adapted to suspension culture enabling scale-up using bioreactors [24]. In addition, they already have approval as a host cell for therapeutic protein production facilitating the approval of new biotherapeutics produced from CHO cells.

1.3.2 Chinese hamster (CH) genome as the reference for CHO cells

Despite the multiple benefits of using CHO cells, there is still the potential to produce safer therapeutics more efficiently through the study and manipulation of CHO cell genetics. To facilitate these studies, a 'gold-standard' reference genome for CHO cells is necessary. However, a variety of different CHO cell lines are used in the biopharmaceutical industry and each cell line has a unique genome because they are subject to frequent, spontaneous chromosomal rearrangements. One CHO cell line's genome cannot effectively represent the others and there can even be differences among cells within a population derived from the same cell line [26, 27].

The Chinese hamster (CH) genome, from which CHO cells were derived, may be an appropriate stable reference genome for CHO cell lines [28]. Therefore, after the sequencing of the CHO-K1 cell line in 2011, efforts have been directed towards the sequencing, assembly, and annotation of the CH genome. Illumina sequencing was used to create the 2013 RefSeq CH genome assembly [27] and the 2013 chromosome-sorted assembly (CSA) [29]. The most recent CH genome assembly, described in Chapter 2,

was built using a hybrid approach that integrated the previous Illumina sequencing data with new PacBio sequencing data [30].

One of the main advantages of using the CH genome assembly as a reference for CHO cells is that it enables the comparison of CHO cell lines. This advantage is demonstrated by Feichtenger et al. [31], where reads from whole genome sequencing, bisulfite sequencing, and CHIP-seq of six CHO cell lines were mapped to the CH genome. SNPs/indels, structural variants, histone modifications, and methylated regions across the six cell lines could then be compared in reference to a stable genome. In addition, the annotated CH genome was able to provide biological and functional insights about the differences, i.e. were the differences in genes, promoter regions, intergenic regions, or transcriptional start sites. Another study used the CH genome to compare patterns of chromosome evolution and instability among CHO cell lines [32]. Overall, examination of differences among the genomes of CHO cell lines and the Chinese hamster can lead to a better understanding of CHO cellular pathways and facilitate the engineering of these pathways to increase CHO cell growth and biotherapeutic protein production and secretion [33, 34, 35].

Other uses of the CH genome assembly for studying CHO cells have been demonstrated. The CH genome has been used for the identification of novel microRNAs in CHO cells that were evolutionarily conserved [36]. In addition, transcriptome and genome data from CH was used to investigate which auxotrophies were present in the Chinese hamster and which were CHO cell-line specific to better understand CHO metabolic processes [37]. Chromosomal and regulatory information about potential gene integration sites has also been gained from the 2013 CH genome [38].

Despite their proven benefits in the study CHO cells, the 2013 CH genome assemblies are still far from complete. CSA, while separated into chromosomes, contains approximately 10% of unknown sequence (gaps) and is split into 28,749 scaffolds. The 2013 RefSeq assembly had less unknown sequence ($\sim$2%), but is split into 52,710 scaffolds and does not contain chromosome assignments for those scaffolds. The 2013 RefSeq assembly has a slightly better N50 (explained in the next section) of 1,558 kbp

than CSA (1,237 kbp) and was subsequently annotated by the NCBI RefSeq pipeline in 2014. A higher quality CH reference assembly would improve the accuracy of mapping sequencing reads from CHO cells and provide a more complete annotation of genes and regulatory elements, enhancing comparisons among CHO cell lines. In addition, a more continuous assembly of the CH genome would further increase the understanding of the chromosomal context of integration sites [39] and could help find possible stable positions for gene insertion through the identification of conserved regions among the genomes of CH and CHO cells.

1.4 Assembly Quality Evaluation

1.4.1 Overview

Comprehensive evaluation of assembly quality is essential to identify the most appropriate reference assembly for a species and be aware of possible limitations regarding the chosen assembly's quality. Many scientists incorrectly assume that reference genomes are complete and correct. However, reference genomes for higher eukaryotes are commonly only high-quality drafts due to their genome size and complexity [40]. It is often difficult to generate a perfectly complete and continuous assembly, i.e. combine all scaffolds and contigs into one sequence, with high confidence. These combined sequences can also contain misassemblies that are difficult to identify and correct.

Metrics that evaluate the completeness and continuity of an assembly reflect how much of the organism's actual genome is represented and in how many separate sequences. Completeness and continuity metrics include the total size of the reference assembly, the number of scaffolds/contigs that make up this assembly, and the N50 metric. The N50 metric is the length of the longest scaffold where the total length of that scaffold plus all longer scaffolds is more than half the length of the genome (Figure 1.4). Another useful metric is the percentage of the assembly that is made up of gaps. Gaps are regions of the assembly where the sequence is unknown, but they are incorporated into assemblies as placeholders to combine contigs into scaffolds when possible.

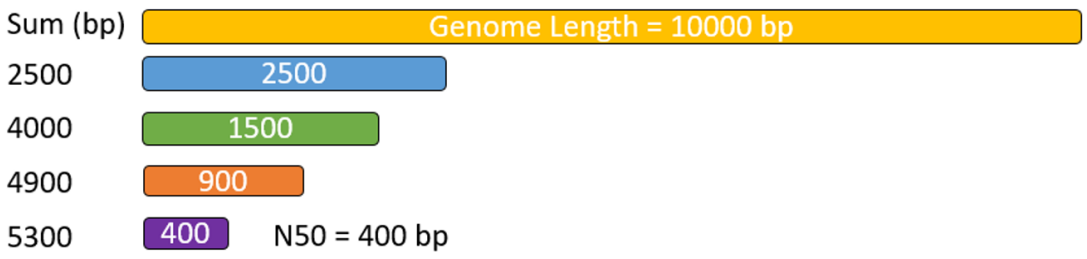

Figure 1.4: An example of the N50 metric for a 10,000 bp assembly.

Assembly accuracy describes how similar the sequence of the assembly is to the true genome sequence of the organism. Measuring the accuracy of an assembly is more challenging than measuring the completeness or continuity, especially when no other assemblies exist for the organism of interest. If a highly accurate reference assembly does exist, accuracy of an assembly can be determined through direct sequence comparison. However, high-quality reference genomes are rare for high-order eukaryotes and would not exist for the sequencing of novel organism genomes.

Several methods have been developed to gain insights into the accuracy of an assembly that do not need a reference assembly. The most commonly used method is to map raw sequencing reads (either reads used in the assembly or reads from the same organism) to the assembly. Error-free reads represent the true sequence of the genome, just in many separate pieces, and they should map perfectly onto a genome if the genome is perfectly assembled. Therefore, errors in read mapping or the lack of read mapping to an area of the assembly often suggest a potential underlying misassembled sequence.

1.4.2 Current tools for assembly evaluation

While single metrics have been developed to describe specific quality aspects of an assembly, there is a need for tools that comprehensively evaluate assemblies based on both accuracy and completeness. Looking at one metric alone can sometimes be

misleading. For instance, a high N50 value means more contigs were joined together into larger scaffolds and suggests a more complete assembly. The N50 is often used to measure the quality of a genome, but this value does not take into account whether the contigs were joined together correctly [41]. Also, most of the single metrics used to evaluate genome assemblies only examine the continuity or completeness of an assembly, while good quality assessment needs to examine all three quality aspects of an assembly; completeness, continuity, and accuracy.

Tools that examine accuracy include QUAST (Quality Assessment Tool for Genome Assemblies) [42] and CQAT (Contig Quality Assessment Tool) [43]. These tools require a high quality reference assembly for the organism of interest. Reference independent tools that use the read-mapping approach described above include Amosvalidate [44], ALE [45], FRCbam [46], and SURankCO [47]. Despite the benefits of these tools, the resulting quality assessments are not comparable across different species. If a standard method for quality assessment of assemblies was to be established, the results should ideally be comparable among all species or at least among species of the same taxonomic class. This would provide a way to quantitatively assess how similar or dissimilar an assembly's quality is to 'gold-standard' assemblies such as the human reference genome.

1.5 Project Goals

The overall purpose of this work is to provide a high quality reference genome for CH and CHO cells to support the manufacturing of biotherapeutic proteins, and to facilitate reference genome selection and comparison through the development of a quality evaluation tool for mammalian genome assemblies. This work has been divided into three main objectives:

1. *To create and annotate genome assemblies for the Chinese hamster and to establish the highest quality assembly as the new reference for CHO cells* (Chapter 2) - A hybrid approach, using Illumina and PacBio sequencing data, was used to build improved Chinese hamster genome assemblies. The quality of the assemblies

was assessed to select the highest quality assembly, PICR, to be the new CH reference assembly. Annotation, gap analysis, and chromatin state analysis of PICR was also completed.

2. *To develop a pipeline for comprehensive genome assembly evaluation that does not require an existing reference genome and apply the pipeline to the CH genomes as proof-of-concept* (Chapter 3) - A novel machine learning-based tool, EvalDNA, was developed to evaluate genome assemblies based on accuracy, completeness, and continuity. EvalDNA does not require a reference genome and the resulting scores are comparable across different species. The meta-assemblies for the Chinese hamster, described in Chapter 1, were scored using EvalDNA and compared against the existing CH genome assemblies as well as the reference genomes of various model organisms.

3. *To use the new high quality CH reference assembly to facilitate the identification of problematic lipases and knock-out target sites* (Chapter 4) - Lipase proteins from the CHO-K1 genome assembly were compared to known problematic lipases to identify additional lipases that could possibly cause problems if they exist in the final drug product. Several sequences/annotations of lipases from the CHO-K1 assembly were corrected using the new CH PICR reference assembly, facilitating the identification of target sites for gene knock-outs.

Chapter 2

A REFERENCE GENOME OF THE CHINESE HAMSTER BASED ON A HYBRID ASSEMBLY STRATEGY

2.1 Preface

This section is adapted from Rupp, MacDonald, Li, Dhiman et al, 2018 with permission (see Appendix C). Existing Illumina sequencing data along with new PacBio sequencing data were used to develop multiple improved genome assemblies for the Chinese hamster (CH). The highest quality assembly was selected to be the reference genome for CH and Chinese hamster ovary (CHO) cells. This work was the result of an international collaboration among members of the CHO genome community. I carried out the annotation and comparison of the PICR and IPCR genome assemblies. Genome assembly and manual quality assessment were completed by Oliver Rupp, gap analysis was completed by Shangzhong Li, and the chromatin state and gene structure analyses were completed by Heenan Dhiman.

2.2 Abstract

Accurate and complete genome sequences are essential in biotechnology to facilitate genome-based cell engineering efforts. The current genome assemblies for *Cricetulus griseus*, the Chinese hamster, are fragmented and replete with gap sequences and misassemblies, consistent with most short-read assemblies. Here, we completely resequenced *C. griseus* using single molecule real time sequencing and merged this with Illumina-based assemblies. This hybrid approach generated a more continuous and complete genome assembly than either technology alone, reducing the number of scaffolds by >28-fold, with 90% of the sequence in the 122 longest scaffolds. Most genes are now found in single scaffolds, including up- and downstream regulatory elements,

enabling improved study of non-coding regions. With >95% of the gap sequence filled, important Chinese hamster ovary cell mutations have been detected in draft assembly gaps. This new assembly will be an invaluable resource for future basic and pharmaceutical research.

2.3 Introduction

For decades, Chinese hamster ovary (CHO) cells have been the primary recombinant protein production host across the biopharmaceutical industry [48]. Characteristics, such as glycosylation, fast growth, and ease of genetic manipulation, help explain their prevalence. The history of CHO cells dates back to the 1950s, when ovarian connective tissue was harvested from the Chinese hamster and derivative cells spontaneously became immortal [49]. Since then, CHO has diverged into different adherent and suspension cell lines, such as CHO-K1, CHO-S, and CHO DG44 [27]. CHO cell protein production capacity has been greatly improved through decades of refinements in bioprocessing strategies, media optimization, and engineering of transgenes and expression vectors. However, little engineering was done on the host cell itself, which remained poorly characterized for decades. Increasing demands on quantities of difficult-to-express proteins, protein quality, and time-to-market now require new strategies that involve cell engineering.

To facilitate CHO cell research and development, the community relies on published genomes for the CHO-K1 cell line and the parent Chinese hamster, sequenced using short-read Illumina technologies [27, 29, 50, 51]. These resources have enhanced the use of transcriptomics, proteomics, genetic engineering, and other technologies [33, 52, 53] to understand and engineer desired traits in cells. However, to improve the accuracy in such endeavors, there is a need for genomic resources with a far more continuous sequence and less pervasive gaps. The acquisition of such continuous sequences is now possible with third-generation sequencing technologies, such as single molecule real time (SMRT) sequencing technology [54], which provide mean read lengths that are more than an order of magnitude larger than earlier sequencing technologies. The

reads can span repetitive elements, resulting in longer contigs and minimal gaps within scaffolds [55, 56, 57]. Long-read sequencing facilitates the creation of mammalian genome assemblies that approach the current quality of the human genome.

To obtain a higher quality reference assembly of the Chinese hamster (CH), we have resequenced CH liver tissue using long-read SMRT technology at 45x coverage. Assemblies generated with Illumina or SMRT sequencing data were merged with the existing publicly available assemblies. Assembly merging yielded four candidate assemblies, which were evaluated for completeness and accuracy using 80 quality metrics. Merging the platform-specific assemblies resulted in a more continuous, accurate, and complete genome assembly than using either technology alone. The final assembly presented is the most continuous Chinese hamster genome to date, with the number of scaffolds reduced to fewer than 3%-6% the number in earlier works, and the mean contig length 16-29 fold longer. The new genome shows substantial improvement in gene completeness and the extent of flanking non-coding DNA, thereby enabling the identification of promoters and enhancers. Finally, 95% of the sequence gaps were filled, exposing hundreds of cell line-specific mutations in coding regions of the genome for several CHO cell lines. For example, an important single nucleotide polymorphism (SNP) in the glycosyltransferase, xylosyltransferase 2 (Xylt2), which impacts glycosylation and which was hidden in gaps in previous assemblies, can now be detected. Thus, this resource will serve as an important reference genome for researchers across the biotechnology industry and scientific community.

2.4 Materials and Methods

2.4.1 Sequencing

2.4.1.1 Illumina sequencing

Short-read data from CH liver tissue were generated using Illumina's sequencing technology in two previously published studies. These included chromosome separated paired-end libraries and mate-pair short-read data [29], and whole-genome libraries

with different insert sizes [27]. The size and coverage of the two sequencing libraries are shown in Table 2.1.

Table 2.1: Overview of the different Illumina sequencing libraries

2013 RefSeq assembly [27]			
Insert size	Bases (Gbp)	Number of read-pairs	Mean read length
180	92.62	497,911,030	93.00
500	61.52	330,723,818	93.00
800	26.64	143,168,236	93.00
2000	30.34	188,590,215	80.55
5000	14.64	88,225,147	83.00
10000	12.48	75,094,089	83.00
20000	2.42	15,063,393	80.00
CSA assembly [29]			
Insert size/chromosome	Bases (Gbp)	Number of read-pairs	Mean read length
180 / chr1	34.30	111,436,801	153.97
180 / chr2	33.10	107,485,431	154.06
180 / chr3	19.96	65,249,415	153.03
180 / chr4	10.34	34,434,540	150.00
180 / chr5	9.26	30,834,375	150.00
180 / chr6	10.66	35,519,916	150.00
180 / chr7	5.22	20,881,069	125.00
180 / chr8	7.62	30,451,947	125.00
180 / chr9,10	18.74	61,114,204	153.37
180 / chrx	11.06	35,424,777	156.00
4,500	10.48	51,509,899	101.73
6,000	16.30	63,722,909	132.70
6,200	14.52	71,488,589	101.50

2.4.1.2 Pacific biosciences SMRT sequencing

Preparation of Chinese hamster tissue - Five female Chinese hamsters (strain 17 A/gy) were raised under certified conditions. At 10 weeks of age, the individuals were euthanized by CO_2 asphyxiation and verified by puncture wound to the abdomen. Livers were removed and cut into multiple pieces, flash frozen in liquid nitrogen, and stored at -80 degrees C until further processing.

High-molecular-weight genomic DNA extraction - High-molecular-weight (HMW) genomic DNA extraction and purification from randomized liver samples were performed using the MagAttract HMW DNA Kit (Qiagen Inc., Venlo, Netherlands) as per the manufacturer's instructions. HMW DNA was confirmed using a Fragment Analyzer (Advanced Analytical Technologies Inc., Ankeny, IA).

SMRT library preparation from genomic DNA samples - HMW DNA (10 μg aliquots) were converted into SMRTbell templates using the Pacific Biosciences RS DNA Template Preparation Kit 1.0 (Pacific Biosciences, Menlo Park, CA) as per the manufacturer's instructions. In summary, samples were end-repaired and ligated to blunt adapters. Exonuclease treatment was performed to remove unligated adapters and damaged DNA fragments. Samples were purified using 0.6x AMPureXP beads (Beckman Coulter Inc., Brea, CA). The purified SMRTbell libraries were eluted in 10 μl of elution buffer. Eluted SMRTbell libraries were size-selected on BluePippin (Sage Science Inc., Beverly, MA) to eliminate library fragments below 5 kbp. Final library quantification and sizing was carried out on a Fragment Analyzer (Advanced Analytical Technologies Inc.) using 1 μl of library. SMRTbell templates were aliquoted, shipped, and prepared for sequencing at the University of Delaware Sequencing & Genotyping Center and the Johns Hopkins University Deep Sequencing and Microarray Core.

SMRT sequencing on the Pacific Biosciences RSII - The amount of primer and polymerase required for the binding reaction was determined using the SMRTbell concentration and library insert size. Primers were annealed and polymerase was bound to SMRTbell templates using the DNA/Polymerase Binding Kit P5 and P6 (Pacific Biosciences). Sequencing was performed using DNA sequencing reagent C3 and C4 (Pacific Biosciences) with Pacific Biosciences RSII sequencers and SMRT Cell V3 (Pacific Biosciences) at the University of Delaware Sequencing & Genotyping Center (DBI) and the Johns Hopkins University Deep Sequencing and Microarray Core (JHU). RSII loading efficiency was optimized for each individual library utilizing

a standardized titration protocol. Over the course of the project, data capture time for the sequencing runs was initially set at 4 hr. This time was extended to 6 hr after software upgrades.

SMRT data metrics - The two sequencing centers ran a total of 202 SMRT cells (92 DBI, 110 JHU). A total of 65 SMRT cells were run using P5/C3 chemistry, whereas 137 SMRT cells were run using P6/C4 chemistry. After filtering and adapter trimming, a total yield of 107.45 Gbp was generated from 13.49 million sequence reads or approximately 45x coverage of the 2.4 Gbp genome. The mean read length calculated from all generated reads was 11.55 kbp. N50 read length calculated from all generated reads was 15.9 kbp.

2.4.1.3 SMRT read error-correction

Before assembly, SMRT reads were error-corrected (SMRT reads have on average 15% errors before correction). As insufficient SMRT coverage was obtained for self-correction of SMRT reads, we used Illumina paired-end reads [27, 29] for SMRT read error correction. The Illumina reads were preprocessed with the ALLPATHS-LG error-correction module for fragment libraries [19]. The reads from the same pair were joined and error-corrected, and gaps were filled if possible. This preprocessing resulted in a longer, single, and error-free read for each read pair.

Two different tools for PacBio read error correction were then tested with different parameters: proovread [58] and LoRDEC [59]. The tools were tested separately and in combination. The best results were achieved when, in the first step, proovread was run on the initial reads with a single iteration. All Illumina reads were mapped to all SMRT reads (allowing for multi-mappings) using the modified version of BWA in the proovread tool. Then, the bam2cns algorithm in proovread was applied to correct the reads based on the majority decision of the Illumina mappings. In the second step, the proovread results were further processed with LoRDEC. Using the corrected reads,

LoRDEC created a de Bruijn graph from the Illumina reads, mapped the nodes (k-mers of size 85 bp) to the SMRT reads, and corrected the unmapped regions following a path in the de Bruijn graph.

2.4.2 Genome size estimation

Genome size was estimated by the k-mer frequency of the Illumina read data using (1) all Illumina whole-genome paired-end libraries with an insert-size of 500 bp, (2) the libraries with an insert size of 800 bp, and (3) a combination of sets one and two. Jellyfish [60] was used to count the frequencies for k-mers of 17, 25, and 31 bp. The GCE tool [61] was used to estimate the genome size.

2.4.3 Genome assembly

The final genome assembly was conducted in two stages. In the first stage, four different assemblies were built with different tools and library combinations using the raw Illumina or the error-corrected SMRT reads. In the second stage, the four primary assemblies were iteratively merged in four different orders using the Metassembler tool [12] (Figure 2.1). Various quality metrics were used to assess the quality of the eight assemblies (four primary assemblies and four meta-assemblies). These metrics were further used to rank the assemblies and select the assembly with the best overall rank. Finally, the PICR was used as the reference assembly after polishing by correcting the single detected misassembly and minor gap filling from the PIRC assembly (see the "PIRC to PICR whole genome alignment" section in the Supplementary Materials).

2.4.3.1 Primary assemblies

Assembly 1: Illumina-based chromosome-sorted assembly (CSA) - The ten chromosome sorted libraries were assembled separately (each also using data from a whole-genome mate-pair library) with the ALLPATHS-LG tool [19]. The resulting scaffolds were filtered for possible contaminations of other chromosomes. The final assembly has been previously published [29] and is available at the NCBI assembly

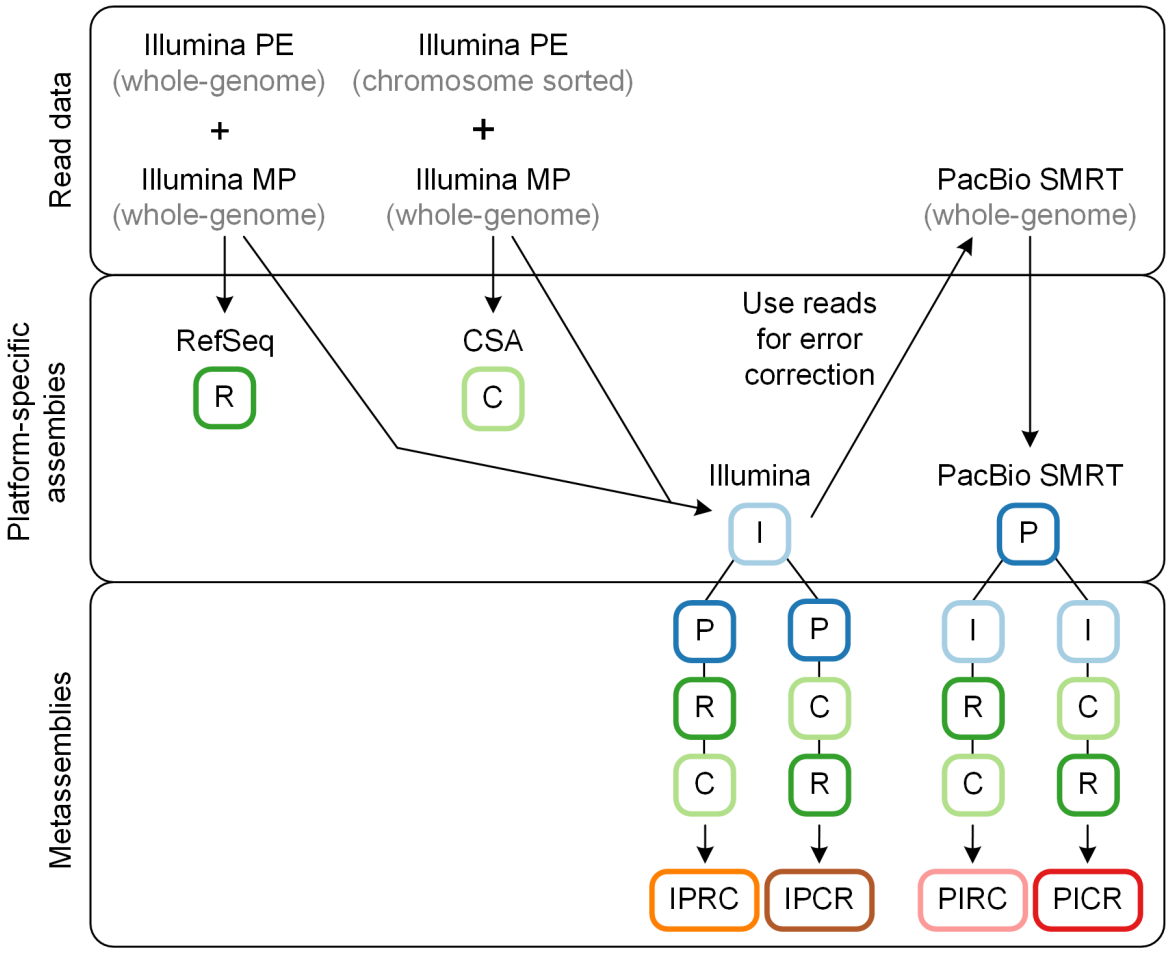

Figure 2.1: Overview of the complete assembly workflow. First, two primary assemblies were built, using all Illumina reads (I) or Pacific Biosciences SMRT reads (P). In addition, previously published assemblies were collected, which were based on chromosome separated reads (C) or the whole genome 2013 RefSeq (R). The four assemblies were then iteratively merged in four different orders, creating four metassemblies (IPCR, IPRC, PICR, PIRC). The best assembly of all eight assemblies was chosen based on a panel of 80 metrics.

archive (accession: GCA_000448345.1).

Assembly 2: Whole-genome Illumina assembly (2013 RefSeq) - The 2013 RefSeq CH genome [27] was built using different paired-end and mate-pair Illumina libraries. The libraries were assembled by SOAPdenovo2 [62]. The assembly is accessible at the NCBI assembly archive (accession: GCA_000419365.1).

Assembly 3: Whole-genome and chromosome-sorted assembly (Illumina) - Sequence data originating from the published chromosome-sorted Illumina libraries [29] and 2013 RefSeq whole-genome Illumina libraries [27] were combined and assembled with the ALLPATHS-LG tool (version 51927) [19].

Assembly 4: Pacific Biosciences SMRT assembly - The ALLPATHS-LG tool was used to merge and error-correct overlapping paired-end Illumina reads, and these reads were further extracted and converted into FASTA format to aid in the SMRT error-correction process. The error-corrected SMRT reads were assembled following the HGAP-3 pipeline [63] without the error-correction step. For better control over the workflow, we used the customizable makefile-based smrtmake workflow [64].

2.4.3.2 Merged assemblies

The four primary assemblies were iteratively merged with the Metassembler [12] tool. For each meta-assembly, one assembly is selected as the primary assembly. The scaffolds of a second assembly are subsequently mapped to the primary scaffolds using NUCmer [65]. A CE (compression/expansion) statistic, based on the distance of mate-pair reads, is computed for both assemblies. Primary scaffolds are joined and gaps are closed with the sequence of the second assembly. If the CE statistics of the primary scaffolds indicate potential errors, the sequence in this area is replaced by the sequence in the second assembly. The resulting scaffolds are then used as primary scaffolds for the next iteration. Changes to the default parameters were applied for the

merging step (asseMerge). The minimal range for finding links between scaffolds was increased to 50,000 and the minimal coverage of the secondary scaffold was lowered to 1x. The minimal gap size for closure was lowered to 1 bp (asseMerge –e 50000 –L 1 –t 1). The order in which the assemblies are merged influences the result of the final meta-assembly, and four different orders were tested (Table 2.2).

Table 2.2: Four different orders were used to merge the four initial assemblies with the Metassembler tool. PICR starts with the PacBio SMRT assembly, after which the Illumina assembly is merged into it, followed by the CSA assembly and the 2013 RefSeq assembly

Base assembly	**Added in step 1**	**Step 2**	**Step 3**	**Name**
PacBio SMRT	**I**llumina	**C**SA	**R**efSeq	**PICR**
PacBio SMRT	**I**llumina	**R**efSeq	**C**SA	**PIRC**
Illumina	**P**acBio SMRT	**C**SA	**R**efSeq	**IPCR**
Illumina	**P**acBio SMRT	**R**efSeq	**C**SA	**IPRC**

2.4.4 Chromosome assignment

Scaffolds were assigned to chromosomes using the chromosome-sorted library coverage, computed for 1 kbp regions. Specifically, for each 1 kbp region of each scaffold, the coverage of each chromosome library was computed. If at least 90% of the 1 kbp region of a scaffold showed a normalized coverage between 0.5 and 2 of the same chromosome, the scaffold was assigned to this chromosome. Scaffolds assigned to the pooled chromosome 9 and 10 library and all unassigned scaffolds were mapped to the mouse genome using NUCmer [65]. Yang et al. [66] and Wlaschin & Hu [67] described the localization of the Chinese hamster chromosomes on the mouse chromosomes. This information was used to assign the mapped scaffolds to a chromosome by manually comparing the mapped position with the localization from Yang et al. [66] and Wlaschin & Hu [67].

2.4.5 Gene prediction and annotation with Maker

We completed the annotation of the PICR and IPCR meta-assemblies using Maker v2.31.8 [68]. Chinese hamster ESTs (40 million reads) from SRA (SRR823966) were assembled using Trinity v2.0.6 [69]. The resulting transcripts were aligned to the previously published CH transcriptome assembly [27], which had used Trinity v. r2011-08-20. NUCmer [65] was used for the alignment with default parameters. A total of 91,027 transcripts were found in both transcriptomes and used as evidence for gene prediction within Maker. In addition, all proteins from the 2014 RefSeq annotation (GCF000419365.1) of the hamster genome were used as evidence.

Repeat masking was done within the Maker pipeline. To identify repeat regions, we used RepeatMasker version open-4.0.6 [70] with Dfam v2.0 (2015-09-23), a database of eukaryotic transposable element and other repetitive DNA sequence alignments, and the RepeatMasker database (release 2015-08-07, derived from RepBase v20.08). Once repeat masking was completed, BLAST v2.2.28 [71] and exonerate v2.2.0 [72] were run within Maker for evidence-based alignments and SNAP v2006-07-28 [73] and Augustus v3.2.2 [74] were run for *ab initio* gene prediction.

The resulting annotation only included genes with more than one type of evidence supporting the prediction, that is, both an *ab initio* prediction and an evidence-based alignment. Functional annotation of Maker's output was done as described in "Support Protocol 3: Assigning putative gene function" of "Genome Annotation and Curation Using MAKER and MAKER-P" [75]. BLAST was used (e-value < 0.001) to search each predicted gene against the Swiss-Prot release-2016-02 database, where the best hit was used to assign a putative function to the gene product.

2.4.6 Gap analysis

2.4.6.1 Identification of the filled-gap sequence

We aligned the 2013 CH RefSeq genome sequence to the PICR CH genome sequence using NUCmer [65] to identify gap sequence (see Figure 2.14). Briefly, NUCmer clusters a set of maximally exact matches as an anchor and then extends alignments

between the clustered matches. Gaps are represented using letters N in the genome, and since they differ between the 2013 RefSeq and PICR meta-assembly, the MUMmer alignments stop at gaps larger than 89 bp. This means that if two fragments that flank both ends of a gap are found on the same PICR scaffold in the same orientation, the sequence between the two matches corresponds to the sequence of the gap. Since sequence errors may occur near gap regions, we consider matches flanking a gap if the distance between the fragment and the gap is less than 10 bp. When the gap is shorter than 90 bp, MUMmer clusters the gap together with the two matches on both ends and only reports the merged long fragment as mapping. In this case, we first used the show-aligns method in MUMmer to output the alignment details between the 2013 CH RefSeq and PICR, and then we extracted the corresponding gap sequence by parsing the alignments. The gap analysis was performed using PICR and 2013 RefSeq hamster assembly, except for the gap in the Xylt2 gene, which was visualized using the 2011 RefSeq CHO-K1 genome assembly [50].

2.4.6.2 Identification of genes with gaps and mutations

We called variants in whole-genome resequencing data from various CHO cell lines [31, 27, 76]. GATK v3.5 [77, 78, 79] was used with the GATK manual recommended parameters. We also called variants using the reads from the 2013 RefSeq assembly project [27] to identify and filter false-positive variants. Pybedtools [80, 81] identified genes with gaps in their coding regions. Gene ontology (GO) term analysis was performed using DAVID [82, 83].

First, to identify classes of genes with gaps in the 2013 RefSeq assembly, we mapped all CH genes to their human homologs. The functional enrichment analysis for all the 2,252 genes with coding gap regions was performed using the human genes with CH homologs as the background gene set. Second, to identify classes of genes with a higher frequency of mutations in gaps, we looked for over-representation of the 132 genes with variants in coding gaps, while using the 2,252 gap-filled genes as the background. GO terms with a p-value smaller than 0.01 were visualized using REViGO

[84]. Code for the gap analysis can be acquired from:
https://github.com/LewisLabUCSD/assembly_gaps.

2.5 Results

2.5.1 Platform-specific assemblies of the Chinese hamster genome

2.5.1.1 Pooled Illumina assembly

In two previous and independent studies, the CH genome assembly was generated using Illumina sequencing data from liver tissue DNA that was acquired from the same hamster colony that was used for deriving CHO cells in 1957. The 2013 RefSeq assembly originated from whole-genome libraries with varying insert sizes [27]. A second assembly, CSA, created from chromosome sorted sequencing libraries is also publicly available [29]. The different libraries combined yielded about two billion read pairs with read lengths from 99 to 150 bp, totaling 442.22 Gbp (Table 2.1). K-mer-based genome size estimations of different libraries and k-mers ranged between 2.55 Gbp and 2.75 Gbp.

We *de novo* assembled the pooled Illumina reads from both previous assemblies using ALLPATHS-LG. This Illumina assembly contained 2.39 Gbp of scaffolds with 2.66% gaps. The scaffold N50 number (the minimal number of scaffolds needed to cover 50% of the assembled genome) was 128, with an N50 length (length of the N50 scaffold) of 5.95 Mbp (Table 2.3), which was much greater than the previously published assemblies.

2.5.1.2 Pacific Biosciences SMRT assembly sequencing assembly

Pacific Biosciences SMRT (PacBio SMRT) sequencing yielded a 107.45 Gbp total sequence from 13.49 million subreads, corresponding to ∼45x coverage of the 2.4 Gbp genome (after filtering and adapter trimming). Pooled and corrected Illumina reads were used to correct sequencing errors of the SMRT reads. Specifically, overlapping paired-end reads were merged and error-corrected as part of the ALLPATHS-LG [19] assembly process. This process created about 836 million single reads, with a

Table 2.3: Assembly metrics of the Illumina scaffolds and PacBio SMRT curated assembly compared to the previously published assemblies.

	2013 RefSeq	**CSA**	**Pooled Illum. scaffolds**	**Curated PacBio SMRT contigs**
Scaffolds [#]	52,710	28,749	17,373	1,659
Length [Gbp]	2.36	2.33	2.39	2.31
Min length [bp]	201	830	898	100,560
Max length [Mbp]	8.32	14.66	25.84	16.08
Mean length [kbp]	44.78	81.14	137.45	1,394.69
Median length [bp]	363	1,927	2,063	693,156
N50 length [kbp]	1,558.30	1,236.52	5,951.71	2,906.73
N50 [#]	450	501	128	223
N90 length [kbp]	395.29	180.69	1,003.29	623.9
N90 [#]	1,558	2,251	468	884
Total N gaps [#]	166,152	290,660	110,314	0
Total N [%]	2.49	10.45	2.66	0

mean size of 171 bp and a total of 143.75 Gbp. These were reused in the SMRT error-correction, which was done in two steps using proovread [58] and LoRDEC [59], leading to a reduction in the indel-ratio (the number of insertion/deletions divided by the number of matches in the alignments against the Illumina contigs) from 0.18 to 0.04 (see Supplementary Materials section for more details). SMRT reads were assembled using HGAP [63], resulting in the assembly hereafter referred as the PacBio SMRT assembly. After removal of duplicate contigs (see Supplementary Materials), the assembly resulted in 2.3 Gbp of non-redundant sequence with an N50 scaffold number of 223 and an N50 length of 2.9 Mbp (Table 2.3).

2.5.2 A highly contiguous meta-assembly is obtained by merging draft assemblies

Recent studies have highlighted the improvements in SMRT-only assemblies compared with Illumina-only assemblies [55, 56, 57, 85, 86]. Here, we found that both the pooled Illumina assembly (with mixed read length) and the PacBio SMRT-only assembly resulted in substantially improved assembly statistics compared with the two published CH genome assemblies (Table 2.3), with an order of magnitude fewer

scaffolds and 2-4x larger N50 values. However, the longer PacBio SMRT reads and the larger Illumina insert libraries should provide unique strengths that can be captured through assembly merging. Therefore, we aligned the scaffolds and contigs from four independent assemblies: the Illumina-based CSA [29], the 2013 RefSeq assembly [27], the pooled Illumina assembly developed here, and our *de novo* uncurated PacBio SMRT assembly. The Metassembler tool [12] uses the first assembly provided as the base and subsequently merges additional assemblies. The tool was applied to the four assemblies using four different orders of merging, resulting in four different meta-assemblies (Table 2.2).

All meta-assemblies showed considerable improvement over all initial draft assemblies (Table 2.4), with far fewer N50 scaffolds (only 32–34 compared with 223 for the PacBio SMRT and 128–501 for the Illumina-based assemblies), and a significant decrease in the gap sequence compared with the Illumina-only assemblies. Improvements in many metrics in all the intermediate merging stages show that all four initial draft assemblies contribute toward the improvement of the final assemblies (Figure 2.2). However, the meta-assemblies starting with the PacBio SMRT assembly outperformed the ones starting with the Illumina assembly in almost all metrics.

Table 2.4: Assembly metrics of the four merged assemblies.

	PICR	**PIRC**	**IPCR**	**IPRC**
Scaffolds [#]	1,829	1,825	2,317	2,304
Length [Gbp]	2.37	2.37	2.36	2.36
Min length [bp]	568	568	915	915
Max length [Mbp]	80.58	80.58	66.35	66.35
Mean length [kbp]	1,295.21	1,298.43	1,019.33	1,024.64
Median length [bp]	37,019	38,181	13,201	14,241
N50 length [kbp]	20,188.72	19,582.71	21,744.88	21,262.79
N50 [#]	32	33	33	34
N90 length [kbp]	4,400.57	4,422.38	3,545.61	3,650.27
N90 [#]	121	122	122	122
Total N gaps [#]	3,237	3,250	72,528	72,536
Total Ns [%]	0.12	0.12	1.13	1.13

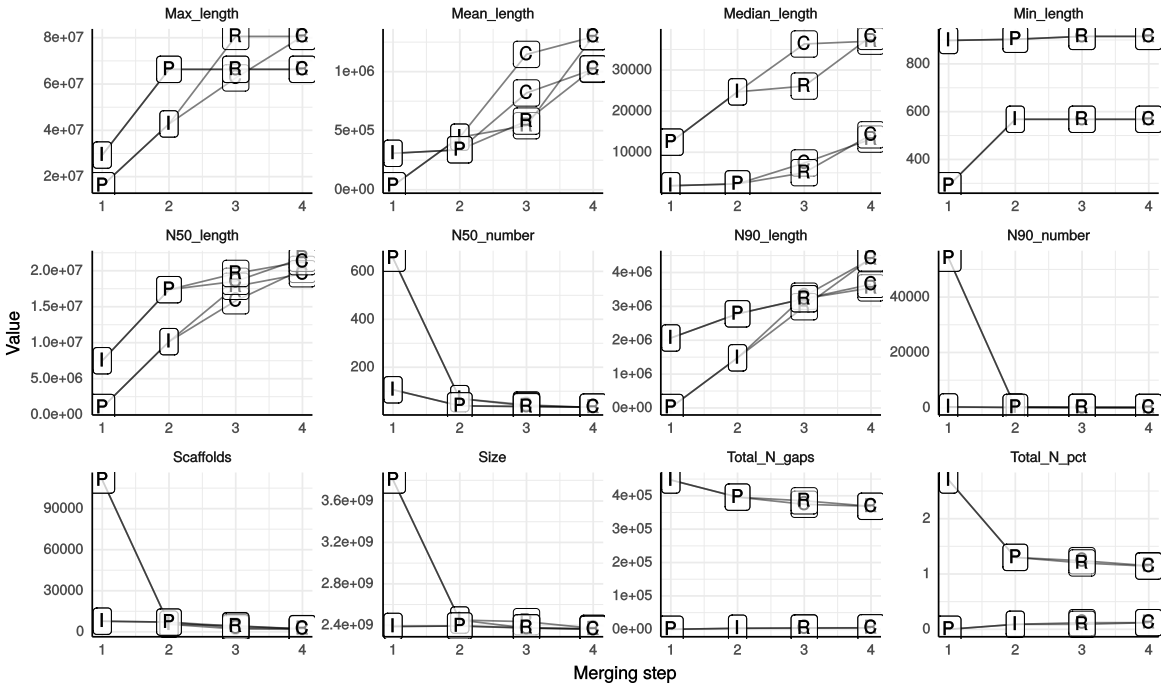

Figure 2.2: Assembly metrics at the different stages of the metassemblies. Improvement in all stages of the metassembly can be seen in most of the metrics.

To validate the accuracy of each assembly, the chromosome-sorted sequencing libraries [29] were aligned to the scaffolds. Misassemblies can be easily identified by decreased read coverage from one chromosome and a rise in coverage from another (Figure 2.3). Manual inspection of all scaffolds larger than 1 Mbp showed only one scaffold with a clear misassembly in the PacBio SMRT-starting (PICR and PIRC) meta-assemblies and 11 in the meta-assemblies starting with Illumina scaffolds (IPCR and IPRC), whereas the 2013 RefSeq assembly has >24 (Figure 2.4). Inspection of the chromosome coverage at the error region in PICR (Figure 2.5) showed a 30 kbp region that contained low and mixed coverage, along with scaffolding gaps. This region was manually cut, and two new scaffolds were created. Ultimately, 96.6% of the sequence could be unambiguously assigned to a specific chromosome (Table 2.5).

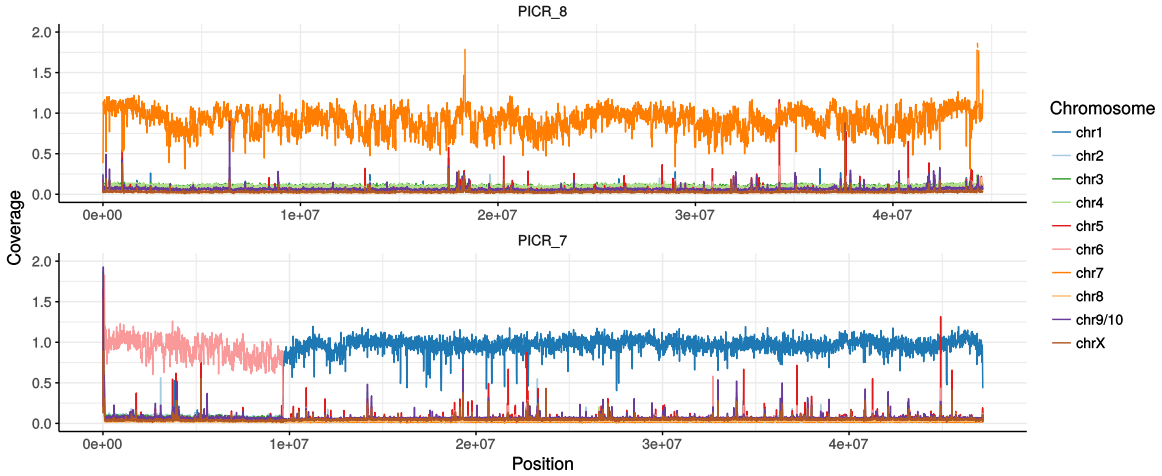

Figure 2.3: Normalized coverage plots identify misassembly sites. CH chromosomes were previously sorted and sequenced separately [29], and reads were aligned to each scaffold. Each color indicates aligned reads from different chromosomes. The top image shows a scaffold in which the normalized coverage shows all reads were from a single chromosome. The bottom plot shows a clear assembly error in which the first 10 Mbp are covered by a different chromosome (pink) than the remaining part (blue).

Table 2.5: Number and size of scaffolds assigned to each chromosome.

Chromosome	**Number of scaffolds**	**Bases [Mbp]**
1	73	549.76
2	35	463.59
3	42	281.86
4	24	231.54
5	56	188.78
6	9	155.90
7	13	134.49
8	24	96.66
9	15	18.79
10	2	32.58
X	48	134.88
unassigned	1,489	80.08

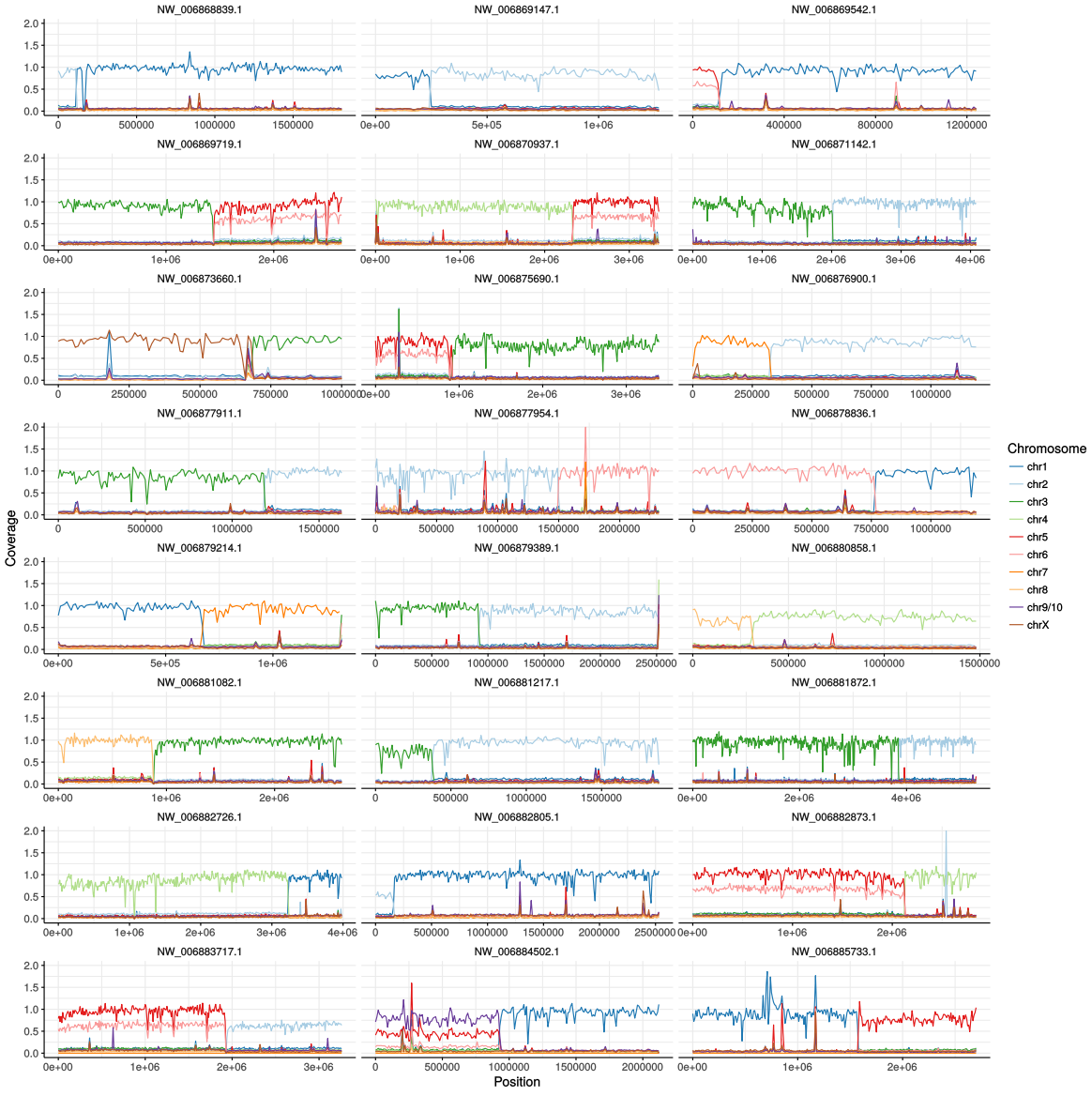

Figure 2.4: Chromosome-sorted reads were realigned to all scaffolds larger than 1 Mbp in the 2013 RefSeq assembly. From these, 24 scaffolds were identified that had more than 5% of the scaffold not associated with the primary chromosome.

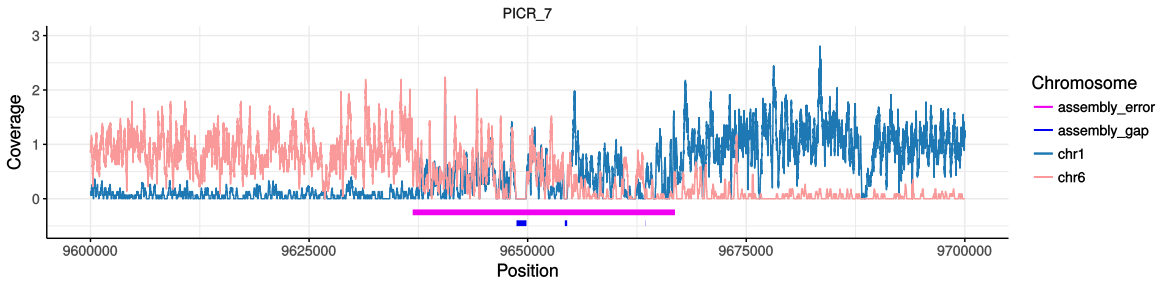

Figure 2.5: Zoomed chromosome coverage plot at the assembly error region of PICR scaffold 7. A 30 kbp region (pink bar) with low and mixed coverage and assembly gap (blue bars) is clearly visible.

2.5.3 The best assembly is identified using 80 assembly metrics

To quantify and compare the quality of the eight assemblies (the four initial assemblies and the four meta-assemblies), we computed 80 different metrics (see Appendix B) that were split into six classes covering different aspects of assembly quality (Figures 2.2, 2.6, and 2.7a). The assemblies were ranked in each class individually. The PICR meta-assembly had the best overall rank in four of the six classes, followed by PIRC with two best overall ranks. Based on this evaluation, PICR was chosen for further analyses.

The PICR meta-assembly has substantially longer contigs (contiguous sequences with "N"-regions smaller than 100 bp) than the previous RefSeq assembly and even assemblies of some model organisms, such as the rat (*Rattus norvegicus*, assembly Rnor_6.0). In addition, PICR is approaching the continuity observed in the mouse reference assembly, GRCm38.p5 (Figure 2.7b and more information in the Supplementary Materials section).

2.5.4 Polishing the final assembly

2.5.4.1 Chromosomes are assigned using reads from flow-sorted DNA

To assign each scaffold to a chromosome, we aligned all chromosome-sorted reads to the PICR meta-assembly. 307 scaffolds were uniquely assigned to a chromosome,

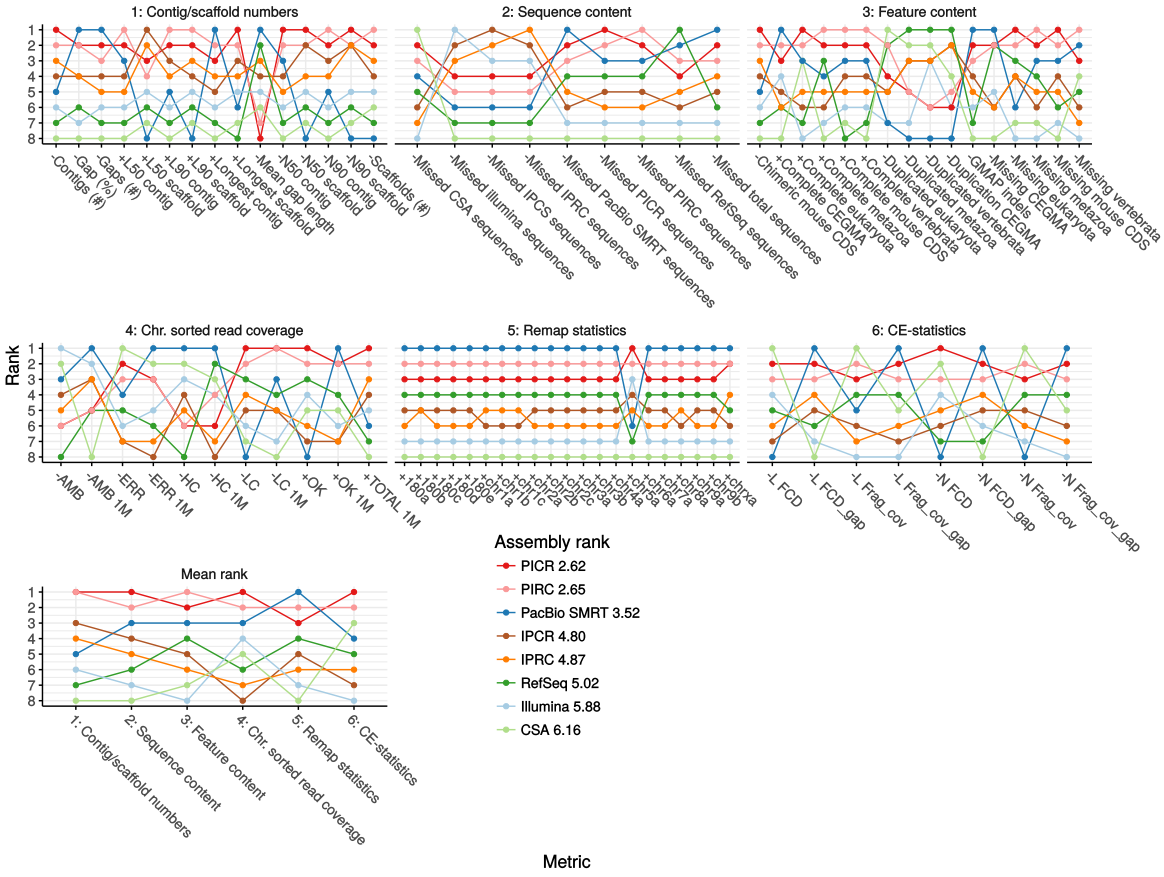

Figure 2.6: Ranks of the assemblies for all metrics in all classes.

accounting for 94% of the genome (or 2.23 Gbp). Unassigned scaffolds and scaffolds assigned to the unseparated CH chromosome 9 and 10 library were instead mapped to the mouse genome. Scaffolds that could be aligned uniquely were assigned to a CH chromosome based on published CH chromosome localization [67, 66]. Fifteen scaffolds (18.79 Mbp) could be assigned to chromosome 9 and 2 scaffolds (32.58 Mbp) to chromosome 10. A detailed list of assigned scaffold numbers and sizes is shown in Table 2.5. The final PICR assembly and the associated raw PacBio SMRT sequencing read data are available under NCBI BioProject PRJNA389969. The existing Illumina assemblies are available under NCBI BioProjects PRJNA167053 (2013 RefSeq) and PRJNA189319 (CSA). Illumina sequencing data for BioProject PRJNA167053 are available from the

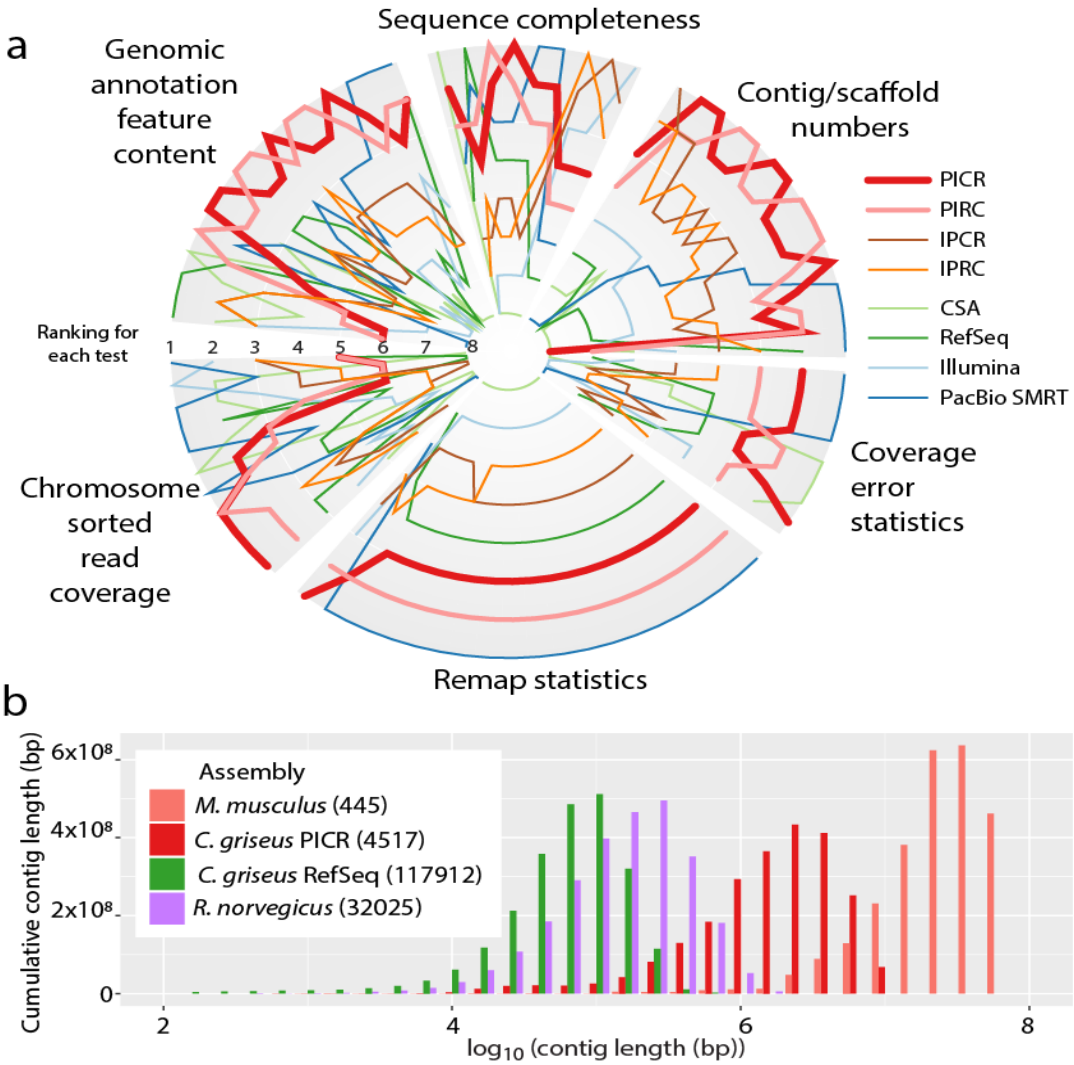

Figure 2.7: The PICR assembly ranked against other mammalian assemblies. (a) The PICR assembly was compared to the other candidate assemblies of *C. griseus* based on 80 different assembly metrics. This shows for each test how the assemblies compare. The best assembly for each test is plotted on the outer rim, while the worst is near the center. Eighty tests were defined in six different categories. On average, the PICR assembly was the most highly ranked with the PIRC assembly closely following. (b) Weighted histogram of the contig lengths for the PICR assembly (red) compared to the Ensemble mouse (salmon), rat (purple), and the prior CH 2013 RefSeq assemblies (green).

Sequence Read Archive under SRP020466.

2.5.4.2 Repeat masking, gene prediction, and annotation

We annotated the PICR and IPCR meta-assemblies using the Maker annotation tool [68]. Due to the similarity of the PICR and PIRC assemblies, we decided to compare the annotation of PICR and IPCR. This comparison demonstrated the impact of using assemblies built from different sequencing methods as the primary assembly in Metassembler. Repeat-masker [70] masked approximately 5.5 million repeats in PICR and 5.7 million in IPCR (Table 2.6).

Table 2.6: Number of repeats by class masked in PICR and IPCR assemblies prior to annotation.

Repeat class	**PICR**	**IPCR**
Simple repeats	2,237,638	2,516,964
Low complexity repeats	271,488	274,248
Long terminal repeats	625,480	601,600
LINEs	882,602	858,268
SINEs	1,282,452	1,227,136
Satellites	8,436	14,714
Retro-tranposons	3,998	4,422
DNA repeat elements	146,430	170,094
RNA repeats	4,940	4,828
Other	23,522	22,470
Unknown	48,586	44,368
Total masked	5,535,572	5,739,112

The Maker annotation yielded ∼1,300 more genes and transcripts in PICR than in IPCR (Table 2.7). Functional annotations were assigned for 23,153 transcripts/proteins in PICR, but only for 21,839 transcripts/proteins in IPCR. The annotations of PICR and IPCR demonstrate that beginning assembly merging with the PacBio SMRT assembly, rather than the Illumina assembly, led to the identification and functional annotation of more genes.

The predicted proteins from PICR were searched using BLAST (e-value < 0.001) against the proteins from IPCR and vice versa to compare the annotation of the two

assemblies. A total of 24,578 proteins in PICR have a BLAST hit in IPCR and 22,970 of these proteins have a functional annotation assigned from the top BLAST hit against the Swiss-Prot database. Only 23,420 proteins in IPCR had a BLAST hit in PICR.

Analysis of the 236 proteins found in IPCR, but not in PICR, showed that most of these genes were not functionally annotated or were duplicates or isoforms of genes in PICR. Some proteins unique to the IPCR assembly include the protease carboxypeptidase Q (Cpq), the histone H3 threonine kinase haspin (Gsg2), the antioxidant sulfiredoxin-1 (Srxn1), and the possible ortholog of DNA-directed RNA polymerase III subunit RPC9 (Crcp). Analysis of the 367 proteins found in PICR, but not in IPCR, showed that about half were not functionally annotated. Proteins of interest unique to the PICR meta-assembly include posphatidylglycerophosphate (pgp or pgs1), which is involved in phospholipid biosynthesis in mammalian cells [87], and two DNA repair-related proteins: breast cancer type 1 susceptibility protein (Brca1) and nonhomologous end-joining factor 1 (NHEJ1). In addition, Bcl-2-like protein 10 (Bcl2l10), a signaling molecule involved in apoptosis, and stress-associated endoplasmic reticulum protein 1 (Serp1) are both in PICR, but not IPCR. MicroRNAs targeting these two proteins in CHO cells have been developed [88].

2.5.4.3 NCBI annotations

To help us select the next CH reference assembly, NCBI completed a 'light' annotation of PICR, IPCR, and 2013 RefSeq assemblies (more information can be found in the Supplementary Materials section). In December 2018 after PICR was submitted to NCBI as the new Chinese hamster reference genome, NCBI ran their full RefSeq annotation pipeline on the PICR (2018 RefSeq) genome. Summary statistics of the 2018 RefSeq annotation of the PICR assembly is provided in Table 2.8.

Table 2.7: Gene and transcript information from the Maker annotation of the PICR and IPCR genome assemblies.

All genes	**PICR**	**IPCR**
Gene count	24,686	23,410
Transcript count	24,948	23,656
Transcripts per gene	1.01	1.01
Avg. length transcript	17,615.04	18,089.17
Total length transcript	439,460,104	427,917,413
Avg. coding length	1,324.93	1,316.11
Total coding length	33,054,355	31,133,905
Avg. exons per transcript	7.49	7.54
Total exons	186,939	178,277
Complete transcripts		
Transcript count	18,476	17,557
Avg. length transcript	18,908.94	19,434.05
Med. length transcript	8,236	8,228
Total length transcript	349,361,499	341,203,668
Avg. coding length	1,334.19	1,317.74
Med. coding length	981	966
Total coding length	24,650,545	23,135,510
Avg. exons per transcript	7.49	7.48
Total exons	138,358	131,262
Incomplete transcripts		
Transcript count	6,472	6,099
Avg. length transcript	13,921.29	14,217.70
Med. length transcript	8,128	8,692
Total length transcript	90,098,605	86,713,745
Avg. coding length	1,298.49	1,311.43
Med. coding length	933	942
Total coding length	8,403,810	7,998,395
Avg. exons per transcript	7.51	7.71
Total exons	48,581	47,015

2.5.4.4 The PICR meta-assembly has more contiguous genes and non-coding regulatory elements

In the previous genome assemblies, many genes were fragmented or separated from their functional genomic elements (e.g. promoters, enhancers, or regions of active or repressed transcription). Thus, efforts to define the chromatin states of genes and

Table 2.8: Gene and transcript information from the December 2018 RefSeq annotation of the PICR genome assembly.

Feature	**PICR (2018 RefSeq)**
Gene count	27,161
Transcript count	56,106
Transcripts per gene	2.08
Avg. length transcript	3,068
Avg. exons per transcript	11.14
Avg. coding length	1,984
Total exons	247,246
Total introns	219,551
Transcript types	
Total mRNA	46,750
Total miRNA	291
Total tRNA	485
Total lncRNA	5,891
Total snoRNA	606
Total snRNA	871
Total guide_RNA	42
Total rRNA	21

their regulatory units were error-prone [31]. We therefore recalculated the chromatin states for the PICR assembly using the ChiPSeq-derived histone mark reads obtained by Feichtinger et al. In comparison with the previously deduced chromatin states, the emission profile of the new chromatin states matched better with those obtained for the well-assembled human epigenome [89] (Figure 2.8a).

To test whether the continuity of genes and their regulatory regions is improved in the PICR meta-assembly, we extracted a shortlist of 1,538 mitochondria-associated genes, localized to 1,654 sites in the mouse genome. We mapped the sequences between the mouse transcription start site (TSS) and the transcription end site (TES) against the PICR meta-assembly, the 2013 RefSeq assembly [27], and CSA [29]. Genes were considered present if both the TSS and TES were found on the same scaffold. Due to the high variance in untranslated regions (UTRs) across species, few genes were identified (Figure 2.8b), demonstrating the importance of a species-specific genome.

We subsequently searched for both the start and the end of the coding sequences on the same scaffold (Figure 2.8c). Of the complete genes found in PICR (1,011), 85% were annotated and localized to 900 unique locations. The corresponding sequences in PICR were elongated to include UTRs, 5 kbp upstream and 1.5 kbp downstream, to capture potential regulatory regions, such as promoters or repressive elements. These elongated sequences were mapped against the previously published Chinese hamster genomes and again checked for presence on a single scaffold (Figure 2.8d).

Several genes had their elongated sequence not properly assembled in earlier assemblies, despite having the coding sequence on a single scaffold in each of the three assemblies (see Supplementary Materials section for details). Examples for three genes, Rab4b, a member of the Ras family of oncogenes, the mitochondrial ribosome protein MRPL27, and TIMM50, a translocase responsible for targeting proteins into the mitochondria, are shown. In all cases, the scaffold in the CSA assembly contained histone marks for active transcription or a genic enhancer, but lacked flanking enhancers and promoter regions. In the new assembly, these are now correctly annotated (Figure 2.8e). The correct assembly of coding and non-coding regions is of increasing importance to better understand their regulatory function and enable engineering applications. A browser with all PICR scaffolds, the preliminary annotation, and the chromatin states throughout a batch culture is available at http://cgr-referencegenome.boku.ac.at/jb/.

2.5.5 Pervasive gaps are filled by SMRT sequencing

The 2013 RefSeq assembly [27] contains 166,152 gaps with a total length of 58.8 Mbp, representing 2.5% of the entire genome. The PICR meta-assembly has eliminated most gaps with only 3,238 remaining (Figure 2.9a). These gaps account for 2.9 Mbp, or 0.1%, of the genome. By aligning the 2013 RefSeq assembly to PICR using MUMmer3.0 [65], we identified the missing sequence for 125,812 (76%) of the RefSeq gaps (Figure 2.9b). The sequence for a subset of RefSeq gaps was not identified in the PICR meta-assembly. Of this subset, 90% could not be unambiguously identified

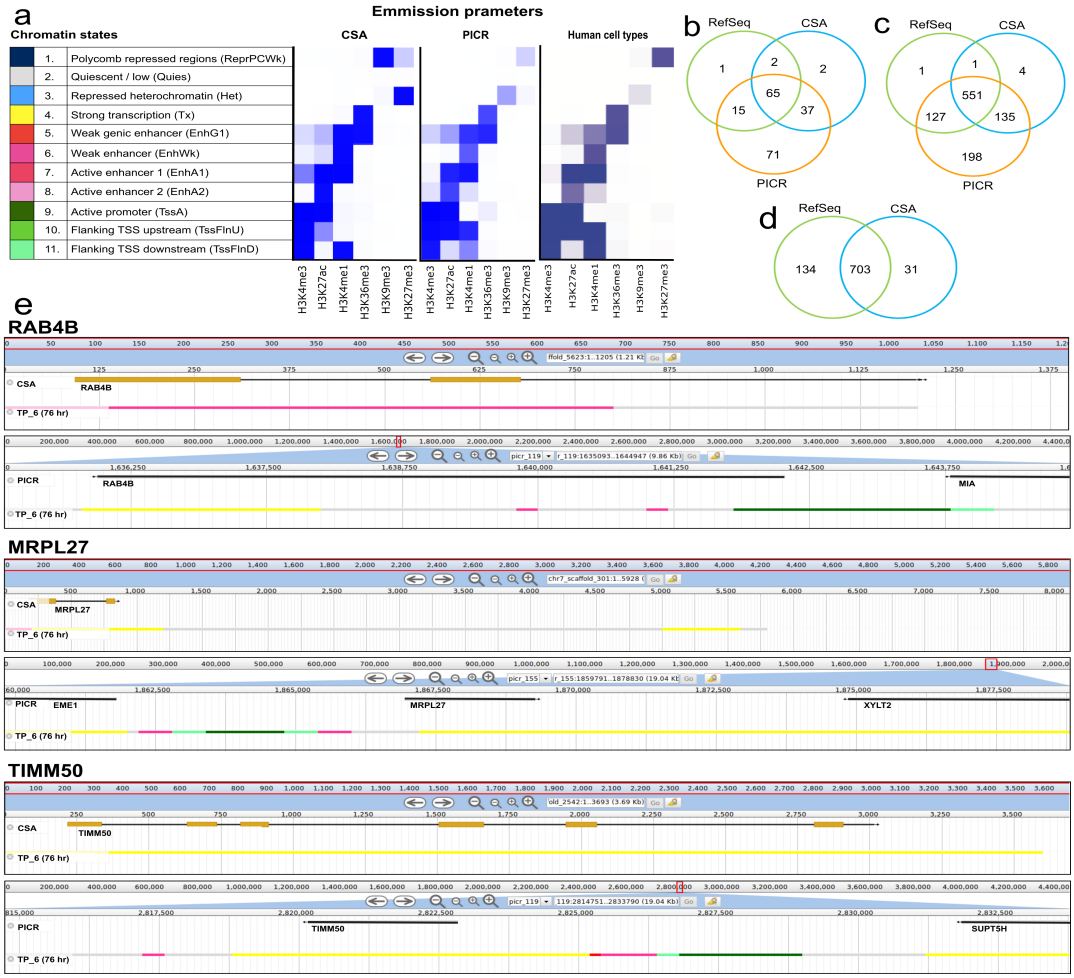

Figure 2.8: Importance of the correct assembly of genes and non-coding regions. (a) Chromatin states defined by histone marks: Left: histone marks for CSA assembly [29], [31]; center: histone marks for PICR assembly; right: histone marks from the Human Epigenome Project [89] (b) 1,538 genes associated with mitochondria were blasted from TSS to TES against the CSA and 2013 RefSeq assemblies. The number of hits completely found on a single scaffold is displayed for each assembly. (c) Mouse coding sequences were blasted against CH assemblies from translation start to end. (d) The 1,011 complete genes found in PICR were extended 5 kbp upstream and 1.5 kbp downstream to include promoters and other regulatory non-coding regions and blasted against existing assemblies. (e) Chromatin states around three genes as found in the previously published CSA-based chromatin state model [31] (top for each gene) and the PICR assembly (bottom for each gene), showing promoter and regulatory elements in addition to active transcription

because the flanking fragments did not both align to the new assembly, likely due in part to misassemblies in the 2013 RefSeq assembly (Figure 2.9).

The elimination of most gaps in the PICR meta-assembly enables more accurate and complete genome editing and genomic analyses since 2,252 genes in the PICR meta-assembly had their 2013 RefSeq assembly gaps filled. We called variants from whole-genome resequencing data for 13 representative resequenced CHO cell lines [31, 27] to identify genes that have newly discovered mutations in the RefSeq coding gaps. Each sample has approximately 300 mutations in coding gaps, 90% of which are SNPs (Table 2.9). Across 13 cell lines, 885 novel variants in coding gaps were found in 134 genes (Figure 2.9c).

Gene classes with the highest gap filling success included genes associated with protein binding, RNA binding, and transcription (Supporting Information Figure 2.15), including genes containing zinc finger motifs and ribosomal genes. Previously, such genes were replete with gaps due to their conserved domains shared across many other genes in the genome. We further explored which genes had coding mutations in their filled gaps. The top GO terms for these 225 genes are also enriched in DNA binding and transcription (Supporting Information Figure 2.16). In summary, the gaps in the previous assembly could potentially confound genomic studies in CHO, especially those involving mutations associated with DNA or RNA binding, including transcription factors.

2.5.5.1 An important mutation in Xylt2 is found within a filled sequence gap

Beyond their importance in biopharmaceutical production, CHO cells were fundamental to cell biology and biochemistry research for many decades. For example, genetic screens of many CHO cell lines were used to identify glycosyltransferases [90, 91, 92, 93] and genetic mapping efforts were deployed to identify causal mutations. The pgsA745 cell line [76] has been used for decades in the glycobiology field due to its deficiencies in glycosaminoglycan synthesis [94], due to a truncation of the Xylt2

Table 2.9: Variant statistics in different CHO cell lines.

	pgsa	**CHOS**	**FCS**	**PFMCB**	**PF6mon**	**1D9MCB**	**1D93mon**	**no_gln**	**C0101**	**CHOS**	**CHOpr**	**DG44**	**ECACC**	**K1SF**
Total variants [10^6]	5.694	4.663	5.948	5.814	5.548	5.901	5.894	5.834	5.566	4.663	5.043	4.728	5.025	5.049
Control variants [10^6]	1.052	0.861	1.123	1.072	0.993	1.134	1.139	1.062	1.051	0.861	0.896	0.853	0.880	0.889
Control filtered variants [10^6]	4.641	3.801	4.826	4.742	4.555	4.768	4.756	4.772	4.515	3.802	4.146	3.875	4.146	4.160
Filtered snp [10^6]	3.764	3.121	3.849	3.758	3.612	3.727	3.726	3.781	3.630	3.121	3.395	3.191	3.384	3.410
Filtered ins [10^6]	0.426	0.327	0.482	0.481	0.467	0.518	0.512	0.491	0.428	0.327	0.363	0.327	0.369	0.362
Filtered del [10^6]	0.451	0.354	0.494	0.503	0.476	0.522	0.518	0.450	0.457	0.354	0.388	0.357	0.393	0.389
Gap variants	20,635	13,997	25,724	25,204	24,327	27,810	27,123	26,304	20,417	13,997	15,434	14,035	15,563	15,586
Gap snps	15,924	11,327	18,778	17,797	17,404	19,386	19,052	18,773	15,623	11,327	12,285	11,389	12,210	12,430
Gap ins	1,788	999	3,201	3,407	3,175	3,969	3,684	3,506	1,828	999	1,158	987	1,220	1,141
Gap del	2,923	1,671	3,745	4,000	3,748	4,455	4,387	4,025	2,966	1,671	1,991	1,659	2,133	2,015
Coding gap variants	268	212	330	333	294	348	339	366	277	212	207	178	195	200
Coding gap snp	249	199	293	285	274	303	290	310	259	199	195	168	186	188
Coding gap ins	6	3	16	25	6	20	23	28	6	3	3	3	4	2
Coding gap del	13	10	21	23	14	25	26	28	12	10	9	7	5	10

protein [95]. However, upon variant calling from whole-genome resequencing data for the pgsA745 cell line using the 2013 RefSeq assembly, we failed to identify the causal mutation, whereas a G->T SNP encoding a premature stop codon was found in exon 1 of Xylt2 when using the PICR genome assembly (Figure 2.9d). This mutation was previously missed since the 2013 RefSeq assembly has a gap of 447 bp that spanned the first exon on scaffold NW_003613846.1. However, this gap was filled in PICR, enabling the identification of the mutation. Thus, filling of the gap sequence leads to a valuable improvement to genomic studies, including the identification of causal variants in CHO cell lines.

2.6 Discussion

For 60 years, CHO cells have been invaluable for biomedical research and fundamental to the study of several biological processes, such as glycosylation [96] and DNA repair [97]. In addition, for >30 years, they have been the host cell of choice for the production of most biotherapeutics. Although the aforementioned research was carried out without genomic resources, new opportunities are arising with published CH and CHO genome sequences [27, 29, 50, 51]. However, the draft nature of these genome sequences poses challenges for many applications. Here, we present a major step forward in further facilitating the adoption of cutting-edge technologies for cell line development and engineering.

The primary outcome here is a substantially improved reference genome sequence for the Chinese hamster. Specifically, the N50 of the PICR meta-assembly is 13x the length of the 2013 RefSeq assembly N50, and we reduced the number of scaffolds to 1/29 the number in 2013 RefSeq. Furthermore, we demonstrated that the initial PICR assembly only had one detected misassembly, whereas the 2013 RefSeq assembly had at least 24 >1 Mbp scaffolds with cross-chromosome misassemblies (Figure 2.4). Finally, we eliminated more than 95% of the gap sequence in the 2013 RefSeq assembly, and provide a more complete and contiguous view of the genomic sequence of the Chinese hamster.

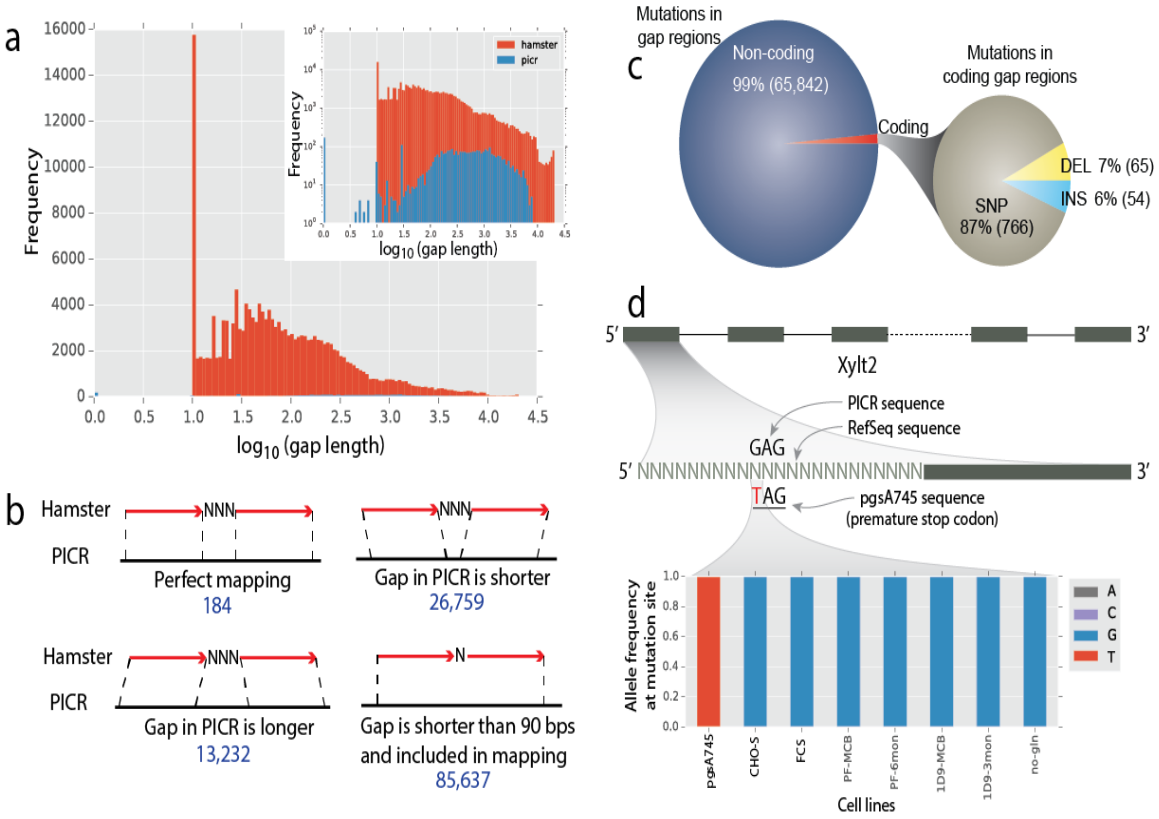

Figure 2.9: Important variants are located in sequence gaps in previous assemblies. (a) >95% of sequence gaps were filled in the PICR metassembly (inset shows the log frequency of gaps to highlight the low frequency of PICR gaps not visible in the normal histogram). (b) The missing sequence in gaps in the 2013 RefSeq assembly was identified by aligning 2013 RefSeq sequence flanking the gaps to the PICR sequence. (c) Across 13 cell lines, we found 65,842 SNP and indel mutations in the 2013 RefSeq gap regions, and 1.3% of these were found in coding regions. (d) A legacy CHO cell line, pgsA745, identified Xylt2 as the glycosyltransferase responsible for the first step in glycosaminoglycan biosynthesis, as this cell line is deficient in this process. Because of a gap in the 2013 Refseq assembly, only in the new PICR metassembly can the causal variant be identified. A G->T mutation introduces an early stop codon in exon 1, resulting in a loss in Xylt2 activity. The genotype is shown for a variety of CHO cell lines with only pgsA745 showing the early stop codon.

Various aspects of the genome assembly were improved by merging the different datasets and data types. First, merging the Illumina reads from two different genome sequencing efforts resulted in a higher quality genome than the starting assemblies. Second, further improvements in the assembly attributes were achieved by merging the single-platform assemblies. Previously, assembly merging with Metassembler was found to modestly improve the starting assemblies [98]. Here, we obtained large gains in the N50, with the PICR meta-assembly being approximately 4-fold more continuous than the starting assemblies. Medium and longer scaffolds were successfully merged, thus reducing the number of N50 and N90 scaffolds. However, by including Illumina-based assemblies, many short scaffolds remained, as seen in the lower median scaffold length in the PICR meta-assembly compared with the curated PacBio SMRT assembly. The merged assembly thus benefited both from the longer reads from the PacBio SMRT contigs and the longer scaffolds from the large insert size libraries used for the Illumina assemblies. It is anticipated that the use of optical mapping and chromatin interaction mapping [55] would further extend the scaffolds and span large repeat regions, resulting in more complete chromosomal maps for the Chinese hamster.

Despite the absence of genomic resources, CHO-based bioprocessing has advanced substantially for ∼30 years. Massive improvements in protein titer were predominantly achieved through media and process optimization. Systematic optimization of CHO cell lines itself has lagged behind *Escherichia coli* and *Pichia pastoris* and has only recovered traction with the comparatively late release of draft genomes. The availability of genomic data now enables improved control over product quality and more predictable culture phenotypes. For example, more contiguous and complete sequences will facilitate the identification of sites for targeted integration of transgenes, enabling more reproducible productivity across clones [99] and reducing the burden of stability testing. In addition, the elimination of gap sequence regions enables the improved identification of genomic variants and design of genome editing tools. Furthermore, by sequencing through repetitive elements, endogenous retroviral elements can be deleted. This could substantially reduce the retroviral particles secreted in

mammalian cell culture [100, 101], increase biopharmaceutical safety, and decrease the burden of adventitious agent testing and purification. Comparable efforts have successfully cleaned up similar elements in the porcine genome [102].

The full benefit of this more continuous genome will become apparent as novel genome-editing tools are applied to control cell phenotypes. These include efforts to delete larger tracts of the sequence, including genes, promoters, and other regulatory elements using paired gRNAs that remove the entire sequence rather than only introducing frameshifts [103]. Thus, genes can be removed or promoters can be replaced with synthetic or inducible elements. Furthermore, with more complete regulatory element sequences, one could use CRISPRa/i to regulate gene expression levels. Finally, tools can be deployed that modify the methylation of endogenous promoters to activate or silence gene expression [104, 105]. Overall, these strategies enhance our control over cell phenotype. As demonstrated, these precision engineering tools are highly dependent on the availability of a continuous and well-assembled genome, as presented here, to the entire scientific and industrial community.

2.7 Acknowledgements

I, along with the other authors of this work, would like to thank George Yarganian for providing hamster tissue, and Valerie Schneider and Françoise Thibaud-Nissen from NCBI who helped to run the "light" version of the NCBI annotation pipeline on the assemblies. This work was supported with generous funding from Biogen, Genentech, Eli Lilly and Company, Dublin City University, University of Queensland, and University of Tokushima. Grant support was provided by the Novo Nordisk Foundation to the Center for Biosustainability at the Technical University of Denmark (NNF10CC1016517 and NNF16CC0021858), NIGMS (R35 GM119850), a FISP fellowship from UC San Diego to S. Li, and NSF (NSF1144726, NSF1412365, NSF1539359, and NSF1736123) to the University of Delaware. H. Dhiman is supported by the EU Horizon 2020 MSCA ITN grant no. 642663. O. Rupp, S. Griep, I. Hernandez, V.

Jadhav, K. Brinkrolf, A. Goesmann, and N. Borth received support from the Austrian Center of Industrial Biotechnology Acib, a COMET K2 competence center of the Austrian Research Promotion Agency. I. Hernandez also received support from the Austrian Science Fund PhD Program "Biotop" (Grant Number W1224). Bioinformatics support by the BMBF-funded project "Bielefeld-Gießen Center for Microbial Bioinformatics-BiGi (Grant Number 031A533)" within the German Network for Bioinformatics Infrastructure is gratefully acknowledged. S. Polson and the computational infrastructure provided by the University of Delaware Center for Bioinformatics and Computational Biology Core Facility is supported through Delaware INBRE, NIGMS (P20 GM103446).

2.8 Author Contributions

I functionally annotated the genome assemblies and contributed toward writing the manuscript. O.R., S.G., and K.B. conducted genome assembly, and contributed toward writing. H.D. and I.H. performed the mitochondrial gene and chromatin state analysis. S.L. conducted gap analysis, prepared CHO pgsA-745 DNA for sequencing, and contributed toward writing. K.H. isolated hamster tissue and prepared DNA for sequencing. M.J.B. conceived of the project and oversaw the hamster DNA preparation. S.P. provided valuable guidance and contributed toward the sequencing and analysis. H.H., B.K., and M.S. contributed toward the sequencing and analysis. V.J. evaluated approaches to separate scaffolds for chromosomes 9 and 10. A.G. oversaw genome assembly efforts. N.E.L. conceived of the project, wrote the manuscript, and guided the gap analysis. N.B. conceived of and coordinated the project, oversaw the chromatin state analysis and wrote the manuscript. K.H.L. conceived of and coordinated the project, and oversaw the genome annotation.

2.9 Supplementary Materials

2.9.1 Additional information on SMRT sequencing and assembly

2.9.1.1 Sequencing and error-correction

Following sequencing of the Chinese hamster, reads were corrected using Illumina reads. To show the correction quality, a sample of 25,000 randomly selected PacBio reads in all three states of correction (raw, after step 1, and final) were mapped to the Illumina contigs. For each read, the indel-ratio (number of indels divided by the number of matches) was computed. The raw reads showed an indel-ratio of 0.18, which could be reduced to 0.06 after the first step and was further reduced to 0.04 in the final step (See Figure 2.10).

2.9.1.2 PacBio SMRT metassembly curation

Assembly of the PacBio reads using HGAP resulted in a final assembly containing 110,954 contigs with 3.80 Gbp total genome sequence. The N50 number was 655 contigs with a N50 length of 995.27 kbp. This assembly was about 50% larger than the expected genome size. The weighted histogram of the contig lengths (i.e. the number of contigs, weighted by the total number of bases in the contigs) in Figure 2.11) shows a clear bimodal characteristic with two peaks at about 20 kbp and 3.5 Mbp and a local minimum at about 100 kbp. The total size of all contigs falling into the second mode (i.e. contigs $\geq$ 100 kbp) is 2.3 Gbp (N50 number: 223 and N50 size: 2.9 Mbp, see Table 2.10). Realigning the complete assembly back to itself, showed that 80,138 contigs (0.97 Gbp) were completely ($\geq$ 90%) contained in larger contigs (some with rearrangements or larger insertions or deletions). These duplicated contigs may have resulted from poorly corrected reads. Nonetheless, the complete assembly was used for further steps and analyses, since the Metassembly algorithm only merges the best-aligning scaffolds (longest alignment, highest percent identity, etc.) while contigs with poorer alignment are discarded. For example, in the first iteration of the merging, about 1.3 Gbp were discarded from the PacBio SMRT assembly, while about 20 Mbp were discarded from the Illumina assembly.

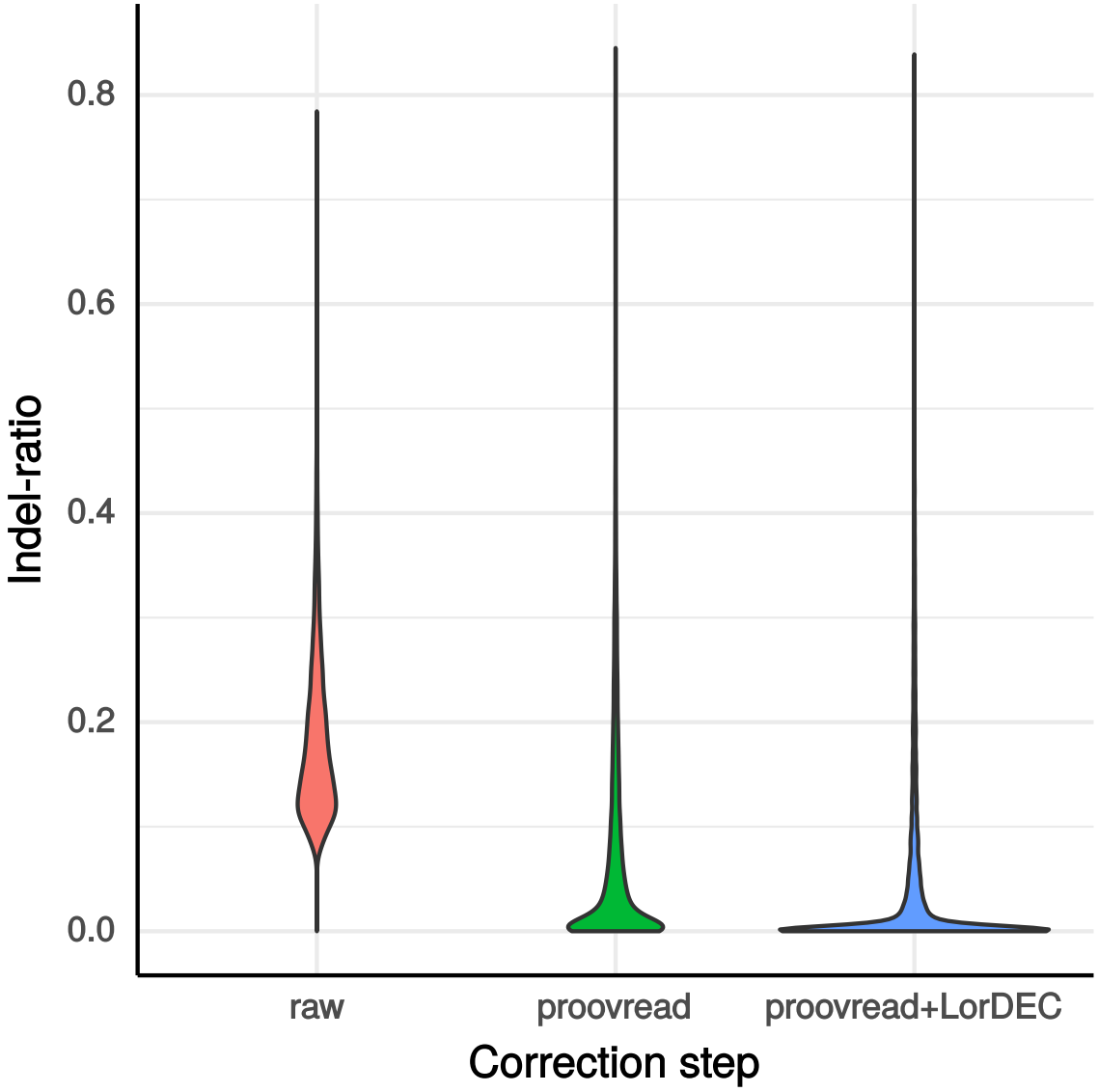

Figure 2.10: Distribution of the indel-ratios (number of indels per matched base) of the raw (red) and error-corrected PacBio reads (green: indel-ratio after first correction step, blue: final indel-ratio).

2.9.1.3 Assembly quality metrics

To identify the best of the eight different assemblies, 80 different metrics were computed, where each metric falls in one of six classes. The assemblies were ranked for each metric and the mean rank over all metrics in each class was computed. For the

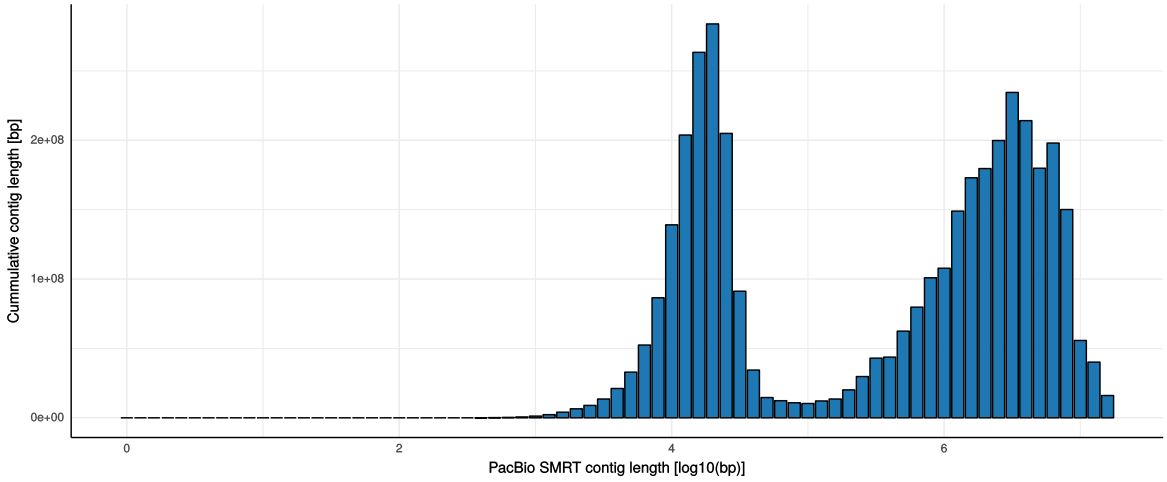

Figure 2.11: Weighted contig length histogram of the PacBio SMRT metassembly shows a clear bimodal distribution (logarithmic length of the contigs is on the x-axis, sum of the length of all contigs per bin on the y-axis). The total size of the second mode is close to the estimated genome size. About 73% of the second mode contigs (about 66% in size) can be completely aligned to the contigs from the first mode.

Table 2.10: Assembly metrics of the complete PacBio SMRT metassembly and the contigs larger than 100 kbp or smaller than 100 kbp.

	PacBio SMRT metassembly	**Mode-2 (contigs $\geq$ 100 kbp)**	**Mode-1 (contigs $<$ 100 kbp)**
Contigs [#]	110,954	1,659	109,295
Length [Gbp]	3.8	2.31	1.49
Min. length [bp]	290	100,560	290
Max. length [Mbp]	16.08	16.08	0.09
Mean length [kbp]	34.28	1,394.69	13.63
Med. length [bp]	12,460	693,156	12,285
N50 length [kbp]	995.27	2,906.73	18.52
N50 [#]	655	223	28,878
N90 length [kbp]	12.76	623.9	8.07
N90 [#]	54,054	884	74,583

final decision, the mean rank of all class-ranks was computed. The assembly with the smallest overall rank was selected as the best overall assembly. Results for each metric in each class can be found in Appendix B.

Class: contig/scaffold numbers. Common assembly statistics, such as N50, L50,

longest contig/scaffold, number of contigs/scaffolds and percentage of gaps in the scaffolds, are used as metrics (Table B.1).

Class: sequence content. For each pair of assemblies, one assembly is mapped to the other assembly using NUCmer [65]. The total number of bases in the first assembly that are not covered by a least one contig of the second assembly is used as a metric of the second assembly. This metric counts the sequence that is missing in the assembly compared to another CH assembly (Table B.2).

Class: feature content. One method to estimate the completeness of an assembly is to count the number of expected features on the genome. Two tools are available that identify the positions of protein-coding genes from a predefined database. The first tool, CEGMA [106], searches 248 core eukaryotic genes on the genome and reports the number of complete and fragmented genes. These number are used as a metric. A second tool, BUSCO [107], uses a similar method. Instead of 248 core genes that could be used for all eukaryotes, BUSCO has a list of genes for different taxonomic levels. The complete and fragmented numbers of the eukaryotic, metazoan and vertebrata were used as metrics. Additionally to these tools, a list of 22,387 mouse coding sequences (CDS) were mapped to the different assemblies using GMAP [108] in chimeric mode. A CDS was classified as "complete" if the coverage reported by GMAP was greater than 95% and the identity was greater than 75%. If the coverage was below 25% or the identity was below 75%, the CDS was classified as "missing". If GMAP reported more than one location for a CDS, it was classified as "chimeric". The rest of the CDSs were classified as "fragmented" (Table B.3).

Class: chromosome sorted read coverage. The chromosome sorted libraries are mapped to the assemblies using smalt [109]. The mean coverage of all chromosome libraries is computed for 1 kbp regions. The mean coverage of each library is computed by finding the peak in the density plot of all 1 kbp regions. This coverage is used to normalize the chromosome coverages for each 1 kbp region. Each region is then classified into 5 classes bases on the normalized coverages. The first class "low coverage" was applied if the normalized coverage of all chromosomes was below 0.5.

The second class "high coverage" was applied, if at least one coverage was above 2. If exactly one normalized coverage was between 0.5 and 2 the 1 kbp region was classified as "normal coverage", and if two or more coverages were between 0.5 and 2, the class "ambiguous" was applied. For each scaffold, the "normal coverage" 1 kbp regions from each chromosome were counted. The chromosome with the most 1 kbp regions was used as the correct chromosome for this scaffold, while all other "normal coverage" 1 kbp regions were re-classified as "false chromosome" (Table B.4).

Class: remap statistics. The mappings from the previous class (chromosome sorted read coverage) were also used to compute the percentage of reads that could be mapped back to the scaffolds. All five whole-genome paired-end libraries with an insert-size of 180 were also included. The libraries were mapped with the same tool and parameter as the chromosome separated libraries. This metric can be used to estimate the completeness of an assembly. The numbers were computed for each library separately (Table B.5).

Class: CE-statistics. To identify potential erroneous regions based on CE-statistics and coverage, the tool REAPR [110] was run with pooled mate-pair libraries that showed a similar insert-size distribution with a mean of about 5.5 kbp. REAPR identifies four kinds of regions, false coverage distribution (FCD) error, FCD error over a gap, low coverage, and low coverage over a gap. The false coverage distribution (FCD) error is the difference between the observed coverage of correctly paired mate-pairs to the expected coverage. If this number is below a threshold, the region is reported. If the overall coverage is too low, the region is reported as "low coverage". The number and total bases of the reported regions were used as metrics (Table B.6).

2.9.2 Comparison to the mouse and rat genomes

Repeat-masked and unmasked *Mus musculus* and *Rattus norvegicus* genomes were downloaded from Ensembl (Release 86). The repeat-masked PICR scaffolds larger than 1 Mbp were mapped to the chromosome sequences of mouse and rat using NUCmer [65] with –maxmatch option. The delta-filter -1 tool was applied to the

mapping results to get the best one-to-one matching of the scaffolds. The results were converted to the DAGchainer input-format, and DAGchainer [111] was applied to compute chains of consecutive mapping. The unmasked genomes were split at each 'N'-stretch larger or equal to 100 bp to construct contig sequences for the contig length comparison.

2.9.2.1 Contig sizes

Although the *Mus musculus* and *Rattus norvegicus* genomes are available at the chromosome level, they differ at contig level. The mouse assembly consists of 445 contigs (continuous sequences with N-stretches less than 100 bp), whereas the rat chromosomes are split into 32,025 contigs. The PICR metassembly shows a continuity between that of the mouse and the rat assemblies, with 4,517 contigs (following splitting contigs at gaps $>$ 100 bp to enable consistent comparison among the assemblies). The same pattern can be seen with other metrics as shown in Table 2.11 and in the weighted histogram of contigs length in Figure 2.7b.

Table 2.11: Contig size metrics for the PICR assembly compared to Ensembl mouse and rat chromosome contigs and to the 2013 RefSeq CH contigs. For all assemblies, chromosomes and scaffolds were split at all 'N'-stretches at least 100 bp long to enable consistent comparison between assemblies.

	PICR	**Mouse**	**Rat**	**2013 RefSeq**
Contigs [#]	4,517	445	32,025	117,912
Length [Gbp]	2.37	2.65	2.65	2,31
Min length [bp]	104	562	84	201
Max length [Mbp]	14.60	91.93	2.18	0.76
Mean length [kbp]	523.81	5,949.87	82.85	19.55
Median length [bp]	52,173	26,251	36,272	1,134
N50 length [kbp]	2,446.66	32,813.18	200.97	84.26
N50 [#]	270	25	3,665	8,297
N90 length [kbp]	418.89	8,381.66	44.49	20.86
N90 [#]	1,148	82	14,494	28,528

2.9.2.2 CH PICR to mouse and rat alignments

The repeat-masked PICR scaffolds larger than 1 Mbp were mapped to the repeat-masked mouse and rat chromosomes with NUCmer and the best one-to-one alignment was computed with delta-filter. The alignments were chained together using DAGchainer. The alignment to the mouse chromosomes contained 295,367 single alignments with a mean length of 592.90 bp and a mean identity of 85.39% covering 175.1 Mbp (6.65%) of the mouse genome. The chaining created 465 chains with a mean length of 5.0 Mbp, covering 2.33 Gbp (88.74%). The alignment to the rat chromosomes contained 277,685 single alignments with a mean length of 584.40 bp and a mean identity of 85.41% covering 162.2 Mbp (5.84%) of the rat genome. The chaining created 1,196 chains with a mean length of 2.0 Mbp, covering 2.41 Gbp (86.91%). Detailed visualizations are shown in Figure 2.12 for the mouse alignment and Figure 2.13 for the rat alignment.

2.9.2.3 Blasting of mouse sequences for UTR and non-coding assembly analysis

The mitochondria mediates a variety of metabolic processes that impact maintenance of cellular physiology and homeostasis. Around 1,538 nuclear genes (at 1,654 unique locations) associated with energy metabolism were extracted from the *Mus musculus* assembly (GRCm38.p5) and checked for homology within CH genome assemblies: CSA Cgr1.0 [29], 2013 RefSeq C_griseus_v1.0 (criGri1) [27] and PICR, using NCBI-BLAST-2.6.0+ [112]. These genes comprise of nuclear genes linked to the mitochondrial Gene Ontology, components of the OXPHOS pathway, and genes in the MitoCarta [113] and QIAGEN (Mouse Mitochondrial Energy Metabolism PCR Array – Version 4.0) lists for mitochondrial energy metabolism.

The analysis was done in three steps to estimate the existence of essential genomic sequences in the three CH assemblies. Primarily keeping the mouse genome as reference, complete genic sequences from TSS-TES were analyzed followed by considering sequences without mouse specific UTRs (start of the first CDS to the end of the

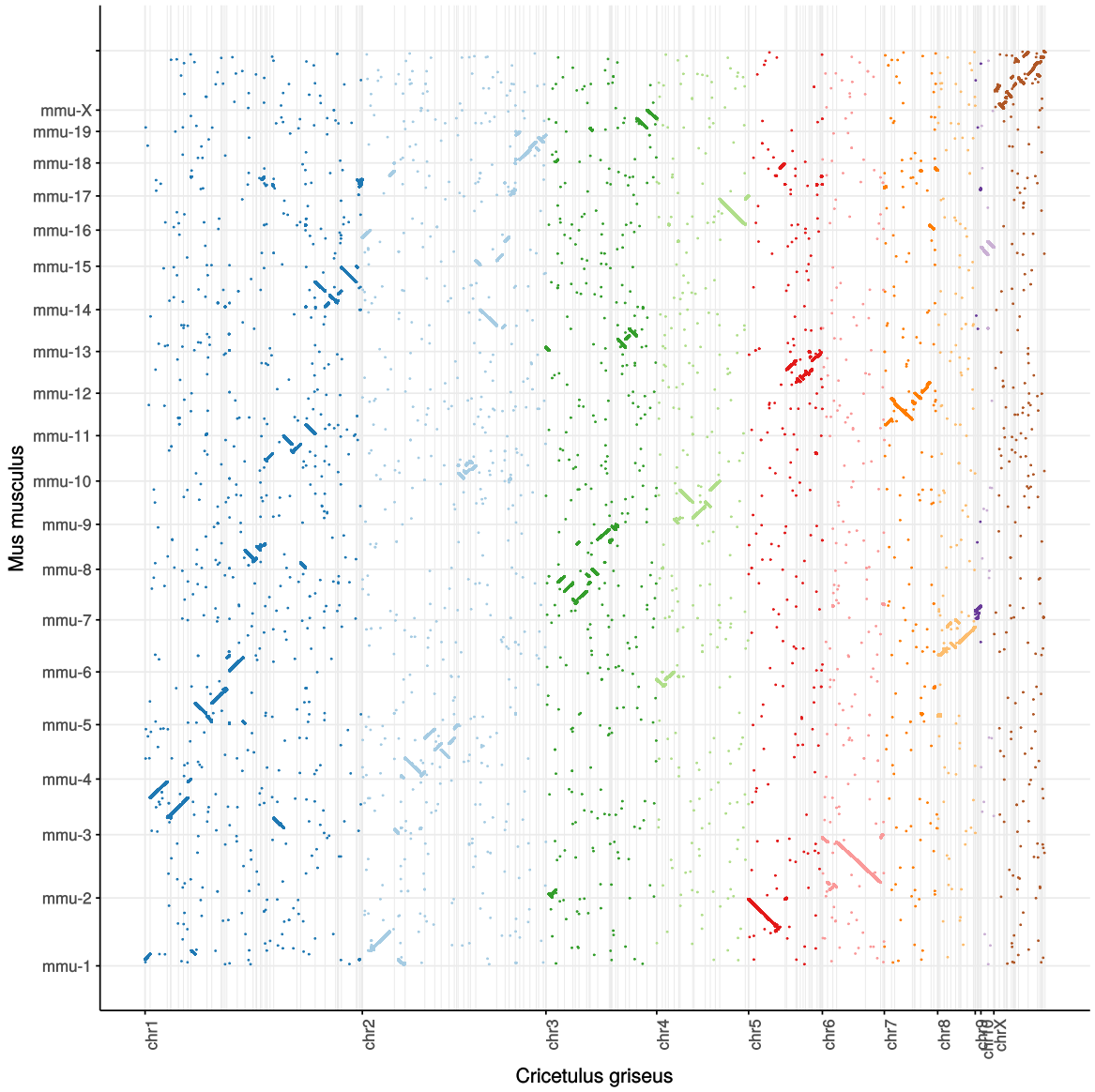

Figure 2.12: NUCmer alignments of the PICR scaffolds to the mouse chromosomes. The PICR scaffolds are ordered by chromosome and position of the longest alignment to the mouse genome.

last CDS). The genes reporting homologous regions with the presence of first and last position of the sequence on the same scaffold in respective assemblies were counted using custom scripts by parsing tabular output of BLASTn reports. For each gene, the hit with first position was flagged as "Start", last position as "Stop" (split genes)

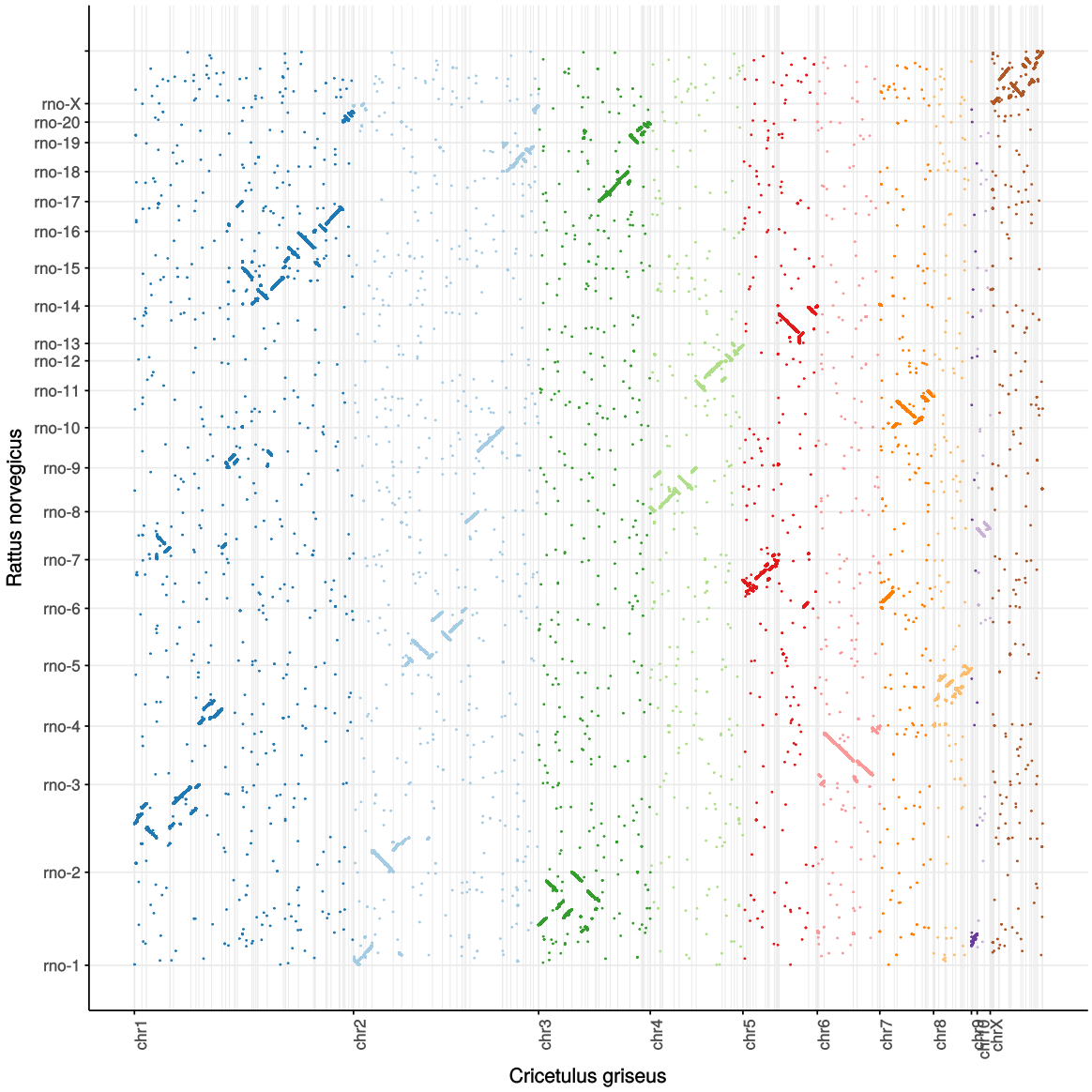

Figure 2.13: NUCmer alignments of the PICR scaffolds to the rat chromosomes. The PICR scaffolds are ordered by chromosome and position of the longest alignment to the mouse genome.

and if both start and end are on the same hit (complete intact gene), it was flagged as "Both". Keeping the scaffold information along with the flags, for all the genes each of the hit with "Start" flag was checked for the scaffold information in "Stop" flagged hits. If a match was found or if the gene had a hit flagged "Both", the gene along with

its coordinates was noted. All such identified genic locations were then compared for different assemblies and venn diagrams were plotted (Figure 2.8).

As the most genic locations were found complete (split or in a single stretch but on the same scaffold) in PICR, it was chosen for further analysis. These 1,011 genic locations correspond to 948 unique genes in the mouse annotation. 858 of those could be found in the PICR annotation at 900 genic locations. Nucleotide sequences were extracted for these 900 locations extending the gene to include the regulatory regions (5 kbp upstream TSS to 1.5 kbp downstream TES). These sequences were then analyzed for sequence homology with the other two CH assemblies to find complete genes along with correctly assembled regulatory regions. For details, see the online supplementary materials for the manuscript (www.ncbi.nlm.nih.gov/pmc/articles/PMC6045439).

The six histone modification marks generated earlier, for PF-MCB CHO-K1 cells sampled twice a day until 9 days, were aligned to the PICR assembly. Based on those alignments an 11-state model was trained as described before [31] and chromatin states were annotated by comparing the emission profiles, displaying enrichment of each histone mark in a state, to the 18-state model deduced from human epigenomes [89]. Genes annotated on the scaffold margins, with plausibly truncated regulatory information, were identified from the CSA annotation [29, 31] and corresponding pattern of chromatin states was observed in both CSA and PICR genomes.

2.9.3 NCBI 'light' annotation

The PICR, IPCR, and the 2013 RefSeq CH assemblies were sent to NCBI to undergo a light version of the NCBI annotation pipeline and to gain further insight into the quality of PICR and IPCR regarding gene content. Evidence used in this pipeline included approximately 500 million RNA sequencing reads from CHO cells (SRP066355 [bioproject PRJNA302601] and SRP073484 [bioproject PRJNA318886]). The transcripts and proteins from GenBank and RefSeq for the Chinese hamster, as well as known RefSeq proteins from human and mouse, were used as evidence.

After the evidence was aligned to the assembly sequences, Gnomon [114] merged overlapping alignments into precursor models. Next, a Hidden Markov Model (HMM) was used to develop *ab initio* gene models and to extend models that were missing a stop or start codon. Gnomon was then run a second time to incorporate alignments of the precursor models to a set of the NCBI nr database. Last of all, some of the resulting gene models were corrected if the evidence from the pipeline strongly suggested that the assembly is wrong.

In addition, external to the pipeline, 30,782 mouse RefSeq transcripts were aligned to PICR, IPCR, and 2013 RefSeq. The best sequence alignments were used to determine the number of transcripts that aligned ($\geq$ 75% identity) to one scaffold, the number split over two or more scaffolds, the number with no alignments above $\geq$ 75% identity, and the number with no alignments that cover $\geq$ 95% of the coding sequences (CDS).

Gene and transcript metrics from the "light" version of the NCBI pipeline are shown in Table 2.12. The PICR and IPCR assemblies have fewer partial genes, corrected coding regions, and genes with premature stops, suggesting that these two assemblies are of higher quality than the 2013 RefSeq assembly. Results of the alignment of mouse transcripts from RefSeq are shown in Table 2.13. PICR has more aligned mouse transcripts than IPCR which suggests a more complete assembly regarding gene content. While PICR and 2013 RefSeq have a similar number of aligned transcripts, PICR has approximately 3-fold fewer split alignments.

2.9.4 GO term analysis of genes with filled gaps

Gap sequences were identified by aligning flanking sequence from the 2013 RefSeq genome to the PICR assembly (Figure 2.14). To identify which biological processes were over-represented among the genes with filled gaps, we performed functional GO term analysis of the 2,252 genes using DAVID [82, 83]. The top GO terms are enriched in protein binding, RNA binding, and transcription molecular functions (Figure 2.15). Gene functional classification results show that these genes are enriched in zinc finger

Table 2.12: NCBI annotation of the 2013 RefSeq (C_griseus_v1.0), PICR, and IPCR assemblies. *The larger number of CDSs for the 2013 RefSeq assembly is a result of the pipeline rules for making alternative variants, which allow more variants for the assembly that NCBI considers the reference (currently 2013 RefSeq). Because NCBI does not create alternative variants for genes that are corrected, this rule does not impact the direct comparison of the corrected CDSs.

	Assembly		
Gene prediction	**2013 RefSeq**	**PICR**	**IPCR**
Total genes	27,982	26,931	25,802
Protein-goding	20,678	20,668	20,074
Partial	2,576	526	516
Total CDS	32,329*	27,205	25,884
Corrected CDS	2,039	963	1,021
have premature stops	747	332	521
have frameshifts	1,583	752	670
Transcripts with no support	105	217	173
Transcript with partial support	5,710	4,192	4,763

Table 2.13: NCBI alignment of mouse coding transcripts from RefSeq (NM_prefix) to the 2013 RefSeq, PICR, and IPCR CH genome assemblies.

	Assembly		
Mouse transcript alignments	**2013 RefSeq**	**PICR**	**IPCR**
Number of coding transcripts	30,782	30,782	30,782
Aligned	29,146	29,131	28,832
Unaligned	1,636	1,651	1,948
Split alignment	1,479	424	486
Corrected CDS	2,039	963	1,021
<95% CDS coverage	10,445	5,342	6,655

and ribosomal genes. These classes of genes were likely to have filled gaps because they often have highly homologous sequence across the gene families, thus leading to difficulties in resolving their sequence in assemblies based on short reads. In addition, some of the genes locate to repetitive regions in the genome. We further explored which classes of genes had gaps with mutations in several representative resequenced genomes. The top GO terms for these 225 genes were also enriched in DNA binding

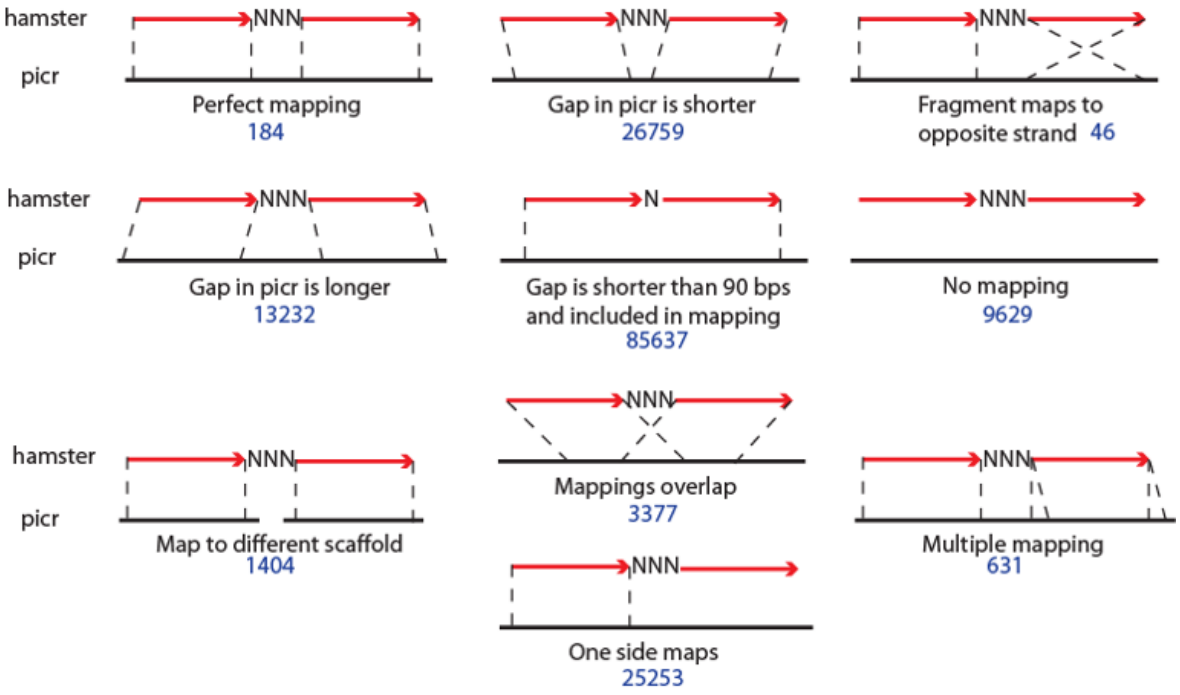

Figure 2.14: Strategy for 2013 RefSeq and PICR gap comparison. Matches around and including gaps were used to identify the corresponding region in the PICR genome. Numbers in the figure represent the amount of the specific gaps we identified.

and transcription (Figure 2.16). In summary, the gaps in the previous assembly could potentially confound genomic studies in CHO, especially if variants are associated with genes involved in DNA or RNA binding, including transcription factors.

2.9.5 PIRC to PICR whole genome alignment

Based on the whole genome alignment of PICR and PIRC using NUCmer [65] and the analysis with dnadiff, 16,305 SNPs and indels and 4,909 structural differences were found. Overlapping structural differences were merged resulting in 4,130 regions for further analysis. These regions with the corresponding regions in the PIRC were further analyzed by identifying gaps and possible errors (based on REAPR and read-coverage analysis). We identified:

1. 17 regions with error in the PICR assembly and correct sequence in the PIRC assembly

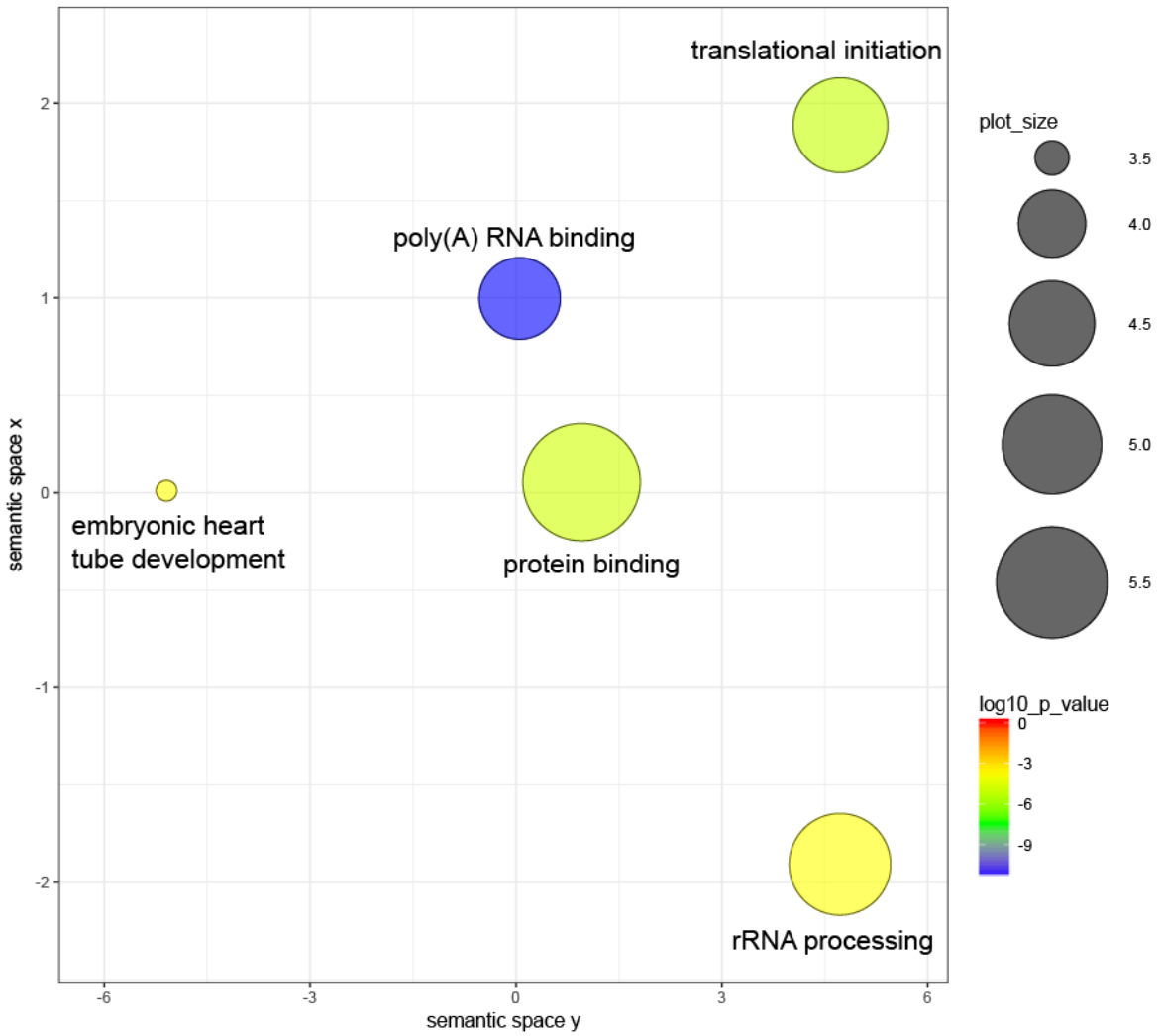

Figure 2.15: GO term analysis of genes with coding gaps. Enriched GO terms were identified using DAVID. 2,252 genes with coding gaps were searched against the whole human GO term sets. The top 5 GO terms with FDR corrected p-value smaller than 0.05 are visualized using REVIGO. Circle size represents the relative gene set size of each GO term compared to the whole human gene sets. This analysis indicates genes related to transcription or translational regulation may be difficult to fully assemble only using Illumina reads, as some classes of transcription factors and other oligonucleotide binding proteins have highly conserved or repetitive sequences (e.g. zinc finger proteins).

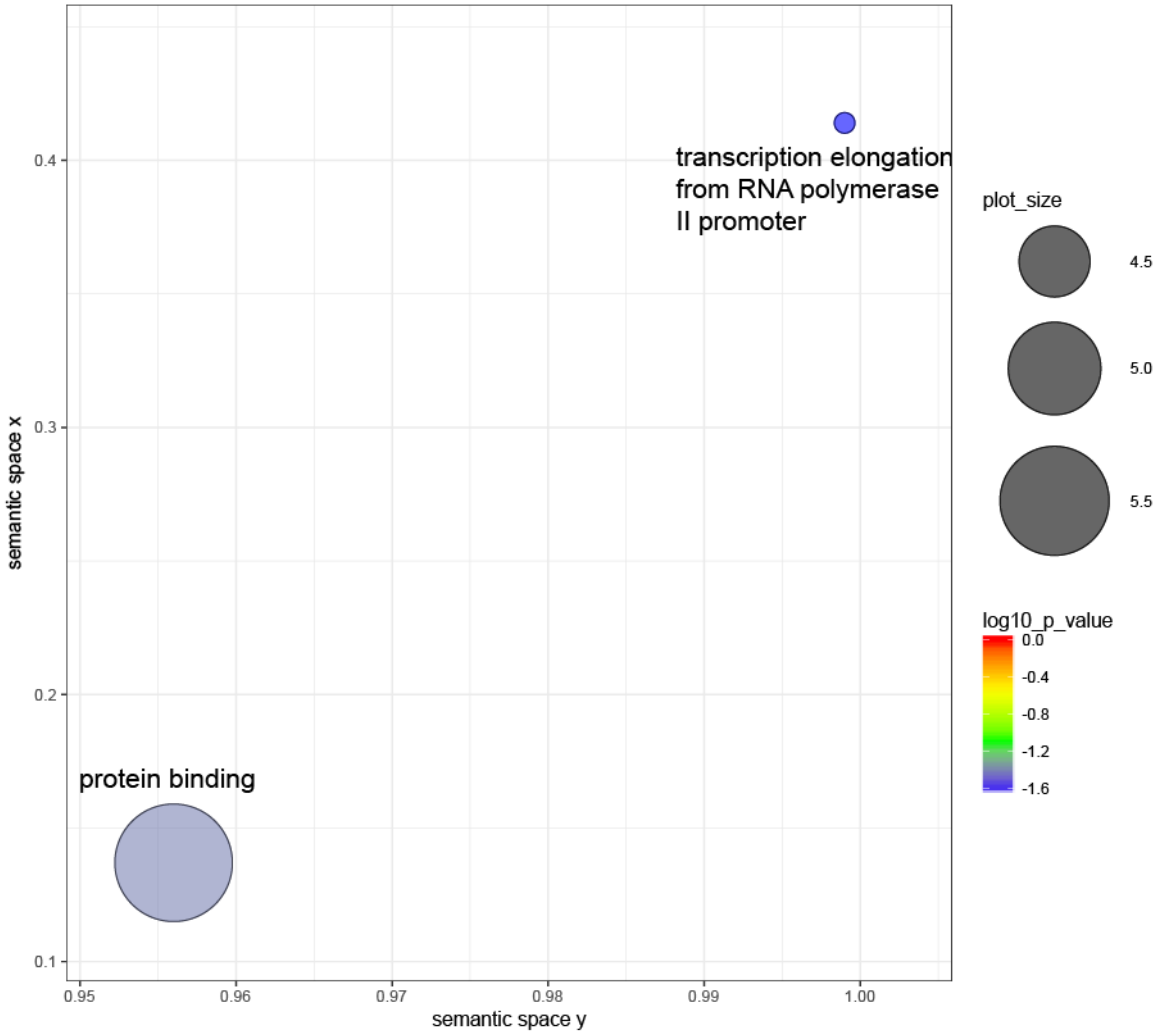

Figure 2.16: GO term analysis of genes with mutations in their coding gaps. Enriched GO terms were identified using DAVID. 134 genes with variants in their coding gaps were searched against all of the 2,252 genes with coding gaps. GO terms with an FDR-corrected p-value smaller than 0.05 were shown in the figure using REVIGO. Circle size represents the relative gene set size of each GO term compared to the whole human gene sets. This analysis indicates genes related to transcription regulation tend to have more mutations than the other genes with coding gaps.

2. 59 regions with gaps in the PICR assembly and correct sequence in the PIRC assembly
3. 388 regions with gaps in the PICR assembly and error in the PIRC assembly
4. 2719 regions with gaps in the both assemblies
5. 122 regions with correct sequence in the PICR assembly
6. 165 regions in the PICR assembly without unique corresponding PIRC regions
7. 665 regions with error in the PICR assembly and errors or gaps in the PIRC assembly

Regions 1) and 2) are possible candidates for error correction in the PICR assembly. For 3), gaps in the PICR could be closed but possible errors could be introduced. The corresponding PIRC regions were identified by aligning the bases 300 bp upstream and downstream of the PICR regions to the PIRC scaffolds. Both upstream and downstream regions needed to match uniquely to the scaffolds to identify the regions.

Additionally, four candidates for scaffolding were found (based on manual inspection of the alignment). The possible scaffolding order is shown below. The chromosome of the scaffolds is shown in brackets. The PIRC scaffold that joins the PICR scaffolds is shown in parentheses.

```
picr_1193[unplaced] + picr_241[unplaced]    (pirc_235)
picr_121[X] + reverse(picr_25[X])           (pirc_16)
reverse(picr_1260[unplaced]) + picr_153[8]  (pirc_154)
reverse(picr_653[unplaced]) + picr_167[9]   (pirc_168)
```

A detailed list of the PICR to PIRC alignment with annotated assembly gaps and errors is included in the online supplementary material for the manuscript (www.ncbi.nlm.nih.gov/pmc/articles/PMC6045439).

Chapter 3

EVALDNA: A MACHINE LEARNING-BASED TOOL FOR THE COMPREHENSIVE EVALUATION OF MAMMALIAN GENOME ASSEMBLY QUALITY

3.1 Preface

This section is adapted from MacDonald and Lee, 2019 (currently in submission to Nucleic Acids Research). In this chapter, we describe the design and development of a novel computational pipeline that evaluates genome assembly quality. The pipeline, EvalDNA, can assess the accuracy of an assembly without requiring a reference genome and also produces results that are comparable across different species, setting it apart from existing assembly evaluation tools. I developed the pipeline and the model, and completed subsequent testing on Chinese hamster genome assemblies and human chromosome 14 assemblies. This project was completed under Dr. Kelvin Lee's guidance.

3.2 Abstract

We present a novel tool, called EvalDNA (Evaluation of *De Novo* Assemblies), which assists in the model development for quality scoring of genome assemblies and does not require an existing reference genome for accuracy assessment. EvalDNA calculates a list of quality metrics from an assembled sequence and applies a model created from supervised machine-learning methods to integrate the various metrics into a comprehensive quality score. A well-tested, accurate model for scoring mammalian genome sequences is provided as part of EvalDNA. This random forest regression model evaluates an assembled sequence based on continuity, completeness, and accuracy, and was able to explain 86% of the variation in reference-based quality scores within the testing data.

EvalDNA with this mammalian model was applied to human chromosome 14 assemblies from the GAGE study to rank genome assemblers and to compare EvalDNA to two other quality evaluation tools. In addition, EvalDNA was used to evaluate several genome assemblies of the Chinese hamster (CH) genome to help establish a better reference genome for the biopharmaceutical manufacturing community. EvalDNA scores also enabled the quality comparison of the selected CH reference genome to the reference assemblies of other organisms at both the full assembly and chromosome levels.

3.3 Introduction

Whole genome assemblies are becoming available for an increasing number of organisms due to the reduced time and monetary costs of DNA sequencing. There has been more than a 3-fold increase in the number of assemblies in NCBI's RefSeq database since August of 2015 [1] (Table 1.1). As of February 2019, there was a total 153,355 assemblies in NCBI RefSeq, consisting of 53,048 unique species. Multiple assemblies are often created for the same species by using different sequencing and/or assembly methods. However, a single genome assembly is typically selected as a reference genome to guide wet-lab and bioinformatics studies. To select the most complete, continuous, and accurate assembly for an organism of interest, comprehensive quality assessment of assemblies is necessary. Researchers should also be aware of any limitations posed by the level of completeness, continuity, and accuracy of their selected reference assembly.

Genome quality is usually assessed by metrics such as gap percent, N50, and the number of scaffolds that make up the assembly. However, these metrics only reflect the completeness and continuity of an assembly, and not the accuracy. For example, the best assembly is often considered the one with the highest N50, but the N50 metric increases even when contigs are joined incorrectly [44, 41].

One way to evaluate the accuracy of an assembly is to compare it to an existing reference assembly for the organism of interest through a direct sequence comparison. The assembly evaluation tools QUAST [42] and CQAT (Contig Quality Assessment Tool) [43] use this method. However, many *de novo* assemblies, those built without

the use of a reference, do not have a suitable assembly available for comparison. In addition, the quality of the alignment between genome assemblies should be assessed, using tools such as ThurGood [115], as errors in alignment can impact the assembly quality assessment.

To overcome this issue, several methods for quality evaluation that do not require an existing reference assembly have been developed. These methods include gene homology methods such as those executed by CEGMA (Core Eukaryotic Genes Mapping Approach) [116] and the more recent BUSCO (Benchmarking Universal Single-Copy Orthologs) [107] programs. The results of these tools reflect the completeness and accuracy of a genome based on expected gene content. However, they only examine the accuracy of well-conserved genes and their copy numbers, rather than the whole genome.

The majority of other reference-independent quality assessment tools use information from mapping sequencing reads back to the genome of interest. Low mapping quality or read coverage can indicate errors in the assembly. Tools using this approach include Amosvalidate [44], ALE [45], FRCbam [46], SURankCO [47], and REAPR [117]. Amosvalidate was the first automated pipeline for misassembly detection that used read mapping information. However, the pipeline, designed in 2008, uses an older assembly format that is not produced by current assemblers. ALE (Assembly Likelihood Estimator) uses Bayesian statistics to determine the probability of an assembly being correct given a set of reads. The resulting ALE score can be used to compare different assemblies of the same genome, but the authors state that ALE should not be used to compare assemblies across organisms [45]. FRCbam provides a feature response curve for an assembly instead of a numeric score. The curve shows the trade-off between the accuracy and the continuity of the assembly. Similar to ALE, FRCbam can only be used to compare different assemblies of the same organism. SuRankCo uses supervised machine learning where the training data includes metrics from read mapping to rank, rather than score, scaffolds/contigs within a single assembly. REAPR examines the quality of an assembly base-by-base and provides multiple quality metrics

derived from read mapping.

Despite the development of these important tools, there is still need for a reference-independent tool that provides a single quality score reflecting the completeness, continuity, and accuracy of an assembly and can be used to compare assemblies from different organisms. Here, we present a novel pipeline called EvalDNA (Evaluation of *De Novo* Assemblies) to address this need. EvalDNA assists in the modeling of genome quality through supervised machine learning, and uses the subsequent model to estimate a single, comprehensive quality score for a given assembled sequence. The quality score being learned is based on the number of differences in an alignment between a training sequence and its reference, calculated using DNAdiff [65].

EvalDNA calculates completeness and continuity metrics, and uses output from SAMtools [118] and REAPR [117] to generate accuracy metrics. A user-specified model, developed from supervised machine learning, is then used to estimate the quality score using a subset of these metrics. We developed and tested a model for scoring mammalian assemblies which is provided as part of EvalDNA. The resulting scores from EvalDNA can be used to directly compare chromosome sequences within a single assembly, compare multiple genome assemblies from the same organism, and even compare assemblies from different organisms as long as each assembled sequence is scored using the same model.

EvalDNA was applied to human chromosome 14 assemblies from the GAGE study [119] to rank genome assemblers and to compare EvalDNA to two other reference-independent quality evaluation tools, ALE and FRCbam. In addition, EvalDNA was run on several existing Chinese hamster (CH) genome assemblies to compare its results to that of a manual ranking of the assemblies described in Rupp et al. [30] as well as rankings from ALE and FRCbam. This comparison provided insight regarding the performance of EvalDNA on organisms that were not used in the training data and confirmed that EvalDNA can be used to select the highest quality assembly. Scores for each chromosome from the 2018 CH PICR reference genome were also estimated using EvalDNA and compared to chromosomes from the previous CH reference assembly and

the reference assemblies for human, mouse, rat, and cow.

Finally, error simulation of PICR chromosomes and scaffolds was done to examine how the EvalDNA score changes as the amount of errors within an assembled sequence increases and to assess EvalDNA's potential to score scaffolds. The mammalian model's potential to score plant genomes was also briefly examined by applying EvalDNA to several versions of the rice genome assembly.

3.4 Methods

3.4.1 Overview of the EvalDNA tool

Users have two options when using EvalDNA. If the sequence of interest is from a mammalian genome, the user can use EvalDNA with the provided mammalian assembly quality scoring model. This option would require the user to run the EvalDNA metric calculation pipeline to collect quality metrics for the sequence of interest and then provide the resulting list of metrics to the 'run model' script to get the final quality score. Here, we focus mainly on this type of usage.

The second option is to create a new scoring model based on a set of assembled sequences each with a reference sequence. These assembled sequences could be derived from organisms with a high quality reference genome that are related to the organism of interest. A script is provided to align each of the training sequences to their corresponding reference sequence to get the target quality score. However, scaling of the scores may still be required (see the 'Reference-based quality scoring' section). A model would need to be trained on this data and finalized in R, and then loaded into the 'run model' script. More on this second type of usage can be found in the EvalDNA documentation.

The EvalDNA metric calculation pipeline is written in Python. General steps are shown in Figure 3.1. The pipeline requires a configuration file, the sequence(s) of interest in FASTA format, and either a set of paired-end DNA sequencing reads in FASTQ format or a BAM file containing the reads mapped to the sequence(s) of interest. If the raw reads are provided, EvalDNA will run SMALTmap [120], which is

the recommended read mapper for REAPR, to map the reads to the provided sequence, creating the BAM file. If the BAM file is provided, EvalDNA can skip the SMALTmap step.

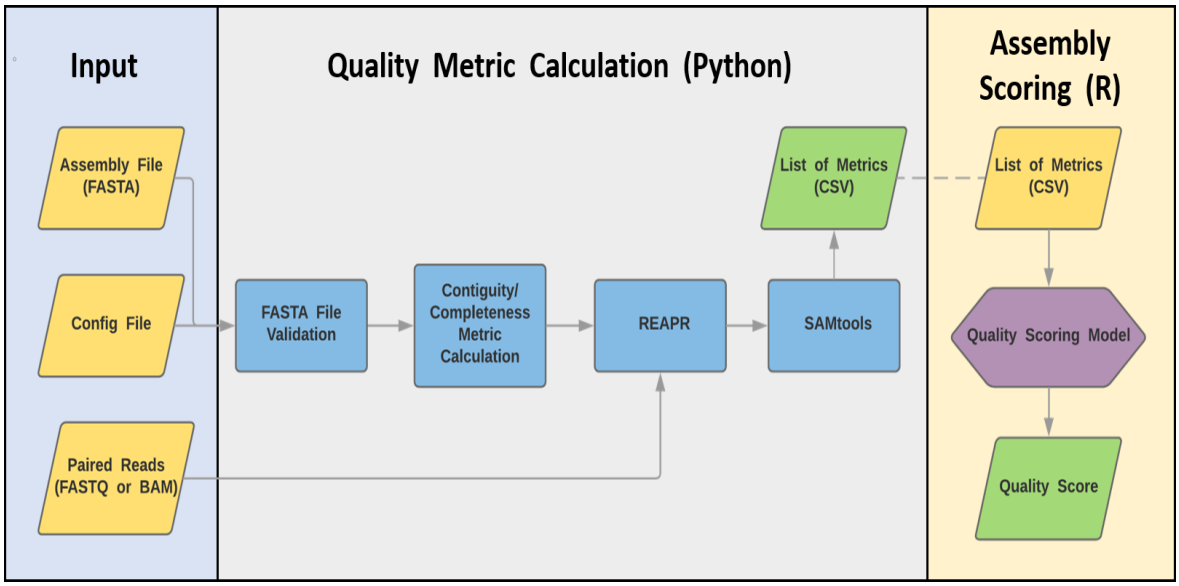

Figure 3.1: The computational workflow of EvalDNA. EvalDNA requires the assembly of interest in FASTA format, a configuration file, and Illumina paired read data in either FASTQ or BAM file format. EvalDNA first calculates contiguity and completeness metrics, and then calculates accuracy metrics based on the output from running REAPR and SAMtools. This part of EvalDNA produces a list of metrics that will be given to the scoring model (written in R) which will estimate the overall quality score for the assembly.

The pipeline calculates the metrics that are used in the mammalian models as well as additional metrics which are still insightful. The selection of metrics used in the model is described in the subsection 'Feature Selection'. The pipeline first calculates a set of commonly used completeness/contiguity metrics, including percent gaps, N50, and the number of scaffolds/contigs. It then executes REAPR [117], followed by SAMtools stats [118] to calculate various metrics reflecting the accuracy of the given sequence(s). Metrics used in model development are normalized by chromosome length.

3.4.2 Training Data

Training data, collected from rat, mouse, and human assembly builds, were used to develop a supervised machine learning model that can estimate reference-based scores for genome assembly quality. Each training instance consists of a set of quality metrics for a single chromosome, $x_i = [x_1, x_2, ...x_d]$ (see Quality Metrics section), and its corresponding quality score, y_i (see Quality Scoring section). Chromosomes from publicly available assemblies and chromosomes with simulated errors were used, totaling 416 training instances. The training data included 276 chromosomes taken directly from publicly available assembly builds, 140 chromosomes with simulated errors, and 17 chromosomes with simulated gaps. The feature set before feature selection consists of 13 assembly quality metrics. Definitions of the metrics in the training data can be found in Appendix B.

3.4.2.1 Assembly versions

Chromosomes from the current and previous builds of the rat (Rnor6) and mouse (GRCm38) reference genomes were downloaded from corresponding organisms' directory on ftp://ftp.ncbi.nlm.nih.gov/genomes/. For instance, mouse chromosome 1 from build37.2 was downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/M_musculus/ARCHIVE/BUILD.37.2/CHR_01/. The FASTA files for each chromosome contained both assembled scaffolds and unplaced scaffolds. Chromosomes from the current human reference genomes (GRCh38) and two assembly builds of another human genome (NA19240) were also used. The NA19240 assembly builds were selected as a training data source because they were built from sequencing reads from a single person. Therefore, differences between reads sequenced from that person's DNA and the assembly are most likely due to errors in the assembly, rather than true differences among individuals. Deciphering what a sequence difference means in an assembly built from a pool of individuals (such as GRCh38) would be more difficult. Assembly build information can be found in Supplementary Materials section.

3.4.2.2 Simulated chromosomes

SINCsimulator [121] was run on a subset of the chromosomes described above to generate errors in the existing chromosomes. Differing levels of single nucleotide polymorphisms (SNPs), insertions/deletions (indels) and copy number variants were provided to generate chromosomes with differing levels of quality. This resulted in 123 chromosomes with simulated errors. A custom script was used to simulate gaps in 17 chromosomes as well. Both of these steps were done to ensure the model was trained on chromosomes of lower quality than ones that would be submitted to NCBI RefSeq or GenBank.

3.4.2.3 Quality metrics

Quality metrics for each training chromosome were calculated using the metric portion of the EvalDNA pipeline. Basic metrics reflecting the completeness and continuity of the chromosome assembly, which include gap percent, N50, N90, scaffold/contig number, and average scaffold length, were collected first.

Several external programs were then run to collect metrics reflecting the accuracy of the assembly. SMALTmap within REAPR maps user-provided reads to the assembly of interest. REAPR then scans the assembly base-by-base identifying possible errors based on the alignment file. The number of bases in each error type, such as bases in clipped reads or low coverage regions, is converted into a percent of the total number of bases to normalize by assembly/chromosome length. Finally, SAMtools is used to calculate the number of read pairs that aligned to the assembly in the expected orientation and distance from one another. These proper read pairs were divided by the number of reads mapped to the assembly to create a proper pair percent metric.

Further details on the complete set of metrics and any corresponding normalization can be found in the Supplementary Materials.

3.4.2.4 Sequencing reads

As described in the previous section, several of the quality metrics in the feature set are derived from mapping sequencing reads to each chromosome sequence. 20.5 giga base pairs (Gbp) of Illumina paired-end read sequencing data from ERR319183, ERR316497, ERR316496, and ERR319170 (Bioproject PRJEB2922) was used as input for the metric calculation portion of EvalDNA for all rat assemblies. The insert size was consistent among these runs, ranging from 473 to 475 base pairs (bp), a requirement for REAPR. 25.7 Gbp of Illumina paired-end read sequencing data from ERR1856364 (Bioproject PRJEB19654, insert size 550 bp) [122] was used for the mouse assemblies. 20.2 Gbp of Illumina paired-end read data from the NA19240 human sequencing run SRR2103647 (Bioproject PRJNA288807, insert size 350 bp) was used for the evaluation of both GRCh38 and NA19240 assemblies.

All reads were trimmed using trim-galore with a quality score cut-off of 26. These reads were used to calculate the accuracy metrics in the training data for the mammalian models. We strongly recommend using at least 10x coverage of reads with an insert size of approximately 350-550 bp when scoring a novel assembled sequence with the mammalian model to stay consist with the amount and insert size of the reads used to create the training data.

3.4.2.5 Reference-based quality scoring

A reference-based quality score for each training instance was calculated based on the NUCmer[65] alignment to the most recent build of the corresponding chromosome. For example, each build of chromosome 1 (query) from the rat genome assembly was aligned to the Rnor6 build chromosome 1 (reference). This method works based on the assumption that more recent builds of an assembly are more accurate. This assumption is supported by the general quality metrics of the assemblies as well as the continuous improvements in DNA sequencing and assembly methods.

For each NUCmer alignment, the number of bases differing between the two sequences were found using DNAdiff. This value was used in the following equation to

get the percent of matching (or correct) bases:

$$percent\_matching\_bases = \frac{length\_of\_reference - total\_differences}{length\_of\_reference} * 100 \qquad (3.1)$$

Using this method, the self-to-self alignment for each chromosome from the most recent assembly will have a percent of matching bases of 100. However, in reality, not all the chromosomes from the most recent build of an assembly will be of identical quality and neither will chromosomes from different organism assemblies. Therefore, this 'percent of matching bases' value cannot directly be used as the score, and instead requires two rounds of scaling; one among the chromosomes in a single assembly (internal scaling) and another to scale between organisms (external scaling).

Internal scale factors were determined by calculating the distance between each chromosome and the ideal set of metrics using Euclidean distance. This ideal set of metrics are metrics that would be produced from a perfectly accurate, continuous, and complete sequence, i.e. gap percent is 0, normN50 is 100, error_free_bases is 100 etc. The best chromosome (i.e. the one with the lowest distance from the ideal chromosome metrics) for each organism kept the score of 100, while the other chromosomes 'percent of matching bases' values were scaled based on differences in the distances from the ideal metrics.

A similar process was used to determine the external scale factors. The distances between each organism's best chromosome and the ideal chromosome metrics were used to determine the best overall chromosome. This best chromosome was chromosome 2 from human NA19240 and kept its score of 100. Again, the distances were used to scale the other organisms' chromosomes to get the final quality score.

This method allows scores from across species to be compared and also provides context for the EvalDNA score. A chromosome with an EvalDNA quality score higher than 100, for instance, is predicted to be of higher quality than the chromosomes from the human reference assembly as well as chromosomes of NA19240, a more recent Illumina/PacBio hybrid assembly of the human genome.

3.4.3 Model development

A randomly selected 20% of the training data was set aside as test data. The quality metrics and the associated reference-based score of each chromosome in the remaining 80% of the training data were used to train several models that can estimate the quality score given the set of quality metrics of an assembled sequences. Model development steps included feature selection, using supervised machine learning to create regression models, and model selection based on performance measures. Regression models rather than classification models were created because the reference-based quality scores (target variable) consist of continuous rather than discrete values. Thus, we were trying to find some function f that maps the input X onto the output Y, where f can be represented by the function: $w_0 + w_1x_1 + w_2x_2 + w_3x_3... + w_dx_d$. The weight, w_i, for each quality metric, x_i, were the values being learned.

3.4.3.1 Feature selection

To determine which of the 13 quality metrics (features) calculated by EvalDNA were correlated with the target quality scores in the training data, the Pearson correlation (r) for each metric was calculated (Table 3.1). Metrics that were not correlated with quality scores in the training data ($-0.1 < r < 0.1$) were removed from the model. These included metrics based on the contig number and REAPR's values for 'fragment coverage distribution (FCD) error within contigs' and 'collapsed repeats'.

The presence of multicollinearity/redundancy among the metrics was identified by calculating the Pearson correlation value between each pair of metrics (Figure 3.2). Since multicollinearity among metrics can reduce the accuracy of a model, metrics were further filtered by calculating the joint mutual information using the 'jmim' function of the Praznik R package [123] and the percent increase in MSE from the importance function of the randomForest R package [124]. Calculation of joint mutual information among the remaining metrics showed that metrics based on REAPR's 'low read coverage' and 'FCD error over gap' shared redundant information with other metrics regarding the target score value and could be removed. The %INCMSE importance

Table 3.1: The Pearson correlation coefficients between each metric and the reference-based quality score in the training data that was used to create mammalian model.

Quality Metric	Pearson Correlation
normN50	0.570
gap_perc	-0.300
prop_pair_perc	0.204
FCD_err_in_contig	-0.099
FCD_err_over_gap	-0.295
low_fc_in_contig	-0.521
low_fc_over_gap	-0.579
links	-0.594
clip	-0.476
coll_repeat	-0.071
low_read_cov	-0.402
error_free_bases	0.701
norm_contig_number	0.025

metric from the randomForest R library [124] was also examined to see which metrics caused the smallest increase in mean squared error (MSE) when replaced by a randomly permuted variable in a random forest regression model. This method suggested that the 'proper pair percent' and 'FCD error over gap' metrics could be removed. Results of these feature selection methods are provided in the Supplementary Materials.

Subsets of the remaining features (normN50, gap_perc, clip, error_free_bases, links, low_fc_over_gap, low_fc_in_contig) were used to generate linear regression models to see which metrics produce high performing models (Figure 3.3). This procedure was carried out by the 'regsubsets' function (exhaustive search) from the R leaps package [125]. The six best performing models all produced an r-squared of 0.74 with the top two models having the smallest residual sum of squares values. NormN50, gap_perc, clip, error_free_bases, low_fc_over_gap, and low_fc_in_contig were chosen as the metrics for subsequent modeling of the quality score. Low_fc_in_contig was chosen rather than links because while links has a negative correlation with quality score (Table 3.1), it is given a positive weight within the linear regression model. This suggests that there

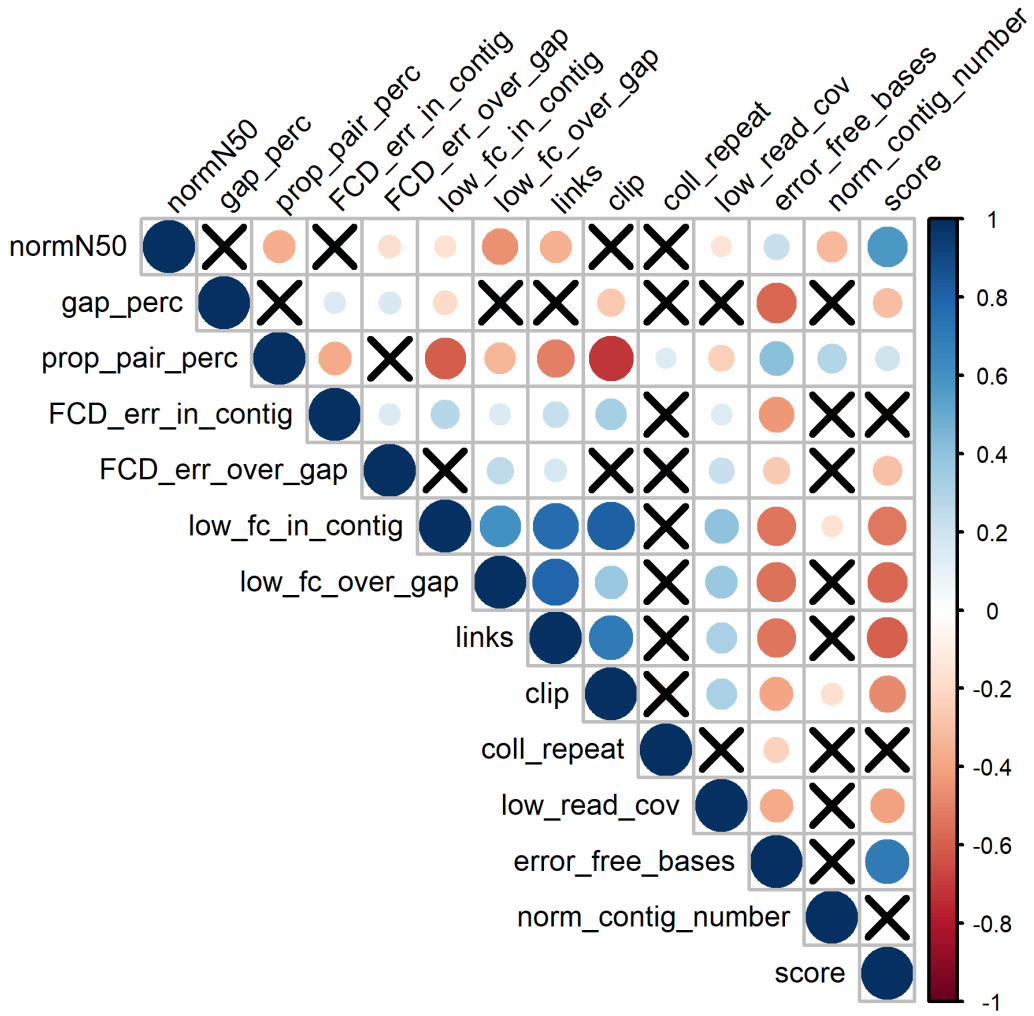

Figure 3.2: Pearson correlation among all metrics. Cells with an X denote metrics with insignificant correlation. Dark blue represents a stronger positive correlation, while dark red represents a stronger negative correlation.

may still be concerns with multicollinearity when using the links metric. Summary statistics and histograms for the selected metrics are shown in Table 3.2 and Figure 3.4.

3.4.3.2 Training of regression models

Model training and testing was carried out using R statistical software [126]. All models tested were from the Caret R package [127]. The full set of training data

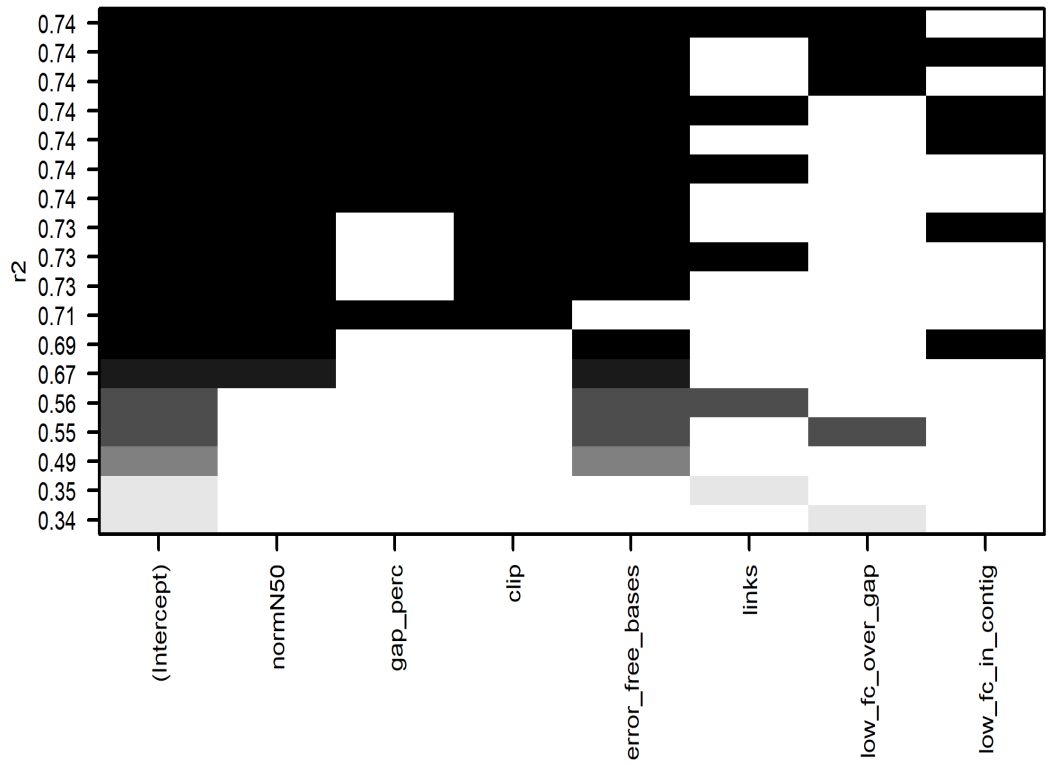

Figure 3.3: Results from regsubsets function for the leaps R package. Each row is a general linear model created from a subset of the features listed along the x-axis. Shaded cells indicated features included in that row's model. The models are ordered and shaded based on their r-squared value given along the y-axis.

was randomly split into two subsets where 80% of the data became the training set and 20% became the testing set. This split resulted in 333 training instances and 83 testing instances.

First, a general linear model using the selected metrics was trained using repeated cross-validation (CV) with 10 folds and 10 repeats. An elastic net model was also trained where cross-validation was used to tune the alpha and lambda hyperparameters. The penalization for elastic net falls between that of Lasso and Ridge regression depending on the alpha and lambda values. The alpha is the elastic net

Table 3.2: Summary statistics for the quality metrics selected to be included in the mammalian genome scoring model along with the reference-based quality score that the model will be estimating.

Metric	Norm. N50	Gap %	Clip	Error Free Bases	Low FC Contig	Low FC Gap	Quality Score
Min.	0.74	0.00	0.01	15.86	0.03	0.00	-79.25
1st Quantile	12.10	0.71	0.025	51.97	0.16	0.004	23.81
Median	22.77	4.04	0.032	55.34	0.59	0.23	51.83
Mean	40.16	5.43	0.08	55.73	1.60	1.65	50.55
3rd Quantile	67.12	7.57	0.06	62.02	2.35	2.17	80.68
Max.	100.00	43.34	0.34	68.04	15.94	11.67	100.00

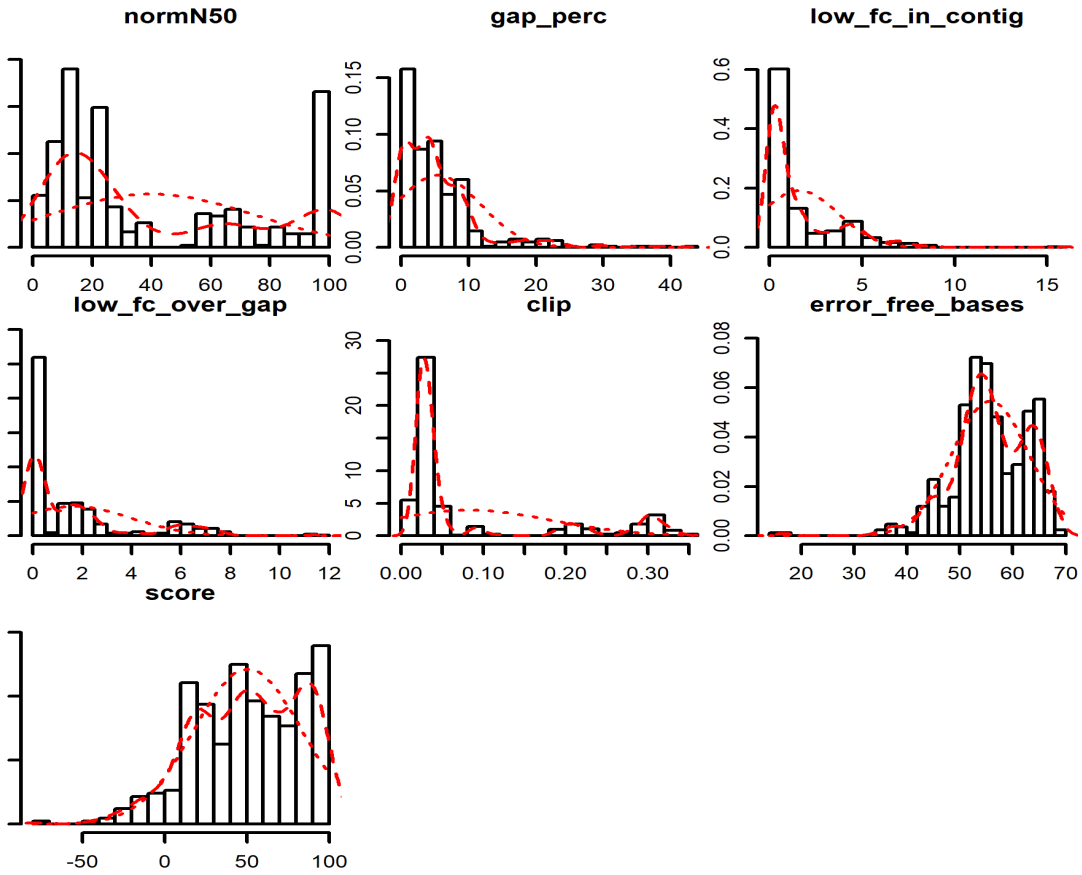

Figure 3.4: Histograms for each selected genome assembly metric as well as the reference-based score (bottom left) from the training data for the mammalian model.

mixing parameter ($0 \geq \alpha \leq 1$), where alpha = 0 would be ridge regression and alpha = 1 would be lasso regression. The lambda parameter determines the amount of coefficient shrinkage (regularization penalty).

In addition, other types of supervised machine learning models were tested. Models tested included K-Nearest Neighbors (KNN) regression, Random Forest (RF) regression, and Support Vector Machines (SVMs) with linear and polynomial kernels. 10-fold cross-validation with 5 repeats was used to tune the KNN model. 10-fold cross-validation was used to tune the RF model and 5-fold cross-validation was used to tune the SVMs. More information about the models and the parameter tuning results are provided in the Supplementary Materials.

3.4.3.3 Model selection

The root mean squared error (RMSE) and r-squared values for each model type were calculated on the test data (Table 3.3). These values reflect each model's performance on the test set. More specifically, the r-squared values reflect the proportion of the reference-based quality score that can be explained by each model, while the RMSE values reflect the differences between the reference-based quality scores and those predicted by each model.

Table 3.3: The r-squared and RMSE values for each type of regression model that was tested to select the best performing model (highlighted in bold) to be the mammalian model.

Regression Model	RMSE	R-squared
General Linear	16.413	0.775
Elastic Net	16.520	0.773
K-Nearest Neighbors	13.615	0.840
Random Forest	**12.697**	**0.860**
SVM (Linear)	17.190	0.774
SVM (Polynomial)	14.363	0.843

The best performing model was random forest regression with 500 trees and an mtry value (number of variables tested at each split) of 2. This model produced a

RMSE of 12.697 and an r-squared of 0.860 when applied to the testing data (Figure 3.5). The random forest model was retrained on the full data to develop the final model that would be used to predict the quality scores of mammalian genome assemblies.

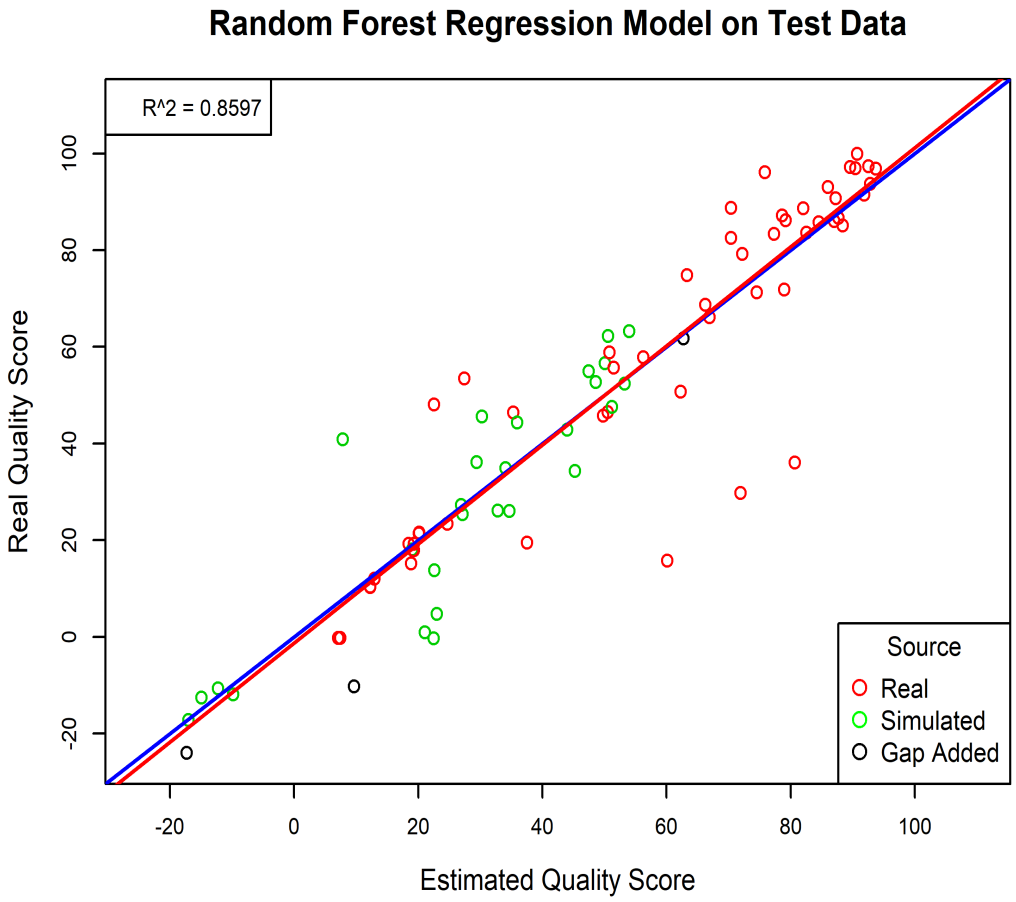

Figure 3.5: Random forest model results on test data. Estimated quality scores for the test instances are plotted against the reference-based quality scores of the test instances. A 100% accurate model would produce the blue line with an r-squared equal to 1. The line of best fit for the plotted data is shown as the red line and has an r-squared of 0.8597.

Once the final model was selected, we wanted to confirm that the source of the assembled sequences in the training set did not impact the ability of the model to predict quality scores. There were no clear patterns regarding the residuals of the scores versus organism source (data point shapes) or regarding the residuals of the scores

versus the generation method of the chromosomes i.e. if they were real, simulated, or had gaps added (data point colors) (Figure 3.6). This observation suggested that the model's ability to predict quality scores of instances within the training/testing data was not impacted by organism or generation methods.

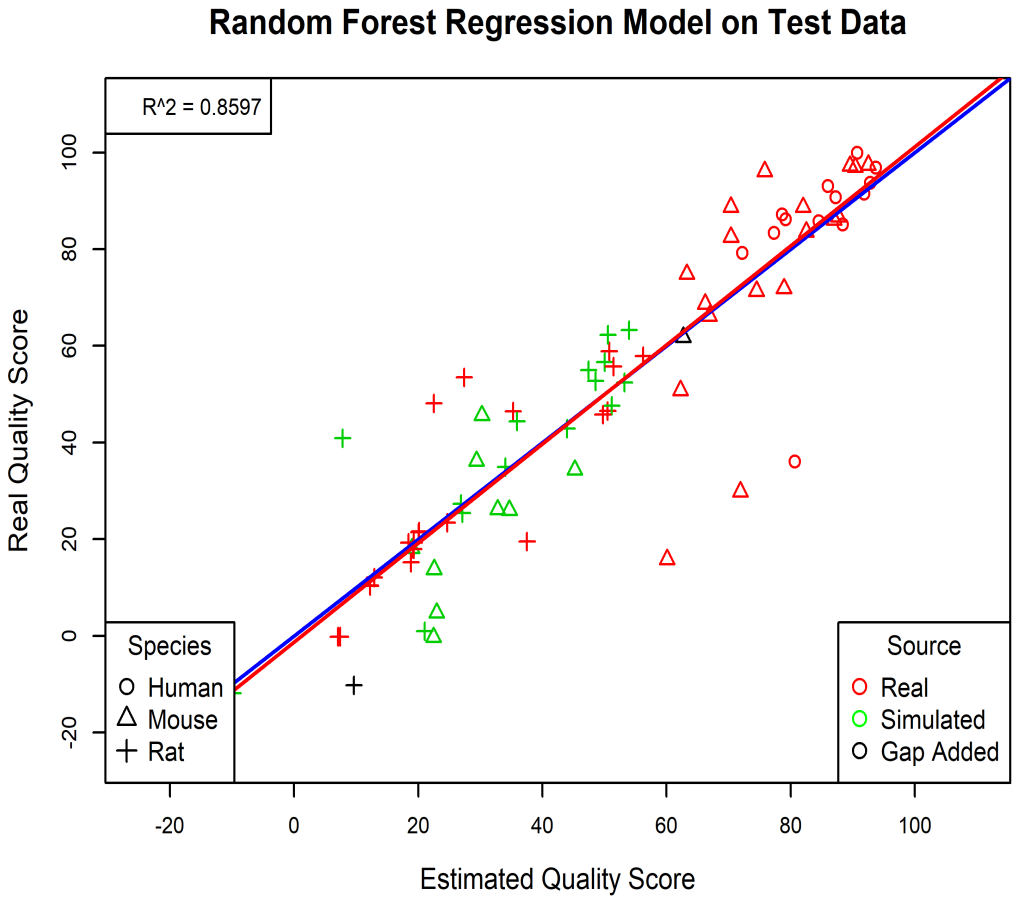

Figure 3.6: Random forest model results on test data with species information. The plot shows the reference-based quality scores versus the EvalDNA quality scores of the test data with species (data point shapes) and sources (data point colors) denoted.

3.4.4 EvalDNA pipeline application

3.4.4.1 Application to Chinese hamster genome assemblies

EvalDNA was applied to new assemblies of the Chinese hamster genome using the mammalian model. Chromosomes from each meta-assembly described in Rupp et al. [30] as well as the previous RefSeq assembly [27] and the chromosome-sorted assembly (CSA) [29] were scored. EvalDNA was also used to score each assembly as a whole with no chromosome separation information provided and including any unplaced contigs.

Illumina reads from SRR954916, SRR954917, and SRR954918 [27] (sequencing project PRJNA167053) were trimmed using a quality cutoff of 26 and a length cutoff of 90 (with the paired option) in Trim Galore [128]. A random subset of trimmed pairs, totaling 20 Gbp, was selected as input for EvalDNA. These sequencing runs were chosen because they had an insert size (500 bp) similar to the reads used in the training data.

3.4.4.2 Comparison to other quality evaluation tools

The manual ranking of the Chinese hamster genome assemblies from Rupp et al. [30] were compared to rankings from EvalDNA, FRCbam, and ALE. Normalized EvalDNA scores, scaled between 0 and 1, for the CH genome assemblies were compared to normalized ALE scores. FRCbam and ALE were run using the same Illumina reads used for EvalDNA (described previously). For ALE, the BAM file was created using Bowtie2 [129] with the '-very-sensitive' parameter instead of SMALTmap. For an unknown reason, ALE was unable to run when given BAM files created with SMALTmap.

FRCbam was run using the BAM files created with the SMALTmap tool within EvalDNA. FRCbam required tuning for the CE-max and CE-min parameters for each set of chromosomes (i.e. chromosome 8 from all assemblies had the same CE-max and CE-min). Estimation of these parameters was done by first graphing the CE-stats distribution provided by FRCbam without specifying the parameters and then using

the 0.95 and 0.05 quantile values from a fitted normal curve as the CE-max and CE-min, respectively. Finally, for each set of chromosomes, the smallest CE-max value was selected to be the CE-max value and the highest CE-min was selected to be the CE-min value.

3.4.4.3 Quality scoring of GAGE assemblies

The human chromosome 14 assemblies were downloaded from the GAGE dataset website (http://gage.cbcb.umd.edu/data/index.html). EvalDNA was run on each assembly to estimate its quality score and subsequently, rank the assemblers. 20.1 Gbp of trimmed paired-end reads from SRR2103647 was given as input. The reads were quality trimmed using Trim Galore (quality cutoff of 26) to ensure high quality reads. The EvalDNA results were used to rank the assemblies and the rankings were compared to those reported in the ALE and FRCbam papers.

ALE was run on the assemblers using identical parameters to those stated in the supplementary information for the ALE paper. We were able to replicate their ranking of the assemblers, and additionally scored the CABOG assembly. We also reran ALE with the same parameters, but with a more recent version of Bowtie2 (version 2.3.3.1). EvalDNA and ALE scores were normalized to be between 0 and 1 for comparison.

3.4.4.4 Scoring of other assemblies

EvalDNA was run on chromosomes from the cow reference genome assembly (ARS-UCD1.2, GCF_002263795.1). Illumina reads from SRR5753530 were trimmed using a quality cutoff of 26 and a length cutoff of 90 (with the paired option) in Trim Galore. These reads, totaling 20.4 Gbp, were selected to use as input for EvalDNA. Read pairs had an insert size of 600 bp.

EvalDNA was also run on several Japanese rice (*Oryza sativa ssp. Japonica*) assemblies as well as the chromosomes from the reference assembly (Os-Nipponbare-Reference-IRGSP-1.0, GCF_001433935.1). The older versions of rice assemblies examined were GCA_000005425.2 and GCA_000149285.1. All assemblies and the sequencing

reads used were from rice of the Nipponbare cultivar. Illumina reads (250 bp long) from SRR547960, SRR547961, SRR547959, SRR547963, and SRR547962 were trimmed using the same parameters for the other organisms. The sequencing reads consisted of 11.5 Gbp and was used as input for EvalDNA. Read pairs had an insert size (450 bp) similar to the reads used in the training data.

3.4.4.5 Error simulation and scoring of PICR chromosomes and scaffolds

Single nucleotide errors were simulated from 5-30%, in increments of 5%, in each chromosome from CH PICR using a custom script. Errors at the same rates were also simulated in scaffolds of various lengths from CH PICR chromosome 1. Errors could be simulated in any location, except for gap regions, across the length of the sequence.

3.5 Results

3.5.1 Evaluating assemblers used in the GAGE study

EvalDNA with the mammalian model was used to score and rank the different assemblies of human chromosome 14 from the GAGE study [119]. The rankings were compared to rankings generated during the original benchmarking tests for ALE and FRCbam [45, 46] as well as the ranking generated by running ALE with an updated version of Bowtie [129] (Figure 3.7 A). Normalized EvalDNA scores (scaled to be between [0,1]) were compared to the two sets of normalized ALE scores (Figure 3.7B).

EvalDNA and FRCbam selected ALLPaths-LG [19] as the best assembler, while the ALE runs ranked ALLPaths-LG as the second best with the CABOG assembler [130] ranking first. EvalDNA ranked the assembly produced by Velvet [131] as the lowest quality assembly, which is not surprising since it is made up of approximately 45% gaps. The Velvet assembly was ranked second to last by the ALE runs and third to last by FRCbam.

One key difference among the rankings is that EvalDNA ranked the ABySS [132] assembly much higher (second place) than either ALE or FRCbam (last place). ALE and FRCbam most likely ranked ABySS the lowest because the assembly is highly

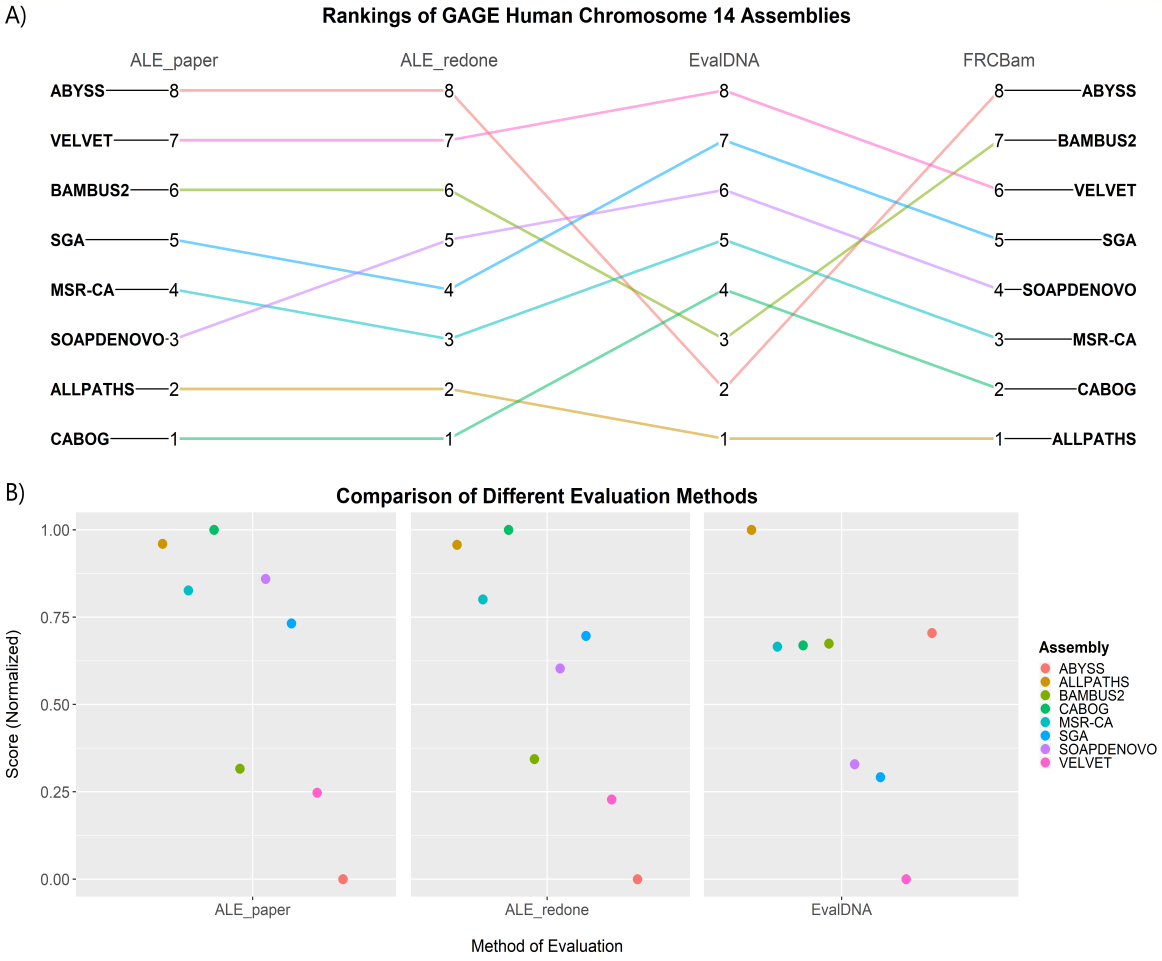

Figure 3.7: Comparison of quality evaluation methods on human chromosome 14 assemblies from the GAGE study. A) The EvalDNA ranking of assemblers used to build the human chromosome 14 assembly are compared to the rankings from ALE and FRCbam. The highest quality assembly is given a rank of 1. B) EvalDNA and ALE scores for the human chromosome 14 assemblies were normalized (scaled to be between [0, 1]). ALE_paper scores were calculated using the same parameters and version of Bowtie described in Clark et al. The ALE_redone scores were calculated with an updated version of Bowtie.

fragmented. However, the ABySS assembly is also one of the more accurate assemblies with fewer scaffold misjoins, inversions, relocations, and indels than the other assemblies [119]. ABySS also has a very low gap percent (0.53%). This observation suggests that EvalDNA's mammalian model may value accuracy and completeness (in regards

to the lack of gaps) over continuity more so than ALE or FRCbam. In addition, examining the normalized EvalDNA scores does show that ABySS, while second in the ranking, scored only slightly better than the CABOG, MSR-CA [133], and BAMBUS2 [134] assemblers.

3.5.2 Scoring of Chinese hamster assemblies for reference assembly selection

In 2018, four assemblies for the Chinese hamster (CH) were built using PacBio sequencing data and existing Illumina data. Manual ranking of these new assemblies as well as two Illumina-only assemblies from 2013 was completed to select the best reference genome for CH and Chinese hamster ovary (CHO) cells [30]. The two Illumina-only assemblies included the 2013 CH RefSeq assembly (GCF_000419365.1) [27] and the 2013 chromosome sorted assembly (CSA, GCA_000448345.1) [29]. EvalDNA results were compared to this ranking to evaluate its performance on real assemblies outside of those used in the training data and if it could be used to select the best assembly to be the new reference genome.

EvalDNA with the mammalian model was used to score the six different CH assembly versions (Table 3.4) as well as each chromosome from the assemblies (Table 3.5). Scaffolds and contigs were assigned to chromosomes based on the coverage of reads mapped from each of the CSA chromosomes. For CSA, sequencing was done on chromosomes after they were individually isolated using flow cytometry. However, chromosomes 9 and 10 could not be separated due to their size similarity [29]. Therefore, for each assembly, scaffolds could be assigned to chromosomes 9 and 10, but not separately, and these chromosomes together are given a single score. The full CH assemblies were also assessed by FRCbam and ALE.

EvalDNA and the manual ranking selected PICR as the CH assembly with the highest overall quality (Figure 3.8A), with PIRC a close second. FRCbam ranked PICR and PIRC as the highest, but the curves were too close to distinguish between them (Figure 3.9). All four evaluation methods agreed that CSA was of the poorest

Table 3.4: The EvalDNA quality scores for the CH genome assemblies.

Assembly	Mammalian Model	Model Without N50
PICR (2018 RefSeq)	70.22	88.47
PIRC	70.20	88.41
IPCR	57.56	59.29
IPRC	57.57	59.22
2013 RefSeq	58.72	64.35
CSA	43.21	40.60

Table 3.5: The EvalDNA quality scores for each chromosome from the CH genome assemblies. The highest score for each chromosome is highlighted in bold.

Chromosome	PICR	PIRC	IPCR	IPRC	RefSeq	CSA
1	**72.04**	71.91	60.11	59.26	59.58	41.51
2	71.00	**71.20**	58.76	57.37	56.92	49.90
3	65.37	**65.42**	55.56	54.51	55.21	39.55
4	**68.75**	68.63	54.71	56.01	56.15	42.49
5	68.02	**68.21**	40.01	63.13	56.38	47.48
6	68.21	**68.24**	62.77	56.70	56.40	50.27
7	**68.95**	68.66	55.89	54.77	55.80	46.44
8	**63.99**	63.84	52.21	56.45	54.47	52.83
9_10	48.31	**52.32**	39.53	51.92	47.62	47.27
X	**53.30**	52.51	49.30	53.14	48.72	30.44

quality. However, EvalDNA and ALE both scored RefSeq higher than IPCR and IPRC, while the manual ranking and FRCbam had this order switched. Examining the ALE and EvalDNA normalized scores more closely (Figure 3.8B) show that these three assemblies are very similar regarding quality (within 0.05 normalized units). The difference in quality may be too small for EvalDNA to meaningfully distinguish between these assemblies.

The accuracy of EvalDNA scores and ranking of CH assemblies was also confirmed by calculating the number of differences between each CH assembly and the 'reference' genome (PICR). This method allows each assembly to get a score, calculated the same way the training instances were scored with the exception of not being

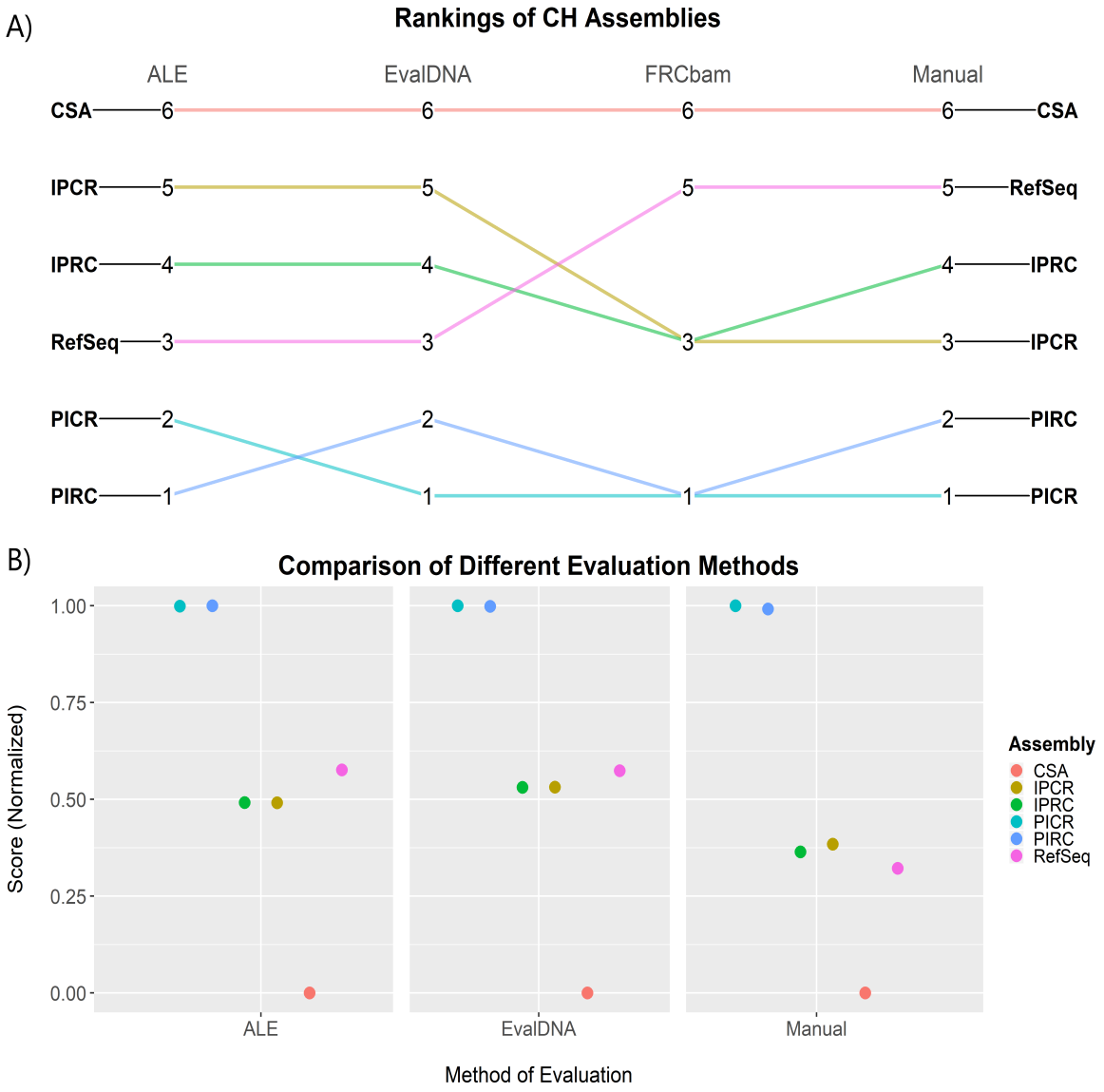

Figure 3.8: Comparison of quality evaluation methods on CH genome assemblies. A) Comparison of the EvalDNA ranking of the multiple CH genome assemblies to a manual ranking, and rankings from ALE and FRCbam. The highest quality assembly is given a rank of 1. B) EvalDNA and ALE scores for the CH assemblies as well as the rankings given in Rupp et al. were normalized (scaled to be between [0, 1]).

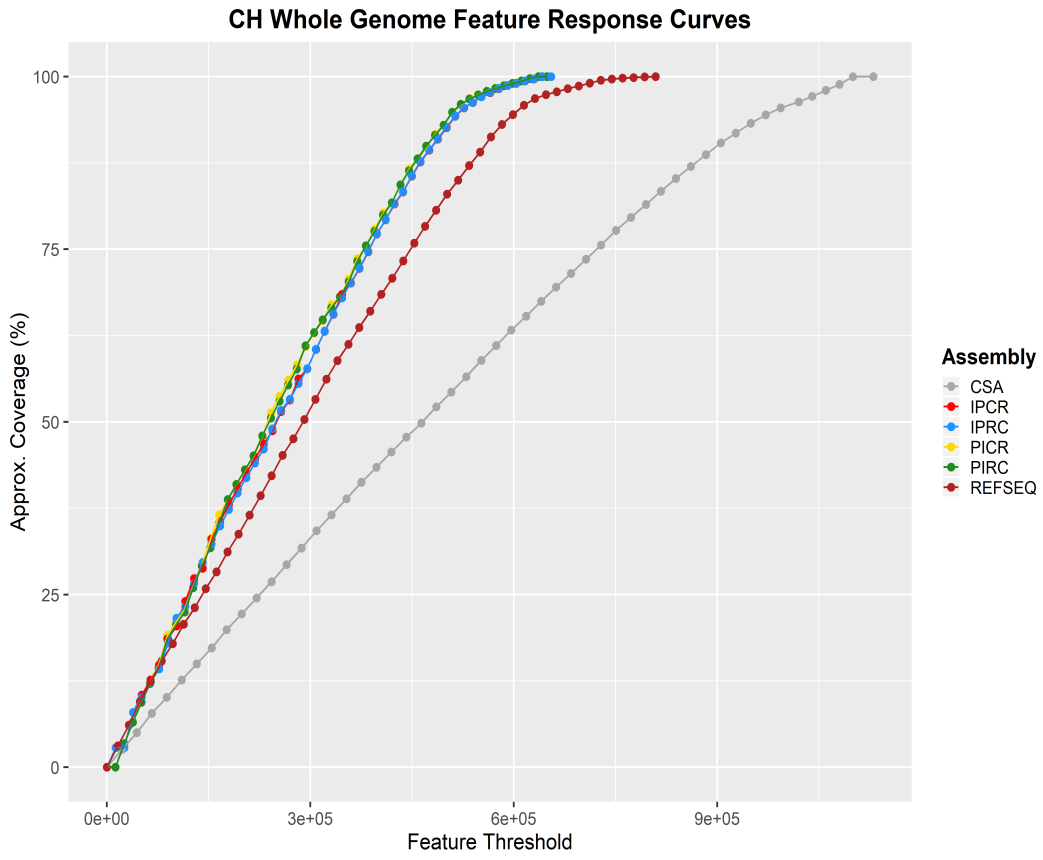

Figure 3.9: The FRCbam results (FRCurves) for the CH genome assemblies. Thresholds of the number of allowed errors (features) are shown along the x-axis. Only contigs (starting with the longest) whose sum of features is less than this threshold can be used to compute the genome coverage, which is shown on the y-axis.

scaled (see "Quality Scoring" section in Methods). The difference between each assembly's score and a score of 100 (PICR's score from aligning PICR to itself) should be similar to the difference between the corresponding assembly's EvalDNA score and PICR's EvalDNA score. The differences were indeed similar (Table 3.6), confirming that EvalDNA can be used to accurately evaluate assemblies from organisms that were not used in the training set.

Table 3.6: Differences between each CH assembly EvalDNA score and PICR's EvalDNA score compared to the differences between NUCmer scores (derived from NUCmer alignments of each assembly to PICR).

Assembly	EvalDNA score	NUCmer score	Difference from PICR EvalDNA score	Difference from PICR NUCmer score
PICR	70.22	100	0	0
PIRC	70.20	99.11	0.03	0.9
IPCR	57.56	85.20	12.5	14.80
IPRC	57.57	85.24	12.64	14.76
RefSeq	58.72	85.31	11.50	14.69
CSA	43.21	63.43	27.00	36.57

3.5.3 Comparing CH assembly quality to other organism reference assemblies

The PICR assembly was selected to be the new Chinese hamster reference assembly (GCF_003668045.1) [30]. EvalDNA scores for the PICR chromosomes were compared to scores from the 2013 CH RefSeq assembly and the reference assemblies for human (GCF_000001405.38), mouse (GCF_000001635.20), rat (GCF_000001895.5), and cow (GCF_002263795.1). The majority of the PICR CH assembly chromosomes are of higher quality than those of the 2013 CH RefSeq assembly and the rat reference assembly (Figure 3.10a). Several chromosomes also scored as high as those from the mouse reference assembly.

EvalDNA was also run on each chromosome from the rice (*Oryza sativa*) reference genome (GCF_001433935.1) (Figure 3.10a). While the model was trained using mammalian data, the results of EvalDNA with this model on rice also seem reasonable. Two older versions of the rice assembly, Build4.0 (GCA_000005425.2) and OrySat_Sep2003 (GCA_000149285.1), were scored. Build4.0 scored within 1 unit of the most recent version, while OrySat_Sep2003 scored significantly lower (more than 30 units). The similar scores between Build4.0 and the most recent reference is not

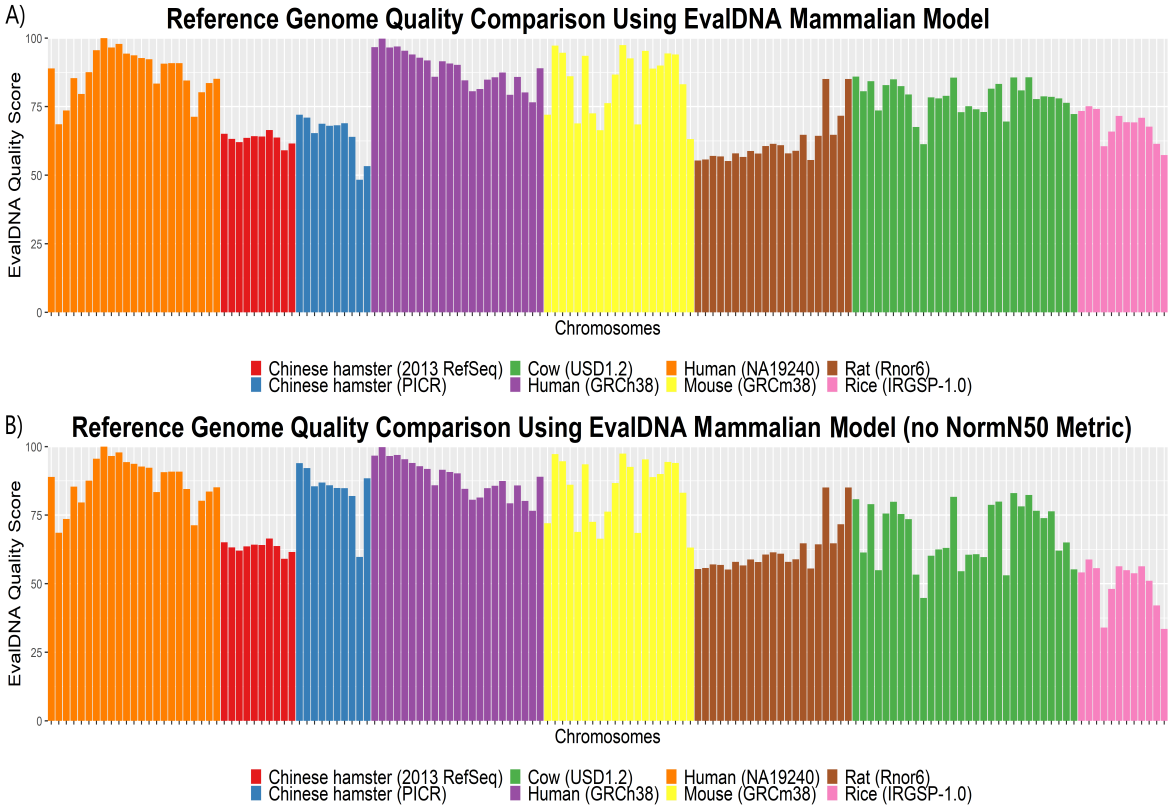

Figure 3.10: EvalDNA quality scores for chromosomes from various genome assemblies. A) EvalDNA quality scores for chromosomes from CH PICR, CH 2013 RefSeq, and the mouse, rat, human, cow, and rice reference genome assemblies. B) EvalDNA quality scores for the same chromosomes but calculated using a model that does not include the normalized N50 metric.

surprising because the accuracy of Build4.0 was already high with an error rate estimated to be less than one per 10,000 nucleotides and possibly as low as 0.15 errors per 10,000 nucleotides [135]. Results of the rice assemblies are given in Table 3.7.

Table 3.7: The EvalDNA scores of various Japanese rice assemblies (all three are of the Nipponbare cultivar).

Assembly	EvalDNA score
IRGSP-1.0 (reference)	81.81
Build4.0	82.44
OrySat_Sep2003	50.82

The scores allow comparison of assemblies across organisms in regards to continuity, completeness, and accuracy. Changing the model to only examine a subset of these categories can give more specific insight into where an assembly excels or needs improvement. For instance, we scored the chromosomes with a different random forest regression model which does not include the normalized N50 metric (Figure 3.10b). This model, described in the Supplementary Materials, enables comparisons across organisms based on completeness and accuracy only. The model shows a large increase in the accuracy and completeness of the 2018 CH PICR reference assembly over the 2013 CH RefSeq assembly. In addition, because each chromosome of the cow and rice assemblies contains just a single scaffold, the original model scored each chromosome from these organisms much higher than the model that does not use the normalized N50 metric (Figure 3.10). The disparity among the scores predicted between these two models does confirm that scores from different models are not directly comparable.

3.5.4 EvalDNA scores correlate with error simulation rates, but not linearly

To examine how changes in the amounts of errors within an assembly affect the EvalDNA score, we ran EvalDNA on versions of the CH PICR chromosomes which contained varying amounts of randomly generated single nucleotide errors. Single nucleotide changes were simulated from 5% to 30% in increments of 5%.

Each simulated chromosome was scored by EvalDNA (Figure 3.11a). Similar trends across all chromosomes are seen, and the scores do not linearly decrease as the amount of errors increase. On average, the quality score decreases slightly (1 unit) between a 0% error rate and a 5% error rate and then decreases an average of 10 units between 5% and 10% error rates. An even larger score decrease (average of 34 units) occurs as the simulated error rates change from 10% to 15%. The scores decrease an average of 17 units from 15-20%, 3 units from 20-25%, and 2 units from 25-30%.

Assessment of the EvalDNA scores with respect to error rates alone is difficult

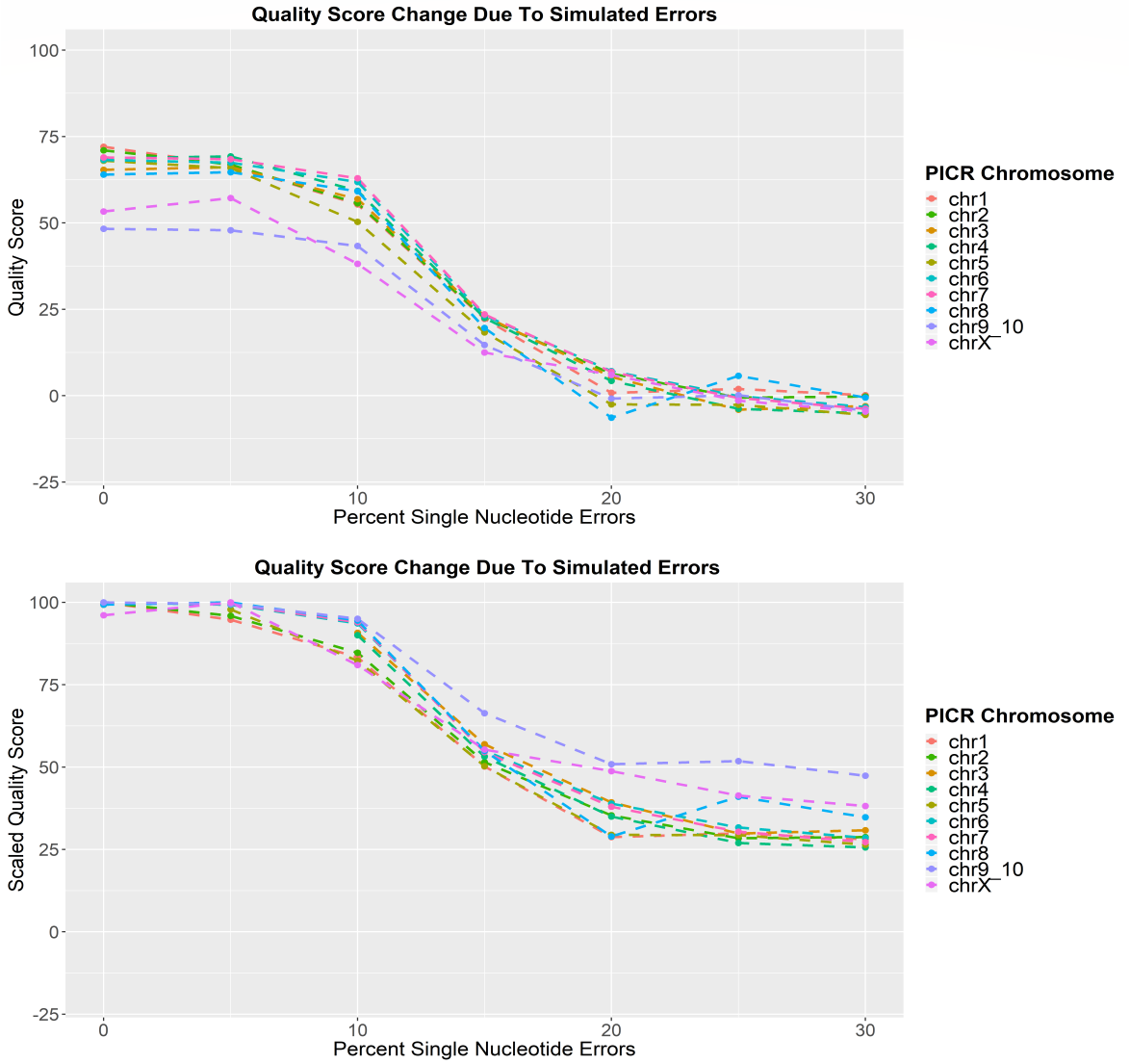

Figure 3.11: The impact of error rates on the EvalDNA quality scores of CH PICR chromosomes. A) Changes in EvalDNA quality scores due to simulation errors. B) Changes in scaled EvalDNA quality scores due to simulation errors. Scores were scaled so that the maximum score for a chromosome became 100.

because none of the PICR chromosomes are perfectly accurate, complete, and continuous before error simulation. The chromosomes with 10% simulated error rate have scores anywhere from 35 to 65 depending on the continuity and completeness of the chromosome. However, a near perfect chromosome or assembly will have a score above

100 and insights can be gained from scaling all the scores so that maximum score for each chromosome is 100 (Figure 3.11b). From scaling, we can see that a perfectly complete and continuous chromosome with a score around 89 corresponds with an error rate of approximately 10%. This means that a chromosome or assembly that is not fully complete and continuous with a score of 89 or above will have a percent error rate lower than 10%. Since most mammalian assemblies are far from being fully continuous and complete, a score of 89 will often mean an error rate of much lower than 10%. Even the chromosomes from the current human reference genome assembly (GRCh38) in the training set have scores ranging from approximately 85 to 100, and GRCh38 has an estimated error rate of 1 in 100,000 bases (0.001%) [136]. Recommended guidelines for how to categorize an assembly based on the reference-based quality scores from the training data are provided in Figure 3.12.

3.5.5 EvalDNA application on scaffolds

Varying levels of single nucleotide errors were randomly generated in several scaffolds from PICR chromosome 1 to examine how well EvalDNA with the mammalian model works on scaffolds. To minimize false mapping, EvalDNA was run using only reads that mapped to the original scaffold with an identity of 0.75 (at least 75% of the bases needed to match).

The error simulation results suggest that EvalDNA's ability to estimate quality scores for scaffolds depends on the amount of errors and the scaffold length (Figure 3.13). The score decreases in a similar manner as the chromosomes did for all length scaffolds with 0-10% errors simulated. As the percent of errors increases beyond 10%, the impact of length on the scores becomes apparent. The scores show the expected decreasing trend for scaffolds longer than 5 Mbp, although at a slower rate than the chromosome scores. The expected decreasing trend is not observed for scaffolds shorter than 1 Mbp and for only some of the scaffolds between 1 Mbp and 5 Mbp long. Therefore, a model specifically trained on scaffolds in these length ranges would be beneficial for short scaffold scoring.

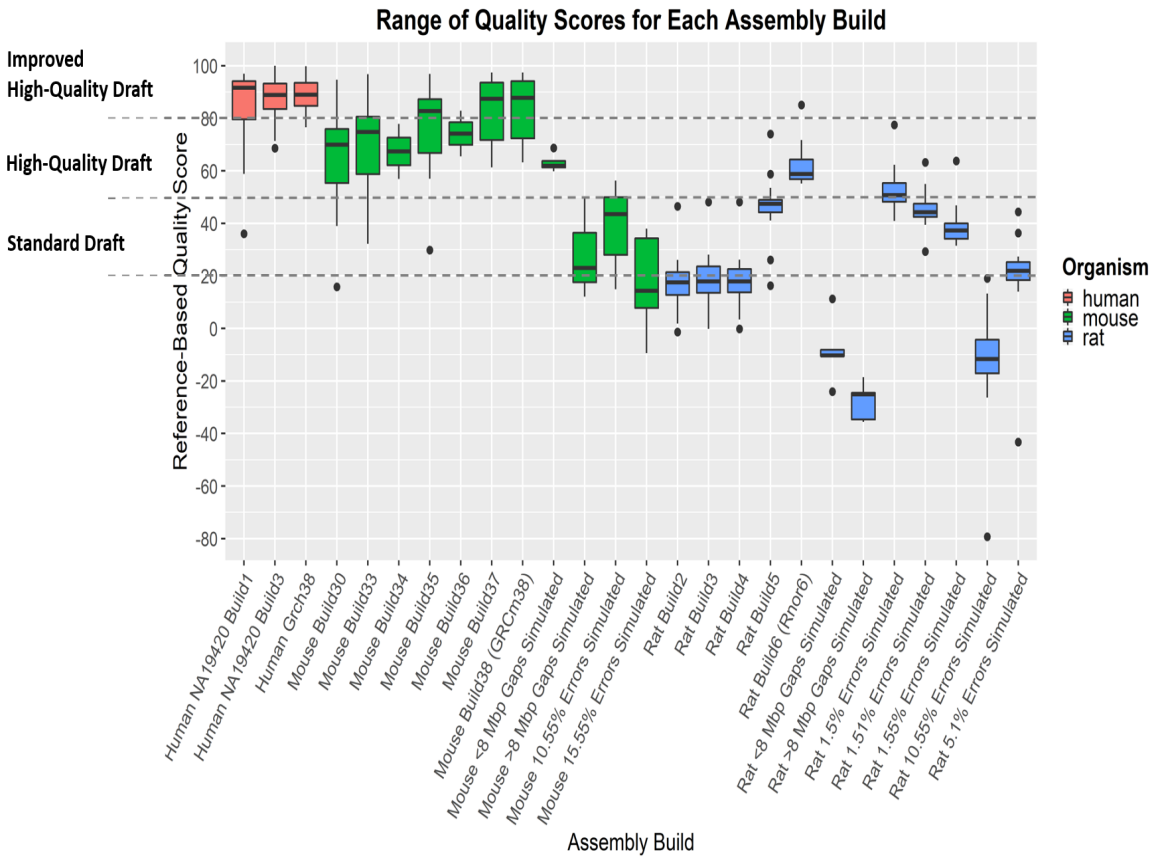

Figure 3.12: Recommended guidelines for EvalDNA quality score interpretation from the reference-based scores of the training data instances.

3.6 Discussion

Here, we presented a novel pipeline, called EvalDNA, for genome quality assessment that does not require a reference genome. We also developed a model, trained on mammalian assembly data, to be used within EvalDNA. The model evaluates an assembly based on completeness, continuity, and accuracy by using the normN50, gap_perc, clip, error_free_bases, low_fc_over_gap, and low_fc_in_contig metrics.

The EvalDNA parison of CH chromosomes to those from other organisms' referencipeline with this mammalian model was able to accurately estimate the quality scores of Chinese hamster genome assemblies and enabled the compe genome assemblies. EvalDNA can also be used to examine the output of different assemblers

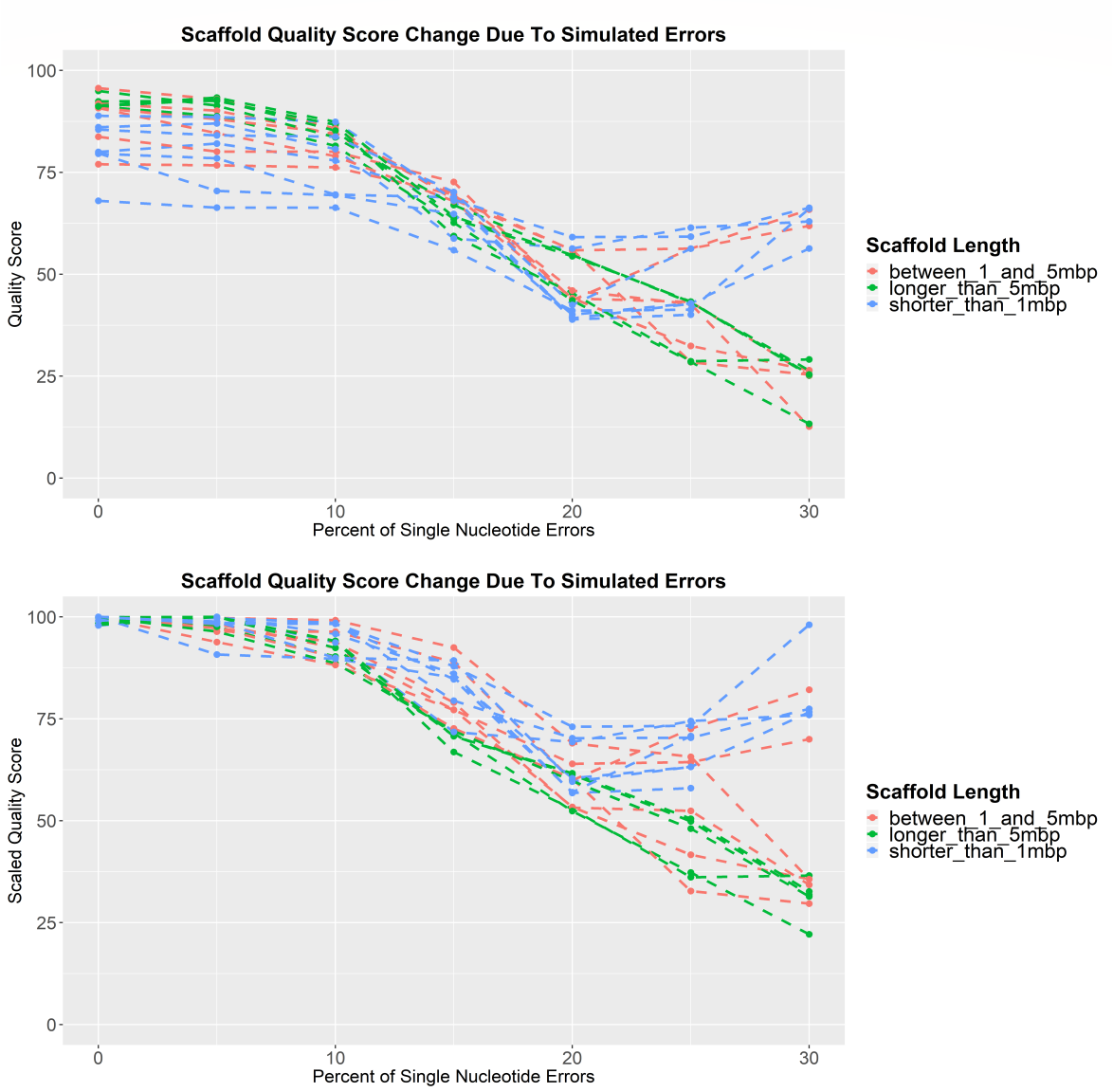

Figure 3.13: The impact of error rates on the EvalDNA quality scores of CH PICR scaffolds. A) Changes in EvalDNA quality scores due to simulation errors. B) Changes in scaled EvalDNA quality scores due to simulation errors. Scores were scaled so that the maximum score for each scaffold became 100.

as demonstrated on the human chromosome 14 data. Often assembler accuracy is tested using sequencing reads simulated from a given assembly [137, 138, 139, 140], but EvalDNA could be used as an additional assembler evaluation method that examines assembler performance on real sequence data.

While EvalDNA with the mammalian model appeared to weigh accuracy over continuity more so than existing tools such as ALE and FRCbam, the model without the normalized N50 metric can be used to score assemblies completely independent of continuity if needed. This model may be useful for situations such as genome annotation, where the accuracy and completeness of an assembly is more important than continuity. A model without the normalized N50 metric could also be useful when comparing chromosomes from different assemblies where the method of how the scaffolds/contigs were assigned to a chromosome may differ and may impact the quality score. However, it is important to note that scores from different models should not be directly compared.

3.6.1 Benefits of a comparable genome assembly score

EvalDNA provides the ability to assign a comprehensive quality score to all assemblies and all chromosomes made available online. A researcher would be able to easily select the best available assembly for their organism of interest from viewing these scores, and even choose the best version of a specific chromosome. More confidence could also be given to findings derived from a high scoring reference genome than findings from a lower scoring reference genome.

The assigned quality score would also be comparable across organisms scored by the same model. The scores would provide insight into how a chosen assembly compares to "gold-standard"genomes, such as the human reference assembly, in terms of overall quality. Because EvalDNA can only be used to compare assemblies from different organisms if the assemblies were scored using the same model, the applicability of the mammalian model across all species should be examined in more depth. Initial results on the rice assembly do suggest that the mammalian model could work to assess plant genome assemblies, but more study is needed.

3.6.2 Applying EvalDNA to scaffolds

The principles used within EvalDNA can be applied to scaffolds as well. However, the mammalian model has been created specifically for whole and chromosome level assemblies. The training data for the mammalian model was generated using the mapping defaults of SMALTmap within REAPR. This only required reads to have at least 50% of bases match the reference to be mapped. For scaffolds, this threshold causes a significant amount of incorrect read mapping as reads from anywhere in the genome could map to the scaffold and therefore, a higher mapping stringency is needed. Initial results of the mammalian model on scaffolds longer than 5 Mbp seemed promising, but did require increasing the mapping stringency to 0.75 (75% bases need to map). Therefore, while the model can be applied to scaffolds longer than 5 Mbp if a higher percent mapping threshold is specified, the resulting quality score will not necessarily be directly comparable to the scores of chromosomes or whole genome assemblies.

3.6.3 Model improvement

The current model on average predicts the score within 13 units of the real score and is able to explain 86% of the variation in quality scores. Therefore, there is potential for model improvement. First, increasing the number of chromosome instances in the training set would help the model become more precise. In addition, the model may benefit from the addition of quality metrics not tested here. The new metrics may be able to capture the remaining 14% of the score quality not captured by the current model.

3.6.4 Long-read sequencing

Currently, high quality paired-end Illumina reads are required to use EvalDNA. A future goal is to extend EvalDNA to use longer reads, such as those from PacBio or Oxford Nanopore sequencing, to assess accuracy either alone or along with Illumina data. This improvement will require the development of metrics that reflect the accuracy of an assembly based on the mapping of long reads. Possible metrics could include

the percent of high quality mapped long reads or the total length of structural variants identified from the long read mapping.

3.7 Conclusions

We developed and tested a novel pipeline, called EvalDNA, for the evaluation of genome assembly quality that does not require a reference genome. A model, which can be used within the pipeline, was created using supervised machine-learning. The model examines the accuracy, continuity, and completeness of either an assembled genome or chromosome, and was able to predict reference-based quality scores of assemblies with an accuracy of approximately 86%.

EvalDNA will allow scientists working with multiple genome assembly versions to identify the most appropriate one to be their reference genome, as well as examine which chromosomes may need to be improved. EvalDNA also enables quality comparison against other organism assemblies, such as high quality reference human and mouse assemblies. EvalDNA scores could become the new standard for the assessment and comparison of genome assembly quality.

3.8 Availability of Data and Materials

EvalDNA and the mammalian model are available on GitHub:

Project name: EvalDNA

Project home page: https://github.com/bioinfoMMS/EvalDNA

Operating system(s): Linux

Programming language: Python v2.7.13 or later, R statistical software v3.5.1 or later

Other requirements: REAPR v1.0.18, SAMtools v0.1.19, R libraries: Caret v6.0-81 and randomForest v4.6-14

License: GNU GPLv3

The genome assemblies and paired-end read data are available from NCBI Assembly and the Sequence Read Archive (SRA) respectively.

3.9 Supplementary Materials

Table 3.8: Build information for each assembly used as a source of the training data chromosomes.

Organism	**Recent Build (RefSeq ID)**	**Recent Build Release Date**	**Previous builds (newest - oldest)**
Human	GRCh38p.12 (GCF_000001405.38)	December 2017	None Used
Human NA19420	NA19240_3.0 (GCA_001524155.4)	July 2017	NA19240_1.0
Mouse (*Mus musculus*)	GRCm38.p1 (GCF_000001635.2)	March 2012	37.2, 36.1, 35.1, 34.1, 33.1, 30
Rat (*Rattus norvegicus*)	Rnor_6.0 (GCF_000001895.5)	July 2016	Rnor_5.0 (5.1), 4.1, 3.1, 2.1

Model Testing

Information about all the models that were examined as well as their RMSE and r-squared values are provided in this section.

1. General linear model with scaling
   (a) Model:
   quality score = 51.5306 + (normN50 * 15.4496) + (gap_perc * -8.2674) + (clip * -11.1747) + (error_free_bases * 8.0580) + (low_fc_over_gap * -1.2670) + (low_fc_in_contig * -2.1865)
   (b) Results on test data:
   RMSE = 16.413
   R-Squared = 0.774

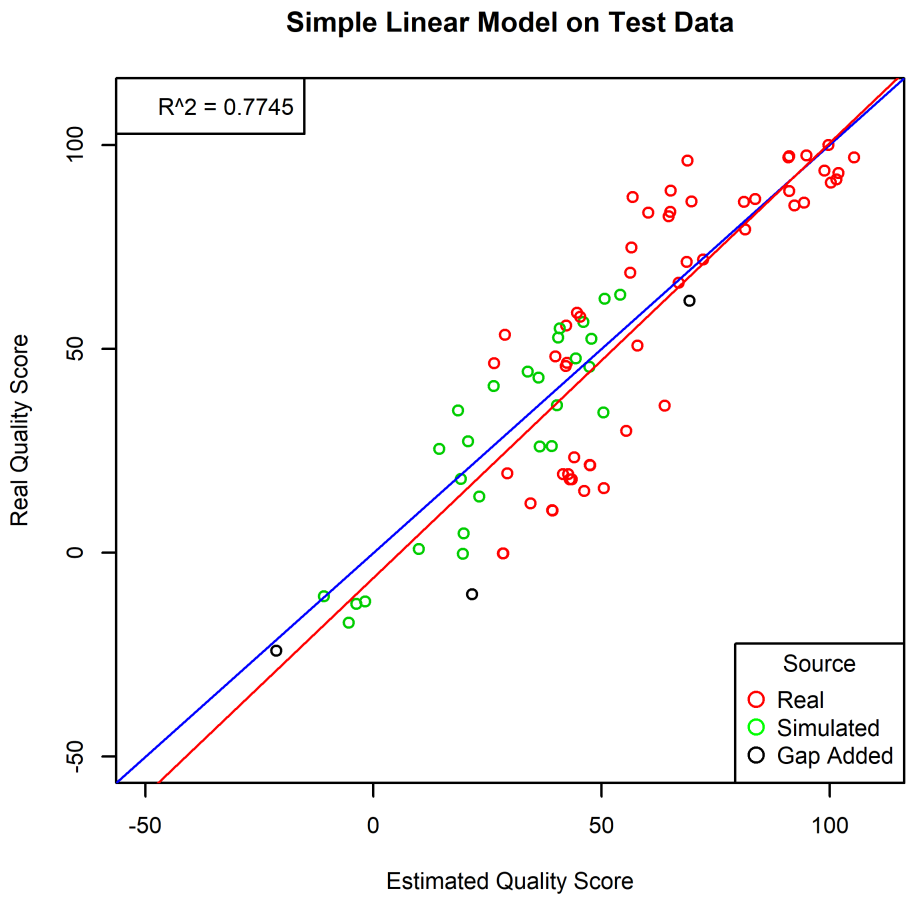

Figure 3.14: Performance of the general linear regression model on test data. The estimated quality scores of the test instances are plotted against the reference-based quality scores of the test instances.

2. Elastic Net - Caret train method 'glmnet' with scaling

   (a) Model:

   alpha = 0.1
   lambda = 0.677
   quality score = 51.5306 + (normN50 * 14.9386) + (gap_perc * -7.6894) + (clip * -10.4590) + (error_free_bases * 8.3534) + (low_fc_over_gap * -1.5006) + (low_fc_in_contig * -2.3295)

   (b) Results on test data:

   RMSE = 16.521
   R-square = 0.773

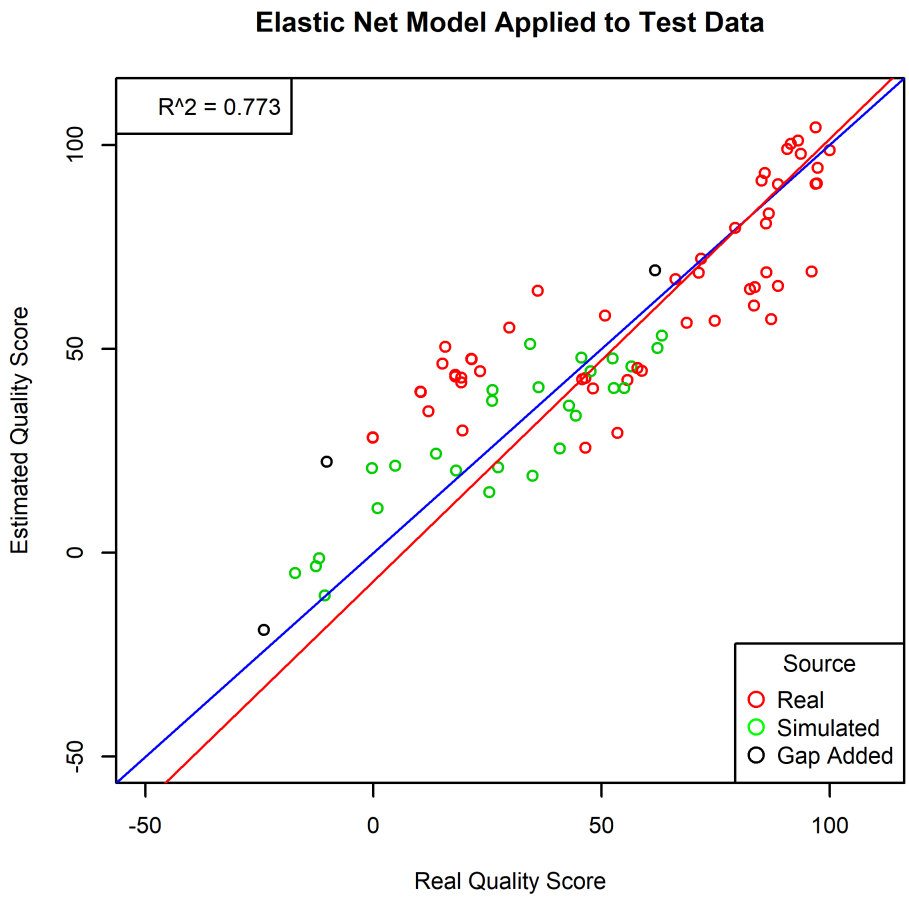

Figure 3.15: Performance of the elastic net regression model on test data. The estimated quality scores are plotted against the reference-based quality scores of the test instances.

3. K-Nearest Neighbors (KNN) regression

   (a) Tuning parameters

   RMSE was used to select the model with the most optimal k value. The final value used for the model was k = 5 (Table 3.9).

   (b) Results on test data:

   RMSE = 13.615
   R-square = 0.840

Table 3.9: Results for tuning the value of k for KNN regression.

k	RMSE	R-squared	MAE
5	11.753	0.873	8.459
7	12.006	0.868	8.598
9	12.463	0.860	8.922

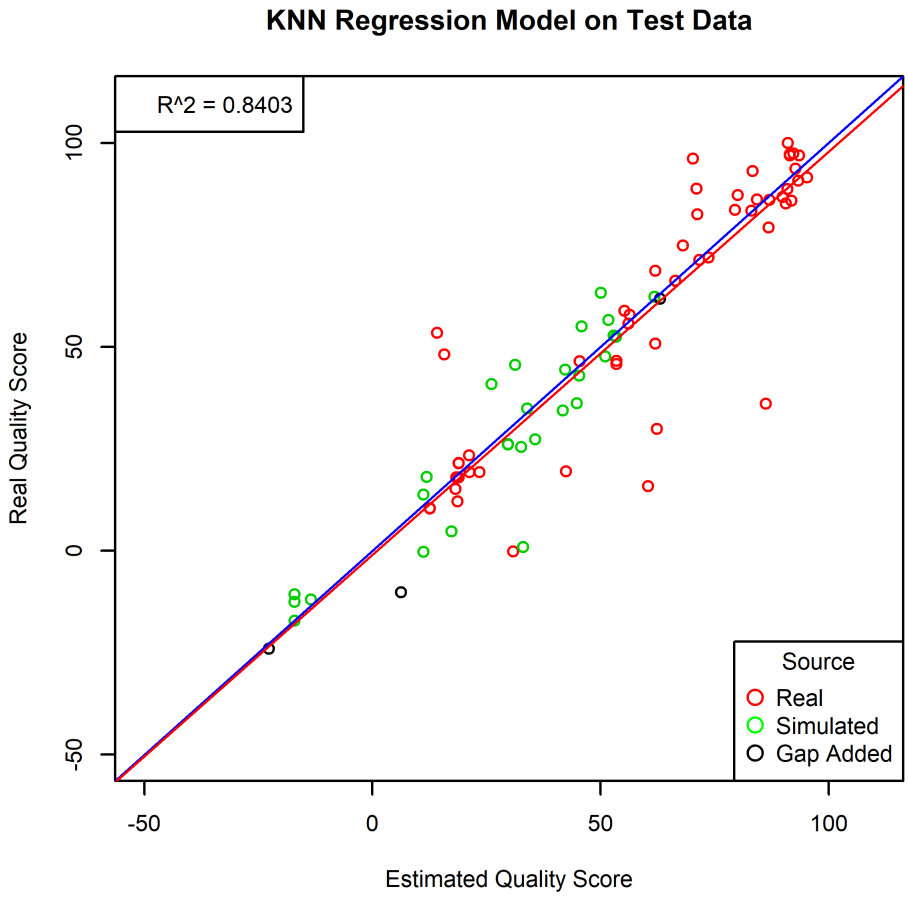

Figure 3.16: Performance of the KNN regression model on test data. The estimated quality scores are plotted against the reference-based quality scores of the test instances.

4. Random forest (rf) regression

   (a) Tuning parameters

RMSE was used to select the model with the most optimal mtry value. The final value used for the model was mtry = 2 (Table 3.10).

Table 3.10: Results for tuning the value of mtry for random forest regression.

mtry	RMSE	R-squared	MAE
2	11.358	0.883	8.099
4	11.500	0.879	8.127
6	11.869	0.871	8.157

(b) Results on test data: See section 3.4.3.3 'Model Selection'.

5. Support Vector Machines with Linear Kernel

   (a) Tuning parameters

   'C' was held constant at a value of 1

   (b) Results on Test Data:

   RMSE = 17.190
   R-square = 0.774

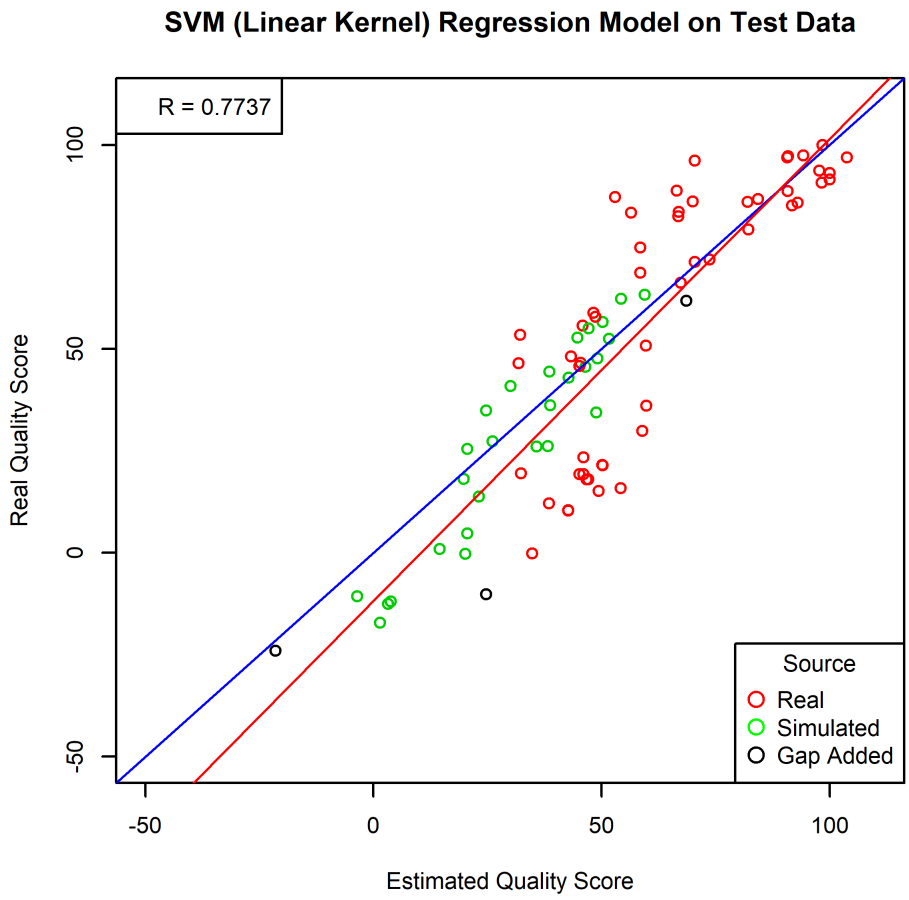

Figure 3.17: Performance of the SVM regression model with a linear kernel on test data. The estimated quality scores are plotted against the reference-based quality scores of the test instances.

6. Support Vector Machines (SVM) regression with Polynomial Kernel

   (a) Tuning parameters:

   RMSE was used to select the optimal value of C, degree, and scale for the Polynomial Kernel SVM. The final values used for the model were degree = 2, scale = 0.1 and C = 0.25.

Table 3.11: Results for tuning the value of C for SVM regression with a polynomial basis function kernel.

Degree	Scale	C	RMSE	R-squared	MAE
1	0.001	0.25	29.499	0.646	24.060
1	0.001	0.50	26.973	0.652	21.646
1	0.001	1.00	24.083	0.670	18.797
1	0.010	0.25	20.392	0.694	15.301
1	0.010	0.50	18.777	0.709	13.829
1	0.010	1.00	18.045	0.715	13.306
1	0.100	0.25	17.470	0.721	12.947
1	0.100	0.50	17.408	0.719	12.853
1	0.100	1.00	17.406	0.718	12.802
2	0.001	0.25	26.966	0.652	21.641
2	0.001	0.50	24.073	0.670	18.789
2	0.001	1.00	21.103	0.688	15.963
2	0.010	0.25	18.558	0.716	13.672
2	0.010	0.50	17.585	0.728	13.017
2	0.010	1.00	16.854	0.743	12.477
2	0.100	0.25	16.039	0.791	11.802
2	0.100	0.50	16.466	0.792	11.842
2	0.100	1.00	16.288	0.794	11.639
3	0.001	0.25	25.233	0.662	19.923
3	0.001	0.50	22.089	0.682	16.894
3	0.001	1.00	19.880	0.699	14.798
3	0.010	0.25	17.643	0.731	13.050
3	0.010	0.50	16.777	0.749	12.363
3	0.010	1.00	16.079	0.769	11.939
3	0.100	0.25	19.824	0.778	11.562
3	0.100	0.50	20.101	0.788	11.267
3	0.100	1.00	18.799	0.800	10.645

(b) Results on test data:
RMSE = 14.363
R-square = 0.843

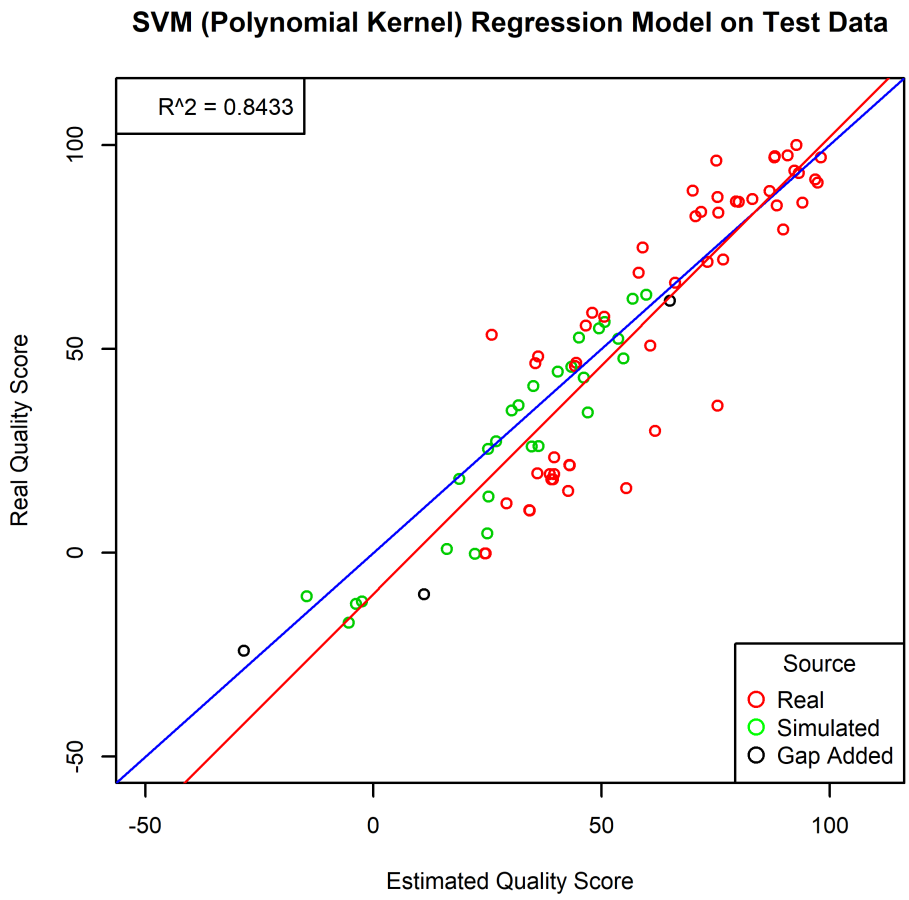

Figure 3.18: Performance of the SVM regression model with a polynomial kernel on test data. The estimated quality scores are plotted against the reference-based quality scores of the test instances.

3.9.1 Model without the normalized N50 metric

A random forest regression model using the same metrics as the main mammalian model, except for normN50, was developed. Parameters were tuned for using 10-fold cross validation. The lowest value of RMSE was used to select the best value of mtry, which was mtry = 3 (Table 3.12) for a random forest with 500 trees. The model was applied to test data and produced a R-squared value of 0.817 and an RMSE of 14.483 (Figure 3.19).

Table 3.12: Results for tuning the value of mtry for random forest regression model with no N50 metric.

k	RMSE	R-squared	MAE
2	12.758	0.852	9.404
3	12.554	0.855	9.181
5	12.793	0.852	9.130

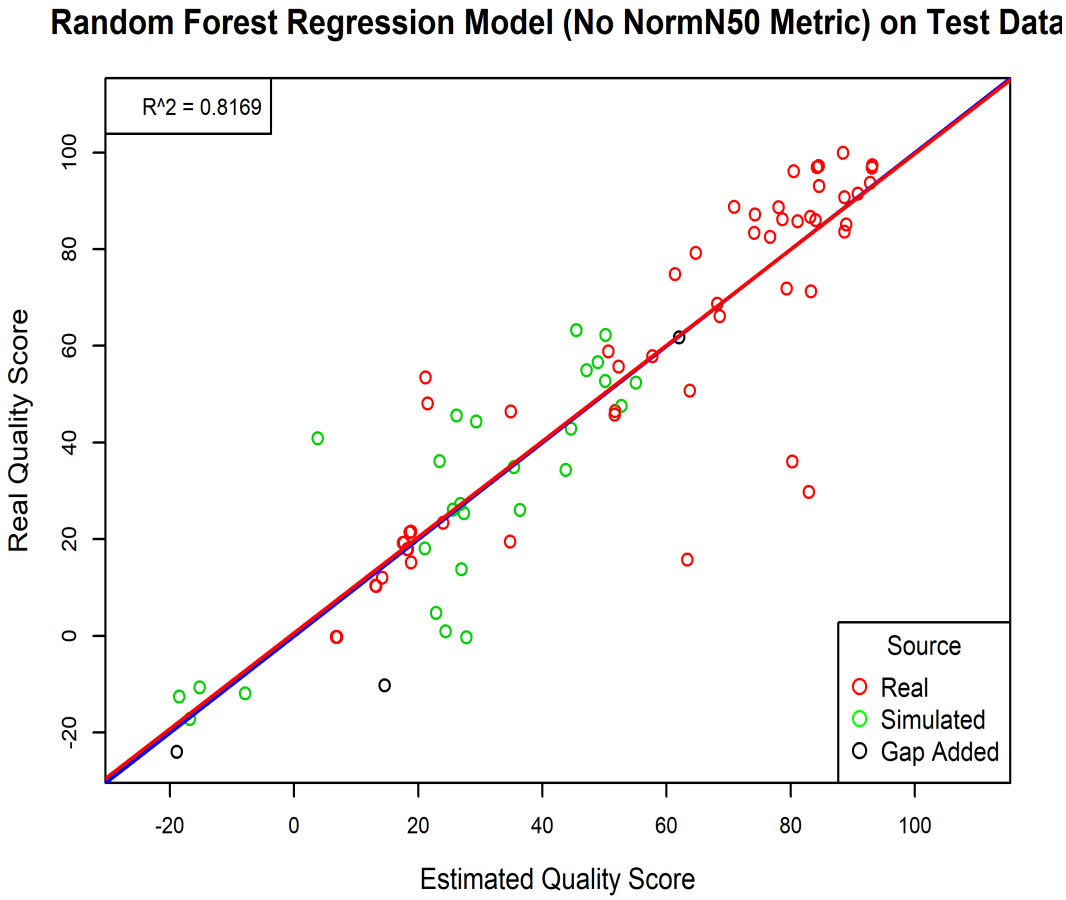

Figure 3.19: Performance of the random forest regression model without the normN50 metric on test data. The estimated quality scores are plotted against the reference-based quality scores of the test instances. A 100% accurate model would produce the blue line with an r-squared equal to 1. The line of best fit for the plotted data is shown as the red line and has an r-squared of 0.817.

Chapter 4

BIOINFORMATIC ANALYSIS OF CHINESE HAMSTER OVARY HOST CELL PROTEIN LIPASES

4.1 Preface

This section is adapted from MacDonald, Hamaker, and Lee, 2018 with permission (see Appendix C). We describe the identification of possible problematic lipases and the correction of several misassemblies and misannotations in CHO-K1 lipase sequences using the most recent CH genome, PICR [30] (described in Chapter 2). Overall, this chapter highlights several benefits of having a high-quality reference genome for CH and CHO cells. I carried out the identification and bioinformatics analyses of the potential problematic lipases and their gene-editing targets. Nathaniel Hamaker carried out Sanger sequencing to confirm the CHO-K1 *Lpl* and *Pnlip* gene family corrections as well as helped with the visualization of gene corrections.

4.2 Introduction

Chinese hamster ovary (CHO) cells are the preferred platform for biotherapeutic protein production. Monoclonal antibodies (mAbs) alone are predicted to reach global sales of 125 billion USD in 2020 [141] and are used to treat many oncological, immunological and cardiovascular diseases. During the production of therapeutic proteins by CHO cells, host cell proteins (HCPs) are also secreted by the cells. Certain HCPs, if not removed during subsequent purification processes, have been shown to cause immunogenic responses in patients [142] and others can shorten the shelf life of the final drug product through a variety of mechanisms including polysorbate degradation [143, 144, 145, 146]. HCPs, therefore, need be reduced to minimal levels, typically 1–100 ppm, in final mAb formulations [147]. While most HCPs are removed from the

therapeutic product during downstream purification steps, certain difficult-to-remove HCPs can remain [148]. Several types of lipases have been identified as problematic HCPs, especially regarding the stability of the mAb product.

Lipoprotein lipase (LPL) has been identified as a particularly difficult-to-remove impurity in CHO cell mAb production that possesses polysorbate 20 (PS-20) and polysorbate 80 (PS-80) degradation activity [143, 149]. PS20 and PS80 are surfactants often added to the drug product as protection from degradation during storage [150, 151]. It has been hypothesized that LPL is able to degrade PS20 and PS80 because polysorbates share structural similarities to triglycerides, the natural substrate of LPL. In particular, they share an ester bond which LPL hydrolyzes within triglycerides to form fatty acids and alcohol molecules [152]. Part of the reason LPL may be especially difficult to remove in a variety of processes producing a variety of products is that LPL has been shown to associate with multiple mAbs in protein A affinity chromatography and also to co-elute in non-affinity polishing columns used in subsequent steps of protein purification [148, 153].

Two other lipases have also displayed polysorbate degrading activity and have been identified in CHO cell-derived drug products. Group XV lysosomal phospholipase A2 (LPLA2 or PLA2G15) was found in the drug product of several mAb-producing cell lines at less than 1 ppm. Even at these low levels, LPLA2 was associated with the hydrolysis of PS20 and PS80 [154]. The rate of polysorbate hydrolysis was shown to be both time and concentration dependent.

Putative phospholipase B-like 2 (PLBL2 or PLBD2) is another difficult-to-remove HCP that has been shown to co-elute with several biotherapeutic antibodies during the protein A chromatography purification process [155]. PLBL2 has been associated with the degradation of PS-20 in a sulfatase drug product [144]. In addition, drug material used in Lebrikizumab clinical trials was found to contain 34-328 ng of CHO PLBL2 per mg of product, and approximately 90% of patients in the clinical trial developed an immune response against PLBL2 [142]. PLBL2 also displayed variable

expression during an extended culture of 136 days [149], a characteristic of difficult-to-remove HCPs because purification processes may not adequately remove the wide range of expression levels reached over time.

Genome editing techniques have been used to knock out a variety of different genes in CHO cell lines. For instance, clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein 9 (Cas9) has been used to knock out methyltransferase genes in CHO cells to stabilize therapeutic protein productivity [156] and a fucosyltransferase gene to prevent the fucosylation of the target biotherapeutic [157]. CRISPR/Cas9 has also been used to knock out a difficult-to-remove HCP impurity. A successful knock-out of *Lpl* using CRISPR/Cas9 was shown to decrease PS80 degradation by 41-47% percent and PS20 degradation by 44-57% [143]. Other genome editing techniques such as transcription activator-like effector nucleases (TALENs) and zinc-finger nucleases (ZFNs) have also been used to effectively knock out genes in CHO cells [158, 159, 160].

While unknown, it is possible that other lipases with similar enzymatic activity to LPL, LPLA2, and PLBL2 could result in polysorbate degradation and/or immunogenic responses if they exist in the final drug product. Here, we identified potentially problematic lipases based on an analysis of the CHO-K1 and Chinese hamster (CH) genomes, and protein sequence similarity to LPL, LPLA2, and PLBL2. Several misassemblies and/or misannotations in the sequences of CHO-K1 lipases were identified and corrected using the most recent CH genome [30], highlighting the importance of accurate and complete reference genomes. The corrected sequences were then examined to identify conserved regions that could be targeted to knock out multiple lipases simultaneously. We also compared the newly corrected CH/CHO-K1 lipase protein sequences to their human orthologs to understand the extent to which any of the lipases may be immunogenic in humans.

4.3 Methods

4.3.1 Protein and CDS alignments of LPL, PLBL2, and LPLA2 from various CH and CHO assemblies

Protein and mRNA sequences of LPL, PLBL2, and LPLA2 were extracted from CHO-K1 Refseq [50], 2013 CH RefSeq [27], and the updated PICR CH assembly [30], to compare sequence differences between CHO-K1 and CH, and to examine changes among different CH assembly versions. Protein alignment was done using MUSCLE [161] and mRNA alignment was done using ClustalO [162] using the default parameters. For mRNA alignments, only the coding sequence (CDS) regions from each transcript were used because untranslated regions (UTRs) are difficult to annotate correctly [163, 164]. An error in the CHO-K1 LPL protein sequence was identified and corrected using the MUSCLE alignment to the CH PICR, mouse, rat, and human orthologs. The corrected CHO-K1 LPL sequence and the original CHO-K1 sequences for PLBL2 and LPLA2 were used in further analyses. The RefSeq IDs for the CHO-K1 transcripts and proteins used in this project are listed in Table 4.1.

4.3.2 Identification of lipases similar to LPL, PLBL2, and LPLA2

An extensive list of lipase enzymes was compiled from searching EMBL-EBI's QuickGO database [165] with the GO term, lipase activity (GO:0016298). Corresponding protein sequences for the identified lipases were extracted from the PICR and CHO-K1 assemblies. BLASTP [166] was used to query LPL, PLBL2, and LPLA2 against this list to identify the most similar proteins with lipase activity. Hits with an E-value < 0.001 were further examined by full sequence alignment with the query (LPL, PLBL2, or LPLA2). For proteins with more than one hit with an E-value < 0.001, a phylogenetic tree of the protein sequences was created using the neighbor-joining algorithm within JalView [167] with PAM250 as the position specific matrix.

Table 4.1: List of all genes/proteins used from the RefSeq annotation of the CHO-K1 assembly, GCF_000223135.1.

Gene Name	Gene Symbol(s)	RefSeq Transcript	RefSeq Protein
Lipoprotein lipase	*Lpl*	XM_003499928.3	XP_003499976.1
Group XV lysosomal phospholipase A2	*Lpla2, Pla2g15*	XM_003504311.3	XP_003504359.1
Lipase H (isoform 2)	*Liph*	XM_016976547.1	XP_016832036.1
Endothelial lipase	*Lipg*	XM_007646318.2	XP_007644508.2
Hepatic triacylglycerol lipase	*Lipc*	XM_003495063.2	XP_003495111.2
Lipase I	*Lipi*	XM_007640048.2	XP_007638238.1
Phospholipase A1 member A	*Pla1a*	XM_007649760.2	XP_007647950.1
Pancreatic lipase related protein 1	*Pnliprp1*	XM_007655147.2	XP_007653337.1
Pancreatic lipase related protein 2	*Pnliprp2*	XM_016963761.1	XP_016819250.1
Pancreatic triacylglycerol lipase	*Pnlip* (LOC100751227)	XM_003515145.3	XP_003515193.2
Phosphatidylcholine-sterol acyltransferase	*Lcat*	XM_003504283.3	XP_003504331.1
Phospholipase B-like 1b	*Plbl1* (*Plbd1*)	XM_003504424.2	XP_003504472.1
Phospholipase B-like 2	*Plbl2* (*Plbd2*)	XM_003510812.3b	XP_003510860.1

4.3.3 Correction of lipase protein and gene sequences

Errors in the CHO-K1 protein sequences were detected by aligning each protein sequence with their orthologs in human, mouse, rat, and CH PICR. The human, mouse, and rat sequences were extracted from UniProt [168] release 2018_1. Once an error was identified, the type and location of the error was characterized by examining the transcript alignment against mouse and PICR using SnapGene (www.snapgene.com). Most errors involved a missing or incomplete exon at the 5' end of the gene. In these cases, the 'correct' exon from mouse that corresponded with the erroneous exon in

CHO-K1 was realigned to the CHO-K1 gene to correct the CHO-K1 gene annotation. Realignment of the newly modified CHO-K1 protein sequence against human, mouse, rat, and CH PICR orthologs was done to validate the correction.

The correction for CHO-K1 PNLIPRP2 was more complicated and benefited significantly from the updated PICR genome. Alignment to PNLIPRP2 mouse, rat, and human protein orthologs showed that the protein sequence for CHO-K1 PNLIPRP2 was missing a segment of amino acids at the 5' end. Visualization of *Pnliprp2* on its scaffold, NW_003617188.1, showed that the gene was incomplete because it was located directly on the end of the scaffold. PICR was used to find the neighboring gene, *Pnliprp1* (XM_007655147.2), which was located on CHO-K1 scaffold, NW_003617412.1. The two scaffolds were realigned to the longer PICR scaffold, picr_24, to confirm that these two scaffolds should be merged in the CHO-K1 genome. The mouse transcript (NM_011128.2) was then aligned to determine that the entire first exon of *Pnliprp2* was located in the NW_003617412.1 scaffold. Finally, the sequence in CHO-K1 that aligned to the mouse exon was used to correct the gene and protein sequence for PNLIPRP2. Exact boundaries were identified using the mouse and golden hamster (XM_005085339.3) exons.

4.3.4 Determination of expression levels of lipases of interest

Each gene was checked for expression in CHO-K1 cells using data from GEO: GSE75094 [169]. The expression levels were visualized on the 'CHO-K1 mRNA expression data' browser on CHOgenome.org [170] using the FPKM (Fragments Per Kilobase of transcript per Million mapped reads) and SAM (Sequence Alignment/Map) coverage tracks. Genes that appeared with any amount of expression in CHO cells were examined further.

4.3.5 Determination of conserved regions in each grouping of lipases

Conserved regions among each of the three groups of BLAST hits (one group per LPL, LPLA2, and PLBL2) were located from the protein alignments. The DNA

sequences underlying the conserved regions were examined as well and used as the query in BLASTn to search against the non-redundant sequence set for CH species ID (10029) to identify possible off-target effects of using the conserved regions as knock-out targets. The PICR assembly was also queried to ensure there were no additional hits when assembly improvements were considered.

4.3.6 Examining the immunogenicity potential of similar lipases

Human protein sequences for LPL, LPLA2, and PLBL2 were extracted from UniProt release 2018_1 and aligned against the corresponding CHO-K1 and PICR protein sequences to calculate percent identity. The percent identity of PLBL2, which is known to be immunogenic, was then used to select a threshold of 80% identity to assess the possible immunogenicity of the BLAST hits of LPL, LPLA2, and PLBL2. The percent identity with their human orthologs were determined, and if the percent identity was lower than that of PLBL2 and its human ortholog, it was flagged as having the potential to cause immunogenic responses.

4.4 Results

4.4.1 Sequence differences among the CH and CHO-K1 Assemblies

Alignments of the protein sequences across the different CH and CHO-K1 assemblies for LPLA2 and PLBL2 showed very little difference. However, the protein sequence for LPL isoform X1 from CH RefSeq has 11 additional amino acids at the 3' end, but the X2 isoform is the same length as the PICR protein (Figure 4.1). The protein sequence of LPL from CHO-K1 is missing one amino acid as shown by the gap in the alignment at position 24, which is then followed by an unknown amino acid at position 25 (Figure 4.1). These errors are also reflected in the mRNA coding sequence alignments (Figure 4.2), which show that the CHO-K1 LPL sequence is missing four guanines. The existence of these nucleotides in CHO-K1 LPL were confirmed by Sanger sequencing. If an sgRNA for a CRISPR/Cas9 knock-out was designed to target this region of Lpl based on the CHO-K1 genome sequence alone, it would be missing

Figure 4.1: Alignment of LPL protein sequence from CHO-K1 RefSeq, 2013 CH RefSeq (isoforms X1 and X2), and the updated CH genome, PICR. Positions 24–25 are in red to highlight the error in the CHO-K1 LPL sequence. The difference between CH RefSeq LPL isoform X1 and X2 is shown in purple.

four nucleotides. This would greatly decrease the binding affinity of the sgRNA to this region in the gene and thus, the knock-out efficiency.

There were no differences in the protein sequences for LPLA2 among the different assemblies (Figure 4.3). Three unknown amino acids exist in the PLBL2 2013 CH RefSeq sequence at base positions 43-46, but alignment to the CHO-K1 and CH PICR sequences suggest that these do not actually exist (Figure 4.4).

Figure 4.2: The beginning (positions 1-125 in *Lpl* CHO-K1) of the coding sequence alignment of *Lpl* from CHO-K1 RefSeq, 2013 CH RefSeq, and PICR.

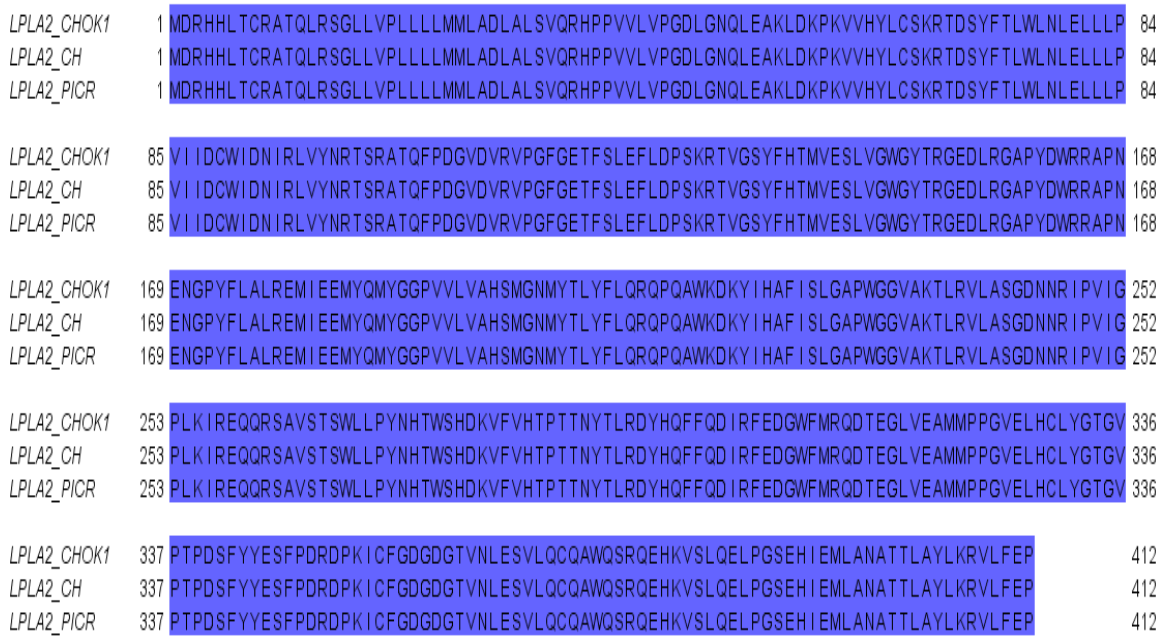

Figure 4.3: Protein alignment of LPLA2 from the CHO-K1 RefSeq, 2103 CH RefSeq, and PICR annotations.

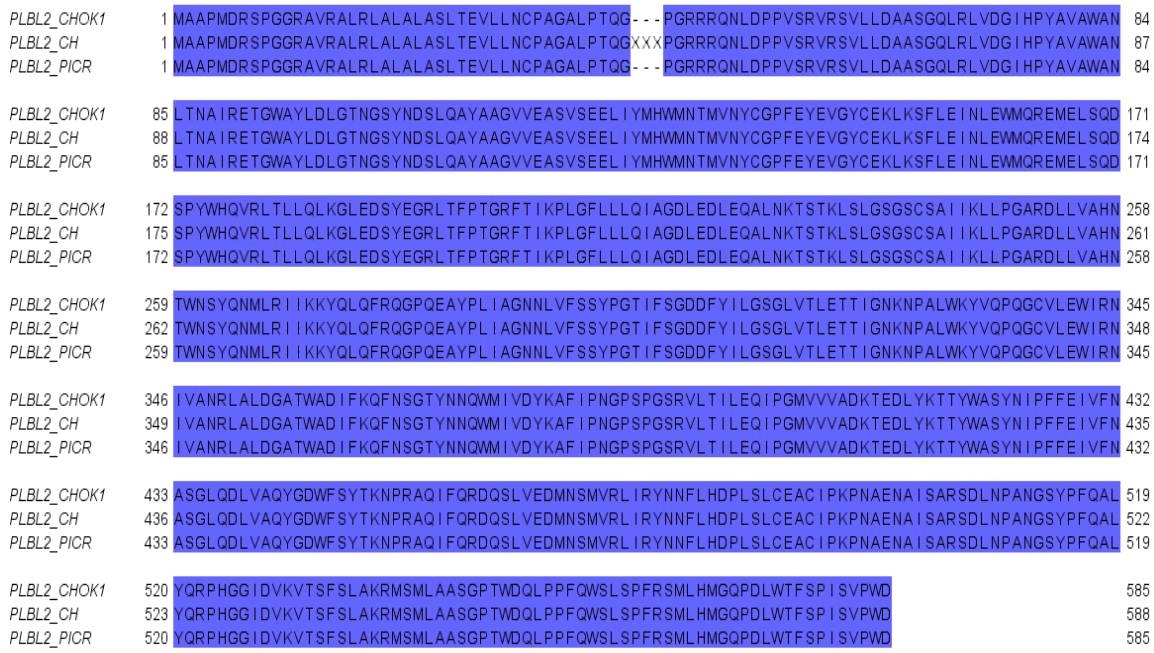

Figure 4.4: Protein alignment of PLBL2 from the CHO-K1 RefSeq, 2013 CH RefSeq, and PICR annotations.

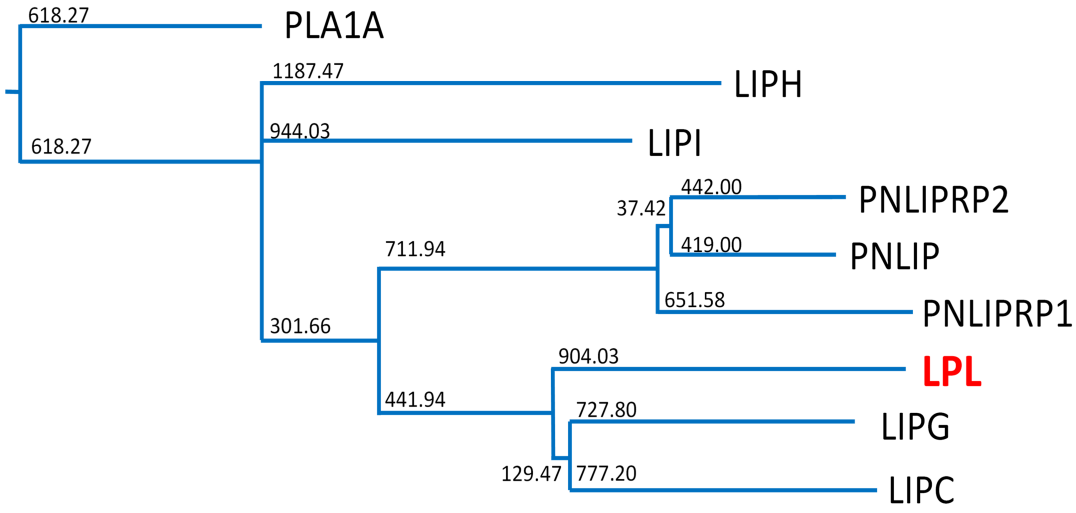

Figure 4.5: Phylogenetic tree derived from the multiple sequence alignment of LPL to its significant BLASTp hits using the neighbor joining algorithm (PAM250) in JalView. Distances of each branch are labeled.

4.4.2 Identification and sequence correction of HCPs related to LPL

BLASTP hits with significant alignment (E-value < 0.001) to LPL from CH and CHO-K1 included LIPC, LIPG, LIPH, LIPI, PLA1A, PNLIP, PNLIPRP1, and PNLIPRP2. All of these lipases belong to the pancreatic lipase gene family [171], which is composed of members with triglyceride lipase activity (EC 3.1.1.3) and the closely related lipoprotein lipase (EC 3.1.1.34) [172]. LPL is most related to the LIPG (endothelial lipase) and LIPC (hepatic lipase) proteins, and then to the PNLIP proteins (pancreatic lipases) (Figure 4.5). The similarity of these eight proteins to LPL at the sequence level suggests that they could potentially degrade PS20 and PS80.

Five of these genes (*Lipi*, *Liph*, *Pla1a*, *Pnliprp2*, and *Pnliprp1*) had evidence that supported their expression in CHO-K1 cells from the CHOgenome.org browser. *Pnliprp1* and *LipH* have also previously been identified as differentially expressed in sodium butyrate treated CHO cells when compared to non-treated cells [173]. In addition, a higher than 1.5 fold change in expression of *Pnliprp1* was observed between a low-producing and a high-producing cell line [174]. Three of the five genes (*Pnliprp2*,

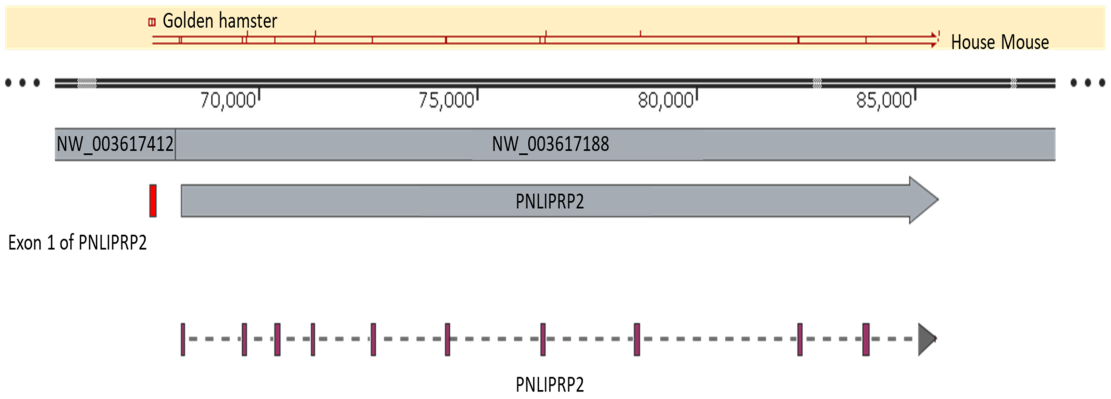

Figure 4.6: View of the *Pnliprp2* gene split over two scaffolds, NW_003617188 and NW_003617412, in the CHO-K1 assembly. This is confirmed with alignment of the mouse *Pnliprp2* gene (NM_011128.2) to the scaffolds. Alignment to the first exon in golden hamster *Pnliprp2* (XM_005085339.3) helped to determine the exact boundaries of the exon in CHO-K1. Alignments are shown at the top of the figure in red.

Pnliprp1, and *Lipi*) had errors in their annotations for the CHO-K1 genome, which could be seen in the multiple sequence alignment against the corresponding protein in mouse, rat, human, and CH PICR. All three genes had a missing exon at the 5' end. *Lipi* was missing the first exon, most likely because the 5' end of the gene overlapped with another gene, Rbm11, located on the antisense DNA strand, which may have complicated the annotation. The CHO-K1 *Pnliprp2* did not contain the first exon because the gene was split over two scaffolds in the CHO-K1 genome. The longer scaffold length in the PICR assembly allowed the two CHO-K1 scaffolds to be merged and the gene to be resolved (Figure 4.6). Not only was PICR able to correct the *Pnliprp2* gene, but the longer scaffold length in PICR enabled the PNLIP family of lipases to be ordered within the CHO-K1 genome. This section of genes was originally split over three scaffolds (Figure 4.7). It is unclear why *Pnliprp1* had a missing exon in its annotation. Sanger sequencing confirmed the joining of the scaffolds.

Once the sequence errors were resolved, the five genes similar to *Lpl* and expressed in CHO-K1 cells were aligned (Figure 4.8). The alignment shows that the six

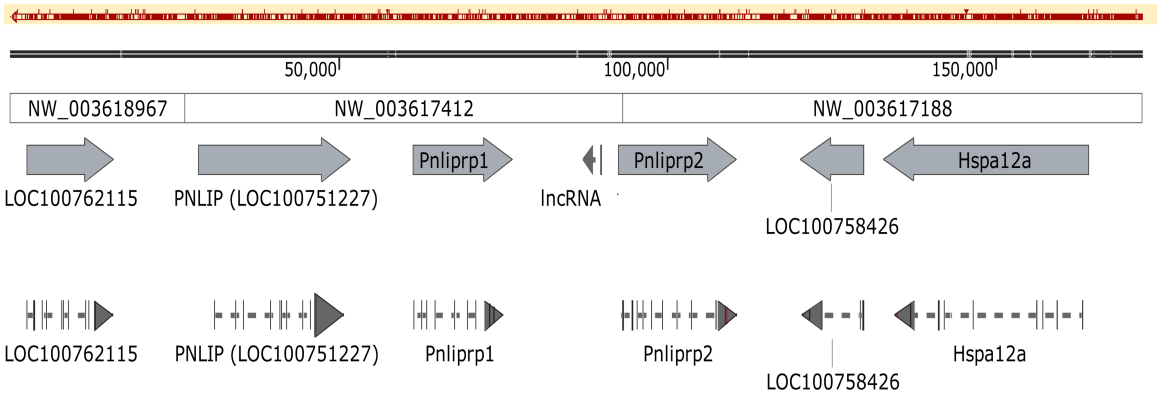

Figure 4.7: View of the *Pnlip* lipase genes, positioned and ordered in a single superscaffold in CHO-K1. Part of the PICR scaffold, picr_24 (2,501,648–2,752,160), is aligned above in red, showing the overlap across the three CHO-K1 scaffolds. The LOC100762115 gene is an uncharacterized relative of the *Pnlip* gene.

proteins all share the same active site residues which make up the well-known catalytic triad [175]. Regions around the first two active site residues are well conserved, particularly the 'RITGLDP' peptide (highlighted in Figure 4.8). This peptide could provide a target location to simultaneously knock down, knock out, or purify the potentially troublesome HCPs. The underlying DNA sequence, however, is not well conserved (Figure 4.9) and therefore, multiple targets will need to be designed and tested for their off-target effects for knock-down and knock-out studies.

4.4.3 Identification of HCPs related to PLBL2

PLBL2 only had a single significant BLASTp hit which was PLBL1. They share 36.33% identity and have 51.76% positive scoring amino acid replacements. Alignment against mouse, rat, and human suggested that there were no errors in either CDS or protein sequence. Alignment of PLBL2 and PLBL1 show that they share the same set of active site residues where five of six align exactly (Figure 4.10). The active sites were identified from the annotation of human PLBL2 and PLBL1 proteins

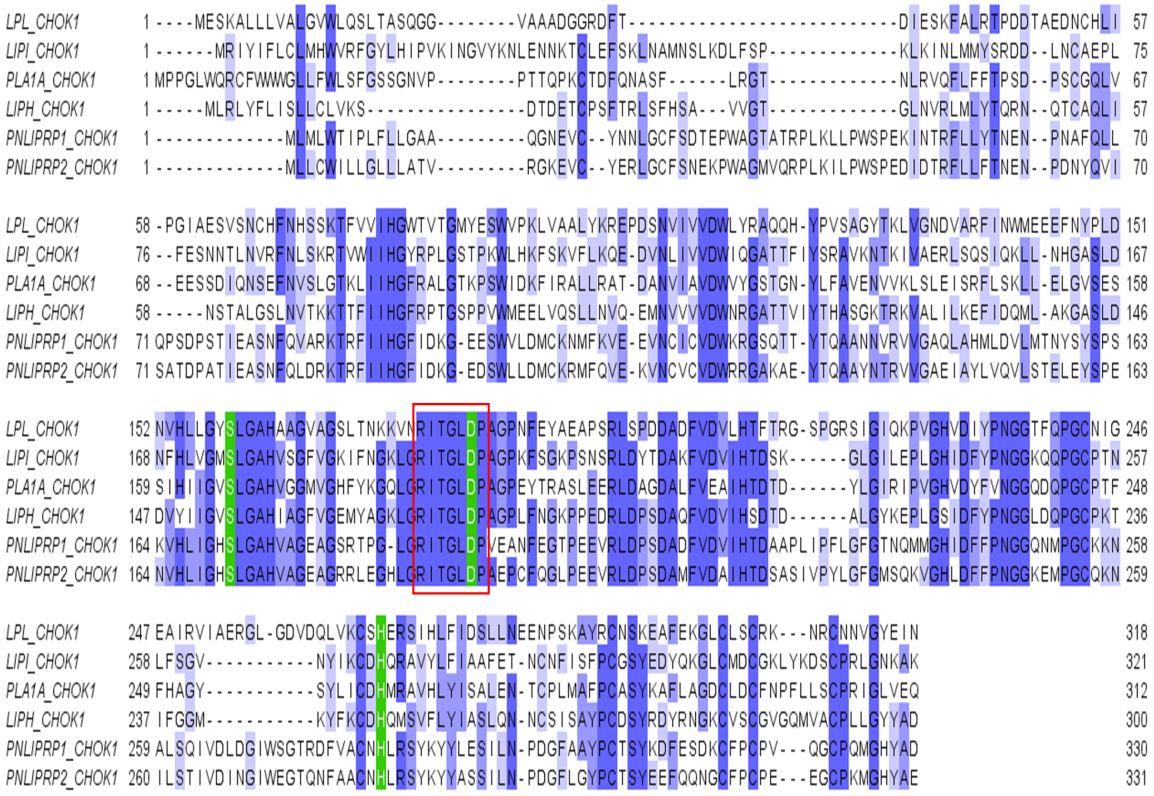

Figure 4.8: Protein sequence alignment of LPL, positions 1–318, to the five similar lipases expressed in CHO-K1 cells. Active sites are highlighted in green (LPL positions: 159 S, 183 D, and 268 H). The potential target conserved peptide, 'RITGLDP', is highlighted within the red box.

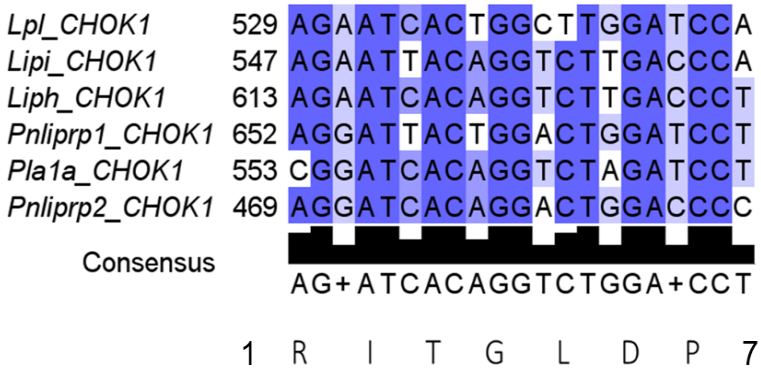

Figure 4.9: The CDS from the various LPL related lipases that code for the conserved 'RITGLDP' peptide.

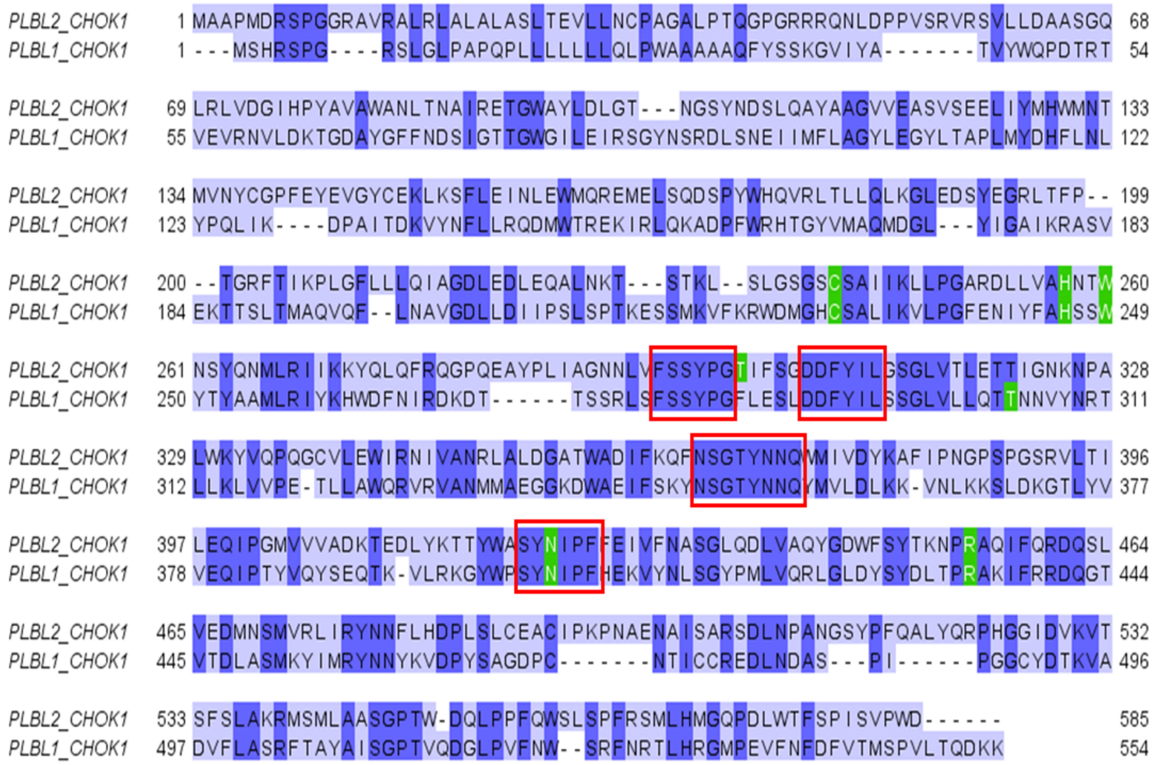

Figure 4.10: Protein sequence alignment of PLBL2 and PLBL1. Active sites (PLBL2 positions: 240 C, 257 H, 260 W, 301 T, 423 N, 454 R) are highlighted in green. The four conserved peptides 'FSSYPG', 'DDFYIL', 'NSGTYNNQ', and 'SYNIPF' are highlighted within the red boxes

(http://genomewiki.ucsc.edu/index.php/Phospholipases_PLBD1_and_PLBD2). However, the exact function or substrates of PLBL2 and PLBL1 are unknown.

Four conserved potential target regions between PLBL2 and PLBL1 are the peptides 'FSSYPG', 'DDFYIL', 'NSGTYNNQ', and 'SYNIPF' (Figure 4.10). The 'DDFYIL' is the most conserved on the DNA level (Figure 4.11a) and can be extended to contain the 'NGG' PAM sequence (Figure 4.11b). This PAM sequence is necessary in the single guide RNA (sgRNA) target site for the most commonly used type of CRISPR/Cas9, *Streptococcus pyogenes*. The extended target site, N'-GATGACTTCTACATCCTNNGCAG-C', also appears to have no off-target hits with less than seven base mismatches when querying the CHO-K1, 2013 CH RefSeq, and

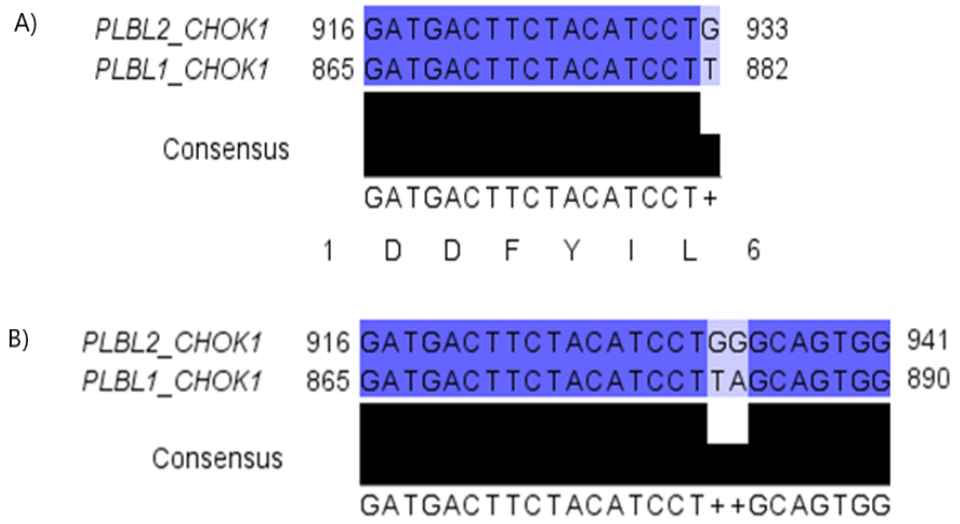

Figure 4.11: Conserved CDS for CHO-K1 PLBL2 and PLBL1 lipases. A) The CDS for the 'DDFYIL' peptide that is conserved in the CHO-K1 PLBL2 and PLBL1 lipases, B) the CDS sequence can be extended in the alignment to contain a 'NGG' PAM sequence at the 3' end that is required for the sgRNA target site of the most commonly used type of CRISPR/Cas9 system.

PICR genome assemblies. However, it should be noted that no evidence to date has been found that suggests the *Plbl1* gene is expressed in CHO-K1 cells.

4.4.4 Identification of HCPs related to LPLA2

The CHO-K1 LPLA2 protein shared significant similarity with the CHO-K1 LCAT (Lecithin-Cholesterol Acyltransferase) protein. This similarity has been described previously: LPLA2 and LCAT are closely related acyltransferases [176] and are members of the $\alpha - \beta$ hydrolase family [177, 178]. LPLA2 transfers fatty acids from glycerophospholipids to lipophilic alcohols, while LCAT transfers fatty acids from glycerophospholipids to cholesterol [176]. The alignment between LPLA2 and LCAT reflects this functional similarity as they share 48.8% protein sequence similarity (68.6% positive scoring amino acid replacements) and the same active site residues making up the catalytic triad (Figure 4.12). LCAT appears to be expressed in CHO-K1 cells

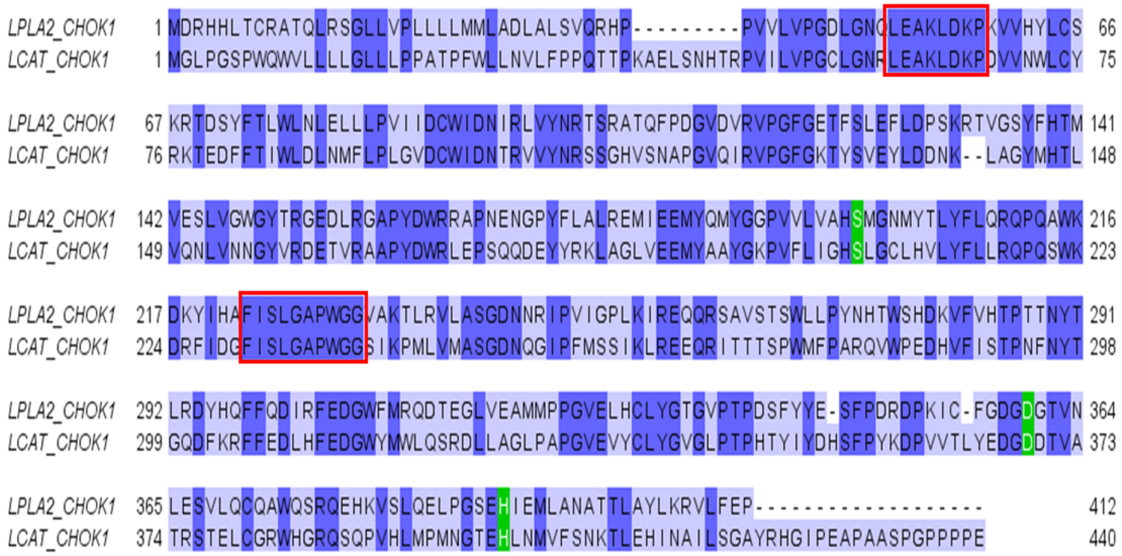

Figure 4.12: Protein sequence alignment of LPLA2 and LCAT. Active sites (LPLA2 positions: 198 S, 360 D, 392 H) are highlighted in green. The conserved peptides 'LEAKLDKP' and 'FISLGAPWGG' are highlighted within the red boxes.

on the CHOgenome.org RNA browser and has been previously described as differentially expressed between sodium butyrate treated and non-treated CHO cells [173]. The peptide 'LEAKLDKP' shared between LPLA2 and LCAT could be used as a gene editing target to knock out the expression of both (highlighted in Figure 4.12). This peptide is also well conserved at the DNA level (Figure 4.13) and no off-target effects were found in the CHO-K1, CH, and PICR genomes, using the sequence N'-CTNGAAGCNAAGCTGGANAAACCA-C' as the BLASTn query. Another potential target is the 'FISLGAPWGG' peptide, but this is not as well conserved at the DNA level.

4.4.5 CHO-K1 lipase similarity to their human orthologs

It has been hypothesized that the more dissimilar the protein sequences of CH are to their human orthologs, the more likely the CH protein can cause an immunogenic response in a patient [179, 180]. LPL and LPLA2 have an identity of 93.47%

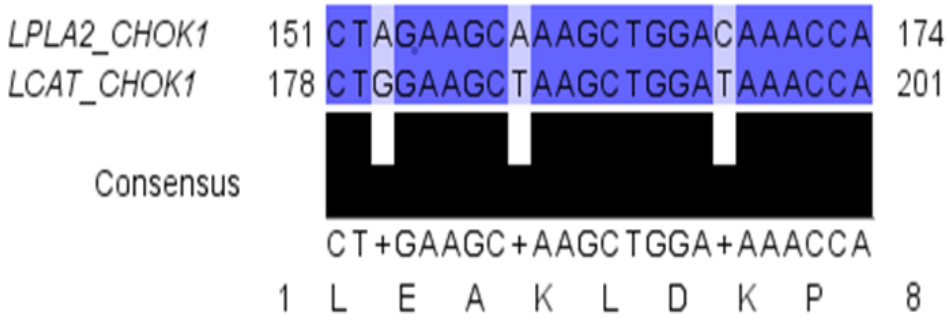

Figure 4.13: CDS for the 'LEAKLDKP' peptide that is well conserved in the CHO-K1 LPLA2 and LCAT lipases.

and 88.59%, respectively, with their human orthologs. There has been no evidence of immunogenic responses against either LPL or LPLA2. PLBL2 is more different from its human ortholog with an identity of 78.95% and it has been shown to cause immunogenic responses in patients [142]. Using 80% identity as the threshold cutoff, we compared the sequences of the other related lipases identified here to their human orthologs. Three proteins did not meet the threshold: LIPI (62.18% identity), PNLIPRP2 (76.97% identity), and PLBL1 (76.35% identity). These differences suggest that these proteins may be more likely to cause immunogenic responses than the other lipases if not removed during the purification process. PLA1A and LIPH isoform X2 were just above the cutoff of at 80.92% and 81.92% identity, respectively. LCAT (89.32%) and PNLIPRP1 (86.30%) were above the threshold.

4.5 Discussion

Sequence errors are a reality for all draft genome assemblies based on existing technology. *De novo* assemblies of human genomes built from short sequencing alone have been shown to be missing millions of bases of duplicated sequence and common repeats, and missing thousands of coding exons [181]. Many of these errors are corrected in later rounds of resequencing which reach an adequate coverage depth, producing more finished assemblies [182]. Reference genomes need to be improved beyond the

draft status to avoid making incorrect inferences in reference-guided studies. For instance, accurate and complete assemblies and annotations are key to perform effective genetic engineering techniques.

Here, we show the advantage of having a significantly higher quality reference genome for CHO cell lines. The new PICR reference assembly enabled the identification and the correction of errors in the sequence of the difficult-to-remove host cell protein, LPL, and the similar PNLIPRP2 protein. The PICR assembly, along with the accurate mouse genome, enabled us to correct two other proteins similar to LPL, PNLIPRP1 and LIPI. The sequences underlying these annotation errors have been corrected in the PICR sequence, indicating that a new NCBI RefSeq annotation for PICR will have the correct sequences and coordinate boundaries for the LPL, PNLIPRP2, PNLIPRP1, and LIPI genes/proteins. Correct gene and protein sequences allowed us to identify significantly similar lipases to three known difficult-to-remove HCPs: LPL, PLBL2, and LPLA2. Our findings are summarized in Table 4.2. Functional and sequence similarity suggest that these related lipases have the potential to cause similar issues if present in the final drug product. Within each grouping of lipases, conserved regions were identified that could serve as targets for the mitigation of the negative impacts of these lipases. An sgRNA could be designed to target the DNA that codes for the 'LEAKLDKP' conserved peptide in LPLA2 and LCAT, knocking out both genes simultaneously. Even if the DNA sequences underlying the conserved peptides are not identical among the genes of interest, multiple different guide RNAs can be applied to target the same location. Multiplexing of CRISPR/Cas9 has previously been successful in CHO cells to knock out three genes simultaneously [157, 183]. It has also been able to target up to 62 retroviral elements in porcine kidney cells [102] and primary purified porcine cells [184]. Targeting the same region would provide consistency in the knockouts, removing any positional impacts of where the inserted or deleted nucleotide(s) occur. For instance, targeting the LPL group of lipases at the 'RITGLDP' peptide identified here would knock out all lipases directly near their second active site residue.

While this approach will prevent the catalytic activity of the targeted lipases,

Table 4.2: Summary of findings for LPL, PLBL2, LPLA2 including the related lipases, conserved peptides, and whether the comparison to the corresponding human ortholog suggests the similar lipases could cause an immunogenic response in patients. Proteins shown to be expressed in CHO-K1 cells are highlighted in bold.

Known HCP	Similar lipases	Conserved peptides (N' to C')	Does protein sequence identity to human ortholog suggest immunogenicity?
LPL	LIPC, LIPG, **LIPH**, **LIPI**, **PLA1A**, PNLIP, **PNLIPRP1**, **PNLIPRP2**	RITGLDP	Yes for LIPI, PNLIPRP2
PLBL2	PLBL1	DDFYIL, FSSYPG, NSGTYNNQ, SYNIPF	Yes
LPLA2	**LCAT**	LEAKLDKP, FISLGAPWGG	No

it is important to note frame-shift mutations from knocking out the lipases can still result in expression of some peptides of the target protein. In theory, these fragments could still have the ability to bind to the product and/or cause immunogenic effects if not purified from the final drug product. A full gene deletion approach as described in Zheng et al. [185] could mitigate this concern. This approach consists of targeting two sequences on either side of the gene of interest, which when cut and repaired can remove the entire sequence between the targets. Multiple, simultaneous gene deletions using a combination of CRISPR/Cas9 and CRISPR/Cpf1 systems have been carried out in CHO cells with no deletion size limitations within 2-150 kb [186]. Thus, it is feasible that sets of several problematic lipase genes could be fully deleted using this multiplex approach.

In addition, alignment of protein sequences to human orthologs can provide insights into the possible immunogenicity of an HCP. Here, we discovered that LIPI, PNLIPRP2, and PLBL1 (if expressed) are the most different from their human orthologs

and may have the most potential to be immunogenic if not removed during purification processes. This work shows the immense value of having access to the highest quality reference genomes and annotations when carrying out genetic engineering studies.

Chapter 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

The Chinese hamster genome was reassembled using existing Illumina data and new PacBio sequencing data through a variety of methods. Manual quality ranking of the candidate assemblies determined that the best assembly was PICR, which was created by merging multiple CH assemblies using a set of contigs assembled from PacBio sequence data as the starting assembly for the merging process. The PICR assembly is a significant improvement over the 2013 CH assemblies, consisting of 43-fold fewer scaffolds and a 13-fold higher N50 length than the 2013 RefSeq CH assembly. In addition, more than 95% of the gap sequence in the CH genome assembly was filled, facilitating the detection of previously unknown CHO cell mutations. This greatly improved reference genome assembly will be a highly beneficial resource for CHO-related research.

In addition, a novel tool that comprehensively evaluates genome assemblies was created to facilitate reference genome selection. The reference-independent quality scoring system in EvalDNA enables comparison of assemblies for a single species and provides insights into how complete, continuous, and accurate a genome assembly is in relation to assemblies from other species. Another benefit of using EvalDNA for accuracy assessment over other existing tools is that quality assessment based on alignments to a reference are biased towards the reference sequence because the reference is assumed to be correct. EvalDNA results, since they are independent of a reference assembly, are not biased in this regard.

EvalDNA uses a model created from supervised machine-learning methods to estimate an assembly's quality score. A model, specifically created to assess mammalian

assembly quality, was built from training data consisting of quality metrics from publicly available mammalian chromosome assemblies, including those from the human, mouse, and rat reference genomes. This random forest model was able to accurately (r-squared of 0.86) estimate the reference-based quality scores of the test data.

EvalDNA and this mammalian model was tested on the CH genome assemblies as well as assemblies of human chromosome 14 from the GAGE study. The results of EvalDNA for the CH genome assemblies agreed with the results from two other reference-independent quality assessment tools (ALE and FRCbam). The highest (PICR) and lowest (CSA) scoring CH assemblies were also in agreement with the manual rankings described in Chapter 2. The applicability of EvalDNA to assess assembler output was also demonstrated by scoring various human chromosome 14 assemblies. Discrepancy in the ranking of assemblers among EvalDNA, ALE, and FRCbam suggested that the mammalian model weighs accuracy and completeness over continuity more so than the other quality assessment tools.

Overall, a tool, such as EvalDNA, that integrates quality metrics in a consistent manner across different species could become the new standard for the assessment and comparison of genome assembly quality. Directly comparable results would enable quick selection of reference assemblies for newly-sequenced organisms as well as quick assessment of quality in comparison to "gold-standard" reference genomes. In addition, a standardized meaning of high quality in the field of genome informatics is still being established as sequencing and assembly technology continues to improve [187, 188, 40]. The single comprehensive quality score produced by EvalDNA per genome assembly could become an easy way to define, or at least partially define, high quality.

Finally, several benefits of having an available, high-quality reference genome assembly for genetic engineering target selection were demonstrated. The newly established CH and CHO cell reference assembly PICR (RefSeq 2018) was able to correct several misassemblies/misannotations in lipase gene sequences from the CHO-K1 assembly. Specifically, PICR enabled the correction of sequence errors in the *Lpl*, *Pnliprp2*, *Pnliprp1*, and *Lipi* genes. The corrected sequences were then used to identify

target sites for the simultaneous knock-out of lipases with possible enzymatic activity similar to three known problematic lipases, LPL, LPLA2, and PLBL2.

5.2 Recommendations for Future Work

5.2.1 Further improvement of the Chinese hamster genome assembly

While the PICR CH genome assembly is significantly more continuous than the 2013 CH RefSeq assembly, PICR has yet to reach the highest level of continuity found in the best mammalian reference assemblies. The final PICR (2018 RefSeq) assembly is split into 1,830 scaffolds and 4,825 contigs, a huge improvement over the 52,710 scaffolds and 218,862 contigs of the 2013 RefSeq assembly. However, more continuous mammalian assemblies have been built. For example, the reference assembly for human is split into 472 scaffolds and 998 contigs, and the reference assembly for mouse consists of only 162 scaffolds and 605 contigs. Human and mouse scaffolds have also been ordered and orientated on a chromosome-scale, a current shortcoming of the PICR genome.

There are several methods that could be used to gain more information about the order, orientation, and distance of contigs/scaffolds in regards to one another, and thus, enable the development of an even more continuous CH assembly. One method is to use data from optical mapping of the CH genome to combine and position contigs. Optical mapping involves digesting long DNA molecules (>100 kbp long) by restriction enzymes. The DNA molecules are then flourescently stained and imaged to enable the estimation of distances between cut-sites [189]. The resulting optical map of fragment sizes can be compared to fragment sizes from a simulated digestion of the genome of interest created using the known cut-site for the restriction enzyme. Each resulting fragment size in the simulated digestion should match a fragment size in the optical map, providing the fragment's location in the genome [190]. However, errors are often found in optical maps including errors in the estimation of fragment size, missing or additional cut-sites, or even missing fragments [189].

A more recently developed approach uses Hi-C sequencing data, which consists of chromatin interactions within and between chromosomes, to scaffold contigs. Genomic regions on the same chromosome are found to interact more often than regions on different chromosomes. In addition, the probability of an interaction occurring decreases exponentially as the distance between the regions on the chromosome increases [191]. Therefore, the amount of interactions between regions of a genome can provide information on the organization of contigs and scaffolds [192, 193].

Hi-C data has been used in combination with other scaffolding techniques to achieve the most continuous assemblies to-date. Hi-C and optical mapping data was used with long-read sequence data to assemble the goat (*Capra hircus*) genome [194]. The resulting assembly is split into 31 chromosome-sorted scaffolds, where chromosomes 1-29 are made up of a single scaffold and chromosome X is made up of two scaffolds. The addition of Hi-C and long-read sequencing data enabled about a 20-fold improvement in continuity over the previous goat genome assembly, which was built using short-read sequencing and optical mapping data [195]. More recently, Hi-C with Chicago [196], a form of Hi-C where chromatin interactions can occur up to several 100 kbp apart, was used to assemble the water buffolo (*Bubalus bubalis*) genome, reaching one scaffold per chromosome [197].

5.2.2 Improvement of EvalDNA

The accuracy of the EvalDNA mammalian model described in Chapter 3 could be increased by adding more chromosome instances to the training data. These chromosomes could be from other well-studied mammalian genomes or be modified versions of the currently used chromosomes through gap or error simulation. The accuracy could also possibly be improved by adding more metrics to the model. These metrics could be derived from the mapping of long reads back to the assembly of interest. For instance, one type of metric that may improve the model would be one that reflects the amount of collapsed or incorrectly assembled repeats. Currently, REAPR's collapsed repeat metric, derived from short-read mapping, is not included in the model because

it did not significantly correlate with the reference-based quality score. Because identifying and resolving misassembled repeats in a genome assembly is easier if longer reads are available [198], a metric representing misassembled repeats derived from long-read mapping may better correlate with the assembly quality score than one derived from short-read mapping. The importance of each metric in estimating the quality score should be confirmed using feature selection methods before it is added to the model.

Another future goal to enhance EvalDNA is to extend its capability to score shorter assembled sequences such as scaffolds. Initial examination of the current mammalian model on scaffolds showed that the model performed reasonably well (similar to its performance on chromosomes) for scaffolds 5 Mbp or longer with up to 10% errors simulated. However, the accuracy metrics for the model needed to be derived from mapping only the reads that met a stringency threshold of 0.75 (75% or more of the bases in a read need to match the reference sequence). Performance of EvalDNA on shorter scaffolds may be improved by creating a scaffold-specific model for EvalDNA that is trained on scaffolds from mouse, rat, and human assembly builds and where a read mapping stringency of $\sim$0.75 is used to calculate the REAPR metrics.

In-depth examination is also needed to determine which taxonomic groups the current mammalian model can accurately score. Ideally, a model would be able to accurately score genome assemblies from species across different taxonomic groups to enable quality comparison among all genome assemblies. In addition to assessing the mammalian model's performance on mammalian genome assemblies, we briefly examined its applicability to plant genome assemblies. EvalDNA with the mammalian model was used to score several rice (*Oryza sativa*) genome assemblies and provided reasonable scores in comparison to the mammalian assemblies and with respect to current knowledge of the assemblies. Scoring more non-mammalian assemblies as well as completing error simulation tests, similar to those done for the Chinese hamster genome, could help determine EvalDNA's applicability to other species.

5.2.3 Simultaneous knock-out of CHO cell lipases

The identification of conserved genetic engineering target sites in possibly problematic lipases described in Chapter 4 facilitates the simultaneous knock-out of these genes. There are a variety of gene-editing tools that could be used, including CRISPR-systems, zinc finger nucleases, and TALENs, each with their own set of target design requirements. Here, we focused on identifying regions within lipase genes that could be targeted by sgRNAs of CRISPR-systems. For lipases with conserved sites at the DNA level, one sgRNA could be designed to simultaneously guide CRISPR-systems to each of those genes. For instance, we found that the peptide "LEAKLDKP" shared between the *Lpla2* and *Lcat* genes is well conserved at the DNA level with a sequence of 5'-CTNGAAGCNAAGCTGGANAAACCA-3' ('N' bases are not conserved). We also found no potential off-target effects for this site in the CHO-K1, CH, and PICR genomes. However, future experiments are required to assess the impact of the non-conserved bases in the target site on knock-out efficiency.

For lipases with conserved sites only at the protein level, a multiplexed CRISPR-system could be designed to contain more than one sgRNA. Each sgRNA would target the underlying DNA sequence of the conserved protein region in each lipase. Multiplexing has successfully been used in CHO cells to knock out three genes simultaneously [141, 183] and it may be beneficial to apply a similar approach to one or more of the groups of lipase genes identified in Chapter 4 to create novel cell lines. Each target site would need to be searched against the CHO-K1 and CH genomes to assess to possibility of off-target effects. In addition, there are a variety of methods available to form the array of sgRNAs, which differ with respect to the number of sgRNAs that can be expressed, editing efficiency, and cloning efficiency [199]. The best method for knocking out each group of CHO cell lipases would need to be investigated.

5.3 Concluding Remarks

The work presented here provides the biomanufacturing community with a significantly improved reference assembly for CHO cells and the genomics community

with a novel tool for the evaluation of genome assembly quality. We demonstrated how the increased continuity, completeness, and accuracy of the new CH reference genome could facilitate future study and genetic engineering of CHO cells, subsequently leading to more efficient and safer production of biotherapeutics. The selection of the new CH reference assembly was completed through the manual quality assessment of multiple draft assemblies, demonstrating the need for automated reference-independent quality assessment methods. The development of the EvalDNA pipeline fulfilled this need, producing comprehensive quality scores that enable the easy comparison of draft genome assemblies from the same species for reference assembly selection. In addition, EvalDNA scoring of the novel CH reference assembly showed that this assembly has surpassed the quality of the rat reference genome and is approaching the quality of the "gold-standard" mouse reference genome. Last of all, the ability of EvalDNA to produce scores comparable across species is a significant step toward establishing a standard method for assembly quality evaluation.

BIBLIOGRAPHY

[1] P. A. Kitts, D. M. Church, F. Thibaud-Nissen, J. Choi, V. Hem, V. Sapojnikov, R. G. Smith, T. Tatusova, C. Xiang, A. Zherikov, M. DiCuccio, T. D. Murphy, K. D. Pruitt, and A. Kimchi, "Assembly: a resource for assembled genomes at NCBI," *Nucleic Acids Research*, vol. 44, no. D1, pp. 73–80, 2016.

[2] K. Wetterstrand, "The cost of sequencing a human genome from the NHGRI Genome Sequencing Program (GSP)," 2016.

[3] M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu, "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers," *BMC Genomics*, vol. 13, no. 341, 2012.

[4] T. Klingstrom, E. Bongcam-Rudloff, and O. V. Pettersson, "A comprehensive model of DNA fragmentation for the preservation of High Molecular Weight DNA," *bioRxiv*, p. 254276, oct 2018.

[5] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. DeWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, "Real-Time DNA Sequencing from Single Polymerase Molecules," *Science*, vol. 323, no. 5910, pp. 133–138, 2009.

[6] S. Koren, G. P. Harhay, T. P. L. Smith, J. L. Bono, D. M. Harhay, S. D. Mcvey, D. Radune, N. H. Bergman, and A. M. Phillippy, "Reducing assembly complexity of microbial genomes with single-molecule sequencing," *Genome Biology*, vol. 14, no. 9, p. R101, 2013.

[7] S. Koren and A. M. Phillippy, "One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly," *Current Opinion in Microbiology*, vol. 23, pp. 110–120, 2015.

[8] K. Berlin, S. Koren, C.-S. Chin, J. P. Drake, J. M. Landolin, and A. M. Phillippy, "Assembling large genomes with single-molecule sequencing and locality-sensitive hashing," *Nature Biotechnology*, vol. 33, no. 6, pp. 623–630, 2015.

[9] S. Koren, M. C. Schatz, B. P. Walenz, J. Martin, J. T. Howard, G. Ganapathy, Z. Wang, D. A. Rasko, W. R. McCombie, E. D. Jarvis, and A. M. Phillippy, "Hybrid error correction and *de novo* assembly of single-molecule sequencing reads," *Nature Biotechnology*, vol. 30, no. 7, pp. 693–700, 2012.

[10] C. Ye, C. M. Hill, S. Wu, J. Ruan, Z. Ma, J. C. Venter, S. Koren, T. Laver, C. Ye, Z. S. Ma, C. H. Cannon, M. Pop, D. W. Yu, S. Koren, A. M. Phillippy, K. Berlin, L. Salmela, E. Rivals, H. Lee, T. Hackl, R. Hedrich, J. Schultz, F. Forster, M. Boetzer, W. Pirovano, C. S. Chin, F. J. Ribeiro, A. C. English, A. Bashir, K. F. Au, J. G. Underwood, L. Lee, W. H. Wong, N. Nagarajan, M. Pop, E. W. Myers, P. A. Pevzner, H. Tang, M. S. Waterman, J. R. Miller, J. T. Simpson, R. Durbin, T. J. Treangen, D. D. Sommer, F. E. Angly, S. Koren, M. Pop, S. Batzoglou, G. Myers, M. J. Chaisson, D. Brinza, P. A. Pevzner, S. Kurtz, M. Chaisson, G. Tesler, T. F. Smith, M. S. Waterman, C. Ye, Z. S. Ma, M. Chakraborty, J. G. Baldwin-Brown, A. D. Long, J. J. Emerson, A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, "DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies," *Scientific Reports*, vol. 6, no. 7, 2016.

[11] M. Chakraborty, J. G. Baldwin-Brown, A. D. Long, and J. J. Emerson, "Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage," *Nucleic Acids Research*, vol. 44, no. 19, 2016.

[12] A. Wences and M. Schatz, "Metassembler: merging and optimizing *de novo* genome assemblies," *Genome Biology*, vol. 16, p. 207, 2015.

[13] A. V. Zimin, D. R. Smith, G. Sutton, and J. A. Yorke, "Assembly reconciliation," *Bioinformatics*, vol. 24, no. 1, pp. 42–45, 2008.

[14] G. Ganapathy, J. T. Howard, J. M. Ward, J. Li, B. Li, Y. Li, Y. Xiong, Y. Zhang, S. Zhou, D. C. Schwartz, M. Schatz, R. Aboukhalil, O. Fedrigo, L. Bukovnik, T. Wang, G. Wray, I. Rasolonjatovo, R. Winer, J. R. Knight, S. Koren, W. C. Warren, G. Zhang, A. M. Phillippy, and E. D. Jarvis, "High-coverage sequencing and annotated assemblies of the budgerigar genome," *GigaScience*, vol. 3, no. 11, 2014.

[15] S. M. Utturkar, D. M. Klingeman, M. L. Land, C. W. Schadt, M. J. Doktycz, D. A. Pelletier, and S. D. Brown, "Evaluation and validation of *de novo* and hybrid assembly techniques to derive high-quality genome sequences," *Bioinformatics*, 2014.

[16] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, "A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity," *Science*, vol. 337, no. 6096, pp. 816–821, 2012.

[17] P. Mali, L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church, "RNA-guided human genome engineering via Cas9," *Science*, vol. 339, no. 6121, pp. 823–826, 2013.

[18] L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang, "Multiplex genome engineering using CRISPR/Cas systems.," *Science*, vol. 339, no. 6121, pp. 819–823, 2013.

[19] S. Gnerre, I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe, "High-quality draft assemblies of mammalian genomes from massively parallel sequence data.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 4, pp. 1513–1518, 2011.

[20] J. Alföldi and K. Lindblad-Toh, "Comparative genomics as a tool to understand evolution and disease," *Genome Research*, vol. 23, no. 7, pp. 1063–8, 2013.

[21] D. Liu, M. Hunt, and I. J. Tsai, "Inferring synteny between genome assemblies: a systematic evaluation," *BMC Bioinformatics*, vol. 19, no. 1, p. 26, 2018.

[22] K. R. Kukurba and S. B. Montgomery, "RNA sequencing and analysis," *Cold Spring Harbor Protocols*, vol. 2015, pp. 951–69, apr 2015.

[23] T. T. Puck, S. J. Cieciura, and A. Robinson, "Genetics of somatics mammalian cells: Long-term cultivation of euploid cells from human and animal subjects," *Journal of Experimental Medicine*, vol. 108, pp. 945–956, 1958.

[24] K. P. Jayapal, K. F. Wlaschin, W.-S. Hu, and M. G. S. Yap, "Recombinant protein therapeutics from CHO cells - 20 years and counting," *CEP Magazine*, vol. CHO Consortium, pp. 40–47, 2007.

[25] G. Walsh, "Biopharmaceutical benchmarks 2018," *Nature Biotechnology*, vol. 36, no. 12, pp. 1136–1145, 2018.

[26] F. M. Wurm and D. Hacker, "First CHO genome," *Nature Biotechnology*, vol. 29, no. 8, pp. 718–720, 2011.

[27] N. E. Lewis, X. Liu, Y. Li, H. Nagarajan, G. Yerganian, E. O'Brien, A. Bordbar, A. M. Roth, J. Rosenbloom, C. Bian, M. Xie, W. Chen, N. Li, D. Baycin-Hizal, H. Latif, J. Forster, M. J. Betenbaugh, I. Famili, X. Xu, J. Wang, and B. O. Palsson, "Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetulus griseus draft genome.," *Nature Biotechnology*, vol. 31, no. 8, pp. 759–67, 2013.

[28] S. Carillo, S. Mittermayr, A. Farrell, S. Albrecht, and J. Bones, "Glycosylation analysis of therapeutic glycoproteins produced in CHO cells," in *Heterologous Protein Production in CHO Cells*, vol. 1603, pp. 227–241, Humana Press, New York, NY, 2017.

[29] K. Brinkrolf, O. Rupp, H. Laux, F. Kollin, W. Ernst, B. Linke, R. Kofler, S. Romand, F. Hesse, W. E. Budach, S. Galosy, D. Müller, T. Noll, J. Wienberg, T. Jostock, M. Leonard, J. Grillari, A. Tauch, A. Goesmann, B. Helk, J. E. Mott, A. Pühler, and N. Borth, "Chinese hamster genome sequenced from sorted chromosomes," *Nature Biotechnology*, vol. 31, no. 8, pp. 694–695, 2013.

[30] O. Rupp, M. L. MacDonald, S. Li, H. Dhiman, S. Polson, S. Griep, K. Heffner, I. Hernandez, K. Brinkrolf, V. Jadhav, M. Samoudi, H. Hao, B. Kingham, A. Goesmann, M. J. Betenbaugh, N. E. Lewis, N. Borth, and K. H. Lee, "A reference genome of the Chinese hamster based on a hybrid assembly strategy.," *Biotechnology and Bioengineering*, vol. 115, no. 8, pp. 2087–2100, 2018.

[31] J. Feichtinger, I. Hernández, C. Fischer, M. Hanscho, N. Auer, M. Hackl, V. Jadhav, M. Baumann, P. M. Krempl, C. Schmidl, *et al.*, "Comprehensive genome and epigenome characterization of cho cells in response to evolutionary pressures and over time," *Biotechnology and Bioengineering*, vol. 113, no. 10, pp. 2241–2253, 2016.

[32] C. S. Kaas, C. Kristensen, M. J. Betenbaugh, and M. R. Andersen, "Sequencing the CHO DXB11 genome reveals regional variations in genomic stability and haploidy," *BMC Genomics*, vol. 16, no. 1, p. 160, 2015.

[33] A. Richelle and N. E. Lewis, "Improvements in protein production in mammalian cells from targeted metabolic engineering," *Current Opinion in Systems Biology*, vol. 6, pp. 1–6, 2017.

[34] D. M. Wuest, S. W. Harcum, and K. H. Lee, "Genomics in mammalian cell culture bioprocessing.," *Biotechnology Advances*, vol. 30, no. 3, pp. 629–38, 2012.

[35] I. Tossolini, F. J. López-Díaz, R. Kratje, and C. C. Prieto, "Characterization of cellular states of cho-k1 suspension cell culture through cell cycle and rna-sequencing profiling," *Journal of Biotechnology*, vol. 286, pp. 56–67, 2018.

[36] A. B. Diendorfer, M. Hackl, G. Klanert, V. Jadhav, M. Reithofer, F. Stiefel, F. Hesse, J. Grillari, and N. Borth, "Annotation of additional evolutionary conserved microRNAs in CHO cells from updated genomic data.," *Biotechnology and bioengineering*, vol. 112, no. 7, pp. 1488–1493, 2015.

[37] H. Hefzi, K. S. Ang, M. Hanscho, A. Bordbar, D. Ruckerbauer, M. Lakshmanan, C. A. Orellana, D. Baycin-Hizal, Y. Huang, D. Ley, V. S. Martinez, S. Kyriakopoulos, N. E. Jiménez, D. C. Zielinski, L.-E. Quek, T. Wulff, J. Arnsdorf,

S. Li, J. S. Lee, G. Paglia, N. Loira, P. N. Spahn, L. E. Pedersen, J. M. Gutierrez, Z. A. King, A. M. Lund, H. Nagarajan, A. Thomas, A. M. Abdel-Haleem, J. Zanghellini, H. F. Kildegaard, B. G. Voldborg, Z. P. Gerdtzen, M. J. Betenbaugh, B. O. Palsson, M. R. Andersen, L. K. Nielsen, N. Borth, D.-Y. Lee, and N. E. Lewis, "A consensus genome-scale reconstruction of chinese hamster ovary cell metabolism.," *Cell systems*, vol. 3, no. 5, pp. 434–443, 2016.

[38] N. Pristovsek, S. Nallapereddy, L. M. Grav, H. Hefzi, N. E. Lewis, P. Rugbjerg, H. G. Hansen, G. M. Lee, M. R. Andersen, and H. F. Kildegaard, "Systematic evaluation of site-specific recombinant gene expression for programmable mammalian cell engineering," *ACS Synthetic Biology*, p. acssynbio.8b00453, 2019.

[39] N. K. Hamaker and K. H. Lee, "Site-specific integration ushers in a new era of precise CHO cell line engineering," *Current Opinion in Chemical Engineering*, vol. 22, pp. 152–160, 2018.

[40] The Vertebrate Genome Project, "A reference standard for genome biology," *Nature Biotechnology*, vol. 36, pp. 1121–1121, dec 2018.

[41] L. Acuña-Amador, A. Primot, E. Cadieu, A. Roulet, and F. Barloy-Hubler, "Genomic repeats, misassembly and reannotation: a case study with long-read resequencing of *Porphyromonas gingivalis* reference strains," *BMC Genomics*, vol. 19, no. 1, p. 54, 2018.

[42] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, "Quast: quality assessment tool for genome assemblies," *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, 2013.

[43] W. Xiao, "Contig Quality Assessment Tool (CQAT)," 2018.

[44] A. M. Phillippy, M. C. Schatz, and M. Pop, "Genome assembly forensics: finding the elusive mis-assembly," *Genome Biology*, vol. 9, no. 3, p. R55, 2008.

[45] S. C. Clark, R. Egan, P. I. Frazier, and Z. Wang, "Ale: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies," *Bioinformatics*, vol. 29, no. 4, pp. 435–443, 2013.

[46] F. Vezzi, G. Narzisi, and B. Mishra, "Reevaluating assembly evaluations with Feature Response Curves: GAGE and Assemblathons," *PLoS ONE*, vol. 7, no. 12, p. e52210, 2012.

[47] M. Kuhring, P. W. Dabrowski, V. C. Piro, A. Nitsche, and B. Y. Renard, "Surankco: supervised ranking of contigs in *de novo* assemblies.," *BMC Bioinformatics*, vol. 16, no. 1, p. 240, 2015.

[48] G. Walsh, "Biopharmaceutical benchmarks 2014," *Nature Biotechnology*, vol. 32, no. 10, pp. 992–1000, 2014.

[49] J. H. Tjio and T. T. Puck, "Genetics of somatic mammalian cells," *The Journal of Experimental Medicine*, vol. 108, no. 2, pp. 259–268, 1958.

[50] X. Xu, H. Nagarajan, N. E. Lewis, S. Pan, Z. Cai, X. Liu, W. Chen, M. Xie, W. Wang, S. Hammond, M. R. Andersen, N. Neff, B. Passarelli, W. Koh, H. C. Fan, J. Wang, Y. Gui, K. H. Lee, M. J. Betenbaugh, S. R. Quake, I. Famili, B. O. Palsson, and J. Wang, "The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line.," *Nature Biotechnology*, vol. 29, no. 8, pp. 735–741, 2011.

[51] F. N. K. Yusufi, M. Lakshmanan, Y. S. Ho, B. L. W. Loo, P. Ariyaratne, Y. Yang, S. K. Ng, T. R. M. Tan, H. C. Yeo, H. L. Lim, *et al.*, "Mammalian systems biotechnology reveals global cellular adaptations in a recombinant CHO cell line," *Cell Systems*, vol. 4, no. 5, pp. 530–542, 2017.

[52] H. F. Kildegaard, D. Baycin-Hizal, N. E. Lewis, and M. J. Betenbaugh, "The emerging CHO systems biology era: harnessing the 'omics revolution for biotechnology," *Current Opinion in Biotechnology*, vol. 24, no. 6, pp. 1102–1107, 2013.

[53] J. S. Lee, L. M. Grav, N. E. Lewis, and H. Faustrup Kildegaard, "CRISPR/Cas9-mediated genome engineering of CHO cell factories: application and perspectives," *Biotechnology Journal*, vol. 10, no. 7, pp. 979–994, 2015.

[54] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, "Real-time DNA sequencing from single polymerase molecules," *Science*, vol. 323, no. 5910, pp. 133–138, 2009.

[55] D. M. Bickhart, B. D. Rosen, S. Koren, B. L. Sayre, A. R. Hastie, S. Chan, J. Lee, E. T. Lam, I. Liachko, S. T. Sullivan, J. N. Burton, H. J. Huson, J. C. Nystrom, C. M. Kelley, J. L. Hutchison, Y. Zhou, J. Sun, A. Crisà, F. A. Ponce de León, J. C. Schwartz, J. A. Hammond, G. C. Waldbieser, S. G. Schroeder, G. E. Liu, M. J. Dunham, J. Shendure, T. S. Sonstegard, A. M. Phillippy, C. P. Van Tassell, and T. P. L. Smith, "Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome," *Nature Genetics*, vol. 49, no. 4, pp. 643–650, 2017.

[56] D. Gordon, J. Huddleston, M. J. P. Chaisson, C. M. Hill, Z. N. Kronenberg, K. M. Munson, M. Malig, A. Raja, I. Fiddes, L. W. Hillier, C. Dunn, C. Baker,

J. Armstrong, M. Diekhans, B. Paten, J. Shendure, R. K. Wilson, D. Haussler, C.-S. Chin, and E. E. Eichler, "Long-read sequence assembly of the gorilla genome," *Science*, vol. 352, no. 6281, 2016.

[57] Y. Jiao, P. Peluso, J. Shi, T. Liang, M. C. Stitzer, B. Wang, M. S. Campbell, J. C. Stein, X. Wei, C.-S. Chin, K. Guill, M. Regulski, S. Kumari, A. Olson, J. Gent, K. L. Schneider, T. K. Wolfgruber, M. R. May, N. M. Springer, E. Antoniou, W. R. McCombie, G. G. Presting, M. McMullen, J. Ross-Ibarra, R. K. Dawe, A. Hastie, D. R. Rank, and D. Ware, "Improved maize reference genome with single-molecule technologies," *Nature*, vol. 546, no. 7659, pp. 524–527, 2017.

[58] T. Hackl, R. Hedrich, J. Schultz, and F. Förster, "Proovread: large-scale high-accuracy PacBio correction through iterative short read consensus," *Bioinformatics*, vol. 30, no. 21, pp. 3004–3011, 2014.

[59] L. Salmela and E. Rivals, "LoRDEC: accurate and efficient long read error correction," *Bioinformatics*, p. btu538, 2014.

[60] G. Marçais and C. Kingsford, "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers," *Bioinformatics*, vol. 27, no. 6, pp. 764–770, 2011.

[61] B. Liu, Y. Shi, J. Yuan, X. Hu, H. Zhang, N. Li, Z. Li, Y. Chen, D. Mu, and W. Fan, "Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects," *arXiv:1308.2012 [q-bio]*, 2013.

[62] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S.-M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.-W. Lam, and J. Wang, "SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler," *GigaScience*, vol. 1, no. 1, p. 18, 2012.

[63] C. Chin, D. Alexander, P. Marks, A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. Eichler, S. Turner, and J. Korlach, "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data," *Nature Methods*, vol. 10, no. 6, pp. 563–569, 2013.

[64] M. Hsieh, "Smrtmake: Hackable smrtpipe workflows using makefiles instead of smrtpipe.py," 2016.

[65] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg, "Versatile and open software for comparing large genomes," *Genome Biology*, vol. 5, no. 2, 2004.

[66] F. Yang, P. C. O'Brien, and M. A. Ferguson-Smith, "Comparative chromosome map of the laboratory mouse and Chinese hamster defined by reciprocal chromosome painting," *Chromosome Research*, vol. 8, no. 3, pp. 219–227, 2000.

[67] K. F. Wlaschin and W.-S. Hu, "A scaffold for the Chinese hamster genome," *Biotechnology and bioengineering*, vol. 98, no. 2, pp. 429–439, 2007.

[68] C. Holt and M. Yandell, "MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects," *BMC Bioinformatics*, vol. 12, no. 1, p. 491, 2011.

[69] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, "Full-length transcriptome assembly from RNA-Seq data without a reference genome.," *Nature Biotechnology*, vol. 29, no. 7, pp. 644–652, 2011.

[70] A. Smit, R. Hubley, and P. Green, "RepeatMasker Open-4.0," 2015.

[71] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, "BLAST plus: architecture and applications," *BMC Bioinformatics*, vol. 10, no. 421, p. 1, 2009.

[72] G. Slater and E. Birney, "Automated generation of heuristics for biological sequence comparison," *BMC Bioinformatics*, vol. 6, no. 1, p. 31, 2005.

[73] I. Korf, "Gene finding in novel genomes.," *BMC Bioinformatics*, vol. 5, p. 59, 2004.

[74] O. Keller, M. Kollmar, M. Stanke, and S. Waack, "A novel hybrid gene prediction method employing protein multiple sequence alignments," *Bioinformatics*, vol. 27, no. 6, pp. 757–763, 2011.

[75] M. S. Campbell, C. Holt, B. Moore, and M. Yandell, "Genome annotation and curation using MAKER and MAKER-P," *Current Protocols in Bioinformatics*, vol. 48, 2014.

[76] X. M. van Wijk, S. Döhrmann, B. M. Hallström, S. Li, B. G. Voldborg, B. X. Meng, K. K. McKee, T. H. van Kuppevelt, P. D. Yurchenco, B. O. Palsson, *et al.*, "Whole-genome sequencing of invasion-resistant cells identifies laminin $\alpha 2$ as a host factor for bacterial invasion," *mBio*, vol. 8, no. 1, 2017.

[77] G. A. Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, *et al.*, "From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline," *Current Protocols in Bioinformatics*, 2013.

[78] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. Del Angel, M. A. Rivas, M. Hanna, *et al.*, "A framework for variation discovery and genotyping using next-generation dna sequencing data," *Nature Genetics*, vol. 43, no. 5, pp. 491–498, 2011.

[79] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, *et al.*, "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Research*, vol. 20, no. 9, pp. 1297–1303, 2010.

[80] R. K. Dale, B. S. Pedersen, and A. R. Quinlan, "Pybedtools: a flexible python library for manipulating genomic datasets and annotations," *Bioinformatics*, vol. 27, no. 24, pp. 3423–3424, 2011.

[81] A. R. Quinlan and I. M. Hall, "Bedtools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010.

[82] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using david bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[83] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic acids research*, vol. 37, no. 1, pp. 1–13, 2009.

[84] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, "REVIGO summarizes and visualizes long lists of gene ontology terms," *PloS One*, vol. 6, no. 7, p. e21800, 2011.

[85] L. Shi, Y. Guo, C. Dong, J. Huddleston, H. Yang, X. Han, A. Fu, Q. Li, N. Li, S. Gong, K. E. Lintner, Q. Ding, Z. Wang, J. Hu, D. Wang, F. Wang, L. Wang, G. J. Lyon, Y. Guan, Y. Shen, O. V. Evgrafov, J. A. Knowles, F. Thibaud-Nissen, V. Schneider, C.-Y. Yu, L. Zhou, E. E. Eichler, K.-F. So, and K. Wang, "Long-read sequencing and de novo assembly of a Chinese genome," *Nature Communications*, vol. 7, 2016.

[86] J. Zhang, L.-L. Chen, S. Sun, D. Kudrna, D. Copetti, W. Li, T. Mu, W.-B. Jiao, F. Xing, S. Lee, J. Talag, J.-M. Song, B. Du, W. Xie, M. Luo, C. E. Maldonado, J. L. Goicoechea, L. Xiong, C. Wu, Y. Xing, D.-x. Zhou, S. Yu, Y. Zhao, G. Wang, Y. Yu, Y. Luo, B. E. P. Hurtado, A. Danowitz, R. A. Wing, and Q. Zhang, "Building two indica rice reference genomes with PacBio long-read and Illumina paired-end sequencing data," *Scientific Data*, vol. 3, 2016.

[87] K. Kawasaki, O. Kuge, S. C. Chang, P. N. Heacock, M. Rho, K. Suzuki, M. Nishijima, and W. Dowhan, "Isolation of a Chinese hamster ovary (CHO) cDNA encoding phosphatidylglycerophosphate (PGP) synthase, expression of which corrects the mitochondrial abnormalities of a PGP synthase-defective mutant of

CHO-K1 cells," *Journal of Biological Chemistry*, vol. 274, no. 3, pp. 1828–1834, 1999.

[88] V. Jadhav, M. Hackl, A. Druz, S. Shridhar, C. Y. Chung, K. M. Heffner, D. P. Kreil, M. Betenbaugh, J. Shiloach, N. Barron, J. Grillari, and N. Borth, "CHO microRNA engineering is growing up: Recent successes and future challenges," *Biotechnology Advances*, vol. 31, no. 8, pp. 1501–1513, 2013.

[89] A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis, "Integrative analysis of 111 reference human epigenomes.," *Nature*, vol. 518, no. 7539, pp. 317–330, 2015.

[90] Y. Maeda, H. Ashida, and T. Kinoshita, "CHO glycosylation mutants: GPI anchor," *Methods in enzymology*, vol. 416, pp. 182–205, 2006.

[91] S. K. Patnaik and P. Stanley, "Lectin-resistant CHO glycosylation mutants," *Methods in Enzymology*, vol. 416, pp. 159–182, 2006.

[92] P. Stanley, "Chinese hamster ovary mutants for glycosylation engineering of biopharmaceuticals," *Pharmaceutical Bioprocessing*, vol. 2, no. 5, pp. 359–361, 2014.

[93] L. Zhang, R. Lawrence, B. A. Frazier, and J. D. Esko, "CHO glycosylation mutants: proteoglycans," *Methods in Enzymology*, vol. 416, pp. 205–221, 2006.

[94] J. D. Esko, T. E. Stewart, and W. H. Taylor, "Animal cell mutants defective in glycosaminoglycan biosynthesis," *Proceedings of the National Academy of Sciences*, vol. 82, no. 10, pp. 3197–3201, 1985.

[95] K. Cuellar, H. Chuong, S. M. Hubbell, and M. E. Hinsdale, "Biosynthesis of chondroitin and heparan sulfate in chinese hamster ovary cells depends on xylosyltransferase ii," *Journal of Biological Chemistry*, vol. 282, no. 8, pp. 5195–5200, 2007.

[96] J. S. Goh, Y. Liu, K. F. Chan, C. Wan, G. Teo, P. Zhang, Y. Zhang, and Z. Song, "Producing recombinant therapeutic glycoproteins with enhanced sialylation using CHO-gmt4 glycosylation mutant cells," *Bioengineered*, vol. 5, no. 4, pp. 269–273, 2014.

[97] L. H. Thompson, E. P. Salazar, K. W. Brookman, C. C. Collins, S. A. Stewart, D. B. Busch, and C. A. Weber, "Recent progress with the DNA repair mutants of Chinese hamster ovary cells," *Journal of Cell Science*, vol. 6, pp. 97–110, 1987.

[98] K. R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W.-C. Chou, J. Corbeil, C. Del Fabbro, T. R. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N. A. Fonseca, G. Ganapathy, R. A. Gibbs, S. Gnerre, l. Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J. B. Hiatt, I. Y. Ho, J. Howard, M. Hunt, S. D. Jackman, D. B. Jaffe, E. D. Jarvis, H. Jiang, S. Kazakov, P. J. Kersey, J. O. Kitzman, J. R. Knight, S. Koren, T.-W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. MacCallum, M. D. MacManes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T. D. Otto, B. Paten, O. S. Paulo, A. M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F. J. Ribeiro, S. Richards, D. S. Rokhsar, J. G. Ruby, S. Scalabrin, M. C. Schatz, D. C. Schwartz, A. Sergushichev, T. Sharpe, T. I. Shaw, J. Shendure, Y. Shi, J. T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B. M. Vieira, J. Wang, K. C. Worley, S. Yin, S.-M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, and I. F. Korf, "Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species," *GigaScience*, vol. 2, p. 10, 2013.

[99] J. S. Lee, T. B. Kallehauge, L. E. Pedersen, and H. F. Kildegaard, "Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway," *Scientific Reports*, vol. 5, 2015.

[100] K. P. Anderson, M. A. Low, Y. S. Lie, G. A. Keller, and M. Dinowitz, "Endogenous origin of defective retroviruslike particles from a recombinant Chinese hamster ovary cell line," *Virology*, vol. 181, no. 1, pp. 305–311, 1991.

[101] D. N. Wheatley, "Pericentriolar virus-like particles in Chinese hamster ovary cells," *The Journal of General Virology*, vol. 24, no. 2, pp. 395–399, 1974.

[102] L. Yang, M. Güell, D. Niu, H. George, E. Lesha, D. Grishin, J. Aach, E. Shrock, W. Xu, J. Poci, R. Cortazio, R. A. Wilkinson, J. A. Fishman, and G. Church, "Genome-wide inactivation of porcine endogenous retroviruses (PERVs).," *Science*, vol. 350, no. 6264, pp. 1101–1104, 2015.

[103] V. Schmieder, N. Bydlinski, R. Strasser, M. Baumann, H. Kildegaard, V. Jadhav, and N. Borth, "Enhanced genome editing tools for multi-gene deletion knock-out approaches using paired CRISPR sgRNAs in CHO cells," *Biotechnology Journal*, vol. 13, no. 3, 2017.

[104] S. Morita, H. Noguchi, T. Horii, K. Nakabayashi, M. Kimura, K. Okamura, A. Sakai, H. Nakashima, K. Hata, K. Nakashima, and I. Hatada, "Targeted DNA demethylation in vivo using dCas9-peptide repeat and scFv-TET1 catalytic domain fusions," *Nature Biotechnology*, vol. 34, no. 10, pp. 1060–1065, 2016.

[105] A. Vojta, P. Dobrinić, V. Tadić, L. Bočkor, P. Korać, B. Julg, M. Klasić, and V. Zoldoš, "Repurposing the CRISPR-Cas9 system for targeted DNA methylation," *Nucleic Acids Research*, vol. 44, no. 12, pp. 5615–5628, 2016.

[106] G. Parra, K. Bradnam, and I. Korf, "CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes," *Bioinformatics*, vol. 23, no. 9, pp. 1061–1067, 2007.

[107] F. A. Simao, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, "BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs," *Bioinformatics*, vol. 31, no. 19, p. 3210–3212, 2015.

[108] T. D. Wu and C. K. Watanabe, "GMAP: a genomic mapping and alignment program for mRNA and EST sequences," *Bioinformatics (Oxford, England)*, vol. 21, no. 9, pp. 1859–1875, 2005.

[109] H. Ponstingl, "Smalt - sequence mapping and alignment tool v0.7.6," 2016.

[110] M. Hunt, T. Kikuchi, M. Sanders, C. Newbold, M. Berriman, and T. D. Otto, "REAPR: a universal tool for genome assembly evaluation," *Genome Biology*, vol. 14, no. 5, p. R47, 2013.

[111] B. J. Haas, A. L. Delcher, J. R. Wortman, and S. L. Salzberg, "DAGchainer: a tool for mining segmental genome duplications and synteny," *Bioinformatics*, vol. 20, no. 18, pp. 3643–3646, 2004.

[112] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[113] S. E. Calvo, K. R. Clauser, and V. K. Mootha, "MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1251–D1257, 2015.

[114] A. Souvorov, Y. Kapustin, B. Kiryutin, V. Chetvernin, T. Tatusova, and D. Lipman, "Gnomon–NCBI eukaryotic gene prediction tool," *National Center for Biotechnology Information*, pp. 1–24, 2010.

[115] H. Shatkay, J. Miller, C. Mobarry, M. Flanigan, S. Yooseph, and G. Sutton, "ThurGood: Evaluating assembly-to-assembly mapping," *Journal of Computational Biology*, vol. 11, no. 5, pp. 800–811, 2004.

[116] G. Parra, K. Bradnam, and I. Korf, "CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes," *Bioinformatics*, vol. 23, no. 9, pp. 1061–1067, 2007.

[117] M. Hunt, T. Kikuchi, M. Sanders, C. Newbold, M. Berriman, and T. D. Otto, "REAPR: a universal tool for genome assembly evaluation.," *Genome Biology*, vol. 14, no. 5, p. R47, 2013.

[118] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, 2009.

[119] S. L. Salzberg, A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marçais, M. Pop, and J. A. Yorke, "GAGE: A critical evaluation of genome assemblies and assembly algorithms," *Genome Research*, vol. 22, no. 3, pp. 557–567, 2012.

[120] H. Ponstingl, "SMALT," 2015.

[121] S. Pattnaik, S. Gupta, A. Rao, and B. Panda, "SInC: an accurate and fast error-model based simulator for SNPs, Indels and CNVs coupled with a read generator for short-read sequence data," *BMC Bioinformatics*, vol. 15, no. 1, p. 40, 2014.

[122] L. Alsøe, A. Sarno, S. Carracedo, D. Domanska, F. Dingler, L. Lirussi, T. Sen-Gupta, N. B. Tekin, L. Jobert, L. B. Alexandrov, A. Galashevskaya, C. Rada, G. K. Sandve, T. Rognes, H. E. Krokan, and H. Nilsen, "Uracil accumulation and mutagenesis dominated by cytosine deamination in CpG dinucleotides in mice lacking UNG and SMUG1," *Scientific Reports*, vol. 7, no. 1, p. 7199, 2017.

[123] M. B. Kursa, *Praznik: collection of information-based feature selection filters*, 2018. R package version 5.0.0.

[124] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[125] T. Lumley, *Leaps: regression subset selection*, 2017. R package version 3.0.

[126] R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[127] J. W. Max Kuhn, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, and T. Hunt., *Caret: classification and regression training*, 2018. R package version 6.0-80.

[128] F. Krueger, "Trim Galore," 2018.

[129] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, no. 4, pp. 357–359, 2012.

[130] J. R. Miller, A. L. Delcher, S. Koren, E. Venter, B. P. Walenz, A. Brownley, J. Johnson, K. Li, C. Mobarry, and G. Sutton, "Aggressive assembly of pyrosequencing reads with mates.," *Bioinformatics*, vol. 24, no. 24, pp. 2818–2824, 2008.

[131] D. R. Zerbino and E. Birney, "Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs.," *Genome Research*, vol. 18, no. 5, pp. 821–829, 2008.

[132] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol, "ABySS: a parallel assembler for short read sequence data.," *Genome Research*, vol. 19, no. 6, pp. 1117–1123, 2009.

[133] A. V. Zimin, G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg, and J. A. Yorke, "The MaSuRCA genome assembler," *Bioinformatics*, vol. 29, no. 21, pp. 2669–2677, 2013.

[134] S. Koren, T. J. Treangen, and M. Pop, "Bambus 2: scaffolding metagenomes.," *Bioinformatics*, vol. 27, no. 21, pp. 2964–2971, 2011.

[135] Y. Kawahara, M. de la Bastide, J. P. Hamilton, H. Kanamori, W. R. McCombie, S. Ouyang, D. C. Schwartz, T. Tanaka, J. Wu, S. Zhou, K. L. Childs, R. M. Davidson, H. Lin, L. Quesada-Ocampo, B. Vaillancourt, H. Sakai, S. S. Lee, J. Kim, H. Numa, T. Itoh, C. R. Buell, and T. Matsumoto, "Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data," *Rice*, vol. 6, no. 1, p. 4, 2013.

[136] G. Spudich, "Ensembl blog: new human assembly coming," 2013. Accessed on 21 January, 2019.

[137] G. Myers, "A dataset generator for whole genome shotgun sequencing," *Proceedings: International Conference on Intelligent Systems for Molecular Biology*, pp. 202–10, 1999.

[138] X. Hu, J. Yuan, Y. Shi, J. Lu, B. Liu, Z. Li, Y. Chen, D. Mu, H. Zhang, N. Li, Z. Yue, F. Bai, H. Li, and W. Fan, "pIRS: Profile-based Illumina pair-end reads simulator," *Bioinformatics*, vol. 28, pp. 1533–1535, jun 2012.

[139] D. Pratas, A. J. Pinho, and J. M. O. S. Rodrigues, "XS: a FASTQ read simulator.," *BMC research notes*, vol. 7, p. 40, 2014.

[140] M. Escalona, S. Rocha, and D. Posada, "A comparison of tools for the simulation of genomic next-generation sequencing data," *Nature Reviews Genetics*, vol. 17, no. 8, pp. 459–69, 2016.

[141] D. M. Ecker, S. D. Jones, and H. L. Levine, "The therapeutic monoclonal antibody market.," *mAbs*, vol. 7, no. 1, pp. 9–14, 2015.

[142] S. K. Fischer, M. Cheu, K. Peng, J. Lowe, J. Araujo, E. Murray, D. McClintock, J. Matthews, P. Siguenza, and A. Song, "Specific immune response to phospholipase B-like 2 protein, a host cell impurity in Lebrikizumab clinical material," *The AAPS Journal*, vol. 19, no. 1, pp. 254–263, 2017.

[143] J. Chiu, K. N. Valente, N. E. Levy, L. Min, A. M. Lenhoff, and K. H. Lee, "Knockout of a difficult-to-remove CHO host cell protein, lipoprotein lipase, for improved polysorbate stability in monoclonal antibody formulations," *Biotechnology and Bioengineering*, vol. 114, no. 5, pp. 1006–1015, 2017.

[144] N. Dixit, N. Salamat-Miller, P. A. Salinas, K. D. Taylor, and S. K. Basu, "Residual host cell protein promotes polysorbate 20 degradation in a sulfatase drug product leading to free fatty acid particles," *Journal of Pharmaceutical Sciences*, vol. 105, no. 5, pp. 1657–1666, 2016.

[145] H. Dorai, A. Santiago, M. Campbell, Q. M. Tang, M. J. Lewis, Y. Wang, Q.-Z. Lu, S.-L. Wu, and W. Hancock, "Characterization of the proteases involved in the N-terminal clipping of glucagon-like-peptide-1-antibody fusion proteins," *Biotechnology Progress*, vol. 27, no. 1, pp. 220–231, 2011.

[146] S. X. Gao, Y. Zhang, K. Stansberry-Perkins, A. Buko, S. Bai, V. Nguyen, and M. L. Brader, "Fragmentation of a highly purified monoclonal antibody attributed to residual CHO cell protease activity," *Biotechnology and Bioengineering*, vol. 108, no. 4, pp. 977–982, 2011.

[147] L. C. Eaton, "Host cell contaminant protein assay development for recombinant biopharmaceuticals," *Journal of Chromatography A*, vol. 705, no. 1, pp. 105–114, 1995.

[148] N. E. Levy, K. N. Valente, L. H. Choe, K. H. Lee, and A. M. Lenhoff, "Identification and characterization of host cell protein product-associated impurities in monoclonal antibody bioprocessing," *Biotechnology and Bioengineering*, vol. 111, pp. 904–912, 2014.

[149] K. N. Valente, A. M. Lenhoff, and K. H. Lee, "Expression of difficult-to-remove host cell protein impurities during extended Chinese hamster ovary cell culture and their impact on continuous bioprocessing," *Biotechnology and Bioengineering*, vol. 112, no. 6, pp. 1232–1242, 2015.

[150] P. A. Marichal-Gallardo and M. M. Álvarez, "State-of-the-art in downstream processing of monoclonal antibodies: process trends in design and validation," *Biotechnology Progress*, vol. 28, no. 4, pp. 899–916, 2012.

[151] B. A. Kerwin, "Polysorbates 20 and 80 used in the formulation of protein biotherapeutics: structure and degradation pathways," *Journal of Pharmaceutical Sciences*, vol. 97, no. 8, pp. 2924–2935, 2008.

[152] J. Mead, S. Irvine, and D. Ramji, "Lipoprotein lipase: structure, function, regulation, and role in disease," *Journal of Molecular Medicine*, vol. 80, no. 12, pp. 753–769, 2002.

[153] N. E. Levy, K. N. Valente, K. H. Lee, and A. M. Lenhoff, "Host cell protein impurities in chromatographic polishing steps for monoclonal antibody purification," *Biotechnology and Bioengineering*, vol. 113, no. 6, pp. 1260–1272, 2016.

[154] T. Hall, S. L. Sandefur, C. C. Frye, T. L. Tuley, and L. Huang, "Polysorbates 20 and 80 degradation by group XV lysosomal phospholipase A2 isomer X1 in monoclonal antibody formulations," *Journal of Pharmaceutical Sciences*, vol. 105, no. 5, pp. 1633–1642, 2016.

[155] B. Tran, V. Grosskopf, X. Wang, J. Yang, D. Walker, C. Yu, and P. McDonald, "Investigating interactions between phospholipase B-Like 2 and antibodies during Protein A chromatography," *Journal of Chromatography A*, vol. 1438, pp. 31–38, 2016.

[156] M. Aga, N. Yamano, T. Kumamoto, J. Frank, M. Onitsuka, and T. Omasa, "Construction of a gene knockout CHO cell line using a simple gene targeting method," *BMC Proceedings*, vol. 9, no. Suppl 9, p. P2, 2015.

[157] L. M. Grav, J. S. Lee, S. Gerling, T. B. Kallehauge, A. H. Hansen, S. Kol, G. M. Lee, L. E. Pedersen, and H. F. Kildegaard, "One-step generation of triple knockout CHO cell lines using CRISPR/Cas9 and fluorescent enrichment," *Biotechnology Journal*, vol. 10, no. 9, pp. 1446–1456, 2015.

[158] K. F. Chan, W. Shahreel, C. Wan, G. Teo, N. Hayati, S. J. Tay, W. H. Tong, Y. Yang, P. M. Rudd, P. Zhang, and Z. Song, "Inactivation of GDP-fucose transporter gene (Slc35c1) in CHO cells by ZFNs, TALENs and CRISPR-Cas9 for production of fucose-free antibodies," *Biotechnology Journal*, vol. 11, no. 3, pp. 399–414, 2016.

[159] Y. Santiago, E. Chan, P.-Q. Liu, S. Orlando, L. Zhang, F. D. Urnov, M. C. Holmes, D. Guschin, A. Waite, J. C. Miller, E. J. Rebar, P. D. Gregory, A. Klug, and T. N. Collingwood, "Targeted gene knockout in mammalian cells by using engineered zinc-finger nucleases," *PNAS*, vol. 105, no. 15, pp. 5809–5814, 2008.

[160] Z. Yang, S. Wang, A. Halim, M. A. Schulz, M. Frodin, S. H. Rahman, M. B. Vester-Christensen, C. Behrens, C. Kristensen, S. Y. Vakhrushev, E. P. Bennett, H. H. Wandall, and H. Clausen, "Engineered CHO cells for production of diverse, homogeneous glycoproteins," *Nature Biotechnology*, vol. 33, no. 8, pp. 842–844, 2015.

[161] R. C. Edgar, "MUSCLE: Multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[162] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.," *Molecular systems biology*, vol. 7, no. 1, p. 539, 2011.

[163] N. J. Schurch, C. Cole, A. Sherstnev, J. Song, C. Duc, K. G. Storey, W. H. I. McLean, S. J. Brown, G. G. Simpson, and G. J. Barton, "Improved annotation of 3' untranslated regions and complex loci by combination of strand-specific direct RNA sequencing, RNA-seq and ESTs," *PLoS ONE*, vol. 9, no. 4, p. e94270, 2014.

[164] S. Shenker, P. Miura, P. Sanfilippo, and E. C. Lai, "IsoSCM: improved and alternative 3' UTR annotation using multiple change-point inference.," *RNA*, vol. 21, no. 1, pp. 14–27, 2015.

[165] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, and R. Apweiler, "QuickGO: a web-based tool for Gene Ontology searching.," *Bioinformatics*, vol. 25, no. 22, pp. 3045–3046, 2009.

[166] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[167] A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton, "Jalview Version 2-A multiple sequence alignment editor and analysis workbench," *Bioinformatics*, vol. 25, no. 9, pp. 1189–1191, 2009.

[168] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O. Donovan, N. Redaschi, and L.-s. L. Yeh, "UniProt : the Universal Protein knowledgebase," *Nucleic Acids Research*, vol. 32, 2004.

[169] N. Lee, J. Shin, J. H. Park, G. M. Lee, S. Cho, and B.-K. Cho, "Targeted Gene Deletion Using DNA-Free RNA-Guided Cas9 Nuclease Accelerates Adaptation of CHO Cells to Suspension Culture," *ACS Synthetic Biology*, vol. 5, no. 11, pp. 1211–1219, 2016.

[170] B. G. Kremkow, J. Y. Baik, M. L. MacDonald, and K. H. Lee, "CHOgenome.org 2.0: Genome resources and website updates," 2015.

[171] F. Carrière, C. Withers-Martinez, H. van Tilbeurgh, A. Roussel, C. Cambillau, and R. Verger, "Structural basis for the substrate selectivity of pancreatic lipases and some related proteins," *Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes*, vol. 1376, no. 3, pp. 417–432, 1998.

[172] B. Persson, G. Bengtsson-Olivecrona, S. Enerback, T. Olivecrona, and H. Jornvall, "Structural features of lipoprotein lipase. Lipase family relationships, binding interactions, non-equivalence of lipase cofactors, vitellogenin similarities and functional subdivision of lipoprotein lipase," *European Journal of Biochemistry*, vol. 179, no. 1, pp. 39–45, 1989.

[173] F. Birzele, J. Schaub, W. Rust, C. Clemens, P. Baum, H. Kaufmann, A. Weith, T. W. Schulz, and T. Hildebrandt, "Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing," *Nucleic Acids Research*, vol. 38, no. 12, pp. 3999–4010, 2010.

[174] C. A. Orellana, E. Marcellin, R. W. Palfreyman, T. P. Munro, P. P. Gray, and L. K. Nielsen, "RNA-Seq highlights high clonal variation in monoclonal antibody producing CHO cells," *Biotechnology Journal*, vol. 13, no. 3, 2018.

[175] J. Emmerich, O. U. Beg, J. Peterson, L. Previato, J. D. Brunzell, H. B. Brewer, and S. Santamarina-Fojo, "Human lipoprotein lipase. Analysis of the catalytic triad by site-directed mutagenesis of Ser-132, Asp-156, and His-241.," *The Journal of Biological Chemistry*, vol. 267, no. 6, pp. 4161–4165, 1992.

[176] A. Glukhova, V. Hinkovska-Galcheva, R. Kelly, A. Abe, J. A. Shayman, and J. J. G. Tesmer, "Structure and function of lysosomal phospholipase A2 and lecithin:cholesterol acyltransferase.," *Nature communications*, vol. 6, p. 6250, 2015.

[177] M. Hiraoka, A. Abe, and J. A. Shayman, "Structure and function of lysosomal phospholipase A2: identification of the catalytic triad and the role of cysteine residues.," *Journal of lipid research*, vol. 46, no. 11, pp. 2441–2447, 2005.

[178] J. A. Shayman, R. Kelly, J. Kollmeyer, Y. He, and A. Abe, "Group XV phospholipase A2, a lysosomal phospholipase A2," *Progress in Lipid Research*, vol. 50, no. 1, pp. 1–13, 2011.

[179] C. Bailey-Kellogg, A. H. Gutiérrez, L. Moise, F. Terry, W. D. Martin, and A. S. De Groot, "CHOPPI: a web tool for the analysis of immunogenicity risk from host cell proteins in CHO-based protein production.," *Biotechnology and Bioengineering*, vol. 111, no. 11, pp. 2170–2182, 2014.

[180] A. H. Gutiérrez, L. Moise, and A. S. De Groot, "Of [hamsters] and men: a new perspective on host cell proteins.," *Human vaccines & Immunotherapeutics*, vol. 8, no. 9, pp. 1172–1174, 2012.

[181] C. Alkan, S. Sajjadian, and E. E. Eichler, "Limitations of next-generation genome sequence assembly.," *Nature Methods*, vol. 8, no. 1, pp. 61–65, 2011.

[182] S. Meader, L. W. Hillier, D. Locke, C. P. Ponting, and G. Lunter, "Genome assembly quality: assessment and improvement using the neutral indel model.," *Genome Research*, vol. 20, no. 5, pp. 675–684, 2010.

[183] J. Shin, N. Lee, Y. Song, J. Park, T. J. Kang, S. C. Kim, G. M. Lee, and B.-K. Cho, "Efficient CRISPR/Cas9-mediated multiplex genome editing in CHO cells via high-level sgRNA-Cas9 complex," *Biotechnology and Bioprocess Engineering*, vol. 20, no. 5, pp. 825–833, 2015.

[184] D. Niu, H.-J. Wei, L. Lin, H. George, T. Wang, I.-H. Lee, H.-Y. Zhao, Y. Wang, Y. Kan, E. Shrock, E. Lesha, G. Wang, Y. Luo, Y. Qing, D. Jiao, H. Zhao, X. Zhou, S. Wang, H. Wei, M. Güell, G. M. Church, and L. Yang, "Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9," *Science*, vol. 357, no. 6357, pp. 1303–1307, 2017.

[185] Q. Zheng, X. Cai, M. H. Tan, S. Schaffert, C. P. Arnold, X. Gong, C.-Z. Chen, and S. Huang, "Precise gene deletion and replacement using the CRISPR/Cas9 system in human cells," *BioTechniques*, vol. 57, no. 3, 2014.

[186] V. Schmieder, N. Bydlinski, R. Strasser, M. Baumann, H. F. Kildegaard, V. Jadhav, and N. Borth, "Enhanced genome editing tools For multi-gene deletion knock-out approaches using paired CRISPR sgRNAs in CHO cells," *Biotechnology Journal*, vol. 13, no. 3, 2018.

[187] P. Chain, D. V. Grafham, R. S. Fulton, M. G. FitzGerald, J. Hostetler, D. Muzny, J. Ali, B. Birren, D. C. Bruce, C. Buhay, J. R. Cole, Y. Ding, S. Dugan, D. Field, G. M. Garrity, R. Gibbs, T. Graves, C. S. Han, S. H. Harrison, S. Highlander, P. Hugenholtz, H. M. Khouri, C. D. Kodira, E. Kolker, N. Kyrpides, D. Lang, A. Lapidus, S. A. Malfatti, V. Markowitz, T. Metha, K. E. Nelson, J. Parkhill, S. Pitluck, X. Qin, T. D. Read, J. Schmutz, S. Sozhamannan, P. Sterk, R. L. Strausberg, G. Sutton, N. R. Thomson, J. M. Tiedje, G. Weinstock, A. Wollam, and J. C. Detter, "Genome project standards in a new era of sequencing," *Science*, vol. 326, pp. 236–237, oct 2009.

[188] K. R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W.-C. Chou, J. Corbeil, C. Del Fabbro, T. R. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N. A. Fonseca, G. Ganapathy, R. A. Gibbs, S. Gnerre, É. Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J. B. Hiatt, I. Y. Ho, J. Howard, M. Hunt, S. D. Jackman, D. B. Jaffe, E. D. Jarvis, H. Jiang, S. Kazakov, P. J. Kersey, J. O. Kitzman, J. R. Knight, S. Koren, T.-W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. MacCallum, M. D. MacManes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T. D. Otto, B. Paten, O. S. Paulo, A. M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F. J. Ribeiro, S. Richards, D. S. Rokhsar, J. G. Ruby, S. Scalabrin, M. C. Schatz, D. C.

Schwartz, A. Sergushichev, T. Sharpe, T. I. Shaw, J. Shendure, Y. Shi, J. T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B. M. Vieira, J. Wang, K. C. Worley, S. Yin, S.-M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, and I. F. Korf, "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species," *GigaScience*, vol. 2, no. 1, p. 10, 2013.

[189] L. Mendelowitz and M. Pop, "Computational methods for optical mapping.," *GigaScience*, vol. 3, no. 1, p. 33, 2014.

[190] N. Nagarajan, T. D. Read, and M. Pop, "Scaffolding and validation of bacterial genome assemblies using optical restriction maps.," *Bioinformatics*, vol. 24, no. 10, pp. 1229–1235, 2008.

[191] J.-M. Belton, R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker, "Hi-C: a comprehensive technique to capture the conformation of genomes.," *Methods*, vol. 58, no. 3, pp. 268–276, 2012.

[192] J. N. Burton, A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman, and J. Shendure, "Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions.," *Nature Biotechnology*, vol. 31, no. 12, pp. 1119–1125, 2013.

[193] J. Ghurye, M. Pop, S. Koren, D. Bickhart, and C.-S. Chin, "Scaffolding of long read assemblies using long range contact information," *BMC Genomics*, vol. 18, no. 1, p. 527, 2017.

[194] D. M. Bickhart, B. D. Rosen, S. Koren, B. L. Sayre, A. R. Hastie, S. Chan, J. Lee, E. T. Lam, I. Liachko, S. T. Sullivan, J. N. Burton, H. J. Huson, J. C. Nystrom, C. M. Kelley, J. L. Hutchison, Y. Zhou, J. Sun, A. Crisà, F. A. Ponce de León, J. C. Schwartz, J. A. Hammond, G. C. Waldbieser, S. G. Schroeder, G. E. Liu, M. J. Dunham, J. Shendure, T. S. Sonstegard, A. M. Phillippy, C. P. Van Tassell, and T. P. L. Smith, "Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome," *Nature Genetics*, vol. 49, no. 4, pp. 643–650, 2017.

[195] Y. Dong, M. Xie, Y. Jiang, N. Xiao, X. Du, W. Zhang, G. Tosser-Klopp, J. Wang, S. Yang, J. Liang, W. Chen, J. Chen, P. Zeng, Y. Hou, C. Bian, S. Pan, Y. Li, X. Liu, W. Wang, B. Servin, B. Sayre, B. Zhu, D. Sweeney, R. Moore, W. Nie, Y. Shen, R. Zhao, G. Zhang, J. Li, T. Faraut, J. Womack, Y. Zhang, J. Kijas, N. Cockett, X. Xu, S. Zhao, J. Wang, and W. Wang, "Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*)," *Nature Biotechnology*, vol. 31, no. 2, pp. 135–141, 2013.

[196] N. H. Putnam, B. L. O'Connell, J. C. Stites, B. J. Rice, M. Blanchette, R. Calef, C. J. Troll, A. Fields, P. D. Hartley, C. W. Sugnet, D. Haussler, D. S. Rokhsar,

and R. E. Green, "Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage.," *Genome Research*, vol. 26, no. 3, pp. 342–350, 2016.

[197] W. Y. Low, R. Tearle, D. M. Bickhart, B. D. Rosen, S. B. Kingan, T. Swale, F. Thibaud-Nissen, T. D. Murphy, R. Young, L. Lefevre, D. A. Hume, A. Collins, P. Ajmone-Marsan, T. P. L. Smith, and J. L. Williams, "Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity," *Nature Communications*, vol. 10, no. 1, p. 260, 2019.

[198] F. J. Sedlazeck, H. Lee, C. A. Darby, and M. C. Schatz, "Piercing the dark matter: bioinformatics of long-range sequencing and mapping," *Nature Reviews Genetics*, vol. 19, no. 6, pp. 329–346, 2018.

[199] B. Minkenberg, M. Wheatley, and Y. Yang, "CRISPR/cas9-enabled multiplex genome editing and its application," *Progress in Molecular Biology and Translational Science*, vol. 149, pp. 111–132, 2017.

Appendix A

GLOSSARY OF TERMS

An alphabetical listing of technical terms and tools used in this dissertation.

1. **ALE (Assembly Likelihood Estimator)** - an assembly quality evaluation tool that uses Bayesian statistics to determine the probability of an assembly being correct given a set of reads.

2. **ALLPATHS-LG** - a whole-genome shotgun assembler that uses short sequencing reads (such as Illumina reads).

3. **Assembly Quality** - the accuracy, continuity, and completeness of an assembly, i.e. how close the assembly is to the organism's true genome.

4. **BLAST (Basic Local Alignment Search Tool)** - a tool that finds regions of similarity between a query (a nucleotide or protein sequence) and a specified database of sequences.

5. **BUSCO (Benchmarking Universal Single-Copy Orthologs)** -– a tool that evaluates an assembly regarding gene content. Output metrics are based on the expectation of finding single-copy orthologs in the organism's genome.

6. **Caret** - a R library for machine learning that includes various regression, classification and clustering algorithms.

7. **CEGMA (Core Eukaryotic Genes Mapping Approach)** - an older tool that is similar to BUSCO and evaluates an assembly regarding gene content.

8. **Clipped Reads** - reads that require bases to be removed to be able to map to the reference sequence. The removed bases are considered 'soft-clipped'.

9. **ClustalO (Clustal Omega)** - a tool for multiple sequence alignment that uses guide trees and HMM profiles to generate the alignments.

10. **Compression-Expansion (CE) Statistic** - a metric based on the comparison between the mean insert size of mate pairs spanning a base in the assembly and the expected mean insert size (the mean of the insert sizes of all the mate pairs). A large and positive CE value indicates an insertion, where a negative value indicates a deletion.

11. **Contig** - one piece of continuous DNA, assembled from DNA reads.

12. **Copy Number Variation (CNV)** - variation in the number of repeat regions (could be a copied gene) in a genome assembly.

13. **CRISPR/Cas (Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated protein)** - CRISPR/Cas-systems were adapted from the genome editing systems of bacteria. The CRISPR/Cas complex includes a guide RNA that targets a specific DNA sequence and a Cas enzyme which cuts the DNA at that site. CRISPR-Cas9, one of the most commonly used CRISPR/Cas systems, uses the Cas9 enzyme to cleave the DNA, but other Cas enzymes can be used such as Cpf1. As the cell's DNA repair machinery repairs the cut site, nucleotides are often added or deleted which can knock-out or knock-down the gene being targeted.

14. **DAVID (Database for Annotation, Visualization and Integrated Discovery)** - a set of web-accessible tools for functional annotation analysis of gene lists including GO term enrichment analysis and the identification of functionally related gene groups.

15. **Delta-filter** - a tool within the MUMmer package that filters the alignment files from NUCmer based on alignment length, identity, uniqueness and/or consistency.

16. ***De Novo* Assembly** - an assembly created without a reference genome.

17. ***De Novo* Metrics** - quality metrics that can be calculated without a reference genome.

18. **Dnadiff** - a tool within the MUMmer package that identifies differences between aligned sequences.

19. **E-value** - a statistic used by BLAST which represents the probability that an alignment (or a BLAST hit) could occur by chance. It depends on the alignment score and the size of the database being searched.

20. **Elastic Net Regression** - a type of penalized linear regression model. The penalization for elastic net falls between that of Lasso and Ridge regression depending on the alpha and lambda hyper-parameters.

21. **EvalDNA (Evaluation of *De Novo* Assemblies)** - a pipeline to evaluate genome assembly quality using machine learning methods and without a reference genome. The pipeline was developed in this project.

22. **FASTA** - file format for representing nucleotide or protein sequences.

23. **Fragment** - refers to paired-end reads and the sequence between them.

24. **Fragment Coverage (FC)** - the number of fragments mapping over a base in the assembly.

25. **FRCbam** - an assembly quality evaluation tool that produces a feature response curve (FRC). The FRC shows the trade-off between the accuracy and the continuity of the assembly. More specifically, the FRC shows the coverage (y-axis) along the genome from contigs (starting with the longest) whose sum of errors (features) is less than a specified threshold (x-axis).

26. **Gap** - unknown sequence/bases in a genome assembly, which are represented as N's in FASTA files.

27. **Gene Annotation** - the identification of the location and structure of genes within an assembly.

28. **GMAP** - a tool for aligning and mapping coding sequences (CDSs) to a genome. The tool can be used to identify the intron–exon structure of genes.

29. **Gene Ontology (GO) terms** - a hierarchical and controlled set of terms about gene and gene product function. GO provides a way to consistently represent gene and gene product features among all species.

30. **HGAP (Hierarchical Genome Assembly Process)** - a whole-genome assembler that uses long-read sequencing data (i.e. PacBio reads). HGAP corrects errors on the longest reads using shorter reads from the same library before assembling the reads. HGAP can also be given already corrected PacBio reads to assemble.

31. **Host Cell Proteins (HCPs)** - proteins expressed from the host cell that is not the biotherapeutic. Most (>99%) are removed during the purification process. Small amounts of certain HCPs can cause safety and stability problems for the drug product if not removed.

32. **Indel** - refers to either an insertion (addition) or deletion in a DNA sequence.

33. **Jalview** - software for the visualization and editing of multiple sequence alignments.

34. **Jellyfish** - a tool that counts k-mers (substrings of DNA sequences of length k) in a genome. Results can be used for genome size estimation.

35. **K-Nearest Neighbors (KNN) Regression** - calculates distances between the features of a new data instance and the features in other training instances to predict the new instance's target score.

36. **Lasso Regression** - a type of penalized linear regression with L1 regularization, meaning that the penalty term is the sum of the absolute value of the coefficients. One feature among highly correlated features will be randomly chosen to stay in the model and the coefficients of the others will be set to 0.

37. **LoRDEC** - a tool for error-correction of long reads from PacBio sequencing. The tool builds a De Bruijn Graph (DBG) representing low error rate short reads. A path that corrects the sequence of the long read is identified for each error found on the long read.

38. **Maker** - an open-source pipeline for genome annotation. Maker annotates genes based on *ab-initio* gene predictions and alignments of protein and EST sequences to a genome.

39. **Metassembler** - a tool that iteratively updates a starting assembly based on pair-wise alignments to other assemblies. Conflicting regions between assemblies are resolved by using the local sequence with the best Compression–Expansion (CE) statistic. Different assembly merge orders produce different final assemblies.

40. **MUMmer** - an open source software package for the fast alignment of large nucleotide and protein sequences.

41. **MUSCLE (Multiple Sequence Comparison by Log-Expectation)** - a tool for multiple sequence alignment, specifically created for aligning three or more protein sequences.

42. **N50 Length** - the length of the scaffold where the total length of that scaffold plus all longer scaffolds is equal to or is more than half the length of the genome.

43. **NUCmer** - a tool within the MUMmer package that is used to align closely related nucleotide sequences.

44. **Ordinary Least Squares (OLS)** - a method for estimating the unknown parameters (coefficients of the features) in a linear regression model. The goal is to minimize the sum of the squares of the differences between the observed responses in the training data and those predicted by the linear model.

45. **Orthologs** - homologous genes found in different species that evolved from the same ancestral gene

46. **Paired Reads** - two reads, one from each end of a fragment.

47. **PAM250** - a PAM (Point Accepted Mutation) substitution matrix is used for scoring alignments between sequences. Each cell in the matrix contains a value reflecting relatedness between the amino acids of each substitution. The alignment score is the sum of the scores for each pair of aligned amino acids. The

PAM250 matrix represents 250% change in the amino acids over a certain amount of molecular evolution.

48. **Penalized Linear Regression** – type of Ordinary Least Squares (OLS) linear regression, where models that are more complex are penalized to avoid overfitting.

49. **Proovread** - a tool for error-correction of long reads from PacBio sequencing using short-reads or unitigs (high-quality assembly fragments). Proovread maps short reads to the long read and identifying the consensus sequence. The short-reads can be remapped to this consensus sequence. Because more short reads may be able to map after the first-round of errors have been corrected, additional error-corrections may be completed.

50. **Properly Paired Reads** - paired reads that are in the expected orientation and distance apart when mapped to the reference genome.

51. **Pybedtools** - a python wrapper for the BEDTools suite. BEDtools include tools that can count, intersect, merge, and complement genomic intervals from BED, BAM, and GFF/GTF files.

52. **Random Forest** - an ensemble machine learning method for either classification or regression that is based on multiple decision trees. The output is either the mode class (classification) or mean prediction/calculation (regression) of the individual trees.

53. **Reads** - pieces of sequenced DNA, which are outputs of the sequencing machine.

54. **Read Coverage** - the number of reads mapping over a certain base in the assembly.

55. **REAPR (Recognition of Errors in Assemblies using Paired Reads)** - a tool that uses mapped paired end reads to produce a variety of metrics that reflect the accuracy of an assembly. The tool does not require a reference genome.

56. **Reference Assembly** - the main set of DNA sequences used to represent the genome of an organism. The highest quality assembly for an organism should be used as the reference.

57. **REViGO** - a web server that reduces long lists of GO terms into representative subset determined by clustering algorithms. REVIGO can also generate a variety of visualizations of this subset of GO terms.

58. **Ridge Regression** - a type of penalized OLS linear regression with L2 regularization, meaning that the penalty term is the sum of the square of the coefficients. The coefficients cannot be zero in Ridge regression (no features are eliminated), but will be shrunk by the same factor.

59. **SAM (Sequence Alignment/Map)/BAM (Binary Alignment/Map)** - the file format used to store large sequence alignments including the mapping of sequencing reads to a reference genome.

60. **SAMtools** - a suite of tools for manipulating and using alignments in the SAM format, including a method to calculate read pair statistics from SAM files.

61. **Scaffold** - a sequence formed from more than one contig that could be connected and ordered, possibly over a gap.

62. **SINC Simulator** - a tool that simulates SNPs, indels, and CNVs in a given FASTA sequence.

63. **Single Nucleotide Polymorphism (SNP)** - a variation in a single nucleotide in a sequence from the reference sequence.

64. **Supervised Machine Learning** - machine learning where the model is learned from labeled training data. Labeled training data consists of a set of features (inputs) and the output value to be predicted/calculated in the model.

65. **Support Vector Machine (SVM) Regression** - a method to learn the optimal hyperplane, given labeled training data, for scoring new samples.

66. **Testing Set** - the set of data instances used to assess the accuracy of a machine learning model.

67. **Training Set** - the set of data instances used to learn parameters of a machine learning model.

Appendix B

QUALITY METRICS FOR CH ASSEMBLIES FOR CHAPTER 2

80 different quality metrics were calculated for each of the Chinese hamster assemblies described in Chapter 2. The descriptions and results for each metric, separated into multiple tables by class type, are provided in this appendix. All data was collected by Oliver Rupp. Rankings derived from this data were used to select the new Chinese hamster reference assembly (Chapter 2) and were also used to evaluate EvalDNA's performance on the CH assemblies (Chapter 3).

Table B.1: Quality metrics for the contig/scaffold numbers class.

Metric	**CSA**	**2013 RefSeq**	**Illumina**	**PacBio**	**PICR**	**PIRC**	**IPCR**	**IPRC**	**Description**
Contigs	319,409	218,862	127,687	110,957	5,066	5,075	74,845	74,840	Number of contigs
Scaffolds	28,749	52,710	17,373	110,956	1,829	1,825	2,317	2,304	Number of scaffolds
Gap %	10.45	2.49	2.66	0.00	0.12	0.12	1.13	1.13	Percent of 'N' bases
Gap Count	290,660	166,152	110,314	1	3,237	3,250	72,528	72,536	Number of 'N'-stretches
Mean gap length [bp]	838	354	577	1	887	884	369	367	Mean length of 'N'-stretches
L50 contig [bp]	11,892	27,129	35,636	995,273	1,894,627	1,954,687	56,166	56,058	N50 length of contigs
L50 scaffold [bp]	1,236,516	1,558,295	5,951,711	995,273	20,188,717	19,582,713	21,744,884	21,262,794	N50 length of scaffolds
L90 contig [bp]	2,811	7,091	9,459	12,762	349,951	354,712	16,356	16,347	N90 length of contigs
L90 scaffold [bp]	180,686	395,288	1,003,287	12,762	4,400,567	4,422,379	3,545,615	3,650,273	N90 length of scaffolds
Longest contig [bp]	166,487	219,443	353,448	16,079,867	9,313,581	11,070,296	625,642	625,658	Length of longest contig
Longest scaffold [bp]	14,658,418	8,324,132	25,843,536	16,079,867	80,584,097	80,583,080	66,348,894	66,348,834	Length of longest scaffold
N50 contig	50,682	25,879	19,661	655	337	342	12,629	12,658	N50 number of contigs
N50 scaffold	501	450	128	655	32	33	33	34	N50 number of scaffolds
N90 contig	186,854	87,321	67,035	54,054	1,392	1,382	41,845	41,904	N90 number of contigs
N90 scaffold	2,251	1,558	468	54,054	121	122	122	122	N90 number of scaffolds

Table B.2: Quality metrics for the sequence content class. Each row contains the number of bases missing from the specified assembly for each of the other assemblies (column). *This metric is the sum of all bases from the assemblies that are missing. Homologuous regions missing in more than one assembly were counted multiple times.

Metric [Mbp]	**CSA**	**2013 RefSeq**	**Illumina**	**PacBio**	**PICR**	**PIRC**	**IPCR**	**IPRC**
Missed CSA sequence	0	23.834	76.106	14.627	14.049	14.162	75.178	75.650
Missed Illumina sequence	268.732	126.750	0	100.688	98.818	99.175	24.497	25.675
Missed IPCR sequence	246.548	97.471	18.698	73.115	70.198	70.626	0	1.400
Missed IPRC sequence	246.037	96.271	18.730	71.996	69.174	69.580	0.286	0
Missed PacBio sequence	1,236.695	362.430	648.566	0	144.958	208.685	437.864	434.222
Missed PICR sequence	212.195	29.862	92.987	3.341	0	0.558	72.567	72.671
Missed PIRC sequence	200.553	29.695	93.346	3.031	0.389	0	72.798	72.877
Missed RefSeq sequence	185.821	0	93.934	3.213	9.836	9.632	79.965	79.926
Missed Total sequence*	2,596.581	766.314	1,042.367	270.010	407.424	472.418	763.155	762.422

Table B.3: Quality metrics for the feature content class. CDSs from mouse were mapped to the different assemblies using GMAP. A CDS is "complete" if coverage $\geq 95\%$ and identity $\geq 75\%$. A CDS is "missing" if coverage $< 25\%$ or identity $< 75\%$. If more than one location was found for a CDS, it was classified as "chimeric".

Metric	**CSA**	**2013 RefSeq**	**Illumina**	**PacBio**	**PICR**	**PIRC**	**IPCR**	**IPRC**	**Description**
Chimeric Mouse CDS	688	574	379	330	195	218	243	241	# of chimeric mouse CDS (GMAP)
Complete CEGMA	206	222	225	230	227	228	223	222	# of complete CEGMA genes
Complete eukaryota	193	189	184	193	206	205	191	192	# of complete eukaryotic genes (BUSCO)
Complete metazoa	585	621	591	620	640	644	610	615	# of complete metazoan genes (BUSCO)
Complete Mouse CDS	14116	14045	15722	17001	17201	17210	16361	16351	# of complete mouse CDS
Complete vertebrata	1302	1311	1332	1396	1407	1414	1353	1346	# of complete vertebrata genes (BUSCO)
Duplicated eukaryota	16	20	24	24	21	20	23	23	# of duplicated BUSCO eukaryotic genes
Duplicated metazoa	25	21	30	34	29	29	27	27	# of duplicated BUSCO metazoan genes
Duplicated vertebrata	20	17	22	38	23	23	22	22	# of duplicated BUSCO vertebrata genes
Duplication CEGMA	1.96	1.91	1.99	2.03	1.99	1.98	1.93	1.93	CEGMA duplication ratio
GMAP Indels	16.04	15.93	12.16	10.79	10.86	10.86	11.82	11.84	Mean # of indels (only "complete" genes)
Missing CEGMA	20	15	16	14	15	15	18	18	# of missing CEGMA genes
Missing eukaryota	189	184	195	188	177	178	187	187	# of missing BUSCO eukaryotic genes
Missing metazoa	183	164	185	157	148	146	175	171	# of missing BUSCO metazoan genes
Missing Mouse CDS	2772	2315	2408	1932	1822	1827	2173	2185	# of missing mouse CDS
Missing vertebrata	1140	1162	1225	1126	1128	1113	1187	1193	# of missing BUSCO vertebrata genes

Table B.4: Quality metrics for the chromosome-sorted read coverage class. Correct "OK" regions have a chromosome read coverage between 0.25 and 2. Low-coverage ("LC") regions are regions where the highest chromosome coverage is < 0.25. High-coverage ("HC") regions are regions where the highest chromosome coverage > 2. Ambiguous ("AMB") regions are regions where the second highest chromosome coverage is at least 90% of the highest chromosome coverage. "ERR" regions are regions that have exactly one chromsome coverage between 0.25 and 2, but the chromosome differs from the assigned chromosome of the scaffold.

Metric	**CSA**	**2013 RefSeq**	**Illumina**	**PacBio**	**PICR**	**PIRC**	**IPCR**	**IPRC**	**Description**
AMB	0.48	0.58	0.47	0.50	0.55	0.55	0.51	0.52	% of ambiguous regions
AMB 1M	0.58	0.53	0.48	0.44	0.53	0.53	0.51	0.51	% of ambiguous regions ($\geq$ 1 Mbp)
ERR	0.22	1.10	1.15	0.78	0.71	0.72	3.75	3.75	% of error regions
ERR 1M	0.22	1.34	1.23	0.14	0.67	0.67	3.89	3.88	% of error regions ($\geq$ 1 Mbp)
HC	1.18	2.22	1.80	0.45	2.04	2.04	1.87	1.88	% of high coverage regions
HC 1M	0.28	0.26	0.29	0.09	0.30	0.29	0.37	0.36	% of high coverage regions ($\geq$ 1 Mbp)
LC	7.23	3.45	5.43	20.26	0.36	0.37	4.15	4.14	% of low coverage regions
LC 1M	4.95	3.04	4.83	0.77	0.29	0.29	4.00	4.00	% of low coverage regions ($\geq$ 1 Mbp)
OK	90.89	92.65	91.15	78.01	96.34	96.32	89.71	89.72	% of correct coverage regions
OK 1M	93.97	94.83	93.18	98.55	98.22	98.22	91.24	91.24	% of correct coverage regions ($\geq$ 1 Mbp)
TOTAL [kbp]	1,361,777	1,633,249	2,149,501	1,898,261	2,266,484	2,265,755	2,249,548	2,249,600	Total size of OK regions ($\geq$ 1 Mbp)

Table B.5: Quality metrics for the remap statistics class. Each row contains the percentages of reads that could be mapped back to each assembly from the specified library.

Metric	CSA	2013 RefSeq	Illumina	PacBio	PICR	PIRC	IPCR	IPRC	Library Description
180a	87.37	96.95	92.57	99.65	98.55	98.66	94.67	94.66	whole genome paired-end
180b	87.29	96.91	92.55	99.64	98.55	98.65	94.65	94.65	whole genome paired-end
180c	87.35	96.95	92.57	99.65	98.56	98.66	94.67	94.66	whole genome paired-end
180d	87.33	96.92	92.54	99.64	98.54	98.65	94.65	94.64	whole genome paired-end
180e	87.27	96.87	92.51	99.63	98.53	98.63	94.63	94.62	whole genome paired-end
chr1a	77.00	84.20	81.70	87.45	86.17	86.21	82.84	82.85	chromosome 1 paired-end
chr1b	73.88	81.14	78.71	84.71	83.25	83.28	79.85	79.86	chromosome 1 paired-end
chr1c	73.93	81.17	78.74	84.74	83.28	83.31	79.88	79.90	chromosome 1 paired-end
chr2a	83.25	88.01	85.36	90.63	89.86	89.91	86.39	86.38	chromosome 2 paired-end
chr2b	82.08	86.94	84.30	89.78	88.93	88.97	85.34	85.33	chromosome 2 paired-end
chr2c	82.20	87.06	84.42	89.90	89.05	89.09	85.46	85.45	chromosome 2 paired-end
chr3a	79.54	88.25	86.09	91.83	90.79	90.85	87.14	87.12	chromosome 3 paired-end
chr3b	77.33	86.08	83.97	90.06	88.75	88.80	84.97	84.95	chromosome 3 paired-end
chr4a	76.79	88.06	84.68	91.85	90.60	90.68	86.19	86.17	chromosome 4 paired-end
chr5a	73.50	75.63	79.87	79.41	80.12	79.95	79.57	79.44	chromosome 5 paired-end
chr6a	82.09	87.74	85.79	91.24	90.46	90.49	87.55	87.49	chromosome 6 paired-end
chr7a	82.82	91.17	86.56	94.76	93.23	93.40	88.94	88.86	chromosome 7 paired-end
chr8a	71.53	78.36	75.75	82.14	80.75	80.86	77.30	77.33	chromosome 8 paired-end
chr9a	68.10	88.04	76.36	95.63	92.05	92.35	85.68	85.60	chromosome 9/10 paired-end
chr9b	66.73	86.89	75.03	94.92	91.03	91.34	84.43	84.35	chromosome 9/10 paired-end
chrxa	66.18	82.23	80.75	88.70	85.91	85.91	82.22	82.24	chromosome X paired-end

Table B.6: Quality metrics for the CE statistics class, which were collected using the REAPR tool. Length metrics are in Mbp.

Metric	**CSA**	**2013 RefSeq**	**Illumina**	**PacBio**	**PICR**	**PIRC**	**IPCR**	**IPRC**	**Description**
FCD length	6.67	14.41	12.03	52.03	8.03	10.43	15.10	15.03	Total size of ungapped FCD errors
FCD_gap length	133.35	44.29	53.57	0	1.80	2.15	41.67	41.47	Total size of gapped FCD errors
Frag_cov length	0.02	0.60	4.92	3.36	0.23	0.22	4.82	4.83	Total size of ungapped fragment coverage errors
Frag_cov_gap length	2.03	1.10	54.78	0	0.27	0.27	54.30	54.18	Total size of gapped fragment coverage errors
FCD	1,916	2,921	2,294	10,616	1,613	2,151	2,764	2,757	# of ungapped FCD errors
FCD_gap	24,994	7,224	7,013	0	227	269	4,597	4,572	# of gapped FCD errors
Frag_cov	84	219	1,061	4,356	207	199	1,039	1,043	# of ungapped fragment coverage errors
Frag_cov_gap	791	139	2,980	0	102	135	2,972	2,974	# of gapped fragment coverage errors

Appendix C

QUALITY METRIC DEFINITIONS FOR CHAPTER 3

All of the quality metrics examined during feature selection in Chapter 3 are described here. Metrics are converted into percentage of bases per assembled sequence or normalized by assembly length if needed. Note that not all of the metrics were included in the final quality scoring model due to multicollinearity or lack of significant correlation with the reference-based quality score in the training data.

1. Normalized N50 length - N50 length normalized by total sequence length.

$$normN50 = \frac{N50\_length}{total\_length} * 100, \;\; 0 \leq N50\_length \leq 100 \tag{C.1}$$

2. Gap percent - percent of total bases which are gaps (N's).

$$gap\_percent = \frac{total\_gap\_length}{total\_length} * 100, \;\; 0 \leq gap\_percent \leq 100 \tag{C.2}$$

3. Normalized contig count - Number of separate pieces (scaffolds/contigs) the sequence of interest is split into normalized by the sequence length in megabases (Mbp).

$$norm\_contig = \frac{total\_contigs}{total\_length} * 1,000,000, \;\; 0 \; < \; norm\_contig \; < \; 1,000,000 \tag{C.3}$$

4. Links - percent of total bases impacted by link errors called by REAPR. These bases are located in regions where a significant proportion of the reads mapped to this region also mapped elsewhere.

$$links = \frac{bases\_in\_link\_regions}{total\_length} * 100, \;\; 0 \leq links \leq 100 \tag{C.4}$$

5. Collapsed repeats - percent of total bases impacted by the collapsed repeat errors called by REAPR.

$$collapsed\_repeats = \frac{bases\_in\_collapsed\_repeats}{total\_length} * 100, \quad 0 \leq collapsed\_repeats \leq 100 \tag{C.5}$$

6. Clips - percent of total bases impacted by the clip errors called by REAPR. These bases are located in regions where a significant proportion of the reads had to be clipped to map to this region.

$$clip = \frac{bases\_in\_clip\_regions}{total\_length} * 100, \;\; 0 \leq clip \leq 100 \tag{C.6}$$

7. Low read coverage - percent of total bases impacted by the low read coverage errors called by REAPR. These bases are in regions with low coverage of proper paired reads.

$$low\_read\_coverage = \frac{bases\_in\_low\_read\_coverage\_regions}{total\_length} * 100, \quad 0 \leq low\_read\_coverage \leq 100 \tag{C.7}$$

8. Properly paired read percent - percent of mapped reads that are properly paired as determined by SAMtools.

$$proper\_pair\_percent = \frac{reads\_in\_proper\_pairs}{total\_reads\_mapped} * 100, \quad 0 \leq proper\_pair\_percent \leq 100 \tag{C.8}$$

9. Error free bases - percent of bases called by REAPR as error free. A base is called error free if it has at least 5x coverage of perfect and unique mapped reads.

$$error\_free\_bases = \frac{total\_error\_free\_bases}{total\_length} * 100, \;\; 0 \leq error\_free\_bases \leq 100 \tag{C.9}$$

10. Fragment coverage distribution (FCD) errors in contig - percent of bases in regions that REAPR marks as an FCD error within a contig (the region does not contain any gaps).

$$FCD\_err\_in\_contig = \frac{bases\_in\_FCD\_error\_contig\_regions}{total\_length} * 100, \quad 0 \leq FCD\_err\_in\_contig \leq 100 \tag{C.10}$$

11. FCD errors over gap - percent of bases in regions that REAPR marks as an FCD error and the region contains a gap.

$$FCD\_err\_over\_gap = \frac{bases\_in\_FCD\_error\_gap\_regions}{total\_length} * 100, \quad 0 \leq FCD\_err\_over\_gap \leq 100 \tag{C.11}$$

12. Low fragment coverage (FC) in contig - percent of bases in regions that REAPR marks as having low fragment coverage and the region does not contain any gaps.

$$low\_fc\_in\_contig = \frac{bases\_in\_low\_FC\_contig\_regions}{total\_length} * 100, \quad 0 \leq low\_fc\_in\_contig \leq 100 \tag{C.12}$$

13. Low fragment coverage (FC) over gap - percent of bases in regions that REAPR marks as having low fragment coverage and the region contains a gap.

$$low\_fc\_over\_gap = \frac{bases\_in\_low\_FC\_gap\_regions}{total\_length} * 100, \quad 0 \leq low\_fc\_over\_gap \leq 100 \tag{C.13}$$

Appendix D

REPRINT PERMISSIONS

D.1 Reprint Permissions for Figure 1.1 from Chapter 1

Publication: Figure from John Eid, Adrian Fehr, Jeremy Gray et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009 Jan 2.

Permission: This is a License Agreement between Madolyn L. MacDonald and The American Association for the Advancement of Science provided by the Copyright Clearance Center.

Table D.1: License agreement for Chapter 1, Figure 1.1

License Number	4558390506267
License Date	Mar 29, 2019
Licensed Content Publisher	The American Association for the Advancement of Science
Licensed Content Publication	Science
Licensed Content Title	Real-time DNA sequencing from single polymerase molecules
Licensed Content Author	John Eid, Adrian Fehr, Jeremy Gray et al.
Licensed Content Date	Jan 2, 2009
Licensed Content Volume	323
Licensed Content Issue	5910
Type of Use	Thesis/Dissertation
Requestor type	Scientist/Individual at Research Institution
Format	Electronic
Portion	Excerpt
Title of Thesis/Dissertation	Development of a Novel, Reference-Free Tool for the Comprehensive Evaluation of Genome Assembly Quality and its Application to Establish a Reference Assembly for Chinese Hamster Ovary (CHO) Cells

D.2 Reprint Permissions for Chapter 2

Publication: O. Rupp, M. L. MacDonald, S. Li, H. Dhiman et al. A reference genome of the Chinese hamster based on a hybrid assembly strategy. Biotechnol Bioeng. 2018 Apr 28.

Permission: This is a License Agreement between Madolyn L. MacDonald and John Wiley & Sons provided by the Copyright Clearance Center.

License: This article is available under the terms of the Creative Commons Attribution License (CC BY) (which may be updated from time to time) and permits use, distribution and reproduction in any medium, provided that the Contribution is properly cited. Permission is not required for this type of reuse.

D.3 Reprint Permissions for Chapter 4

Publication: M. L. MacDonald, N. K. Hamaker, K. H. Lee. Bioinformatic analysis of Chinese hamster ovary host cell protein lipases. AIChE Journal. 2018 Sep 11.

Permission: This is a License Agreement between Madolyn L. MacDonald and John Wiley & Sons provided by the Copyright Clearance Center.

Table D.2: License agreement for Chapter 4

License Number	4533740704989
License Date	Feb 21, 2019
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	AIChE Journal
Licensed Content Title	Bioinformatic analysis of Chinese hamster ovary host cell protein lipases
Licensed Content Author	Madolyn L. MacDonald, Nathaniel K. Hamaker, Kelvin H. Lee
Licensed Content Date	Sep 11, 2018
Licensed Content Volume	64
Licensed Content Issue	12
Licensed Content Pages	8
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Title of Thesis/Dissertation	Development of a Novel, Reference-Free Tool for the Comprehensive Evaluation of Genome Assembly Quality and its Application to Establish a Reference Assembly for Chinese Hamster Ovary (CHO) Cells