DNA SEQUENCE ANALYSIS: NEW APPLICATIONS WITH HIGH THROUGHPUT SEQUENCING AND NEW METHODS IN STUDYING GENE FAMILIES AND HUMAN HAPLOGROUPS

A Dissertation in
Genetics
by
Yazhou Sun

© 2012 Yazhou Sun

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2012
The dissertation of Yazhou Sun was reviewed and approved* by the following:

Stephen Schaeffer
Professor of Biology
Chair of Committee

Hong Ma
Distinguished Professor of Biology
Dissertation Advisor

Gong Chen
Associate Professor of Biology

Anton Nekrutenko
Associate Professor of Biochemistry and Molecular Biology

Naomi Altman
Professor of Statistics

Robert F. Paulson
Associate Professor of Veterinary and Biomedical Sciences
Chair, Intercollege Graduate Degree Program in Genetics

*Signatures are on file in the Graduate School
Understanding the sequential information coded in DNA, RNA and proteins is important for both basic and applied researches in life sciences. Extensive efforts have been devoted to the research and development of DNA sequence analysis methods. The studies described in this dissertation explored new applications of existing methods in the context of the recent development of ultra-high throughput sequencing technologies. This dissertation also included new methods developed for studying gene families and human haplogroups. The theories, algorithms and tools for analyzing DNA sequence information concerning these studies are reviewed in Chapter 1 of this dissertation.

With the recent development in DNA sequencing technologies, came many new research opportunities. Great challenges also came along, mainly because of the large data size of the latest high throughput sequencing technologies. The potential of these new technologies was exploited to complete a 100,000 years old ancient polar bear mitochondrial genome. With this and some additional modern bear data, the matrilineal polar bear’s divergence time was estimated to be around 130,000 years ago, which is significantly older than some recent estimates. This estimate indicated that modern polar bear matrilineal ancestors adapted to the niche polar environment within 30,000 years after the speciation event and propagated along the entire Arctic Circle for the next 100,000 years. This recent speciation and rapid expansion process is analogous to the evolution and migration of modern humans. The lineage characteristics of the latter were also briefly studied using the same technologies. (Chapter 2)

Because of the increased efficiency from the latest sequencing technologies, more and more complete human mitochondrial genomes have been generated at an increasingly faster
speed. Although mitochondrial haplogroups, and their classification and identification were widely used in human evolution and population studies, the current tools could not fully take advantage of the rapidly growing number of new mitochondrial genomes. An updated mitochondrial haplogroup classification system was thus developed with evolutionary models that incorporate the mitochondrial genomic variations within the human population. These variations have not been considered by previous methods, which could lead to incorrectly classified haplogroups. The variation parameters, including the whole-genome substitution rate (0.013 - 0.1 substitutions per generation), the rate heterogeneity among sites (Gamma distribution shape parameter $\alpha = 0.7078$) and the percentage of invariant sites (64%), were estimated based on 7985 full-length human mitochondrial genome sequences. Haplogroups were then classified based on the corrected genetic distance estimation and modeled with position specific matrices. A new haplogroup identification system was developed based on the resulting matrices and the maximum-likelihood estimation (MLE) method, permitting fast and accurate haplogroup assignment for both known and new mitochondrial genomes. The entire system is available through the HapSearch web application (http://hapsearch.synblex.com). (Chapter 3)

The latest sequencing technologies also allowed a more thorough study of stage-specific transcriptional activities. To elucidate the transcriptomic profiles and new transcriptomic activities in neural development, nine recent RNA-seq datasets corresponding to tissues/organs ranging from stem cell, embryonic brain cortex to adult whole brain were analyzed. The global similarities between the neural and stem cell transcriptomes were found on both genic and chromosomal levels. A previously undocumented high level of unannotated expression was found in mouse embryonic brain cortices, the intronic part of which was found to be strongly associated with gene ontology (GO) categories that are important for synaptogenesis and neural circuit
formation. This suggested potentially novel genes, gene functions and regulatory mechanisms in early brain development. (Chapter 4)

Although the speed of generating genomic sequences was increasing rapidly, the development of genome annotation was lagging behind. This slowed down or prevented a broader utilization of the newly sequenced genomes. To partially mitigate this situation, a new tool, called Phoenix, was developed for retrieving homologues of a given gene or gene family from unannotated genomes. Phoenix exhibited fast and accurate performance in simulation using known gene families’ data. Its advantage was further demonstrated by correctly retrieving homologues of a gene family that has a known complex evolutionary history. This tool allows gene family studies in unannotated genomes or even partially assembled genomes. (Chapter 5)

Finally, this dissertation concluded with a discussion of the intrinsic limitations and advantages of the DNA sequence analysis, along with its current and future application potentials. (Chapter 6)
TABLE OF CONTENTS

LIST OF FIGURES .................................................................................................................. x

LIST OF TABLES .................................................................................................................... xii

ACKNOWLEDGEMENTS ......................................................................................................... xiv

Chapter 1 ................................................................................................................................. 1

Background ............................................................................................................................. 1

A brief history on biological sequence analysis................................................................. 1
Background on pairwise DNA sequence alignment and analysis................................. 5
  Basic pairwise sequence alignment algorithms ............................................................ 5
  Score matrix and scoring alignment ................................................................................. 7
  Heuristics algorithms for performance considerations .................................................. 9
  Heuristic algorithms for ultra high-throughput sequencing data mapping ................ 10
Background on multiple DNA sequence alignment and analysis .................................. 11
  Heuristic algorithms for multiple sequence alignment ................................................ 12
  Scoring a multiple sequence alignment ....................................................................... 13
  Profiling a multiple sequence alignment ..................................................................... 14
Estimate evolutionary distances from sequence alignment ........................................... 17
  Theories of estimating the number of substitutions between two sequences .......... 17
  Estimate the substitution rate variations among sites ................................................... 20
Phylogeny reconstruction ................................................................................................. 22
  Phylogeny reconstruction from evolutionary distances ................................................. 22
  Phylogeny reconstruction from discrete sites by maximum parsimony method ...... 24
  Phylogeny reconstruction from maximum likelihood (ML) method ........................ 25
  Reconstructing phylogeny by sampling (Bayesian method) ........................................ 25
  Phylogeny reconstruction without sequence alignment ................................................. 27
Molecular dating .................................................................................................................. 28
Information analysis between biological sequences of different types ......................... 29
  Between DNA and RNA sequences ........................................................................... 29
  RNA-seq data mapping .............................................................................................. 30
  Between RNA and protein sequences ....................................................................... 31
  Between DNA and protein sequences ...................................................................... 31
Applications of DNA sequence analysis in genome assembly and annotation .......... 33
  Genome assembly problem ......................................................................................... 33
  Algorithms and tools for de novo genome assembly .................................................. 34
  Genome annotation problem ....................................................................................... 37
  Homology-based annotation ......................................................................................... 39
  Model-based annotation ............................................................................................. 40
  Integrated annotation .................................................................................................. 42

Chapter 2 Information analysis of mammalian mitochondrial DNA sequence in molecular dating applications .................................................. 44
Chapter 3 Probabilistic modeling of major human mitochondrial haplogroups and parameterization based on population

Project background ............................................................... 73
Introduction ........................................................................... 74
Analyses and Results ............................................................... 78
Retrieval of mitochondrial genome sequence data and initial analysis .......... 78
Estimation of mitochondrial genome variation parameters from human population data ........................................................... 79
Variable regions outside D-loop and distribution of disease-related SNPs .......... 86
Identifying major haplogroups by clustering based on estimated parameters .......... 87
Not including substitution rate variation among sites led to different phylogenies from the same dataset ........................................................... 89
Probabilistic haplogroup model and Maximum-Likelihood profile search .......... 91
High accuracy in haplogroup identification supported using simulated new sequences modeled after human mitochondrial genome-wise substitution patterns ........................................................................... 92
Materials and methods ................................................................ 93
Data collection ........................................................................... 93
Methods for genetic distance estimation in the presence of substitution rate variation between nucleotides and among sites ........................................................... 94
Correction for invariant sites with invariant sites removal (ISR) .................. 96
Parameter estimation .................................................................. 96
Software implementation ............................................................... 97
Summary and discussion .............................................................. 97
Newly estimated population variation parameters and the implications .......... 97
Limitations in haplogroup resolution and modeling ............................ 100
The effect of differences between intraspecific and interspecific data on rate and divergence time estimation.................................................................101
The effect of differences between intraspecific and interspecific data on the pan-genome concept and disease SNP discovery.................................104

Chapter 4  DNA-based information analysis of RNA-seq data.................................107

Project background........................................................................................................107
Introduction.......................................................................................................................108
The C-value paradox and the extra information provided by transcriptome ..........108
Further transcriptome complexity revealed by RNA-seq ...............................................109
Assessing extra information conveyed by neurodevelopmental transcriptomes ....110
Analyses and Results.........................................................................................................111
Mapping gigabytes of RNA-seq data...........................................................................111
Medium and high resolution genome-wide transcription activity profile ...............112
Genome-wide comparison revealed similarity between neural and stem cell transcriptomes .................................................................117
Expression pattern of selected neurodevelopmental disorders associated genes ....128
High level of unannotated expression at E18 and P7 stages ....................................130
Concordant expression changes between intronic TARs and flanking exons .........133
Few unannotated TARs connected with known exons ..............................................134
Few newly discovered intronic TARs were present in known mRNA and EST databases..................................................................................................................136
Few intronic TARs present in miRbase and lncRNAdb ............................................138
Cases of high sequence conservation and coding potential of intronic TARs .....138
Association of intronic TARs in neurodevelopmental processes .......................140
Materials and Methods..................................................................................................147
RNA-seq mapping.........................................................................................................147
Data normalization .........................................................................................................148
Unannotated transcriptionally active region (TAR) calling ....................................148
RT-PCR experiment for intronic TARs .....................................................................149
Modeling connection between unannotated TARs and known transcripts .........150
Testing model effectiveness .........................................................................................152
Gene Ontology (GO) analysis ......................................................................................153
Project summary and discussion..................................................................................155
Similarity between neural and stem cell transcriptomes ........................................155
Intronic TARs as regulators in early brain development .........................................156
Time-sensitive sampling and its application in revealing time-dependent complexity .................................................................................................157

Chapter 5  Short-take: DNA-based protein sequence retrieval in unannotated genomes ......160

Project background........................................................................................................160
Introduction.......................................................................................................................160
Challenges of genome annotation in the era of ultra high-throughput sequencing..160
Method of retrieving a certain gene or gene family based on homology in genomes .................................................................161
Current homologue retrieving tools and their limitations ........................................163
LIST OF FIGURES

Figure 1-1: Example of LCS computation between two DNA sequences......................... 6

Figure 1-2. Gamma probability density function distribution with $\beta = 1/\alpha$, $\alpha \in (0.3, 2)$ on $X \in (1, 10)$............................................................................................................ 21

Figure 2-1: Sampling site locations. ................................................................................. 49

Figure 2-2: Estimated composition of the recovered DNA ............................................. 50

Figure 2-3: Coverage of mitochondria genome assembled from ancient polar bear DNA. .... 51

Figure 2-4: Genome variation pattern among all known complete ursine mitochondria. ...... 52

Figure 2-5: Polar bear evolution history reconstructed with full length mitochondrial sequence and leaf node calibration. .................................................................................. 54

Figure 2-6: Phylogeny from BEAST analysis on control region sequences. ...................... 57

Figure 2-7: BEAST analysis of 162 individuals based on complete Mitochondrial sequences................................................................................................................................. 66

Figure 2-8: Dataset haplogroup composition..................................................................... 68

Figure 2-9: Sequence differences of the assembled mitochondrial genomes. .................... 69

Figure 3-1. Geospatial distribution of number of mitochondrial genome sequences from countries around the world................................................................. 78

Figure 3-2. Number of invariant sites decreases as more genomes are sampled.............. 79

Figure 3-3. Estimating substitution rate and variation parameters. .................................. 84

Figure 3-4. Mitochondrial site-specific substitutions estimated from 7975 modern human... 85

Figure 3-5. Distribution of estimated substitutions and reported diseases-related SNPs in human mitochondrial genome......................................................................................... 87

Figure 3-6. Haplogroup clusters inferred from genetic distance based on Neighbor-Joining guide tree. ................................................................................................. 89

Figure 3-7. Haplogroup clusters inferred based on the model that does not consider rate variation among sites and the presence of invariant sites. ........................................ 90
Figure 3-8. Percentage of new sequences matched with original haplogroup classification with introduced mutations following known mitochondrial genome-wide substitution patterns. ................................................................. 93

Figure 3-9. Genome length variation in collected data................................................................. 94

Figure 3-10: Distribution of sample variance at 20 bp resolution of human mitochondrial genome. ................................................................. 106

Figure 4-1: Genome-wide expression map at 1 Mb resolution for E18 and P7......................... 114

Figure 4-2: 100 Kb resolution expression map of chromosome X ............................................. 115

Figure 4-3: Genomic expression analysis in 1 Kb interval.......................................................... 116

Figure 4-4: Chromosomal expression level between stages....................................................... 119

Figure 4-5: Sample/Stage clustering based on Correspondence Analysis of chromosomal expression profiles. .................................................................................. 121

Figure 4-6: Sample/Stage clustering based on Orthologous gene expression pattern......... 126

Figure 4-7: Selected gene expression level at different stages. ................................................. 126

Figure 4-8: Expression pattern of genes associated with neurodevelopmental disorders...... 128

Figure 4-9: Unannotated transcript expression. ........................................................................ 132

Figure 4-10: Neural signaling GO terms significantly overrepresented by genes with intronic TARs......................................................................................... 141

Figure 4-11: Neurogenesis GO terms significantly overrepresented by genes with intronic TARs ......................................................................................... 142

Figure 4-12: Regulation GO terms significantly overrepresented by genes with intronic TARs. ......................................................................................... 143

Figure 5-1: Phoenix program algorithm. .................................................................................. 167

Figure 5-2: Elongation factor 2 gene tree based on sequences retrieved by Phoenix......... 171

Figure 5-3: An example profile HMM model structure............................................................ 179
### LIST OF TABLES

Table 1-1: Genome size of milestone genome projects .................................................. 4

Table 1-2: DNA score matrix example estimated from alignment of 9000 human mitochondria data based on normalized log-odds ratio ........................................... 7

Table 2-1: Genome sequencing and assembly information ........................................... 52

Table 2-2: Comparison of recent ursine clade molecular dating results ....................... 55

Table 2-3: Statistics of assembled human mitochondrial genomes .............................. 63

Table 2-4: Comparison of molecular dating results for modern human lineages ............ 67

Table 2-5: Number of SNPs identified between mitochondrial genomes ..................... 69

Table 3-1: Positive-strand substitution probability between nucleotides in human mitochondrion ........................................................................................................ 85

Table 3-2: Statistics of assembled human mitochondrial genomes .............................. 88

Table 4-1: RNA-seq mapping result using TopHat ..................................................... 112

Table 4-2: Chromosomal expression across stages (read count) .................................. 120

Table 4-3: Principal inertias (eigenvalues) of Correspondence Analysis .................... 120

Table 4-4: Intronic and exonic expression level changes for genes with intronic TARs (from E18 to P7) ................................................................. 134

Table 4-5: Physical connection between unannotated TARs and known exons .......... 135

Table 4-6: E18 intronic TARs with hits in miRbase ..................................................... 137

Table 4-7: E18 intronic TARs with hits in lncRNAdb .................................................. 137

Table 4-8: Exonic and intronic RNA-seq read number comparison between E18 and P7 for exon guidance related genes ......................................................... 144

Table 4-9: Manually validated expressed exons physically connected with other exons .... 154

Table 4-10: Manually validated expressed single exon genes (SGEs) ......................... 154

Table 5-1: Performance of JmjC family homologue retrieval in 10 genomes by Phoenix .... 169
Table 5-3: Sample configuration file for running Phoenix ........................................ 173
Table 5-4: Phoenix command line argument with a configuration file .......................... 173
Table 5-5: Command line arguments of Phoenix ..................................................... 174
Table 5-6: Relation between configuration file parameters and command line arguments... 174
ACKNOWLEDGEMENTS

I want to first thank my PhD advisor, Professor Hong Ma for supporting me and providing me the unique opportunities to study genomics, bioinformatics and computational biology in a broad spectrum of research projects. I’d like to thank the chairman of my thesis committee, Professor Stephen Schaeffer, for his support and advice, especially during my transitions in my PhD. I’d also like to thank members of my thesis committee, Professor Gong Chen, Professor Naomi Altman and Professor Anton Nekrutenko, for their support, inspiration and helpful input. I want to specially thank Professor Chen for his helpful and constructive discussions in my neuroscience-related studies. I also want to thank Professor Altman for her critical and invaluable input from a statistics point of view for many of my research projects. I also thank the Intercollege Graduate Program in Genetics, Chairman of the program Professor Robert Paulson and former Chairman Professor Richard Ordway for giving me the opportunity for pursuing my PhD study in genomic and bioinformatics at Penn State. I’d like to express my most sincere gratitude to Professor Ordway for his support and belief at the most difficult times of my PhD study. I also need to thank Professor Stephan Schuster for his advice, support and collaboration at the earlier half of my PhD study. I want to thank Professor Charlotte Lindqvist in Department of Biological Sciences at University of Buffalo, for her leadership, help, hard work and collaboration in the polar bear mitochondria genome project. I sincerely thank Prof. Mike A. Steel from University of Canterbury for his invaluable help regarding the understanding and implementation of GTR + Gamma + Invariant Site Removal model in my research. I also need to thank Prof. Masatoshi Nei for discussion about the understanding and calculation of genetic distance in general. I’d like to thank Dr. Ji Qi, who introduced me to the field of bioinformatics.
and genomics. I’d also like to thank Dr. Xiaofan Zhou, who collaborated with me on various gene family and evolution related studies when he was a graduate student in Professor Ma’s lab. I want to thank Xinwei Han, my colleague, classmate and officemate, for his input and involvement in neural transcriptome related studies. Finally I’d like to thank members of Ma lab, especially Xuan Ma, for discussions on various data analysis techniques.

I also want to thank my family members for their encouragement and support in my life. I’d like to give my special thanks to my wife, who is also my colleague in various research projects I was involved in, for her care, her love and her belief in me. Finally, I’d like to thank my friends who have helped me and cared about me throughout the years.
Chapter 1

Background

A brief history on biological sequence analysis

Just as why Sir Isaac Newton developed calculus, the development of biological sequence analysis was driven by challenges encountered in life sciences researches. The first biological sequences ever obtained by human were from peptides in the 1950s [1]. Immediately, scientists tried to study these sequences by comparing them from different species. This was achieved by manually aligning different amino acid positions according to similarity. This is the beginning of biological sequence analysis. Although very short, these peptide sequences provided the first glimpse of the power brought by sequence analysis. For example, when studying hemoglobin, Zuckerkandl and Pauling in 1962 found that the amino acid substitution rate is almost linear with the divergence time suggested by paleontological evidence [2]. This directly led to the concept of molecular clock, which then became the founding theory of modern molecular and evolutionary biology.

However, the observed amino acid changes between sequences, as used by Zuckerkandl and Pauling, did not represent the actual number of substitutions between them. Multiple substitutions at the same site could only be experimentally observed as one substitution at most. The longer the evolutionary distance is, the greater the underestimation will be. With the most abstraction and the simplest assumption, Jukes and Cantor in 1969 developed a evolutionary distance measure that extrapolates the hidden substitutions based on observation [3]. Although
the conditions of this method are too restrictive to fit any realistic data, it started a long and proliferating journey of evolutionary model development. Studies involving evolutionary distance estimation to date, which range from ecology to biomedicine, depend on one of the models developed upon the one that Jukes and Cantor wrote in 1969.

With the advent of rapid and reliable DNA sequencing methods [4-6], the available DNA sequences increased in both length and quantity. Determining ancestral sites manually, which was feasible when the sequences were short and the number of sequences is small, became increasingly intractable. Without sites properly aligned, sequence comparison is impossible and no information could be inferred. Needleman and Wunsch in 1970 provided an algorithm to find the global optimal alignment between two protein sequences [7]. This is the first sequence alignment algorithm specifically designed for biological sequences. However, algorithms for DNA sequence alignment did not appear until ten years later, when Smith and Waterman introduced the first general algorithm that works both on DNA and proteins. More importantly, it finds the local optimum instead of a global one, which makes possible searching from a large database of heterogeneous sequence origin. By applying heuristics to this algorithm, faster and more efficient sequence alignment tools were built that formed the foundation of modern day sequence comparison studies. Among these tools, BLAST is the most notable because it has become an indispensable tool to any biologist [8, 9].

Fossil records and paleontological analyses were once the only way to study evolution. With the accumulation of sequence data, especially when fossils are unavailable or inconclusive, molecular evolutionary genetics has become increasingly important and informative. However, without properly inferred phylogeny, information derived from alignment and evolutionary distance estimation could not be directly related to evolution. The first usable phylogenetic
reconstruction method for biological sequence data, UPGMA, was developed by Sokal and Michener in 1958 based on a simple hierarchical clustering algorithm [10]. However, not until the 1970s and 1980s when the parsimony method [11], neighbor-joining method [12] and maximum-likelihood method [13] were developed specifically for biological sequence data, did the molecular phylogenetics field start to thrive. Human and chimpanzee split was first determined using mitochondrial genomes in 1985 by Hasegawa, Kishino and Yano [14]. The African origin hypothesis was approved also using human mitochondrial genomes in 2000 by Ingman et al [15]. The same molecular phylogenetics methods were also used in epidemiology studies, for example, studies of the H1N1 swine-origin influenza virus responsible for the 2009 flu pandemic [16, 17].

In the 1990s, with the help of high throughput Sanger capillary sequencing, scientists began a multinational endeavor to sequence the entire human genome. This is the Human Genome Project (HGP). Their initial divide-and-conquer strategy of cloning and ordering sequences, which had been used in previous genome projects, was accurate but slow at the scale of human genome. Table 1-1 lists the milestone genome projects to illustrate their scale differences. The Celera Corporation led by J. Craig Venter later joined the competition for mapping the first human genome, relying on private funds and more importantly, on the shotgun sequencing method. This method provided huge speed improvements over the strategy then used by the public coalition. However, the chromosomal layout and orientation information of the shotgun reads were lost as a sacrifice. Without piecing together these randomized reads, the data generated at a higher speed is useless. Celera Assembler, an automated sequence assembler specifically designed for human data, was used to solve this problem. The advantages brought by the shotgun sequencing method coupled with the automated assembly made the HGP adopt the same strategy in the end, leading to a rapid progress of the public project. The human reference
A genome draft was published in 2001[18]. All genome projects have adopted the same strategy since then.

Table 1-1: Genome size of milestone genome projects

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome Size</th>
<th>Sequenced Date</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>bacteriophage MS2</td>
<td>3.5 Kb</td>
<td>1976</td>
<td>First sequenced genome (RNA)</td>
</tr>
<tr>
<td>bacteriophage phi X174</td>
<td>5.3 Kb</td>
<td>1977</td>
<td>First DNA genome</td>
</tr>
<tr>
<td><em>Haemophilus influenzae</em></td>
<td>1.8 Mb</td>
<td>1995</td>
<td>First bacterial genome</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>12.1 Mb</td>
<td>1996</td>
<td>First eukaryotic genome</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>100 Mb</td>
<td>1998</td>
<td>First multicellular organism genome</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>157 Mb</td>
<td>2000</td>
<td>First plant genome</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>3.2 Gb</td>
<td>2001</td>
<td>First human genome</td>
</tr>
</tbody>
</table>

Although the most reliable and definitive way of determining gene function is through direct experiments, with thousands of genomes and millions of gene records currently available (NCBI and Genomes Online Database statistics), it is much easier and faster to predict gene function based on homology. The theories and algorithms for sequence alignment, genetic distance calculation and phylogeny inference constituted the very foundation of a new generation of tools that were designed for homology search and function prediction. These tools were used widely from genome annotation to medical research. The methods and applications of biological sequence analysis enable geneticist to harness the power residing in the vast amount of sequence information. This dissertation describes new applications of existing methods utilizing the large quantity of sequence information from ultra-high throughput sequencing technologies, and new methods exploiting sequence information for studying gene families and human haplogroups. The theories, algorithms and tools concerning the studies described in this dissertation are introduced first:
Background on pairwise DNA sequence alignment and analysis

One sequence alone, without comparison with any other sequence, provides little biological insights. Information could be obtained by comparing sequences. The simplest case is the comparison between two sequences. This requires putting sites descended from the same ancestry at the same position, i.e., sequence alignment. If no insertion or deletion has happened along the evolution route for two sequences, they should be of the same length as the ancestral sequence. As a result, the \( i \)th position at sequence 1 should correspond to the same \( i \)th position at sequence 2. However, insertion and deletion do happen and identifying corresponding sites between the two sequences are not straightforward. Different algorithms have been developed to find the optimal sequence alignment based on different criteria.

Basic pairwise sequence alignment algorithms

The sequence alignment problem is directly related to the Longest Common Subsequence (LCS) problem in computer science. An example DNA alignment using LCS algorithm is provided in Figure 1-1. When the backtracking pointer moves horizontal or vertical instead of diagonal, it means one site from one sequence is matched to more than one sites in the other sequence. In this case, the extra sites are aligned to gaps.
The first alignment algorithm specifically for biological sequence, which is extremely similar to the LCS algorithm, was developed by Needleman and Wunsch in 1970 [7] and further improved by Gotoh [19]. The recurrence function describing the improved algorithm is

\[
F(X_i, Y_j) = \max \begin{cases} 
  F(X_{i-1}, Y_{j-1}) + s(x_i, y_j) & (1), \\
  F(X_{i-1}, Y_j) + g & (2), \\
  F(X_i, Y_{j-1}) + g & (3)
\end{cases}
\]

where \( X_i \) is sequence \( X \) with \( i \) symbols in length, \( Y_j \) is sequence \( Y \) with \( j \) symbols in length, \( x_i \) is the \( i \)th symbol in sequence \( X \) and \( y_j \) is the \( j \)th symbol in sequence \( Y \). The \( s(x_i, y_j) \) is the similarity function that obtains the score for matching symbols \( x_i \) and \( y_j \) from a known score matrix. The \( g \) in the above formula is the gap penalty function [20].

The Needleman-Wunsch algorithm aligns sequences from end to end. This is called global alignment. However, in order to identify exons inside genomic DNA sequences, or domains inside protein sequences, an algorithm that is capable of finding the matched sections
between two sequences is needed. This is called local alignment. With a slight modification of
Needleman-Wunsch algorithm, a new one is proposed by Smith and Waterman in 1981 for local
alignment [21]. The Smith-Waterman algorithm is

\[
F(X_i, Y_j) = \begin{cases} 
0 & \text{(1)} \\
F(X_{i-1}, Y_{j-1}) + s(x_i, y_j) & \text{(2)} \\
F(X_{i-1}, Y_j) + g & \text{(3)} \\
F(X_i, Y_{j-1}) + g & \text{(4)} 
\end{cases}
\]

All the symbols in formula (2) have the same meaning as in formula (1). The only difference is
that at each step Smith-Waterman compares the current maxima with 0. This means Smith-
Waterman algorithm can only take 0 as the minimal value for any given cell, while Needleman-
Wunsch permits negative scores. As a result, the backtracking route in Smith-Waterman
algorithm is different. Instead of backtracking from the bottom right corner of the entire \(i \times j\) score
matrix (assuming the two sequences to be aligned is of length \(i\) and \(j\)), it starts from the cell with
the highest score in the entire matrix. The backtracking stops when it reaches a cell with a 0
value.

Score matrix and scoring alignment

The score matrix is used to reward matched symbols between the two sequences while
penalize unmatched symbols. However, it does not reward all matches and penalize all
mismatches equally. For example, the penalty for transition is less severe than that for
transversion. A score matrix could also reflect the GC bias. An example DNA score matrix is
provided in Table 1-2. Unlike in the LCS algorithm, where the score is binary (matched or
unmatched), a score matrix thus allows fine-tuning for each specific kind of substitution.

Table 1-2: DNA score matrix example estimated from alignment of 9000 human mitochondria
The first matrices used by biologists were actually hand-built from their experience. However, the score matrix could be built from a probabilistic perspective, comparing a nonrandom model with a random model. Assuming there are two residues $a$ and $b$ at an aligned site of a pairwise alignment, if these two residues are aligned at random, the probability for this event is $p_a \times p_b$. $p_a$ is the probability of getting a residue $a$ at random. If these two residues are truly related, the probability is then $p_{ab}$, representing the change from $a$ to $b$ in an evolutionary process. The **odds ratio** (the ratio of odds from one model compared with that of another model) between these two probabilities is [20]

$$\frac{P(a,b \mid \text{Nonrandom})}{P(a,b \mid \text{Random})} = \frac{p_{ab}}{p_a \times p_b} \quad (3).$$

Assuming site independence, the odds ratio of the entire alignment under the random and nonrandom models is

$$\frac{P(\text{alignment} \mid \text{Nonrandom})}{P(\text{alignment} \mid \text{Random})} = \frac{\prod p_{x_iy_i}}{\prod p_{x_i} p_{y_i}} \quad (4),$$

where the $x_i$ is the $i$th symbol of sequence $X$ and $y_i$ is $i$th symbol of sequence $Y$ at the $i$th column of the alignment [20]. The higher the odds ratio, the more probable it is that the alignment is truly nonrandom. It is obvious that formula (4) is a product of formula (3) from all sites. A good alignment algorithm essentially tries to maximize formula (4) from formula (3). In practice, the

<table>
<thead>
<tr>
<th>From</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>57</td>
<td>-124</td>
<td>-125</td>
<td>-73</td>
</tr>
<tr>
<td>T</td>
<td>-118</td>
<td>68</td>
<td>-79</td>
<td>-109</td>
</tr>
<tr>
<td>C</td>
<td>-125</td>
<td>-84</td>
<td>57</td>
<td>-10</td>
</tr>
<tr>
<td>G</td>
<td>-52</td>
<td>-93</td>
<td>-95</td>
<td>100</td>
</tr>
</tbody>
</table>
log odds ratio is normally used to avoid underflow problem. For example, the widely-used BLOSUM matrix is derived in like this [22].

**Heuristic algorithms for performance considerations**

The algorithm complexity for the above three algorithms are all $O(nm)$ for two sequences with length of $n$ and $m$. This is true for both the computation time and memory usage. As a result, the above three algorithms could only handle pairwise alignment for moderate-sized sequences. In the case of alignment between two mammalian genomes, it is computationally impossible to use the above algorithms, even though these algorithms guarantee to find the optimal alignment. So it is impractical to use them in modern applications except for special cases. For example, currently there are more than 1000 genomes publicly available, and NCBI GenBank database contains more than $10^{11}$ bases. Even for alignment between two bacterial genomes, the complexity for the above algorithms is on the order of $10^{12}$. Any meaningful database search with GenBank would have a complexity over $10^{12}$, assuming a meager 10 bp query. Under these circumstances, faster heuristic methods were developed. Among them, BLAST [8, 9] and FASTA [23] are the best known two.

Almost all modern heuristic methods start the alignment process with *n-tuple* word matching, or “seeding”. Word matching means an exact symbol match between two sequences. The idea of seeding is that true significant alignment normally contains short stretches of residues that are matched exactly. Normally seeding length is kept short to maintain the sensitivity of the alignment algorithm. For example, BLASTN seeding length is 11 and BLASTP seeding length is 3 by default. After first identifying these seeds, the subsequences between two neighboring seeds are then aligned using Smith-Waterman algorithm. By using seeding, the search space is greatly
reduced from an $n \times m$ matrix. However, it is possible that a method like this overlooks the true optimal alignment. As a result, they do not guarantee finding the global optimal. Reducing the seeding length will increase the sensitivity, but at the same time greatly increase the search space, and thus the computational cost. Increasing the seeding length will decrease the sensitivity, but with a prior knowledge that the two sequences in question are highly similar, these heuristic methods will produce results comparable to the full Smith-Waterman algorithm with greatly reduced time.

**Heuristic algorithms for ultra high-throughput sequencing data mapping**

Since 2005, ultra high-throughput sequencing technologies, including 454 [24], Illumina/Solexa [25], became available. In 2011, the 454 sequencing platform can generate up to 1 million 700-bp reads per run. Illumina sequencing platform could generate over 2 billion 100-bp reads per run. Hundreds of millions of sequencing reads could be produced in relatively short time. Although BLAST is already a heuristic algorithm for local alignment, it still could not efficiently handle data at this scale. More specialized mapping tools were developed to rapidly map large amount of data to a reference genome.

Compared to BLAST, these specialized mapping tools rely exclusively on word matching. MAQ by Li, Ruan and Durbin [26] uses 8-bp word length and indexes the reads by hash tables. This method does not allow gapped alignment. SOAP by Li et al [27] uses a similar method but allows gap. Bowtie algorithm by Langmead et al [28] uses Burrows-Wheeler transform (BWT) algorithm and the FM index for exact string matching as the alignment tool. These two algorithms in Bowtie allow great increase in mapping speed, even compared with MAQ and SOAP, but with more sacrifices in terms of accuracy and sensitivity. Bowtie
guarantees to find at least one exact match if the read has exact matches in the reference, but it doesn’t guarantee to find the optimal alignment if there is no exact match. BWA by Li and Durbin [29] combines the strength of MAQ with BWT. The read mapping speed of these specialized tools is significantly faster than BLAST, but the trade-off is that they are far less sensitive. Word matching normally could only afford at most 2 mismatches for 8-tuple seed, if the read length is about 30 to 40 bp. It does not allow for long gap between parts of the same read. Because of splicing, RNA-seq reads spanning exon-exon junctions are expected to have long gaps over the introns when mapped. This feature makes the aforementioned tools useless when this type of reads is mapped. These tools are thus mainly used for re-sequencing projects that target new SNP and structural variation (SV) discoveries.

**Background on multiple DNA sequence alignment and analysis**

Sequences may come as a group, for example, as a gene family or a population. The collective information, or a profile, of the sequences from the same group is useful in many cases, for example, gene family assignment and human haplogroup assignment. Building a profile requires multiple sequence alignment. The algorithms for pairwise sequence alignment can be generalized to include multiple sequences. However, the performance of such an algorithm renders itself useless. For example, assuming there is $k$ number of sequences to be aligned, the two-dimensional matrix for pairwise alignment becomes $k$-dimensional. If the overall alignment length is $n$, the algorithm complexity, for both time and memory usage, becomes $O(n^k)$. For a moderate-sized gene family with only 30 members, the complexity is well above $10^{30}$. This means this algorithm could only take very few short sequences before reaching the limit of even the most advanced hardware. Although this algorithm guarantees to find the global optimal of a
multiple sequence alignment under a given score matrix, its exponential complexity makes it impractical.

**Heuristic algorithms for multiple sequence alignment**

Modern multiple sequence alignment programs all rely on a progressive alignment method. The progressive alignment method is to build the complete multiple sequence alignment by building successive pairwise alignments. The simplest procedure is to first choose two sequences out of $N$ total sequences and build a pairwise alignment for the chosen two. A third sequence is picked from the remaining $(N-2)$ sequences and aligned to the pairwise alignment of the first two. This continues until the last sequence is picked and aligned to the alignment of the $(N-1)$ sequence alignment. After that the complete alignment is done.

Algorithms implementing this progressive alignment concept differ greatly in their actual alignment building steps. The order of picking sequences is different among different programs, but most algorithms use a guide tree. Commonly the closest nodes in the tree are aligned first. Different algorithms also have different ways of scoring an alignment. Alignment scores are needed to distinguish optimal alignments from suboptimal ones, and different scoring mechanisms will have different effect on the resulting multiple alignment.

Any progressive method of building multiple sequence alignment is heuristic and doesn’t guarantee the global optimal. Any errors that happened during pairwise alignment stage, either between sequences or between alignments, are propagated to the final multiple-alignment. Different iterative methods were later developed to compensate this problem. The concept of iterative method is to systematically realign the sequence from the sequence pool after the initial
multiple sequence alignment is done, so that the errors accumulated in the initial multiple sequence alignment process could have a chance to be corrected. Most modern alignment software includes this iterative feature, for example, MUSCLE [30]. The iterative method helps to improve alignment further, but it still does not guarantee the global optimal. However, the software implementations of the progressive and iterative method could handle up to 1000 sequences with acceptable speed, and thus are widely used.

**Scoring a multiple sequence alignment**

Because alignment algorithm relies on a given scoring mechanism to find the optimal alignment, the final result is thus determined by how the score is calculated. Almost all alignment algorithms treat each alignment column independently, and the overall score $S$ for the entire alignment of $n$ columns could be written as

$$S = \sum_{i} S_i + g$$

where $S_i$ is the score for the $i$th column and $g$ is the scoring function for gaps. The $g$ function is implemented differently in different algorithms. A commonly used $g$ function is similar to that of pairwise alignment, $o + (l-1)e$, where $o$ is the gap opening penalty and $e$ is the unit gap elongation penalty. Two most widely used methods to calculate alignment score for a given column are minimum entropy method and sum of pairs (SP) method.

The idea behind minimum entropy method is that the optimal alignment is also the one minimizes the total entropy for the alignment. The entropy of each column is calculated in the Shannon entropy. The idea behind SP method is that the optimal alignment for each column will maximize the sum of all pair-wise substitution score $s(a, b)$ between any two of $N$ residues in that
column. Because SP score scheme treats each sequence as a descendent from each one of the rest (N-1) sequences, the true log odds ratio should be \( \log(p_1 \ldots / p_1 \ldots p_n) \) instead [20]. However, SP score is still widely used for its simplicity and direct correlation with substitution score matrix derived from pairwise alignment [20].

**Profiling a multiple sequence alignment**

An appropriately built alignment for a group of sequences captures the collective features of them. Normally these features include the conserved sites, the variable sites and the extent of variation.

**Position specific score matrix model**

In the case of a single column inside an alignment, if there are total \( k \) sequences in the alignment and total \( c_a \) number of residue \( a \) in this column, then the estimated emission probability at this site for symbol \( a \) is

\[
e(a) = \frac{c_a}{k}
\]

(6).

Because sometimes the number of sequences in a given alignment is not large enough to include all possible symbols, the ones not included would have zero emission probability by formula (6). A model with zero probability for a specific symbol at a given site would reject all future sequences with this symbol at this site. This is overly restrictive and unfavorable. Some adjustment is needed to give them a non-zero probability. The simplest method is called Laplace’s rule, which is adding 1 count for each symbol that is not included in the original
alignment. More sophisticated methods could consider draw the unrepresented symbols from a background distribution [20].

Assuming the emission probability for a symbol \( a \) at the \( i \)th column is \( e_i(a) \), for an alignment of length \( l \), the probability of getting a sequence \( X \) from the alignment is given by

\[
P(X \mid \text{nonrandom}) = \prod_{i=1}^{l} e_i(x_i) \quad (7),
\]

where \( x_i \) is the \( i \)th symbol of the sequence \( X \). The probability of getting a sequence \( X \) from a random model is given by

\[
P(X \mid \text{random}) = \prod_{i=1}^{l} p_{x_i} \quad (8),
\]

where \( p_{x_i} \) is the probability of getting the symbol \( x_i \) randomly from a background (random) distribution. The log odds ratio between these two models for this sequence is thus

\[
\frac{P(X \mid \text{nonrandom})}{P(X \mid \text{random})} = \sum_{i} S(x_i, i) \quad (9).
\]

\( S(x_i, i) \) could be considered as a score similar to that introduced in the context of pair-wise alignment. \( S(x_i, i) \) is the joint score between position and symbol, and thus could be called a position specific score. A matrix over all possible positions and symbols for the given alignment is a position specific score matrix (PSSM) [20].

**Hidden Markov model for multiple sequence alignment**

The PSSM approach treats all columns the same. However, insertion and deletion columns could be considered differently from matched columns. The matched, insertion and
deletion columns are different states. To reflect the above considerations, a modeling method needs different emission probabilities for different sites from different states. It also needs to allow transitions between different states. Determining the state of a column is challenging, since the ancestral sites cannot be observed directly. In addition, the position of the matched columns and insertion/deletion columns are different from alignment to alignment. A Hidden Markov model (HMM) could be used in this case to form a unified framework for profiling multiple sequence alignment.

A hidden Markov model is a Markov model with hidden (unobservable) states. As in a normal Markov model, the output is generated based on the emission probability from different states conditioned on the previous state or states. However, in a hidden Markov models, the states cannot be directly observed, and only the output is observable. It is thus necessary to infer the state from the output. Therefore this process could be abstracted into the emission of two sequences: one is the output sequence with symbols we could observe; the other is the state sequence that is not directly observable. Each state in the state sequence corresponds to a symbol in the output sequence.

If the output sequence is $X$, the state path is $S$, $x_i$ is the $i$th symbol of output $X$ and $s_i$ is the $i$th state of path $S$. When the state of $s_i$ is $l$ and the state of $s_{i-1}$ is $k$, the transition probability between states $s_i$ and $s_{i-1}$ could be defined as

$$t_{s_i, s_{i-1}} = P(s_i = l | s_{i-1} = k) = t_{kl}$$

This makes the state sequence a first order Markov chain. The emission probability of state $k$ is $e_k(x)$. The joint probability of getting an output sequence $X$ from a state sequence $S$ is
\[ P(X,S) = t_{0S} \prod_{i=1}^{k} e_{x_i}(s_i)x_{i+1}t_{s_is_{i+1}} \]  

(11),

where 0 is a special start state to initialize the Markov chain. A standard Viterbi algorithm, first introduced in 1967 by Andrew Viterbi, could be used to decode the state sequence [31]: given the most probable path \( V_k(i) \) up to the \( i \)th position with observation \( x_i \) from state \( k \) is known, the most probable path for the \((i+1)\)th position is given by [20]

\[ V_l(i+1) = e_{l}(x_{i+1}) \max(V_k(i)t_{kl}) \]  

(12).

**Estimate evolutionary distances from sequence alignment**

The differences between two sequences in a pairwise alignment are considered to be a measurement of evolutionary distance between them.

**Theories of estimating the number of substitutions between two sequences**

Nucleotide changes of DNA sequences (or amino acid changes of protein sequences) are memory-less and could be modeled as first-order Markovian process. Because of this property, the observed number of substitutions (as in Hamming distance [32] or p-distance[33]) is always an underestimation of the actual number of substitutions happened between two sequences, since multiple substitutions could occur at the same site and only the effect of the last substitution could be observed experimentally. The larger the genetic distance between the sequences, the more severe the underestimation would be. A correction of this underestimation could be achieved mathematically assuming a certain substitution model. Many substitution models have been proposed for correcting this. Jukes and Cantor proposed the first such model in 1969 that
considers all stationary base frequencies (also called base composition, $q_A, q_C, q_G$ and $q_T$) the same and all substitution rates ($r_{ij}, i$ and $j \in \{A, C, G, T\}$) between bases also the same [3]. Kimura in 1980 proposed another model that allows two different substitution rates for transitions and transversions [34]. Felsenstein in 1981 proposed a model that allows different base frequencies but only allows the same substitution rate between bases [13]. Hasegawa, Kishino and Yano in 1985 proposed a method (HKY model) combining Felsenstein’s 1981 model and Kimura’s 2-parameter model together, which has total five free parameters [14]. Tamura and Nei in 1993 proposed an improvement of HKY model that allows two different transition rates for purines and pyrimidines, which increased the number of free parameters to six [35].

Since 1984, as first proposed by Lanave et al, a model that allows different substitution rates between each pair of nucleotides as well as different base frequencies for all four nucleotides has been proposed [36]. This kind of model assumes time reversibility (thus called general time reversible, GTR, model), which means a substitution from base $i$ to base $j$ is the same as a substitution from $j$ back to $i$ ($i, j \in \{A, C, G, T\}$). Under this model, given a rate matrix $R$, at the time point $T$ since the two sequence diverged, the transition probability matrix $P(T)$ at time $T$ is given as

$$P(T) = e^{RT}$$

(13),

where the rate matrix $R$ is given as

$$R = \begin{bmatrix} r_{AA} & r_{AT} & r_{AC} & r_{AG} \\ r_{TA} & r_{TT} & r_{TC} & r_{TG} \\ r_{CA} & r_{CT} & r_{CC} & r_{CG} \\ r_{GA} & r_{GT} & r_{GC} & r_{GG} \end{bmatrix}$$

(14).
This model has been re-proposed in different forms, but of mathematical equivalence, by Barry and Hartigan in 1987 [37] and Rodriguez et al in 1990 [38]. Rodriguez et al provided a general formula for calculating the genetic distance with multiple hits correction:

\[ d = -\text{tr}[X \log(X^{-1}F)] \] (15),

where \( d \) is the estimated number of substitutions per site between the two sequences in question, \( X \) is the base frequency matrix and \( F \) is the observed divergence matrix (substitution per site per type of substitution). \( \text{tr} \) is the trace operation. The forms of \( X \) and \( F \) are given as

\[
X = \begin{bmatrix}
q_A & 0 & 0 & 0 \\
0 & q_T & 0 & 0 \\
0 & 0 & q_C & 0 \\
0 & 0 & 0 & q_G
\end{bmatrix} \quad (16),
\]

\[
F = \begin{bmatrix}
x_{AA} & x_{AT} & x_{AC} & x_{AG} \\
x_{TA} & x_{TT} & x_{TC} & x_{TG} \\
x_{CA} & x_{CT} & x_{CC} & x_{CG} \\
x_{GA} & x_{GT} & x_{GC} & x_{GG}
\end{bmatrix} \quad (17),
\]

where \( x_{ij} \) represents the average probability at a given site of a pair-wise alignment that in the first sequence the nucleotide is \( i \) and in the second is \( j \), \( (i, j \in \{A, C, G, T\}) \). The total number of free parameters of this model is 9. The GTR model allows the most flexibility for accommodating base frequency variations between nucleotides and substitution rate variations commonly observed in actual data. Actually all the aforementioned substitution models could be derived by restricting the number of free parameters in the GTR model.
Estimate the substitution rate variations among sites

However, besides assuming site independence, the GTR model also assumes constant substitution rate among sites. The latter condition is frequently violated in real experimental data. Uzzell and Corbin first pointed out the importance of this rate variation in evolutionary inference in 1971 [39]. They have also used statistical distributions to model rate variations among sites. Their study and another one by Holmquist et al in 1983 [40] have concluded that the substitution rate is not uniform among sites, and thus the substitution process should not be modeled by a Poisson process, which assumes constant rate. Instead, they pointed out that a negative-binominal distribution should be used. It is known that a negative-binominal distribution could be generated with a Poisson process whose expected number of occurrence parameter $\lambda$ follows a gamma distribution (a gamma-Poisson mixture) [35].

Because of its shape flexibility, the gamma distribution has been widely examined and used to model this rate variation among sites in genetic distance calculation [35, 40-46]. The probability density function of a gamma distribution is given as

$$g(x;\alpha,\beta) = \frac{\beta^{\alpha}}{(\alpha - 1)!} x^{\alpha-1} e^{-\beta x}$$  \hspace{1cm} (18),

where $\alpha$ and $\beta$ are the shape parameters of the gamma distribution, with $\alpha\beta$ as mean. The current model applies a correction factor $g$ following gamma distribution at each site of the sequence to the mean substitution rate $r$, and as a result the mean of this gamma distribution needs to be 1. In this case, $\beta = 1/\alpha$, so there is only one free shape parameter to estimate (Figure 1-2).
Waddell and Steel gave a formula to calculate the genetic distance corrected for rate variations among sites, based upon the GTR distance formula given by Rodriguez et al in 1997 [46]. This formula allows any distributions to be used for modeling rate variations among sites. The general form of this formula is

\[ d = -tr[X \cdot M^{-1}(X^{-1}F)] \]  

(19),

where \( M^{-1} \) is the inverse function of the moment generating function \( M \) of the selected distribution. In the above case of using a gamma distribution:

\[ M^{-1}(x) = \frac{\alpha}{\alpha(1 - x^\beta)} \]  

(20).
The most extreme case of rate variation among sites is that at some sites the substitution rate is zero, which means these sites are invariable in the context of the studied evolution process. These sites are referred as invariant sites [46]. Shoemaker and Finch in 1989 first identified the influence of invariable sites on genetic distance estimation [47]. Although ambiguities in determining these invariant sites are expected, the percentage of invariant sites $p_{inv}$ could be estimated. It is thus possible to first estimate the expected number of substitutions per site for variable sites by adjusting the $F$ matrix to $F_{var}$:

$$F_{var} = \frac{F - p_{inv}X_{inv}}{1 - p_{inv}}$$

(21),

where $X_{inv}$ is the base frequency of the invariant sites [46].

**Phylogeny reconstruction**

A tree structure (phylogram) could be used to represent the relative evolutionary distance between individuals.

**Phylogeny reconstruction from evolutionary distances**

A phylogenetic tree could be built based on a given distance matrix. A tree-building algorithm based on similarity or distance matrixes is essentially a clustering algorithm. Some of the well-known distance based algorithms are
**Unweighted pair group method using arithmetic averages (UPGMA) method**

UPGMA is a simple hierarchical clustering algorithm first introduced by Sokal and Michener in 1958 [10]. This method builds a phylogeny by: 1) clustering together two nodes with the shortest distance in the matrix, 2) at the same time replacing these two nodes with a single new node. A node may include only one sequence or multiple sequences. The distance between two nodes $A$ and $B$ is defined as

$$d_{AB} = \frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} d(x,y)$$  \hspace{1cm} (22),

where $|A|$ is the cardinality (number of elements) of set $A$.

**Neighboring-joining method**

UPGMA assumes a strict molecular clock on all branches. Neighboring-joining method proposed by Saito and Nei in 1987 works on trees with additivity alone but not necessarily with a strict molecular clock. The neighborliness distance $D$ between the $i$th and $j$th sequence of the set of sequences $A$ is given as

$$D_{ij} = d_{ij} - \left( \frac{1}{|A| - 2} \sum_{k \in A} d_{ik} + \frac{1}{|A| - 2} \sum_{k \in A} d_{jk} \right)$$  \hspace{1cm} (23),

where $k$ is another sequence in set $A$ other than $i$ or $j$. Two sequences with the shortest $D$ distance are defined as neighbors. As in UPGMA method, two neighboring sequences are merged together and replaced by a single node. The distance between another sequence $k$ and the newly formed node $n$ is defined as

$$d_{nk} = \frac{1}{2} (d_{ik} + d_{jk} - d_{ij})$$  \hspace{1cm} (24).
Neighbor-joining tree is a minimal evolution tree. Given the branch length of a tree represents evolutionary distance, a minimal evolution tree is a tree that minimizes the total branch length of the tree. For a collection of $n$ sequences, there are total $(2n-3)$ branches. But the total number of possible unrooted topologies is $(2n-5)!$ (i.e., $3\times5\times\ldots\times(2n-5)$). The total number of possible rooted topology is $(2n-3)!$ [20]. For a set of sequences with only 20 members, the possible unrooted topologies are $2\times10^{20}$. It is obvious that finding the best topology from all possible ones is quite challenging. The neighbor-joining algorithm is a heuristic method that does not guarantee the global optimal tree. However, in practice, the neighbor-joining method could produce trees close to the optimal tree with vastly reduced complexity, if the divergence between sequences is not too great.

**Phylogeny reconstruction from discrete sites by maximum parsimony method**

One major drawback of distance-based method is that once distance is calculated, the information of each individual substitution is lost. A plethora of methods have been developed to treat each site discretely and build the phylogeny based on each individual site. The philosophy behind these methods is maximum parsimony. Its objective is to find a phylogenetic tree that could explain the differences between the sequences with the least number of substitutions. A standard search algorithm was given by Fitch in 1971 [11]. The tree generated by parsimony method is still a tree of minimal evolution. However, this method severely underestimates the number of substitutions, as implied by the parsimony methodology.
Phylogeny reconstruction from maximum likelihood (ML) method

The likelihood of a tree is given by $P(\text{data}|\text{tree})$. A tree includes a topology $T$ and a collection of the branch lengths. The latter could be expressed as $t^\bullet$, which stands for $t_1…t_n$ for a tree with $n$ branches. The maximum likelihood tree is the tree that maximizes $P(\text{data}|T, t^\bullet)$ [20]. This likelihood could be computed following method introduced by Felsenstein [13].

Given the method for calculating the likelihood of a given tree, the tree that maximizes the likelihood of obtaining the observed data could be searched among all possible trees. However, the search space size is $(2n-5)!!$ for all unrooted trees of $n$ sequences and $(2n-3)!!$ for all rooted tree. Felsenstein showed that it is sufficient to only search in the unrooted trees for the optimal topology if the given substitution matrix is multiplicative $(A(x)A(y) = A(x+y)$ for substitution matrix $A$) and reversible. However, the number of trees to be searched is still $(2n-5)!!$ [20].

Reconstructing phylogeny by sampling (Bayesian method)

The search space for the optimal tree using ML method increases explosively when the number of sequences increases. Many optimization algorithms have been proposed to reduce the search space, but the computation for large data sets is still demanding [20]. An alternative strategy is to sampling the tree space. Sampling allows direct calculation of the posterior probability of the phylogenetic tree (defined by topology $T$ and the set of branches $t^\bullet$) conditioned by the observed data (the collection of aligned sequences $X^\bullet$). According to Bayes’ Theorem, the posterior probability (conditional probability) of a hypothesis $H$ (the probability of $H$ after evidence $E$ is observed) is given by [20]
The posterior probability of a phylogenetic tree is given under Bayesian framework as

\[ P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} \quad (25). \]

This method was first proposed by Rannala and Yang in 1996 \[48\]. Mau, Newton and Larget gave an improved algorithm in 1999, which samples using Markov chain Monte Carlo (MCMC) method \[49\]. The latter was proven to be able to handle large data sets with many sequences and many sites. This method was integrated into the BEAST software package \[50\] and used in my study on molecular dating using mitochondria sequences \[51\].

The philosophy of this method is to sample trees randomly from a probability distribution based on given data. The posterior probability of a certain topology is estimated by the fraction of this specific topology in the sample (given the sample is large enough). Similarly, the branch length could also be estimated in this manner. This method depends greatly on the sampling procedure and the distribution model used.

The commonly used sampling procedure is MCMC, which is a random process that generates new trees depends on the previous tree at each step. This makes the process Markovian. The initial tree is normally generated using UPGMA or NJ method. If the current tree is \( \Psi = (T, t^\star) \), the next tree \( \Psi' = (T', t'^\star) \) is generated based on the current tree according to a model distribution \( Q(\Psi', \Psi) \). This process was first shown by Hastings in 1970 \[52\] and the evaluation expression is given as

\[ r = \frac{P(X^\star \mid \Psi')Q(\Psi', \Psi)}{P(X^\star \mid \Psi)Q(\Psi, \Psi')} \quad (27). \]
When \( r > 1 \), the new tree \( \Psi' \) is accepted, and when \( r < 1 \), the current tree \( \Psi \) is retained. This process is thus called Metropolis-Hasting algorithm.

The distribution model \( Q(\Psi', \Psi) \) for generating a new tree \( \Psi' \) from \( \Psi \) has two parts, the first part is to adjust the branch length, the second part is to switch leaf nodes. However, the a priori information for the branch length adjustment and leaf node switching could not be obtained. Mau et al assumed a flat distribution in their study.

**Phylogeny reconstruction without sequence alignment**

With the rapid growth of genome sequencing technology, a large number of genomes have become available. Phylogenetic studies based on whole genome alignment have become more and more attractive and feasible, but only in terms of available data. As discussed earlier, multiple sequence alignment is computationally expensive even with heuristic algorithms. There is another collection of methods that calculate evolutionary distances without using full multiple sequence alignment [53-55]. The philosophy is to model the sequence inheritance process in a random model, in which the descendents randomly sample the ancestor sequence. In this model, the frequencies of oligomer sequences of a certain length \( K \) (\( K \)-mer) are calculated for each lineage. The total number of possible \( K \)-mers for DNA is \( 4^K \) and for protein is \( 20^K \). The distances between different lineages are then calculated based on these \( 4^K \) or \( 20^K \) variables using dimension reduction. This method has been implemented in CVtree web server [56, 57]. However, this method discarded the order and orientation information of the sampled hexamers. Conserved sites are also weighted equally as non-conserved sites. Biologically meaningful evolutionary rates cannot be directly derived from this model. This makes the phylogeny only meaningful in terms
of its topology. But its computational simplicity, speed and capability of handling genome-sized data are undisputed and desirable.

**Molecular dating**

The molecular clock hypothesis was first proposed by Zuckerkandl and Pauling in 1962 [2]. They found that the number of amino acid differences between hemoglobins of different species is roughly linear with their species divergence time (based on fossil records). They hypothesized that the evolutionary rate is constant over time and over species. Later, Kimura derived a molecular clock from the neutral theory of molecular evolution under the assumption that most observed changes are neutral [58]. Based on the above hypothesis and assumption, the branch length of a phylogenetic tree is thus linearly correlated with time. Therefore it is possible to date evolutionary events through phylogenetic trees, with the help of proper fossil calibrating.

However, studies have shown that evolutionary rate varies along the genome and among different lineages [18, 59-62]. It is then proposed that each branch could have an independent rate of evolution [13]. This model has been implemented in several widely used phylogenetic software packages, including PHYLIP, PAUP* and MrBayes [63]. Under this model, it is only possible to infer topology but not possible to infer evolutionary rate or divergence time without extensive fossil calibration on every branch.

Several studies later have shown that an intermediate between these two extremes could be used [64-67]. This intermediate theory is called a relaxed molecular clock. Currently, commonly used relaxed molecular clock model is Bayesian relaxed clock model [65, 67, 68]. In this model, the rate of a given branch is drawn from a parametric distribution whose mean is
derived from the rate of the parental node or branch [65]. However, the relaxed molecular clock model is phylogeny dependent. The rate distribution depends on the topology of the tree, and sometimes is parameterized based on the parental branch length. Drummond et al in 2006 proposed a method for co-estimating the clock and the phylogeny at the same time [65]. The correctness of this method depends heavily on a prior parameters selected by users. When a prior information is scarce, a flat or uniform distribution is more appropriate. This method is implemented in BEAST [50], which is used in my study in Chapter 2.

**Information analysis between biological sequences of different types**

Although the information flow between DNA, RNA and protein is not entirely reversible, sequences of different molecular types could be compared directly in silico.

**Between DNA and RNA sequences**

RNA sequences are routinely mapped to genomic DNA sequences to help identify genes. The relation between DNA and RNA nucleotides is one-to-one. The difference between DNA and RNA in terms of chemistry is that DNA comprises deoxyribose and RNA comprises oxyribose. The complementary nucleotide in RNA to thymine (T) is Uracil (U). Current RNA sequencing technologies normally produce the complementary DNA (cDNA) sequence for a given RNA molecule, so the comparison between DNA and RNA sequences commonly seen is still between the four DNA nucleotides (A, T, C and G).

Because of post-transcription modifications, especially in eukaryotes, the corresponding RNA and DNA sequences are not exactly the same. Post-transcription modifications include 5’
capping, 3’ cleavage and polyadenylation, and splicing. Because of these modifications, an RNA sequence is normally a non-contiguous subset of the corresponding genomic DNA sequence. This makes local alignment the appropriate algorithm for aligning RNA sequence to DNA sequence. For typical EST [69] and full cDNA sequence reads generated by Sanger sequencing, BLASTN [8, 9] could be directly used to map RNA sequence to DNA.

**RNA-seq data mapping**

An experimental method (RNA-seq) was developed that uses ultra high-throughput sequencing technologies to sequence RNA transcripts [70, 71]. The high data amount requires tools even faster than BLAST. ERANGE is among the first tools developed specifically for RNA-seq mapping [71]. It requires an annotated genome reference. The annotated exons are concatenated together to allow read mapping over exon-exon junctions. Han et al used a similar method, but their method allows new splicing variants of known exons to be mapped [72]. However, if there were previously unknown exons, the above methods would not be able to reliably identify these exons. TopHat by Trapnell, Pachter and Salzberg [73] is able to map exon-exon junction spanning reads over previously unknown exons. Their method is that: 1) map RNA-seq reads to the reference genome to detect transcriptionally active regions, 2) concatenate these regions, 3) use these concatenations to identify exon-exon junction spanning reads from initially unmappable reads with relaxed settings. TopHat was used in my study on transcriptomes in brain development [74].
Between RNA and protein sequences

Messenger RNAs (mRNAs) translates to amino acid sequence based on a given genetic code table. 3 RNA nucleotides code one single amino acid. There are total 64 possible combinations for a 3-bp long RNA sequence, but there are only 20 common amino acids. The information content of a 3-tuple RNA sequence is 1.7 bits higher than a 1-tuple amino acid sequence. The genetic code table from RNA to amino acid is redundant. This also makes the reverse conversion from amino acid to RNA ambiguous. This makes aligning protein to RNA sequence ambiguous. Yockey argued that this information content difference supports the theory that life originated from RNA molecules instead of proteins [75].

Matured mRNAs could be easily translated to protein sequence given the correct genetic code table and reading frame. There are currently 23 different genetic code tables for different organisms and different organelles [76, 77]. For example, the vertebrate mitochondrial code is different from the standard code for eukaryotes, and is also different from invertebrate mitochondrial code. Given a RNA sequence and a genetic code table, there are three possible ways of translation. Normally only the longest open reading frame (ORF) is considered. Therefore, RNA sequence could be compared with protein sequence after it is converted to amino acid sequence.

Between DNA and protein sequences

Given a DNA sequence and a genetic code table, there are six possible ways of translation, since a DNA molecule has two strands. However, because of introns and untranslated regions of exons, the open reading frame is not contiguous on DNA. Same as RNA, a DNA
sequence could be compared with a protein sequence after it is converted to amino acid sequence. BLASTX [9] is one of the tools that is capable of mapping DNA against protein by converting DNA sequence to amino acid sequence.

Sometimes, it is helpful to compare different DNA sequences at amino acid sequence level, especially when they are from remotely related organisms. First, because the redundancy of genetic code table, some DNA level substitution will not change the translated amino acid sequence. These substitutions are called synonymous substitution. DNA substitutions that change amino acid sequences are called nonsynonymous substitutions. Changes at the third codon position of some genetic codons are synonymous, for example, CCT, CCC, CCA and CCG all code Proline. Changes at the first codon position are normally nonsynonymous. Second, amino acid with similar chemical properties could replace each other in a protein without much impact on the normal function. This provides redundancy at amino acid level. Because of these two redundancies, similar proteins could arise from significantly different DNA sequences. Between genes from remotely related species, there might be substantial substitutions at DNA level that make DNA sequence alignment impossible. However, the protein sequences might still be easily aligned. This makes DNA comparison at amino acid level attractive. TBlastX and TblastN provide this functionality [9]. Phoenix, a protein homologue retrieval tool developed by me, exploited this concept.
Applications of DNA sequence analysis in genome assembly and annotation

Genome assembly problem

As introduced at the beginning of this chapter, genome size varies among species and most mammalian genome size is on the order of $10^9$ bp. The current Sanger sequencing technology can only produce read length up to 1000 bp, and the read length of the ultra high-throughput platforms is even shorter. Each read can only capture a small fraction of the entire genome, but a complete genome could be sequenced with many such short reads by the shotgun sequencing method [78]. In the shotgun sequencing method, genomic DNA is first fragmented randomly, and then these fragment are randomly sampled and sequenced. Oversampling is required to produce enough reads to cover the entire genome, since there are coverage depth variations along the genome because of the random sampling. Empirically, a minimal coverage of 10X (average ten reads per site) is needed to completely cover the genome.

Because the sampling process is random, the order and orientation information between reads is lost during shotgun sequencing. In order to reproduce the genome, these reads need to be assembled. This process is called genome assembly. The ultimate goal of genome assembly is to reconstruct the entire genome from the randomly oversampled sequencing reads. However, because of the intrinsic complexity of the genome (for example, repeats and duplicated genes) and technical limitations (for example, sequencing error and bias), complete reconstruction directly from reads is normally impossible. Genome assembler may instead produce several long contiguous genomic fragments (contigs) that could be further joined together by subsequent experiments.
If there is a complete reference genome, genome assembly process could be reduced to read mapping problem, which has been briefly introduced earlier. This is called reference based genome assembly, or genome mapping. If there is no reference genome, this is called \textit{de novo} genome assembly. If there are extensive differences between the reference and the new genome, the genome mapping approach will be unsuccessful because reads with significant differences from the reference won’t be successfully mapped. This is especially true when the reads are mapped by specialized mappers that trade sensitivity for speed. In contrast, \textit{de novo} assembly does not require references. Instead, all reads are compared with each other. Under this circumstance, \textit{de novo} assembly is a better choice.

\textbf{Algorithms and tools for \textit{de novo} genome assembly}

The philosophy of genome assembly algorithm is to find the shortest common supersequence (SCS) between two reads. This is directly related to the LCS algorithm introduced in pairwise alignment. Given two input sequences, if there is a LCS for them and this LCS is not in the middle of either sequence, then the SCS is the merge of the two input sequence with LCS as the overlapping part. For example, if sequence $X_1 = \text{ATCGAAGTCG}$, $X_2 = \text{AGTCGCATTGCAT}$, then the LCS satisfying the above condition is AGTCG. The SCS in this example is ATCGAAGTCGATTGCAT. The boundary conditions of the algorithms introduced for pairwise alignment could be easily modified to identify such LCS. For example, this could be achieved by defining $F(i, 0) = 0$ and $F(0, j) = 0$ in Needleman-Wunsch algorithm. This modification removes penalties for aligning to sequence ends.

A heuristic similar to BLAST could reduce the search matrix. For example, perfect matches of string length $K$ between two reads are first calculated, and then Needle-Wunsch
alignment is triggered between matched words. In the realm of genome assembly, a word of length $K$ is commonly referred as $K$-mer, which equals the meaning of $K$-tuple in information sciences.

For a collection of millions of reads, the simple solution is a heuristic expansion of the above algorithm. First, the longest LCS is identified through pairwise alignment between these reads. Then, the SCS is constructed from the pair with the longest LCS. This pair of sequences is removed from the collection and the newly constructed SCS is added. The above steps are repeated until there is only one sequence left in the collection. The complexity of this algorithm is $O(N^2)$ (or $O(N\log(N))$, depends on implementation) for a collection of $N$ sequences. Similar to pairwise alignment heuristic, a hash table of all possible $K$-mers from all the reads could be established to speed up assembly. If a sequence has length $L$, the maximum possible number of $K$-mers for this sequence is $(L-K+1)$ for $L> K$. The total number of $K$-mers for a large number of sequences could be huge. But because no redundant $K$-mers are stored in the hash table, this approach still consumes less memory than storing all the original reads. This also increases assembly speed in the sense that only two sequences with perfect $K$-mer matches are picked for pairwise alignment.

Algorithms involving pairwise alignments between all reads at each step are called Overlap/Layout/Consensus (OLC) algorithms by Miller, Koren and Sutton [79]. Software implementing OLC algorithm is typically designed for Sanger sequencer data. Because of the relative fewer number of reads, localized multiple sequence alignment is used to call consensus sequences from sequencing reads. This is the most sensitive method with the least heuristics applied, and thus is more likely to reach the global optimal. Celera Assembler [80] is one of the
software implementing OLC algorithms. Similar ones include ARACHNE [81, 82] and CAP [83].

Even with heuristics, an assembly process like the one just introduced could exhaust the resources of the most capable computers rapidly if there are billions of reads. This is now the normal data output for Illumina sequencers. A greedy algorithm could be used to reduce the complexity. Instead of looking for the best alignment from all the reads at each step, a greedy algorithm only focuses on the current contig (or read). It always finds another sequence $X$ from the pool that has the best LCS with the current sequence. The assembly of sequence $X$ and the current sequence might prevent a better solution between $X$ and another sequence in the pool. It is thus possible that the algorithm reaches local maxima instead the global one. The first short-read assembler designed for the ultra high-throughput sequencers SSAKE [84] uses this strategy. It also relies exclusively on word matching for speed. SHARCGS [85] uses a similar approach but in addition considers the quality score of the input reads. VCAKE builds upon the same philosophy but allows mismatches.

Currently, the most widely used approach to assemble short reads from the latest ultra high-throughput sequencers is based on de Bruijn graph. A $K$-mer hash table constructed from all the reads could be linked by edges that represent overlaps between $K$-mers. This linked $K$-mer graph is a form of de Bruijn graph. By completing a de Bruijn graph, the assembly is implicitly completed, because by calculating overlaps between $K$-mers, the alignment between $K$-mers has been done. Under ideal conditions, the neighboring $K$-mers are connected by one single link and there will be only one path connecting all $K$-mers. By walking along this path, the entire assembled genome could be reproduced. However, because of complex features of the genome and sequencing errors in real genome projects, there might be more than two edges connecting
each $K$-mer node. Many studies tried to reduce a complex de Bruijn graph from real data to the one under ideal condition. Euler software, although developed for Sanger data, is the first assembler based on this principle [86]. Velvet [87] is the first such assembler specifically for short reads from ultra high-throughput sequencers. ABySS [88] software is the first such assembler that is capable of memory parallelization in cluster computing by partitioning de Bruijn node computation. SOAPdenovo [89], AllPath [90, 91] and CLC Genomics Workbench (CLC bio) are all based on de Bruijn graph.

**Genome annotation problem**

A fully assembled genome sequence is still just a string of symbols. Functional units such as genes and cis-regulatory elements need to be identified inside this string. This process is called genome annotation. In order to identify protein-coding genes, the most basic step of annotation is to find open reading frames (ORFs). An ORF is a stretch of DNA sequence that could be translated to proteins uninterrupted by stop codons. This is relatively easy in prokaryotes since their genes possess few and very short introns. However, in eukaryotes, a protein-coding region is interrupted by many (from a few to a dozen) introns, normally several kb in length. Because of these long introns, eukaryotic genes are much longer than their prokaryotic counterparts. The longest gene in human genome, dystrophin gene, is 2.3 Mb in length and has 78 introns [92]. In addition, there are post-transcription modifications involved in tRNA, rRNA and other RNA gene product synthesis.

Theories, algorithms and tools for genome annotation had a rapid development in the late 1990s and early 2000s. This rapid development coincided with the completion of major model organism genomes. All gene prediction tools are based on a prior knowledge about the gene. In
the simplest case of ORF detection, the knowledge of the appropriate genetic code table is required. A general annotation of protein-coding genes requires the knowledge about the patterns of promoters, transcription start sites and transcription termination sites. The annotation of exons and introns requires the knowledge about the splicing mechanisms.

In a gene-oriented annotation, for example, a hemoglobin gene annotation requires at least the knowledge about the sequences of known hemoglobin genes. In this case, identifying new hemoglobin genes could be reduced to mapping genomic regions based on sequence similarity. If there is enough sequence information about all genes, genomes could be annotated by simply mapping out all regions similar to known ones. This does not require any abstraction of known knowledge. This does not require mathematic modeling either, except the assumption that similar sequences are related. However, sometimes known information is limited, other information, for example, sequence conservation, domain structure and 3D structure, is needed. This extra information will help to produce a more precise prediction than just based on limited sequence similarity. A process like this necessitates abstraction and modeling.

The methods for genome annotation could be divided into two categories: homology-based and statistical model-based [93]. The latter is also called *ab initio* method [94]. Borodovsky called the former extrinsic methods and the latter intrinsic methods [95]. There are no clear boundaries between these two categories as sometimes they are used together in an integrated manner. The statistical model-based method relies heavily on training data for proper parameterization. This makes it also extrinsic, because external evidence is required. At the same time, sophisticated homology-based methods also consider models.
Homology-based annotation

A homology-based method searches the genome against databases of known genes. Similarity between a stretch of genomic DNA sequence and a known gene is the annotation evidence. Significant similarity indicates stronger evidence. BLASTX program could be used to search protein databases like UniProt, PIR or NCBI nr. BLASTN program could be used to search cDNA databases like UniGene. Many tools were developed to facilitate this process [96-100]. Expressed sequence tags (ESTs) could be directly used, but because they are partially sequenced transcripts, further analysis is required to deduce a full gene model from EST data. EbEST [101], EST_GENOME [102] and TAP [103] are tools for this purpose. Recent RNA-seq data could also be used directly, and this is demonstrated in Chapter 4.

However, it is estimated that only 50% of the genes in a genome could be identified by homology with known genes [93]. If a gene has a low expression level or is only transcribed under a very specific condition, it has a reduced probability of being detected by previous experiments, and thus is less likely to be present in the database. Homology-based prediction for genes like this would have a poor performance. What’s more, for a newly sequenced genome that is distantly related to all known species in the database, the homology-based prediction will be greatly hampered because fewer matches would be found due to greater divergence between them.

The experimental solution to the above challenges is to sequence more transcriptomes from more species and more developmental stages. The high-depth data of RNA-seq technology could also help to detect low-level expression by over sampling, or “deep sequencing”. One of the objectives of my study in Chapter 4 was to find novel transcripts specific to certain neural
developmental stages. RNA-seq technology was used for gene expression detection, and a considerable number of transcripts previously unannotated in the well-studied mouse genome were found.

Homology-base method alone could not precisely determine the boundaries of an exon or a gene. If based on cDNA sequence, 3’UTR and 5’UTR could not be distinguished from protein-coding exons. If searched against a protein sequence, 3’UTR and 5’UTR could not be picked up. Small exons are also easily missed because of lower significance by shorter alignment length.

**Model-based annotation**

Model-based gene prediction finds genes based on models that are parameterized on known sequence characteristics. For example, exons generally have a higher GC content than introns. Different patterns could be identified for exonic and intronic DNAs. Fickett and Tung have found that a good discrimination between them could be achieved by simply counting the frequency of DNA oligomers of a certain length (K-mers) [104]. They also found that the best resolution is from hexamers. This length was also reported by Qi, Luo and Hao to have the best resolution for inferring evolutionary distance from K-mer composite vector technique [56]. Because of the functional constrain is placed on protein-coding exons, the $Ka/Ks$ ratio between exons and introns are also different, which could be used to facilitate gene prediction [105].

However, in order to precisely determine the boundaries of a certain feature, more sophisticated and dedicated modeling is required. Exon-intron boundaries could be determined by identifying splice sites. Under the canonical spliceosome-splicing model, the 5’ end of an intron has a conserved donor site whose sequence is normally C(orA)AG/GTA(orG)AGT, and 3’ end
has a less conserved acceptor site whose sequence is normally T(orC)AG/G [106-109]. The pattern (C/T)T(A/G)A(T/C) of the branch point for the lariat formation is also not as well conserved as the 5’ donor site in terms of both sequence similarity and location [108]. Similarly, sequence features could be summarized for promoters, transcription start sites, transcription termination sites and translation start sites. Gene prediction algorithms looking for these patterns are called signal sensors [93]. The original GenScan software is heavily based on the concept of signal sensor [110].

Patterns of splicing donors, branch points and acceptors for a given species could be further refined by consensus calling from a multiple sequence alignment of experimentally defined and well-documented training sequences. Consensus sequences normally do not reflect base distribution among sites. Position weighted matrix, similar to PSSM introduced previously, could be used to modeling base distribution [96]. This could be considered as a 0-order Markov model [93]. Higher order Markov models could be used to reflect site dependency in the pattern. Neural network (NN) is also used to improve parameterization and pattern recognition, as first proposed by Brunak, Engelbrecht and Knudsen in 1991 [111]. Because the majority of an intron is allowed to vary without impact on splicing, a hidden Markov model (HMM) similar to the one introduced in multiple sequence alignment modeling is needed to allow insertions and deletions of any length [112]. In order to model after site dependency between non-adjacent sites, for example, between donor, branch point and acceptor sites, a decision tree algorithm called maximal dependence decomposition (MDD) was proposed by Burge [113, 114]. An auxiliary method for proofreading exon-intron splicing junction prediction is to check whether the exons are in the same reading frame and not interrupted by premature stop codons.
HMM could also be used for general-purpose gene prediction besides splicing pattern recognition. Unlike the profile HMM introduced for modeling after multiple sequence alignment whose states representing match, insertion and deletion, here hidden states represent instead either exon or intron regions. Viterbi algorithm again is used to find the optimal state path, at the same time assign regions as exons or introns. The first HMM gene finder is developed by Krogh, Mian and Haussler for *Escherichia coli* gene prediction [115]. GenScan also uses HMM in its later versions. Genie [116] and GRPL [117] uses a modified HMM called Generalized Hidden Markov Model (GHMM) by taking into consideration the length of the region.

Model-based annotation is extremely sensitive to parameterization in training. A common issue with model-based approach is over-fitting. It means that the model is overly optimized for the training set, and performs poorly on new genomes that do not closely resemble the characteristics of the training set. Sometimes this could be overcome by including sequences from a more diverse background. But this could cause reduced sensitivity since the pattern might be less significant when a mixer of different lineage-specific patterns is introduced. This is also directly related to difficulties in optimizing multiple sequence alignment from a very diverse background. Sometimes when there is little information about the genome, proper parameterization is almost impossible.

**Integrated annotation**

Instead of parameterization using training data, the model-based method can be parameterized based on new genes predicted by homology-based method. The homology-based method and model-based method could thus be combined. GenomeScan [94] integrates the model-based method from GenScan with homology-based method. It combines exon-intron
HMMs and splicing pattern recognition models with similarity search. In contrast to GenScan, the parameterization is not based on a training set, but on BLASTX search hits against an appropriate protein database. This is in principle similar to estimation maximization (EM): the best hits from the genome against the database parameterize the probabilistic model, which is then used to predict the most probable exon-intron assignment. In this process, the user does not need to explicitly identify a training set. Similarly, TwinScan [118] is a gene prediction tool based on GenScan method for genome-genome homology comparison. Its model is parameterized based on hits to another genome by BLASTN.

For researchers interested in only a few gene families, many tools were developed based on the same concept. GFscan [119] is a gene family search tool based on a motif recognition model but parameterized based on input gene family. SLAM [120] does gene prediction based on HMM, but also parameterizes based on input gene family. GeneWise and GenomeWise [121] are gene prediction tools based on a merged HMM (from DNA directly to protein) that parameterizes based on the homology of input single protein or protein family. These tools bypass the steps of training set selection, training, parameter testing and training set optimization, and thus are widely used for their operation simplicity. This allows the possibility of analyzing gene families before a genome is fully annotated. This also makes annotation of genes from a given family more sensitive and accurate since the parameterization is based on this specific gene family instead of the average of the entire genome. FGF [122], BLASTO [123] and TARGeT [124] are tools designed specifically for this purpose. This leads to the possibility of community or crowd-sourced genome annotation with parameters tailored for each gene family.

(End of Chapter 1)
Chapter 2

Information analysis of mammalian mitochondrial DNA sequence in molecular dating applications

The results in this chapter are included in published literature coauthored by this dissertation author [51, 125]. Used under permission from PNAS and Nature (see Appendix).

Introduction: mitochondrial DNA and its applications in molecular dating

As introduced in Chapter 1, differences between sequences could be used for their divergence time estimation. Mitochondrial DNA sequence is widely used for this purpose. Mitochondrial genome size is relatively small: animal mitochondrial genome size is between 10 and 20 kb, and mammalian mitochondrial genome size is about 16 to 17 kb. It is widely accepted that the origin of eukaryotic mitochondria is an endosymbiotic process during which a prokaryotic organism was introduced into an ancestral eukaryotic cell. This theory that was introduced by Lynn Margulis in 1967 for the first time with microbiological support [126]. Because of mitochondria’s prokaryotic origin, its genome is circular, unlike the linear eukaryotic nuclear genome. A circular genome is less susceptible to DNA degradation as it has no free ends as those in a linear genome. The number of mitochondria inside a single cell is estimated to be around 100 to 1000. Because of the above features, mitochondrial DNA could be relatively easily retrieved from ancient samples and then sequenced [127, 128]. In comparison, nuclear DNA from ancient samples is normally in less quantity, poorer quality and more fragmented.
Brown, George and Wilson in 1979 discovered through restriction endonuclease digestion experiments that primate mitochondrial genomes exhibit a 10 times faster evolution rate than that of their nuclear genomes [129]. Mitochondria control region (CR) rate was also estimated through human pedigree sequencing data by Parsons et al in 1997 and Sigurgardottir et al in 2000 [130, 131]. Both studies concluded that mitochondria control region exhibits rapid mutation. Direct estimation of whole mitochondria mutation rate was done in Drosophila melanogaster in 2008, and the researchers also found that the rate is 10 times faster than nuclear genome [132].

Because of this rapid evolutionary rate, the mitochondrial genome has been used to study the phylogeny and divergence time of recent speciation events. For example, the first molecular dating of the split between human and chimpanzee was done with complete mitochondrial sequences by Hasegawa, Kishino and Yano in 1985 [14]. It has also been extensively used for population studies. For example, the origin of modern human was also estimated using non-CR mitochondrial sequences by Ingman et al in 2000 [15]. Concluded by that study, it is now widely accepted that anatomically modern humans appeared in Africa around 200–160 thousand years ago (kya). They then expanded to most of the habitable parts of the Old World between 90 and 40 kya [133]. The history of modern polar bears is another example because of its recent speciation and rapid population expansion along the Arctic Circle and adjacent lands.

In this chapter, the questions surrounding human and polar speciation and migration were investigated with the help of newly discovered ancient remains, samples within the deep clades of human population, Next Generation Sequencing (NGS) technologies and new molecular dating methods and tools.
The assembly and annotation of a 100,000 years old mitochondrial genome and its application in molecular dating of a Pleistocene polar bear (Ursus maritimus Phipps, 1744) jawbone from the Svalbard islands

Project background

In 2008, a geologically dated 120,000 years old mammalian jawbone was discovered on the Svalbard islands in Norway [134]. Based on its morphology, this jawbone was determined to be from an ancient polar bear. If this is correct, this would be the oldest polar bear remains ever discovered at that time. What’s more intriguing, previous estimation [135-137] of the modern polar bear divergence time is also around or even younger than the age of this sample. Could this jawbone be from, or extremely close to, the ancestors of all modern polar bears? Professor Charlotte Lindqvist from University of Buffalo, whose research group was then using molecular techniques to determine the origin of this jawbone, visited my then research group led by Professor Stephan Schuster. Both groups saw the opportunity to use the NGS technologies to answer this question. I was deeply interested, offered my knowledge and experience in NGS data analysis and joined this project.

Challenges: conflicting estimations of polar bear divergence time

Polar bear belongs to the Ursidae family, which includes several different bear species [138, 139]. Polar bear (Ursus maritimus) is the largest bear as well as the largest land carnivore. Their habitats are along the entire Arctic Circle, on both ice caps and the adjacent lands. Kurtén in 1964 suggested that polar bear is evolutionarily related to brown bear (Ursus arctos) [135]. Studies based on partial mitochondrial sequences supported the notion that polar bears evolved from within brown bears, and are more closely related to brown bear populations that live on the
Admiralty, Baranof, and Chichagof (ABC) islands of southeastern Alaska than to other brown bears [136, 140].

Divergence time estimations between polar bears and brown bears differed considerably between different studies. An estimated divergence time of 200 to 250 thousand years ago has been proposed based on partial mitochondrial DNA sequences by one study [136]. Another study based on complete mitochondrial genome sequences estimated this divergence time to be as early as 1,320 thousand years ago [137]. Krause et al in 2008 estimated the time to be about 660 to 1,170 thousand years ago based on analyses of 10 mitochondrial genomes [138]. Some of the studies used fossil calibration on deep internal node. This might not be appropriate for recent speciation event [141]. Most importantly, they did not include data from brown bears on the ABC islands, which is believed to be the brown bear clade closest to polar bear. Ho et al tried to use internal substitution rate calibration and time-calibrated leaf node samples to minimize deep internal node fossil calibration, and they have also included the brown bear clade from the ABC islands [141]. Their divergence time estimation between polar bear and brown bear was between 48 and 72 thousand years ago, which is much younger than previous estimations. However, in that study the authors used only the control region sequence, which is less than one tenth of the entire mitochondrial DNA length. This reduced the amount of information that could be extracted. But most importantly, the control region may not be the ideal region to apply molecular clock, as Ingman et al have found in human mitochondria [15].
Opportunity: deep leaf node calibration made possible by Pleistocene jawbone

The main habitat for polar bear is over arctic ice caps. With their seasonal freeze and thaw cycle, it is expected that polar bear remains are easily taken away by the Arctic Ocean and are harder to discover than other land animal remains. As a result, their fossils or remains are rare [134, 135, 142, 143]. However, a morphologically identified lower jawbone (left mandible) of an ancient polar bear was discovered on the Svalbard islands in Norway (See map in Figure 2-1) [134]. In that report, carbon dating estimated this jawbone’s age to be older than 45 thousand years and the geological dating put the age between 110 and 130 thousand years [144]. All other known polar bear remains are at a much younger age (no older than 70 thousand years ago) [134]. If the geological dating is accurate, this evidence supports that the polar bear has diverged from brown bear, and then expanded to Arctic Circle 110 to 130 thousand years ago, tens of thousands of years before the estimation by Ho et al [141]. A canine tooth attached to the discovered jawbone was selected for DNA extraction. The protective outer layer of this tooth (enamel), together with the low environmental temperature, it might be possible to extract high quality DNA for sequencing from this 110 thousand years old specimen. This would be a rare opportunity of deep leaf node calibration with relatively precise geological dating, instead of deep internal node fossil calibration.

As introduced in Chapter 1, it is possible through ultra high-throughput sequencing technologies to recover a large amount of sequence data from a relatively small amount of DNA. This allows DNA sequencing of ancient remains whose DNA quantity is limited [128, 145]. Because previous ancient DNA recovery techniques rely on specific primers, they could only recover a few genes or DNA regions that have been well studies in the current species. In contrast, the NGS technologies could amplify and sequence DNA molecules without gene or
region specific primers. The endogenous DNA sequences could be also distinguished from environmental DNA sequences, as they will show up as separate reads. This is unlike Sanger sequencing data, whose endogenous signals normally overlap with contaminant signals if the latter is present. If DNA sequences from this 110 thousand years old specimen were recovered, it would be much older than any known DNA sequence from mammalian sub-fossil remains [146].

Figure 2-1: Sampling site locations.
The assembly and annotation of a complete 100,000 years old mitochondrial DNA

The recovered DNA, extracted by Professor Lindqvist’s group, was sequenced using 454 GS FLX platform by Professor Schuster’s group. Total 77 Mb data were generated. Professor Webb Miller aligned these reads against dog genome to estimate the percentage of bear DNA, as no complete bear nuclear genome was available at that time. 40% of the data could be mapped to the published dog genome, indicating possible endogenous bear data (Figure 2-2). Professor Lindqvist’s group also conducted research on the feeding patterns of the ancient polar bear through isotope analyses, as introduced in [51]. The findings concluded that this jawbone was from a bear individual sharing similar feeding patterns with modern day polar bears [51].

Figure 2-2: Estimated composition of the recovered DNA

The remaining data were then aligned with human genome and NCBI nt database. It was estimated that about 4.5% of the data were human origin, indicating a very low human contamination. However, the remaining 55% of the sequences were estimated to be from bacteria. Although only 40% of the total data, the endogenous DNA could still provide sufficient information. In addition, they had long read length and high signal quality, largely due to low temperature and glacial/permafrost environment as suggested by previous studies [147, 148].
A total of 482,364 reads were generated. Of those, 1290 were determined to be of ursine mitochondria origin. This gave 14X-coverage over the length of a typical bear mitochondrial DNA molecule. These reads were fully assembled into a complete mitochondrial genome (Figure 2-3). The length and annotated genome content of this assembled genome were similar to those of the known Ursidae mitochondrial genomes. In addition, 10 modern bear genomes were sampled using 454 multiplex sequencing technology. Among them, four were brown bears from the ABC islands (2 from Admiralty, 1 from Baranof, and 1 from Chichagof island). Two are brown bears from Kodiak Island (west of the ABC islands). The rest were three polar bears from Alaska. From the resulting data, total six complete mitochondrial genomes were assembled, including three ABC islands brown bears, one Kodiak Island brown bear and two Alaska polar bears (Table 2-1 and Figure 2-1).

Figure 2-3: Coverage of mitochondria genome assembled from ancient polar bear DNA.
Figure 2-4: Genome variation pattern among all known complete ursine mitochondria. The black bars of each genome represent SNPs with NC003428 as the reference.

Table 2-1: Genome sequencing and assembly information

<table>
<thead>
<tr>
<th>MID</th>
<th>ID</th>
<th>Species</th>
<th>Tissue</th>
<th>Location</th>
<th>454 reads</th>
<th>Mitochondria reads</th>
<th>Complete (Non-VNTR)</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>503950</td>
<td>U. arctos</td>
<td>muscle</td>
<td>Alaska (Chilkoot Lake)</td>
<td>83842</td>
<td>2</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>14KB</td>
<td>U. arctos</td>
<td>muscle</td>
<td>Alaska (Baranof Island)</td>
<td>82790</td>
<td>3425</td>
<td>Yes</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>12WH</td>
<td>U. arctos</td>
<td>muscle</td>
<td>Alaska (Admiralty Island)</td>
<td>108532</td>
<td>2571</td>
<td>Yes</td>
<td>36</td>
</tr>
<tr>
<td>4</td>
<td>A91-05</td>
<td>U. arctos</td>
<td>muscle</td>
<td>Alaska (Admiralty Island)</td>
<td>10428</td>
<td>4866</td>
<td>Yes</td>
<td>70</td>
</tr>
<tr>
<td>5</td>
<td>84688</td>
<td>U. arctos</td>
<td>muscle</td>
<td>Alaska (Chichagof Island)</td>
<td>44266</td>
<td>5</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>6</td>
<td>76946</td>
<td>U. arctos</td>
<td>muscle</td>
<td>Alaska (Kodiak Island)</td>
<td>146</td>
<td>18</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>7</td>
<td>76824</td>
<td>U. arctos</td>
<td>muscle</td>
<td>Alaska (Kodiak Island)</td>
<td>78018</td>
<td>891</td>
<td>Yes</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>542</td>
<td>U. maritimus</td>
<td>homogenate</td>
<td>Alaska (Barrow)</td>
<td>120195</td>
<td>261</td>
<td>No</td>
<td>3.5</td>
</tr>
<tr>
<td>9</td>
<td>2495</td>
<td>U. maritimus</td>
<td>muscle</td>
<td>Alaska (Little Diomede Island)</td>
<td>70862</td>
<td>4685</td>
<td>Yes</td>
<td>68</td>
</tr>
<tr>
<td>10</td>
<td>2629</td>
<td>U. maritimus</td>
<td>muscle</td>
<td>Alaska (St. Lawrence Island)</td>
<td>25560</td>
<td>1855</td>
<td>Yes</td>
<td>23</td>
</tr>
<tr>
<td>NA</td>
<td>NA</td>
<td>U. maritimus</td>
<td>tooth</td>
<td>Svalbard (Poolepynten)</td>
<td>482364</td>
<td>1290</td>
<td>Yes</td>
<td>14</td>
</tr>
</tbody>
</table>

1MID: Multiplex Identifiers in 454 Multiplex sequencing.
2VNTR: Variable Number Tandem Repeat.

Mitochondrial genome variation patterns, including SNPs and indels between all known complete ursine mitochondrial genomes, were analyzed. Complete ursine mitochondrial genomes
from other studies [139, 149, 150] were also included to better represent the genetic clusters previously identified using microsatellite data [151]. Figure 2-4 showed that the variation pattern of the ancient polar bear was between those of the modern polar bears and modern brown bears, closer to the brown bears ones from the ABC islands.

**Polar bear divergence time and evolutionary position estimation**

Several different phylogenetic reconstruction methods introduced in Chapter 1 were used in analyzing polar bear’s evolution history. Maximum parsimony, maximum likelihood and Bayesian method all inferred the same topology, placing the Pleistocene polar bear at the branching point between modern polar bears and brown bears (Figure 2-5). This topology agrees with those reported by several recent studies [138, 139, 152]. More importantly, the ancient polar bear node has an extremely short branch length. This indicates that this bear uncovered on the Svalbard islands is genetically and evolutionarily very close to the common ancestors of all modern polar bears. This finding makes the uncovered jawbone a special specimen reflecting the starting point of polar bear evolution. The strongly supported phylogeny and the short branch length allow a robust estimation of polar bear divergence time from within brown bear. The inclusion of complete ABC islands brown bear mitochondrial genomes helps to increase the precision on the upper limit estimation of this divergence time, and the inclusion of polar bears mitochondrial genomes from multiple genetic clades increases the precision on the lower limit estimation.
Using the method first proposed by Drummond et al [65] as introduced in Chapter 1, polar bear’s divergence time is estimated to be about 134 thousand years ago (Table 2-2). This estimation is within the Eemian interglacial period (between 114 to 130 thousand years ago). This analysis has also concluded that the divergence time for all sampled modern polar bear clades is only 50 thousand years ago. This finding agrees with recent studies using deep fossil calibration [139], which is not used here. Using only mitochondrial control region sequences, Professor Lindqvist assembled another data set representing more ursine individuals (39, including several carbon dated ancient bear remains [153]). Although with less information content per sequence, this set included polar bears on the Svalbard islands, where the ancient polar bear jawbone was discovered. A similar analysis was done based on this set of data, and a comparable time estimate was produced (Figure 2-6). All these pieces of evidence support a very recent polar bear
divergence time from within brown bear. This further indicates a very rapid adaptation and expansion of polar bear population in the Arctic.

Table 2-2: Comparison of recent ursine clade molecular dating results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sloth + Sun/Black/Brown/Cave/Polar</td>
<td>6.34 (5.95 – 6.73)</td>
<td>5.39 (4.2 – 6.86)</td>
<td>3.0 (2.2 – 3.8)</td>
<td>NA</td>
<td>NA</td>
<td>2.59 (1.32 – 4.05)</td>
</tr>
<tr>
<td>Sun/Black + Cave/Polar/Brown</td>
<td>6.13 (5.54 – 6.72)</td>
<td>5.05 (3.9 – 6.48)</td>
<td>2.8 (2.1 – 3.5)</td>
<td>NA</td>
<td>2.60 (1.85 – 3.42)</td>
<td>2.49 (1.26 – 3.86)</td>
</tr>
<tr>
<td>Cave + Brown/Polar</td>
<td>NA</td>
<td>2.75 (2.1 – 3.57)</td>
<td>1.6 (1 – 2.1)</td>
<td>NA</td>
<td>1.43 (1.03 – 1.88)</td>
<td>1.39 (0.67 – 2.16)</td>
</tr>
<tr>
<td>Sun + Black</td>
<td>5.673 (5.09 – 6.26)</td>
<td>4.58 (3.51 – 5.89)</td>
<td>2.4 (1.7 – 3)</td>
<td>NA</td>
<td>2.35 (1.58 – 3.16)</td>
<td>2.12 (0.88 – 3.43)</td>
</tr>
<tr>
<td>Asiatic + American Black</td>
<td>5.19 (4.6 – 5.78)</td>
<td>4.08 (3.11 – 5.27)</td>
<td>2.1 (1.4 – 2.7)</td>
<td>NA</td>
<td>NA</td>
<td>1.85 (0.72 – 2.97)</td>
</tr>
<tr>
<td>Brown + Polar²</td>
<td>1.32 (0.93 – 1.71)</td>
<td>0.88 (0.66 – 1.17)</td>
<td>0.6 (0.3 – 0.8)</td>
<td>east brown; 0.4 (0.2 – 0.5) west brown</td>
<td>0.44 (0.14 – 0.97) east brown; 0.31 (0.24 – 0.40) west brown</td>
<td>0.51 (0.29 – 0.76) east brown; 0.33 (0.20 – 0.49) west brown</td>
</tr>
<tr>
<td>ABC + Polar</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0.23 (0.11 – 0.50)</td>
<td>0.152 (0.131 – 0.177)</td>
<td>0.167 (0.132 – 0.214)</td>
</tr>
<tr>
<td>Ancient + Modern</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0.17 (0.11 – 0.32)</td>
<td>0.134 (0.122 – 0.149)</td>
<td>0.138 (0.122 – 0.162)</td>
</tr>
<tr>
<td>Modern</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>68.6k (55 – 166.0)</td>
<td>43.5k (29.0 – 59.0)</td>
<td>46.7k (34.0 – 71.2)</td>
</tr>
<tr>
<td>Method</td>
<td>Other Info</td>
<td>(5.5 – 166.0k)</td>
<td>(29.8 – 59.0k)</td>
<td>(24.0 – 71.2k)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREE-PUZZLE: molecular clock on coding regions. One fossil calibration point</td>
<td>Speciation Yule process, MCMC chain length = 30,000,000, log every 1000, Tracer burn-in 5,000,000 states,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEAST: strict clock, GTR + Γ4. Three fossil calibration points.</td>
<td>Speciation Yule process, MCMC chain length = 10,000,000, log every 1000, burn-in 10,000 states,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEAST: relaxed uncorrelated lognormal clock, GTR + I + Γ4. Two fossil calibration points.</td>
<td>Speciation Yule process, MCMC chain length = 50,000,000, log every 1000, Tracer burn-in 5,000,000 states, all ESS &gt; 210, Sampling 50,001 trees, TreeAnnotator burn-in 1000 trees</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEAST: relaxed uncorrelated lognormal clock. GTR + I + Γ4. No fossil calibration.</td>
<td>Speciation Yule process, MCMC chain length = 50,000,000, log every 1000, Tracer burn-in 5,000,000 states, all ESS &gt; 240, Sampling 50,001 trees, TreeAnnotator burn-in 1000 trees</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEAST: relaxed uncorrelated lognormal clock. GTR + I + Γ4. No fossil calibration.</td>
<td>Speciation Yule process, MCMC chain length = 80,000,000, log every 1000, Tracer burn-in 8,000,000 states, all ESS &gt; 230, Sampling 80,001 trees, TreeAnnotator burn-in 1000 trees</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1Unit: million years ago, unless labeled otherwise. Mean and 95% HPD.
2East brown bear: NC_003427, west brown bear: EU_497665.
3Using HKY model instead of the GTR + I resulted in similar time estimates, but all ESSs are > 1100.
Figure 2-6: Phylogeny from BEAST analysis on control region sequences. Internal node numbers indicating time (unit: million years).
Materials and methods

Method of DNA extraction and initial ancient DNA amplification

The polar bear jawbone sample was processed by the research group led by Professor Charlotte Lindqvist in University of Oslo, Norway, using previously suggested protocols [154]. The details of this procedure were also included here for completeness. Based on [51], a canine tooth was selected and cleaned with 10% chlorine for surface decontamination. About 0.1 grams of dentine powder was extracted under the enamel layer. DNA was then extracted from the dentine powder using previously published protocol [154]. Initial PCR amplification confirmed a fragmented DNA source, and typical ancient DNA damages (for example, C-to-T mutation caused by cytosine deamination).

Method of 454 sequencing of ancient DNA and modern bear DNA

Sequencing libraries were constructed following standard GS FLX protocol (454 Life Sciences). 454 sequencing was done on a GS FLX sequencer (454 Life Sciences) at the Center for Comparative Genomics and Bioinformatics, Pennsylvania State University. Two fragment runs were completed solely for the ancient polar bear sample. This yielded 77 million bp data. One of these two runs had an average length of 156 bp (of total 347,660 reads), and the other had an average length of 168 bp (of total 134,704 reads).

To minimize modern samples’ contamination, all sequencing procedures involving modern samples were not performed until all analyses of the ancient sample were done. DNA extractions from the modern polar and brown bear tissue samples were done at University of
Buffalo, as introduced in [51]. For completeness and legality, the following statement from the published paper [51] is quoted here:

“The modern polar bear samples were obtained from the U.S. Fish and Wildlife Marine Region 7 Mammal Management Office and harvested in 1998, 2001, and 2002 before listing of the species in the Endangered Species Act.”

Multiplex identifiers (MIDs) were used in sequencing of the modern day bear samples. MID is a 10-bp nucleotide identifier unique to each sample, chemically attached to each DNA fragment of that sample. Total 10 modern bear samples were sequenced (MID1 to MID10), 6 of which produced sufficient mitochondrial reads for assembly. Total 624,921 reads (mean length = 186 bp) were produced, with an uneven distribution among samples (Table 2-1).

Method of mitochondrial genome assembly

Reads were first size-selected with a threshold of 100 bp to ensure assembly specificity. Putative ursine mitochondrial reads were then selected by using BLASTN [9] search. NCBI data entry accession no. NC003428 was used as the reference genome. Search threshold was set at E value = e−20. The assembler used here was SeqMan 8.0 (DNASTAR). Manual inspection and correction were then carried out for the entire assembled genome.

Ursine mitochondria genomes normally contain a variable-number tandem repeats (VNTR) region located inside D-loop. Its average length (about 500 bp) is longer than the typical 454 read length at that time. To determine VNTR repeat number, Yeting Zhang first performed PCR amplification over VNTR regions for all genomes (primers: 5′-
CGCACTAAATCGAACGAAC-3', 5'-GGGGTTTGATTAAGCTAAGTT- 3') and then analyzed their length on Agilent 2100 bioanalyzer using DNA 7500 kit (Agilent Technologies). Repeat number was then determined based on the overall VNTR length and the repeat unit length.

**Estimations of divergence times**

Estimation was done in BEAST [50] v. 1.4.8 with Bayesian “relaxed molecular clock” model [65]. The parameters used are listed in Table 2-2. The time calibration of the polar bear jawbone sample is 120 ky. Tracer v. 1.4.1 (http://beastbioedacuk/Tracer) was used to analyze BEAST output and perform statistical testing. TreeAnnotator v. 1.4.8 was used to produce maximum clade credibility trees from BEAST output.

**Project summary and discussion**

The molecular dated polar bear divergence time, along with geological dating of the discovered jawbone from the Svalbard islands, depicted a very recent speciation event followed by rapid adaptation and expansion of the polar bear. Based on the MCRA dating, polar bears had adapted to the harsh Arctic environment within only about 10 to 30 thousand years. For the next 100 thousand years, they expanded along the entire Arctic Circle. This rapid adaptation and expansion is an example of mammalian evolution to niche environment. However, the genetic and molecular mechanism of such a rapid process is still unknown.

The maximum information $I$ mitochondrial genome could provide is only about $32 \times 10^3$ bits, but the nuclear genome has maximum $6 \times 10^9$ bits information, assuming a 3 Gb genome.
Although the actual information \( I \) contained by a genome is related to the number of variable sites instead of the total length, it is obvious that nuclear genome has much more information than mitochondrial genome. The extra information might be able to answer some of the reasons for the rapid adaptation and expansion. Considering the percentage of endogenous bear DNA and its quality, it might be possible to recover the entire nuclear genome through further deep sequencing. However, nuclear genome is of only one copy per cell, and the linear nuclear genome is highly susceptible to hydrolysis and other types of DNA damages. This makes the recovery of high coverage nuclear genome information a great challenge.

**Investigating the lineage characteristics of sampled Khoisan and Bantu genomes**

**Project background**

In 2008 and 2009, my then research group led by Professor Schuster and a collaborating group led by Professor Vanessa M. Hayes from University of New South Wales collected blood and hair samples from several Khoisan (also known as San or Bushmen) people resided deep inside Kalahari desert in Botswana. It has been shown by other studies that this group of people is among the oldest lineages of all modern human [155-157]. However, the lineage status and characteristics of the collected samples could not be determined without sequencing and proper data analysis. The methods I used in the earlier polar bear project were able to answer to this question, and I joined this project.
Modern human’s African origin and current deep African lineages

As reviewed earlier in this chapter, Ingman et al in 2000 estimated the age of the most recent common ancestor (MRCA) of modern human to be around 171,500 years based on non-CR mitochondrial genome sequences [15]. Also shown in that study, African populations have higher intra-population diversity than non-African populations, and San population in Southern Africa has the largest diversity. That study confirmed the “African-origin” hypothesis, instead of the “multi-origin” hypothesis, of human evolution. Although the ancestor of modern human, Homo erectus, migrated out-of-Africa and expanded to Southern and Central Asia, modern non-African populations are all direct descendents of the African L haplogroup, or more precisely the L3 haplogroup, and consisting only two major haplogroups, M and N. The “out-of-Africa” event of modern human, Homo sapiens sapiens, is thus a separate event from that of Homo erectus. The decrease in between-individual diversity in non-African populations is thus the result of the “bottleneck-effect” of the aforementioned migration events.

Previous global haplogroup studies have shown that the L0 group of the L haplogroup is the oldest lineage of modern human [158, 159]. Although it is reported that modern human originated from Eastern Africa [157, 160, 161], the Southern Africa regions have also been shown to be the habitats of ancestors of modern human [133, 162]. Actually, the arid climate currently typical of this region may still resemble the climates tens of thousands of years ago. Kalahari Desert in Botswana, where the Khoisan samples were collected, belongs to this region. Historical linguistics studies have also shown the long history of these people. Khoisan people have a very special form of “click-like” language, which was considered a very old lineage of human language [161]. Are these people sampled in the Kalahari Desert among the oldest
lineages of known human population? Their mitochondrial genomes could be used to answer this question.

Assembling human mitochondrial genomes from a mixture of technologies

There are total seven individuals sampled, six of them were from Botswana. The other sample, ABT, was collected from a South Africa individual to represent the Bantu lineage, a different lineage from Khoisan. The other six samples were KB1, KB2, NB1, NB8, MD2 and MD8. KB1 and KB2 were individuals from Southern Kalahari, while NB1, NB8, MD2 and MD8 were individuals from Northern Kalahari. The two collection sites inside Kalahari Desert are about 800 miles apart, separated largely by desert.

A mixture of 454, Illumina and Sanger sequencing technologies was used to assemble all seven full-length human mitochondrial genomes (Table 2-3). The read coverage of these mitochondrial genomes was greater than 17X. The largest length variation of these sequences compared with the rCRS was only 3 base pairs. Manual inspections were carried out to ensure the quality and accuracy of the assembly.

Table 2-3: Statistics of assembled human mitochondrial genomes

<table>
<thead>
<tr>
<th>Name</th>
<th>Length bp</th>
<th>454 Coverage</th>
<th>Total Number of 454 Reads</th>
<th>Illumina Coverage</th>
<th>Total Number of Illumina Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>KB1</td>
<td>16569</td>
<td>134.05</td>
<td>4630*</td>
<td>17.81</td>
<td>8221</td>
</tr>
<tr>
<td>NB1</td>
<td>16566</td>
<td>139.61</td>
<td>4568*</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>ABT</td>
<td>16570</td>
<td>NA</td>
<td>NA</td>
<td>29.71</td>
<td>13695</td>
</tr>
<tr>
<td>NB8</td>
<td>16567</td>
<td>29.00</td>
<td>1067</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>MD2</td>
<td>16568</td>
<td>41.30</td>
<td>1438</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>KB2</td>
<td>16567</td>
<td>16.69</td>
<td>662</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>MD8</td>
<td>16566</td>
<td>26.20</td>
<td>972</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
Lineage information revealed by phylogenetics analyses

In order to determine their lineage characteristics, the assembled genomes were first compared with known haplogroups. Because of the large number of novel SNPs inside these genomes, current tools including mtDB [163] and PhyloTree [164] databases could not classify these new genomes satisfactorily to any known haplogroups. In order to resolve their haplogroup designation, the phylogeny of these sequences needed to be reconstructed in the context of sequences from all major haplogroups. A total of 162 high-quality full-length mitochondrial genomes were collected from NCBI Nucleotide database. Among them, about 50% (88) were African. Many of these sequences were published by Sarah Tishkoff group. The over-representation of African haplogroups in this dataset provided fine resolution within L and L0 haplogroups. A phylogenetic tree based on these seven new genomes and other previously published ones was built (Figure 2-7). The haplogroup composition of this dataset is shown in Figure 2-8. According to Ingman et al 2000 [15] and our own data, the D-loop region was excluded in order to maintain valid molecular clock properties for divergence time estimation. Using similar method as in polar bear project, the divergence time of the major haplogroups included in this study was estimated (Figure 2-7 and Table 2-4).

All seven new genomes in this study all belong to the L0 haplogroups. However, they reside in two major subgroups of L0. KB2, NB1 and MD8 are from the L0k subgroup, and ABT, MD2, KB2 and NB8 are from the L0d subgroup. These two subgroups split about 170 thousand years ago. Although the two Kalahari sampling sites are separated by hundreds of miles of desert,
the mitochondrial genome flow did not appear to be interrupted. Actually, the KB1-containing group was related to Tanzania L0d group in Eastern Africa, which was suggested to be the origin of modern human by Gonder et al 2007 [155]. Especially for the ABT sample, although he was selected from South African as a representative of Bantu people, later it was discovered that his maternal lineage has contributions from Khoisan people.

Some of the divergence times between clades within L0d were as old as that between M and N haplogroups. This is also reflected by the large number of SNPs/substitutions between these seven genomes. Although the average number of substitutions between two European individuals is about 20 to 30, this number between these seven genomes is about 80 to 100 (Figure 2-9 and Table 2-5).
Figure 2-7: BEAST analysis of 162 individuals based on complete Mitochondrial sequences. Individuals from this study are shown in red.
Table 2-4: Comparison of molecular dating results for modern human lineages

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>H. sapiens sapiens with Neanderthal</strong></td>
<td>NA</td>
<td>NA</td>
<td>706.0 (468.0 – 1015.0)</td>
<td>NA</td>
<td>660.0 (520.0 – 800.0)</td>
<td>NA</td>
<td>Calibration point: 660.0 (520.0 – 800.0)</td>
<td>Calibration point: 660.0 (520.0 – 800.0)</td>
</tr>
<tr>
<td><strong>H. sapiens sapiens</strong></td>
<td>171.5 (121.5 – 221.5)</td>
<td>&gt; 195.0 (190.0 – 200.0)</td>
<td>NA</td>
<td>194.3 (161.8 – 226.8)</td>
<td>NA</td>
<td>136.1 (94.93 – 178.7)</td>
<td>204.9 (116.8 – 295.7)</td>
<td>237.2</td>
</tr>
<tr>
<td><strong>Neanderthal</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>109.8 (84.63 – 138.5)</td>
<td>130.3 (89.7 – 174.3)</td>
<td>141.2</td>
</tr>
<tr>
<td><strong>L0</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>146.4 (121.3 – 171.5)</td>
<td>NA</td>
<td>NA</td>
<td>158.7 (88.6 – 231.5)</td>
<td>187.9</td>
</tr>
<tr>
<td><strong>L0d</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>106.0 (85.8 – 126.2)</td>
<td>NA</td>
<td>NA</td>
<td>107.2 (56.8 – 159.9)</td>
<td>129.8</td>
</tr>
<tr>
<td><strong>Tanzanian L0d</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>30.6 (12.8 – 48.4)</td>
<td>NA</td>
<td>NA</td>
<td>32.0 (14.6 – 51.6)</td>
<td>36.5</td>
</tr>
<tr>
<td><strong>San L0d</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>90.4 (71.5 – 109.3)</td>
<td>NA</td>
<td>NA</td>
<td>92.8 (47.6 – 139.1)</td>
<td>113.0</td>
</tr>
<tr>
<td><strong>L0k, L0f, L0a</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>139.8 (115.2 – 164.4)</td>
<td>NA</td>
<td>NA</td>
<td>148.2 (NA)</td>
<td>176.9</td>
</tr>
<tr>
<td><strong>L0k</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>70.9 (51.2 – 90.6)</td>
<td>NA</td>
<td>NA</td>
<td>78.6 (37.2 – 122.0)</td>
<td>94.5</td>
</tr>
<tr>
<td><strong>L0f, L0a</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>100.1 (87.6 – 112.6)</td>
<td>NA</td>
<td>NA</td>
<td>118.8 (63.1 – 175.9)</td>
<td>144.0</td>
</tr>
<tr>
<td><strong>L0f</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>94.9 (85.5 – 104.3)</td>
<td>NA</td>
<td>NA</td>
<td>93.8 (46.9 – 139.7)</td>
<td>117.2</td>
</tr>
<tr>
<td><strong>L0a</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>54.6 (48.9 – 60.3)</td>
<td>NA</td>
<td>NA</td>
<td>61.1 (29.2 – 95.1)</td>
<td>71.4</td>
</tr>
<tr>
<td><strong>L1, L2, L3, M, N</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>142.3 (104.1 – 180.5)</td>
<td>NA</td>
<td>NA</td>
<td>172.7 (95.2 – 249.7)</td>
<td>200.9</td>
</tr>
<tr>
<td><strong>L3, M, N</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>94.3 (84.4 – 104.2)</td>
<td>NA</td>
<td>NA</td>
<td>93.3 (50.7 – 136.5)</td>
<td>96.0</td>
</tr>
<tr>
<td><strong>Note</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>Range in parentheses is the 95% HPD interval</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td><strong>Methods</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>GTR + Gamma4 + I, Relaxed</td>
<td>GTR + Gamma4 + I, Fixed Clock</td>
<td></td>
</tr>
</tbody>
</table>

Note: The range in parentheses is the 95% HPD interval.
Time unit: thousand years

Figure 2-8: Dataset haplogroup composition
Table 2-5: Number of SNPs identified between mitochondrial genomes

<table>
<thead>
<tr>
<th></th>
<th>rCRS</th>
<th>ABT</th>
<th>KB1</th>
<th>NB1</th>
<th>NB8</th>
</tr>
</thead>
<tbody>
<tr>
<td>rCRS</td>
<td>-</td>
<td>100</td>
<td>89</td>
<td>96</td>
<td>95</td>
</tr>
<tr>
<td>ABT</td>
<td>100</td>
<td>-</td>
<td>51</td>
<td>84</td>
<td>55</td>
</tr>
<tr>
<td>KB1</td>
<td>89</td>
<td>51</td>
<td>-</td>
<td>77</td>
<td>38</td>
</tr>
<tr>
<td>NB1</td>
<td>96</td>
<td>84</td>
<td>77</td>
<td>-</td>
<td>75</td>
</tr>
<tr>
<td>NB8</td>
<td>95</td>
<td>55</td>
<td>38</td>
<td>75</td>
<td>-</td>
</tr>
</tbody>
</table>

Materials and methods

Research participant protection, consent forms, practice ethics, and IRB approvals

The following statement has been included in the published literature [125]. For legal and regulatory considerations, it is included here:
“Using guidelines approved by the Institutional Review Board of Penn State University, USA, (IRB #28460 and IRB#28890), the University of Limpopo, South Africa, (permit on file), and the University of New South Wales, Australia, (HREC #08089 and HREC #08244) all participants were consented by either writing or video consent. The collection of human DNA in Namibia was conducted under a permit by the Ministry of Health and Social Services (MoHSS) of the Namibian Government.”

Methods of 454 and Illumina sequencing

454 sequencing library construction for the extracted blood and hair DNA was carried out following 454 Life Sciences standard GS FLX protocol, and shotgun sequencing runs were performed on a Roche GS FLX sequencer (454 Life Sciences) at the Center for Comparative Genomics and Bioinformatics, Pennsylvania State University. Illumina sequencing library construction was done following Illumina Genome Analyzer (GA II) protocols. Illumina sequencing runs were also done at the same location.

Methods of mitochondrial genome assembly

454 sequencing reads were first filtered by size at 100 bp or larger for quality control. Illumina reads were not subjected to this screening as they have equal length, but they were filter according to their Fastq scores (Q-scores) provided by the sequencer. These filtered reads were then screened for those associated with human mitochondrial DNA using BLASTN [9] against the revised Cambridge Reference Genome [167] from NCBI (accession number J01415.2). The resulting reads were then screened against a set of custom-compiled nuclear mitochondrial genes. These mitochondrial genomes were assembled in SeqMan 8.0 (DNASTAR. Manual inspection and correction were then carried out for the assembled genomes.
Methods of divergence time estimation

The time estimation of divergence events within human lineages were done in the software BEAST [50] v. 1.4.8 with a Bayesian “relaxed molecular clock” model [65]. Parameters used are provided in Table 2-4. One internal calibration point, between *Homo sapiens sapiens* and *Homo sapiens neanderthalensis*, was used based on Green *et al* 2008 results [127]. In addition, six leaf node calibration points for Neanderthal remains were used based on Briggs *et al* 2009 carbon dating results[166]. BEAST output was then analyzed in Tracer v. 1.4.1 package (http://beastbioedacuk/Tracer). Maximum clade credibility trees were calculated in TreeAnnotator v. 1.4.8 package based on the same BEAST output.

Project summary and discussion

Based on the above results, these seven individuals from Kalahari are from the oldest known haplogroup L0. The expected high within-population divergence has also been observed. The two groups led by Professor Schuster and Hayes continued to use massive whole genome shotgun sequencing on KB1 and ABT samples to obtain enough coverage on the nuclear genomes. As mentioned in the published paper [125], high within-population divergence has also been observed in their nuclear genomes. The number of novel SNPs observed in these two nuclear genomes far exceeded the numbers provided by any previously published human nuclear genomes.

However, the higher number of substitutions observed in these seven mitochondrial genomes has also created problems in determining their haplogroup assignment using traditional
tools. Current haplogroup classification systems were based on previous known sequences. The set of inferred haplogroup-determining SNPs was modeled in a deterministic fashion. It does not allow novel mutations different from known patterns to be considered. For situations similar to this study, the only solution currently available is to reconstruct the global phylogenetic tree with sufficient sample representation. If the studied samples were all from a similar geological or phylogenetic location, like in this study where all sequences are from Southern African and presumably all from the L0 subgroup, sufficient number of known-haplogroup sequences were only required for local clades. However, if the composition of the studied samples were diverse, sufficient sampling would be required for the entire human mitochondrial phylogeny. This means reconstructing the entire global mitochondrial tree as in [158, 164]. It will be time consuming and computationally intensive. A new system that could tolerate unknown SNPs is needed for rapid and accurate haplogroup classification.

(End of Chapter 2)
Chapter 3

Probabilistic modeling of major human mitochondrial haplogroups and parameterization based on population

Project background

As mentioned in Chapter 2, current tools failed to recognize the newly sequenced mitochondrial genomes from seven Southern African individuals in the known haplogroup system. In another unpublished study, these tools also failed to properly assign several non-African individuals’ mitochondrial genomes to known haplogroups. Clearly there are improvements needed for human mitochondrial haplogroup identification. Another fact I have observed, intriguing at that time, was that the estimated divergence time (Table 2-2) within the Ursidae clade became younger in every new published study. The newer studies used more intraspecific samples to study the Ursidae evolution at a finer scale. The increasingly younger divergence time estimation was thus likely due to the fact that the parameters used in the evolution models changed when the percentage of intraspecific samples was increased. This echoed what Ho et al had pointed out in 2005 [168]. The currently known population variation parameters for human evolution models were also largely derived from interspecific comparisons. Since there are now a large number of publicly available human mitochondrial genome sequences, it is thus possible and desirable to re-evaluate the parameters for human mitochondrial genome variations.
Introduction

Human mitochondrial haplogroup information has been used for studies of human matrilineal genealogy and population evolution since early 1990s [169-187], especially in tracing human migration events (see the 2007 review by Underhill and Kivisild [188]). The term haplogroup (haplotype group) was first used by Torroni et al in 1993 in their study of the genetic composition of Native American populations using mitochondrial genome information [169]. Four mitochondrial haplogroups were defined by that study and named as haplogroup A, B, C and D, each representing a group of individuals clustered together based on their DNA sequence similarities. Various researchers later introduced other haplogroups based on their own studies; however, these studies were all drawn from selected subpopulation(s). Therefore, they do not represent the global subpopulation distribution. Richards et al in 1998 introduced a cladistic naming scheme for haplogroups [170]. Ruiz-Pesini et al in 2007 unified the classification and nomenclature of haplogroups described by different research groups by constructing a global phylogenetic tree based on thousands of complete mitochondrial DNA sequences [158]. This tree, with revision made by von Oven and Kayser in 2009 [159], still serves as the human mitochondrial haplogroup reference to date [189].

However, the methods used for constructing the 2007 mitochondrial phylogenetic reference tree did not fully consider the mitochondrial genome variation parameters in human populations including the percentage of invariant mitochondrial sites and substitution rate variation among sites [158, 159]. The substitution rate variations between different nucleotides and among sites are well known in human mitochondrial genome [15, 35, 130, 132, 190-192]. Models without these parameters could give incorrectly estimated number of substitutions and thus genetic distances [43, 46, 193]. More accurate estimates can be achieved by employing the
GTR + Gamma + Invariant Site Removal model [43, 46], which was developed to correct for such variations and has been widely accepted and used in many recent phylogenetics studies [51, 65, 125, 137-139, 141]. However, the correctness of this model depends on the accuracy of the aforementioned genome variation parameters estimated from human population.

Parameters reflecting different aspects of genome variation include the percentage of invariant sites in the genome, the substitution rate heterogeneity between different nucleotides and among different sites along the genome, and the average substitution rate for the entire genome. Human mitochondrial genome variations have been widely used for studies in human evolution [15, 35, 194-196], migration [169, 188, 197-200], population genetics [170, 190, 201, 202], heritable diseases [203-207] (large amounts of literature were found in each aforementioned category, the references provided are representative examples). However, parameters of these variations for the human population have not been estimated or only estimated for local regions (for example, hyper-variable region I and II in D-loop/control region). These parameters are important in interpreting and understanding the data in the aforementioned studies, for example, in calculating genetic distances, identifying genetic subpopulations and inferring past evolution and migration events.

These parameters have been previously estimated from interspecific data; however, it is now known that differences exist between estimations from phylogenetic and population (intraspecific) data [130, 131, 208]. For example, the evolutionary rate is normally much higher when sampling within a population than sampling between species. Most of the polymorphic sites generated at this higher rate are subject to selection and genetic drift and subsequently lost from the population, and this loss likely have caused the observed differences between intraspecific and interspecific rates [168, 209]. However, it is also possible, but rarely observed in real data,
that ancestral polymorphic sites could be inherited by descendent species and the observed fixed substitutions between the descendent species would be higher than otherwise expected without considering ancestral polymorphic sites. The time-dependency of evolutionary rate was reported first by Ho et al in 2005 [168]. Following that study, Penny in 2005 called for reconsideration of the techniques in studying population genetics and phylogenetics [209].

As a result, it is necessary to estimate these genome-wide variation parameters in the context of human population. For example, the interspecific substitution rate heterogeneity among sites has been estimated before [42, 60], but not from intraspecific or population data. In addition, the intraspecific overall substitution rate has been estimated from pedigree data for the control region (CR, the most polymorphic region in the mitochondrial genome that harbors the replication origins [210]) [130, 131], but no reports have yet extended this to the whole genome. The percentage of invariant sites could greatly influence the calculation of genetic distance, as demonstrated by Gu and Li in 1996 and Waddell and Steel in 1997 [43, 46]. The inclusion of these parameters could also have an impact on the estimation of random match probability from DNA sequences collected in forensics, as they reflect different numbers of variable sites and different degrees of per-site variation.

Based on different mitochondrial genome variation patterns, the human population could be categorized into subpopulations. Correctly identified subpopulation structures are of great importance in studies of human migration, population and in analysis of disease-associated SNPs data. For example, only by sampling from representatives of clearly identified subpopulations, SNPs could be linked with disease instead of population stratification. The most widely used current human subpopulation classification system is the aforementioned haplogroup system.
In addition, the Ruiz-Pesini et al. (2007) and van Oven and Kayser (2009) studies defined each haplogroup in a deterministic fashion using the consensus substitutions in each group. As a result, each haplogroup could be defined by the consensus nucleotides at a few key or signature sites. However, this approach can only assign a new sequence to an existing haplogroup if there are no mutations at these key sites and the sequence does not contain partial matches to patterns of key site nucleotides for two or more haplogroups. When these conditions are violated, a new mitochondrial DNA sequence could only be accurately assigned to a haplogroup by rebuilding an entire global phylogenetic tree.

Therefore, a probabilistic model describing the features of a haplogroup, instead of a deterministic motif, is thus needed for quick and accurate haplogroup identification. Position specific score matrices (PSSMs) and profile hidden Markov models (profile HMMs) were widely used to describe the sequence features of gene and protein families, both of which are essentially sequence clusters [211-213]. Such probabilistic tools could be used to describe the sequence features of haplogroups. However, one distinction between the current human haplogroup structures and gene clusters needs to be point out: because of human migration and evolution history, different haplogroups are not mutually exclusive clusters. For example, all non-African haplogroups are part of one single large African haplogroup L3. This feature dictates that a direct application of common gene clustering methods would not be sufficient or appropriate for inferring the current haplogroup systems.
Analyses and Results

Retrieval of mitochondrial genome sequence data and initial analysis

A total of 7985 full-length quality-screened mitochondrial genome sequences were retrieved from the NCBI database (http://www.ncbi.nlm.nih.gov/nuccore). Of these, 7975 were from modern humans; 7 were from Neanderthal [127, 166] and 3 from Denisovan [214]. Among the modern human genomes, 3711 (46%) have known location (country) information and 2893 (36%) have previously assigned haplogroup information. 2204 (28%) of them have both types of information and 3585 (45%) of them have neither type of information. These 3711 sequences with known location information included individuals from 86 countries. Among these countries, Japan has the largest number of samples (948), followed by the Russian Federation (558), the United States (297), Italy (227) and Finland (206). The geospatial distribution of the number of samples from these countries is shown in Figure 3-1.

Figure 3-1. Geospatial distribution of number of mitochondrial genome sequences from countries
Estimation of mitochondrial genome variation parameters from human population data

**Percentage of invariant sites**

It is expected that the percentage of variable sites will increase as the number of sampled genomes increases. We observed that the number of invariant size decreases in a logarithmic fashion through sampling from the 7975 modern human sequences (Figure 3-2). The total number of variable sites when all 7975 genomes were sampled is 7901 (35.61%). By including the Neanderthal and Denisovan genomes, this number is 7960 (35.97%), only increasing by 59 (0.36% of the genome) nucleotides. The percentage of invariant sites in the mitochondrial genome was estimated to be about 64% for modern human population.

Figure 3-2. Number of invariant sites decreases as more genomes are sampled.
Substitution rate and rate heterogeneity

It is widely known that the substitution rate is not distributed uniformly among sites [39, 40, 47]. This rate variation can be modeled by a Gamma distribution, and thus reflected by the Gamma distribution shape parameter $\alpha$ [35, 39, 41]. It is necessary to estimate $\alpha$ for the human mitochondrial genome based on these 7975 modern human mitochondrial sequences. However, the large sample size of this data set posed a computational challenge for the estimation procedure. Yang in 1993 has previously developed a Maximum-Likelihood (ML) estimation method for inferring $\alpha$ [41]. However, this method could only handle a small number of sequences with limited length [41]. A couple of simplified or approximation methods have been proposed later [45, 215] to reduce the computational cost, but these methods still could not handle large datasets with more than a dozen sequences of moderate length. A similar Maximum Parsimony (MP) method has also been used [35, 216] with much improved computational efficiency. However, because the MP method still requires the inference of ancestral sequences based on a prior knowledge of the phylogeny, and subsequent comparison along the branches, it still could not handle 8000 full-length mitochondrial sequences. Pesole and Saccone in 2001 have proposed an $\alpha$ parameter estimation method based on pairwise sequence comparison, which does not depend on known phylogenies [217]. However, this tree-free method does not use a phylogeny as the reference of the evolutionary process, and thus treat all evolutionary routes between any two sequences the same. As a result, this method may not correctly estimate the actual number of substitutions at a site when compared with the ML or MP methods.

The method of Pesole and Saccone, however, is the only method capable of handling the data size of this study. Additionally, when population data are considered, tree-based methods may be inappropriate because individuals could be sampled from different generations and treated
as leaf nodes from the same generation when the phylogeny is constructed. The influence of cross-generational sampling may be trivial when only a few individuals are used and one is interested in understanding at the level of species. However, in the case of the human population, the effect of inter-generational sampling should be considered. It is thus desirable to investigate the possibility of correcting the underestimation of this pairwise method in comparison to MP or ML method.

In order to make sure the validity of phylogenetic reconstruction, one random sample from each country is selected, based on the assumption that the generation differences between different current people of different countries are of little impact on the evolutionary distances between them. Sample sizes from 10, 20, … to 80 were used, since the total number of countries in this dataset is smaller than 90. For each sample size, 30 random datasets were selected. Then the differences between the above methods were compared. A program using method similar to Tamura and Nei 1993 method [35] was developed. This tree construction method (tree method) is based on the MP method but also minimizes local likelihood of internal nodes. Another program (pairwise method) using a method similar to the Pesole and Saccone 2001 method [217] was written to reduce underestimation of substitution (Figure 3-3A).

For a dataset of N sequences, the phylogram representation of these N sequences has N leaf nodes and (N - 1) internal nodes. The tree method estimates through 2(N - 1) sequences comparisons. The pairwise method estimates through (N² - N)/2 number of comparisons. If the probability of observing a substitution in tree method is a, and in pairwise method is b, the substitution rate (or count, since genome size and sample size are the same) by tree method \( R_T \) and by pairwise method \( R_P \) are given as:
\[ R_T = 2(N - 1)a \]
\[ R_p = \frac{N^2 - N}{2} b \]
\[ R_T = \left(\frac{a}{b}\right) \frac{4N - 4}{N^2 - N} R_p \]
\[ R_T = \left(\frac{a}{b}\right) R'_p, R'_p = \frac{4N - 4}{N^2 - N} R_p \]

By applying a correction factor of \((4N - 4)/(N^2 - N)\) to the number of substitutions estimated by the pairwise method, \(R_T\) and \(R_p\) are linearly related. Our data supported this hypothesis (Figure 3-3B), showing a clearly (Coefficient of determination \(R^2 = 0.998\)) linear relationship between estimations by tree method \(R_T\) and pairwise method \(R'_p\). This relationship means that it is possible to correctly estimate substitution rates \(R_T\) in a pairwise fashion without a tree.

However, this has only shown the overall substitution rate along the entire genome, but not the rate variation among sites. The Gamma distribution shape parameter \(\alpha\) reflects such variations (see the Methods section for more information). Since GTR + Gamma + Invariant Site Removal model estimates first only the variable sites, only the non-zero substitution rate sites were considered in estimating \(\alpha\). The Maximum-Likelihood estimator of \(\alpha\) was estimated by optimizing the likelihood function following Uzzell and Corbin [39] with Broyden–Fletcher–Goldfarb–Shanno (BFGS) method. Since sample size less than 30 would give inconsistent and sometimes hugely variable \(\alpha\) estimation, only sample sizes larger than 30 were used. We then studied the relationship between shape parameter \(\alpha_T\) estimated by the tree method and \(\alpha_p\) by the pairwise method (correction factor \((4N - 4)/(N^2 - N)\) applied). Our data (Figure 3-3C) showed that \(\alpha_T\) has a strongly supported \((R^2 = 0.91)\) exponential increase relationship with \(\alpha_p\). This means that...
the substitution rate variation of the tree method could also be estimated from that of the pairwise method. $\alpha_T$ thus roughly equals one half of the value of $e$ raised to the power of $\alpha_P$. The Gamma shape parameter $\alpha$ is estimated as 0.3597 using the pairwise method for all variable sites. By applying the correction for the tree method, the $\alpha$ parameter is 0.7078 (See Figure 3-7 for example).

The substitution rate of the mitochondrial control region has been previously estimated using pedigree data. Howell et al. in 1996 [218] estimated this rate to be 0.025 substitutions per generation, and Parsons et al. in 1997 [130] estimated this rate to be 0.03. In contrast, the estimate of Sigurgardottir et al. in 2000 [131] was only 0.0043, and the latest estimate by Howell et al. in 2003 [208] was 0.01. These estimates varied, but they are only estimates based on the control region. With a total of 7975 sequences, it is possible to estimate the rate for the entire mitochondrial genome by comparing the number of substitutions between control region and full mitochondrial genome. The estimated number of substitutions based on the tree method and the pairwise method both gave clearly linear correlation between number of substitutions in the control region and that of the full genome (Figure 3-3D). In summary, there would be in total 3 substitutions for the mitochondrial genome if only one is observed in the control region. This gave an upper estimated rate for the full genome at about 0.1 substitutions per generation with the estimate of Parsons et al. The lower bound for the estimate is about 0.013 substitutions per generation. These numbers mean that only one mitochondrial genome substitution is expected between two individuals about 10 to 77 generations apart.
Figure 3-3. Estimating substitution rate and variation parameters.
The strand-specific human mitochondrial genome-wide substitution rate variation among sites and between nucleotides

The pairwise method allows the estimation of the substitution rate variation among sites and between different nucleotides from 7975 modern human mitochondrial genome sequences. The strand-specific substitution rate variation between different nucleotides is summarized in Table 3-1. The base composition of the analyzed strand (heavy-strand) of the human mitochondrion (rCRS) is A% = 31%, C% = 32%, G% = 13%, and T% = 25%. Given the tree height is fixed evolutionarily or biologically, the substitution rate of a single site is represented by the estimated number of substitutions at that site. The substitution rate variation among different sites is summarized in Figure 3-4.

Table 3-1: Positive-strand substitution probability between nucleotides in human mitochondrion

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>X</td>
<td>0.19</td>
<td>0.65</td>
<td>0.16</td>
</tr>
<tr>
<td>C</td>
<td>0.17</td>
<td>X</td>
<td>0.11</td>
<td>0.72</td>
</tr>
<tr>
<td>G</td>
<td>0.75</td>
<td>0.14</td>
<td>X</td>
<td>0.11</td>
</tr>
<tr>
<td>T</td>
<td>0.15</td>
<td>0.76</td>
<td>0.09</td>
<td>X</td>
</tr>
</tbody>
</table>

Figure 3-4. Mitochondrial site-specific substitutions estimated from 7975 modern human. The bottom panel shows the annotation of the human mitochondrial genome. Window size = 20 bp.
Variable regions outside D-loop and distribution of disease-related SNPs

The mitochondrial hyper-variable regions have long been used as a marker for measurement of human genetic diversity. However, we found that, besides the two hyper-variable regions, there are additional “mutation hotspots” in the mitochondrial genome from the analysis of the 7975 modern human mitochondrial genome sequences (Figure 3-5). Some of them are between genes and some are located inside protein coding genes. We further compared this distribution of within-population variation with the distribution of known disease-related SNPs [158], and we found that the disease-related SNPs accumulated on the plot in conserved regions.
Figure 3-5. Distribution of estimated substitutions and reported diseases-related SNPs in human mitochondrial genome.

From top to bottom: the first panel shows SNPs that were reported to be associated with diseases by at least two published reports from two independent research groups; the second panel shows SNPs that were reported to be associated with diseases by at least one published report; the third panel shows the number of substitutions estimated for each site based on the 7975 modern human mitochondrial genomes; the fourth panel shows the annotation of the human mitochondrial genome (disease associated SNPs data were obtained from MITOMAP database (www.mitomap.org) and then processed). Window size = 20 bp.

Identifying major haplogroups by clustering based on estimated parameters

Based on the estimated genome variation parameters and GTR + Gamma + Invariant Site Removal model, the genetic distance between each pair of the 7985 sequences were calculated. A Neighbor-Joining (NJ) tree and ultra-metric tree were built upon the resulting distance matrix. The NJ tree resulted in a haplogroup relationship that is in better congruence with current haplogroup system. Hierarchical clustering method based on the NJ tree was used to classify haplogroups. Only haplogroups with more than 30 sequence members were retained in order to ensure the validity of statistical properties. Most of the currently known large haplogroups were identified in this method, with a few modifications (Table 3-2). For example, haplogroup E is not part of haplogroup D. Further deep clustering could not satisfactorily separate E from D. In addition, several new haplogroups that have not been reported before were formed through this clustering process.

The relationship between these identified haplogroups were inferred as a dendrogram in Figure 3-6. The newly identified haplogroups were in gray color. Based on the composition of haplogroup members’ country of origin, one possible migration pattern was drawn. As haplogroups become smaller in scale, the intra-haplogroup genetic distance also decreases. For a
haplogroup, an average intra-cluster genetic distance of 10 means that all the members of this haplogroup on average shared a common female ancestor between 50 and 384 generations ago.

Table 3-2: Statistics of assembled human mitochondrial genomes

<table>
<thead>
<tr>
<th>Haplogroups</th>
<th>C</th>
<th>P</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>L0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L3 (L3-X)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td>Part of D</td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td></td>
<td>N &lt; 30</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td>Part of F</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td>Part of H</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HV-X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td></td>
<td></td>
<td>Part of W or X</td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
<td>Part of W</td>
</tr>
</tbody>
</table>

Note: C = Current result, P = Previous result
Figure 3-6. Haplogroup clusters inferred from genetic distance based on Neighbor-Joining guide tree.

(Left) Relation according to Neighbor-joining based hierarchical clustering (phylogeny A). Numbers inside the box represent the average intra-cluster genetic distance. The newly identified haplogroups were in gray color. (Upper right corner) Possible migration/expansion path according to non-major-immigrant country composition of haplogroup clusters

**Not including substitution rate variation among sites led to different phylogenies from the same dataset**

To assess the effect of considering substitution rate variation among sites and the presence of invariant sites in phylogenetic reconstruction and subsequent haplogroup inference, the same set of data was reanalyzed using the same method as above except that the rate variation
among sites and the presence of invariant sites was not included this time. The resulting phylogeny (phylogeny B) in Figure 3-7 showed obvious differences from the one estimated including the rate variation among sites and the presence of invariant sites (phylogeny A). The position of haplogroup I, J, T and F were different in phylogeny B from phylogeny A. A new haplogroup, R8, was identified in phylogeny B but not in A. The sequence composition of haplogroup D is different between these two phylogenies. This methodological difference contributed to on average 16.08% differences between haplogroups in phylogeny A and B, or 1.33% differences when the location of the haplogroups was not considered. This result is consistent with previous studies, which suggested that failing to consider rate variation could lead to incorrectly estimated phylogeny [43, 46, 219-222].

Figure 3-7. Haplogroup clusters inferred based on the model that does not consider rate variation
among sites and the presence of invariant sites.

(Left) Relation according to Neighbor-joining based hierarchical clustering (phylogeny B). Numbers inside the box represent the average intra-cluster genetic distance. (Upper right corner) Possible migration/expansion path according to non-major-immigrant country composition of haplogroup clusters. Haplogroups labeled in red are the ones that have different positions between phylogeny B and A. Haplogroups labeled in cyan are the ones that have different sequence composition between phylogeny B and A.

**Probabilistic haplogroup model and Maximum-Likelihood profile search**

The nucleotide distribution probability of each site was estimated based on the sequences in the given haplogroup ($p(a)$ for having nucleotide $a$ at this site). Sites were defined by the revised Cambridge reference sequence (rCRS) [167]. For each haplogroup, a probabilistic profile containing such information was built. The likelihood of a sequence generated from a given profile is given as a position-specific matrix (PSM):

$$L = \prod p_i(a)$$  \hspace{1cm} (29),

where $p_i(a)$ is the probability of having nucleotide $a$ at the $i$th site. This is essentially a profile HMM with only matched states [20]. To avoid underflow problem, log-likelihood was used in actual calculation. The haplogroup with the profile that has the maximal likelihood was selected as the haplogroup given rise to the sequence. To assess the significance of the match, a significant score was calculated by dividing each profile likelihood value with the average likelihood of all profiles

$$S = \frac{L_k}{L_{\text{average}}}$$  \hspace{1cm} (30),

where $L_k$ is the likelihood of the sequence coming from the $k$th haplogroup. Through simulation, this method has a correct assignment rate of 98% (98% of sequences could be correctly assigned to a haplogroup). A web server was developed based on this method to offer easy access to
High accuracy in haplogroup identification supported using simulated new sequences modeled after human mitochondrial genome-wise substitution patterns

In order to assess the accuracy of the ML profile search method, as well as the haplogroup modeling, implemented in the HapSearch system, a simulated search test was performed. First, new sequences were simulated by introducing random substitutions based on the human mitochondrial genome-wide substitution pattern, namely the variation between different nucleotides and among different sites. Obviously the number of newly introduced substitutions will have an impact on the haplogroup assignment of the newly generated sequence. In this simulation, the range for the number of introduced substitutions was from 1 to 20. For a given number of substitutions, 150 sequences were sampled from the real dataset and the given number of substitutions was introduced to generate new sequences. 30 replicates were done for each given number of substitutions. The haplogroup assignment for each newly generated sequences were compared with the original sequence’s haplogroup designation. The percentage of matched haplogroup assignment as a function of the given number of substitutions is shown in Figure 3-8. The HapSearch system is 98% accurate with only one or two substitutions introduced, and could tolerate up to 10 substitutions while still maintain 96% accuracy.
Figure 3-8. Percentage of new sequences matched with original haplogroup classification with introduced mutations following known mitochondrial genome-wide substitution patterns.

**Materials and methods**

**Data collection**

Complete mitochondrial genome DNA sequences of *Homo sapiens* were collected from NCBI Nucleotide database (http://www.ncbi.nlm.nih.gov/nuccore). A total of 9064 sequences were downloaded in July 2011. These were fully assembled full-length genome sequences without any attached quality information. To ensure sequence quality, sequences with length variations larger than 20 bp from the revised Cambridge reference sequence (rCRS) [167] were discarded, as length variations larger than this amount are rare (Figure 3-9). Full-length genome sequences shorter than this amount are normally product of incomplete assembly, while genome sequences longer than this amount are more likely due to incorrect assembly. We also discarded sequences with more than three undetermined nucleotides, as this is a sign of insufficient sequencing coverage. A total of 7985 sequences were retained after this screening.
Methods for genetic distance estimation in the presence of substitution rate variation between nucleotides and among sites

*General time-reversible (GTR) model for substitution rate variations between nucleotides*

The stationary base frequencies, also called base composition, are represented as \( q_A, q_C, q_G \) and \( q_T \). The substitution rates between bases are represented as \( r_{ij}, (i \text{ and } j \in \{A, C, G, T\}) \).

Under the GTR model first proposed by Lanave *et al* in 1984 [36], given a rate matrix \( R \), at the time point \( T \) since the two sequence diverged, the transition probability matrix \( P(T) \) at time \( T \) is given as

\[
P(T) = e^{RT}
\]  

where the rate matrix \( R \) is given as
This study used the formula provided by Rodriguez et al in 1990 [38] for calculating genetic distance $d$ with GTR model:

$$d = -tr[X \log(X^{-1}F)]$$  \hfill (33),

The forms of $X$ and $F$ are given as

$$X = \begin{bmatrix} q_A & 0 & 0 & 0 \\ 0 & q_T & 0 & 0 \\ 0 & 0 & q_C & 0 \\ 0 & 0 & 0 & q_G \end{bmatrix}$$  \hfill (34),

$$F = \begin{bmatrix} x_{AA} & x_{AT} & x_{AC} & x_{AG} \\ x_{TA} & x_{TT} & x_{TC} & x_{TG} \\ x_{CA} & x_{CT} & x_{CC} & x_{CG} \\ x_{GA} & x_{GT} & x_{GC} & x_{GG} \end{bmatrix}$$  \hfill (35),

where $x_{ij}$ represents the average probability at a given site of a pair-wise alignment that in the first sequence the nucleotide is $i$ and in the second is $j$, $(i, j \in \{A, C, G, T\})$.

**Correction for substitution rate variations among sites with gamma distribution**

Following Uzzell and Corbin [39], the rate variation among sites was modeled using the gamma distribution, whose probability density function is:

$$g(x; \alpha, \beta) = \frac{\beta^\alpha}{(\alpha - 1)!} x^{\alpha - 1} e^{-\beta x}$$  \hfill (36),
where $\alpha$ and $\beta$ are the shape parameters of the gamma distribution, with $\alpha \beta$ as mean. The method used here equals a model that applies a correction factor $g$ from a distribution at each site of the sequence to the mean substitution rate $r$. As a result, $\beta = 1/\alpha$.

The genetic distance calculation formula considering rate variation used in this study was from Waddell and Steel in 1997 [46]:

$$d = -\text{tr}[X \cdot M^{-1}(X^{-1}F)]$$  \hspace{1cm} (37),

where $M^{-1}$ is the inverse function of the moment generating function $M$ of the selected distribution. In the above case of using a gamma distribution:

$$M^{-1}(x) = \alpha(1 - \frac{1}{x^\alpha})$$  \hspace{1cm} (38).

**Correction for invariant sites with invariant sites removal (ISR)**

After estimating the percentage of invariant sites $p_{inv}$, the $F$ matrix was adjusted to $F_{var}$ following the ISR model provided by Waddell and Steel in 1997 [46]:

$$F_{var} = \frac{F - p_{inv}X_{inv}}{1 - p_{inv}}$$  \hspace{1cm} (39),

where $X_{inv}$ is the base frequency of the invariant sites.

**Parameter estimation**

All parameters were estimated under the Maximum-Likelihood framework using R. The sampling was done using the 7985 mitochondrial genome sequences with replacement. All estimations were based on data of sample size larger than 30 to ensure valid statistical property.
The tree method was written in java according to Tamura and Nei 1993 method [35]. The pairwise method was written based on Pesole and Saccone 2001 method [217]. However, the actual distance calculation method used here does not normalize against the overall genetic distance as in formula (1) of their 2001 paper. The philosophies of these two methods are illustrated in Figure 3-3A.

**Software implementation**

The above models were implemented in Java language. The matrix operation was based on JAMA package from NIST (http://math.nist.gov/javanumerics/jama/). The web service implementing this method was developed using Java Servlet technologies.

**Summary and discussion**

**Newly estimated population variation parameters and the implications**

This study estimated the percentage of invariant sites, genome-wide substitution rate and rate heterogeneity among sites for the human mitochondrial genome based on human population data. The estimated percentage of invariant sites (0.64) is larger than that estimated by Waddel and Steel in 1997 (0.59) [46], as they included non-human primates from data by Horai et al [223]. The variable portion of the human mitochondrial genome should not include interspecific differences, as they are different in essence and origin and normal humans probably could not tolerate such interspecific changes.
Two methods were used to estimate the substitution rate heterogeneity, which is reflected by the Gamma distribution shape parameter $\alpha$. The relation between these two methods (pairwise and tree methods) was analyzed. The tree method could only operate on smaller data sets with the number of sequences far less than the nearly eight thousand sequences analyzed in this study. While the pairwise method allows operation on data set even much larger than this number, it does not consider ancestral substitution events as in the tree method. Here in this study it was found that the result of the tree method could be derived from pairwise method by applying a relative simple correction function. A strongly supported logarithmic correlation was found between the results of these two methods, which allows accurate analyses of extremely large data sets with consideration of ancestral substitution events.

The pairwise method gave $\alpha = 0.36$, which was converted to the value 0.71 with the consideration for ancestral substitutions. The rate heterogeneity was calculated only on the variable portion of the mitochondrial genome, which was different from previous studies [35, 46]. This is because of the substitution model (GTR + Gamma + Invariant Site Removal) applies the Gamma distribution correction only on the variable portion of the genome. In addition, this reflects the notion that a subset of mitochondrial genome sites is allowed to mutate within the human population without dramatically altering the fitness or viability of the carrier. Including the invariable portion of the genome would significantly decrease the $\alpha$ estimation.

This study also found a very strong relationship between the number of substitutions within the D-loop and that of the whole genome. This allows the estimation of the genome-wide substitution rate based on previously published D-loop rate from pedigree data. However, the current results from pedigree data varied widely from study to study, as much as seven fold (between Parsons et al 1997 (0.03 substitution per generation/transmission event) [130] and
Sigurgardottir et al 2000 (0.0043) [131]). This wide difference may result from the different pedigrees they have chosen. Parsons et al have included people from a broader genetic background than Sigurgardottir et al, which only included families from Iceland. The population history between these two data sets might also contribute to this discrepancy. The latter population is on an island close to the Arctic Circle and created by migration and settlement events estimated no earlier than the 9th century. Other estimations by Howell et al were larger than this number [208, 218]. There is currently no consensus on this rate estimation, and the true value is likely to be between 0.03 and 0.01. This suggests that people within ten to thirty matrilineal generations apart should have the same mitochondrial genome. This also suggests that mitochondrial genome in full-length alone could not be used to fully index the general population without collision.

The newly estimated population parameters allow correct genetic distance calculation using the GTR + Gamma + Invariant site removal model. Based on the estimated parameters, the genetic distance between any two of the 7985 individuals was calculated. The average distance between Denisovan and modern human is 448 substitutions. The average distance between Denisovan and Neanderthal is 465 substitutions, similar but slightly larger than that between Denisovan and modern human. The average distance between Neanderthal and modern human is about half of these two numbers, 226 substitutions. Denisovan people were thus further apart from human than Neanderthal people. The average distance between the L0 haplogroup and the remaining human samples (Non-L0) is about half of the distance between all modern human and Neanderthal, 102 substitutions. The Non-L0 ground is slightly closer in genetic distance from Neanderthal (226 substitutions), compared with L0 from Neanderthal (229 substitutions). The L1 and Non-L1 haplogroups are on average 95 substitutions apart, and the L2 and L3 haplogroups are on average 63 substitutions apart. The average distance between two major non-African
haplogroups, M and N, is only 46 substitutions apart. This number means that the M and N haplogroups on average are only about 460 to 1380 generations apart. If we assume the generation time of early human population is determined by the time of reaching puberty (15), the most recent estimate of the split between M and N is less than 7000 years old.

**Limitations in haplogroup resolution and modeling**

Same as all the results from a hierarchical clustering method, the average genetic distance within haplogroup decreases as a haplogroup is defined on a finer scale. If the sampling within the haplogroup in this dataset was random, some haplogroups were only formed less than 100 to 300 generations ago, for example, V and K2. However, it is quite possible that the sampling is not random, as pedigree data were preferentially collected by researchers studying inheritable diseases. Nevertheless, some previously defined haplogroups are formed at a very recent time, probably correlated with recent migration and settlement events. This also raises the question, what is the minimal distance for defining a haplogroup? As the haplogroup goes finer, the boundaries between these haplogroups become unstable as new data are introduced. The ratio of the within-group average distance over the between-group average distance could be used as a measurement, similar with the SDR concept used in HIV virus subtyping by Rambaut *et al* 2001 [224]. An expectation maximization method could be used for the implementation of this concept. However, the challenge lies within the biological interpretation of such defined thresholds.

Finally, this study constructed a probabilistic model for haplogroup typing. As large structural variation of human mitochondrial genome is rare, this study opted for a mechanism akin to position-specific matrix (PSM) rather than a full-fledged hidden Markov Model (HMM).
Each site was defined by a reference genome, which is the widely used human mitochondrial reference rCRS. A beneficial side effect of this choice is the complexity of the search is minimal. As a result, the speed is much higher and computational cost is much lower than HMM search algorithms like Viterbi algorithm at the length of about three times of 17000 states. However, the simple model used here indeed treats insertions and deletions, although rare, as special substitution events. To minimize the impact, special heuristics have been built into the system to model known insertions and deletions for each haplogroup PSM. The indels from new sequences are compared with known indels in the PSM first. If there is no exact match, the new indel is inferred by introducing mutation events to its best LCS counterpart in the known indels. The last resort is to treat the new indel as a new rare event with the probability of one in the known data size.

The effect of differences between intraspecific and interspecific data on rate and divergence time estimation

It is estimated that the number of nuclear genome SNPs between two individual human on average is about 3 million [125], which is 0.1% of the genome. Currently, there are 7 million RefSNPs in NCBI dbSNP [225], which is 0.23% of the genome. These 0.23% differences are presumably polymorphic sites between individuals of the same species, and do not represent interspecific differences. As an example, the observed percentage of differences in nuclear genome between human and chimpanzee is 1.23% [226]. These 1.23% differences include not only the fixed substitutions between human and chimpanzee, but also the polymorphic sites in these two genomes. The percentage of fixed differences could only be less than this observed 1.23% [226].
As introduced at the beginning of Chapter 2, the evolutionary rate of human mitochondrial genome was estimated from human pedigree data in studies [130, 131]. Those studies also found that the rate estimated from pedigree/genealogy data is much higher than that from phylogenetic data. If assuming the same 0.23% of polymorphic sites in human mitochondrial genome, the total number of possible polymorphic sites is 37. However, this is only close to the average number of differences between two European individuals. More than 100 differences between one African and one European individual have been observed [125], which is about 0.625% of the mitochondrial genome. Actually, the number of variable sites directly estimated is over 6000. Although some of these differences were possibly originated from sequencing errors [227], the actual number of variable sites is obviously significantly more than the estimated 37. In comparison, the total number of differences between human mitochondrial genome (NCBI accession number: NC_012920) and chimpanzee mitochondrial genome (NCBI accession number: NC_001643) is only 1473. This number also includes the polymorphic sites in these two mitochondrial genomes.

It is obvious that the evolutionary rate is much higher when sampling within a population than between species. Most of the polymorphic sites generated at this higher rate are subject to selection and genetic drift and many of them may subsequently be removed from the population [168, 209]. The rate estimated from fixed substitutions between species is thus lower. This could be illustrated by the following hypothetical (thought) experiment. We could first consider a study of $K$ number of species. If we could only sample one individual in each of the $K$ number of species, there is only $K$ number of sequences in this study. The overall differences ($N$) between two sequences include both fixed substitutions ($N_f$) between species and polymorphic sites ($N_p$) within species. And since there is only one sequence for each species, the overall difference ($N$) is also the evolutionary distance by default. However, if $x$ number of individuals were sampled for
each species, the overall differences \((N)\) between any two individuals of the same species would only come from \(N_p\). The evolutionary rate from within the population is a higher rate from \(N_p\) as we have just discussed. The distance \((N)\) between two species is \(N_f + \lambda x N_p\), where \(\lambda\) is a number between 0 and 1 to reflect the possibility that there might be shared polymorphic sites between individuals of the same species. As \(x\) becomes larger, the distance between species will get further. In practice, this means a longer divergence time between speciation events if an internal node calibration is used, or a shorter divergence time if a leaf node calibration is used. Even if using a rate sampling method as introduced by Drummond et al [65], this will still have some effect on the estimation of rate and time.

This effect is also demonstrated by my study in Chapter 2. From table 2-2, it is clear that the estimated divergence time decreased from Yu et al 2007 study, to Krause et al 2008 study, to Bon et al 2008 study and to my study. Yu et al 2007 study only included 8 sequences, Krause et al 2008 study included 10 sequences and Bon et al 2008 study included 13 sequences. And these three studies all included fossil calibration points. Considering that modern polar bear species diverged from one lineage of brown bears, and also the fact that there are well-known fertile hybrids between polar bear and brown bear, the study of polar bear evolution is more similar to population studies rather than phylogenetic studies. This also means that deep fossil calibration is inappropriate in this case and previous studies have overestimated the divergence times along this lineage.
The effect of differences between intraspecific and interspecific data on the pan-genome concept and disease SNP discovery

The genome of a given species should thus be considered as a core sequence plus a collection of polymorphic sites. This concept is similar to the pan-genome concept first introduced by Tettelin et al in 2005 when studying the genomic diversity of multiple isolates of the pathogen *Streptococcus agalactiae*. In that study, the authors proposed that the pan-genome of a given species should include a core genome that is shared by all isolates (or strains), a set of dispensable genomes that shared by a few isolates and a collection of unique genomes that only present in single isolates [228]. Recently, Li et al introduced the pan-genome concept to human genomics [229]. In that study, the authors found lineage specific genomic sequences that were not found in the human reference genome [18].

Following the previous hypothetical experiment, assuming in a population there are \( N_p \) number of polymorphic sites in the genome that are allowed to vary from individual to individual and \( N_f \) number non-polymorphic sites that are not variable. Under ideal conditions, there would be no overlaps between polymorphic and non-polymorphic sites and the emission probabilities at different polymorphic sites are uniform. In this case, the information \( I \) for a given population or species from the non-polymorphic sites is zero, since they are not variable. The total information for population study is from polymorphic sites. Therefore for population studies, only the \( N_p \) number of polymorphic sites should and need to be considered. In contrast, for phylogenetic studies, only the \( N_f \) number of non-polymorphic sites should be compared since the complexity from polymorphic sites contributes to the population level. However, in practice, there would be overlaps between polymorphic and non-polymorphic sites. Non-polymorphic sites for a given species could be considered as sites under strong selection. Mutations at these sites would influence the fitness of the individual bearing these mutations. We could call these sites \( N_{fp} \).
However, slightly deleterious or deleterious but non-lethal mutations could be present in the population for a long period of time. This situation, from the perspective of an observer, increases $N_p$ and decreases $N_f$. Emission probabilities at different polymorphic sites are thus very likely not uniform. A systematic bias in the distribution of the emission probabilities could represent selection, but also ancestral polymorphism as mentioned before. In addition, just by genetic drift, polymorphic sites could appear to be under strong selection. All these make distinguish $N_p$ from $N_f$ sites more difficult.

Based on this point of view, most of the human disease-related SNPs should come from $N_{fp}$, which should also be very rare compared with polymorphic SNPs. These sites should also be complete or highly penetrant. As shown in Figure 3-5, reported disease-associated SNPs appeared to accumulate in regions where the mitochondrial genome has lower variance from individual to individual. However, microarray-based methods normally survey with SNPs that are well represented in the population, which are mostly $N_p$ sites. In comparison, sequencing-based methods could identify all the SNPs, which include $N_r$, $N_f$, and $N_{fp}$. The $N_f$ and $N_p$ should be both considerably higher than $N_{fp}$. In order to uncover the rare $N_{fp}$ sites, at least $N_f$ sites need to be surveyed. For $N_p$ sites, as they are variable among healthy individuals of the studied population, they may not need to be surveyed for the purpose of discovering disease-related SNPs.

If we assume that the cutoff between conserved region and non-conserved region is 0.248 variance at 20 bp resolution as shown in Figure 3-10, we could categorize SNPs as being from conserved region or otherwise. With this cutoff, of total 829 20-bp regions of the mitochondrial genome, 802 regions could be determined as conserved. Of total 462 reported disease-associated SNPs, 452 of them are in conserved regions. This number is slightly higher than the expected number of disease-associated SNPs being from the conserved regions under the assumption that
the distribution of disease-associated SNPs is uncorrelated with regional sequence conservation in the population. But the difference is not statistically significant. If we use the data of disease-associated SNPs reported by multiple sources, the trend is similar but the Chi-square test statistics are still not significant. This might result from the non-randomness of the used dataset. There are many mitochondrial genomes in the dataset that are from disease related studies and thus contain disease-associated mutations. As a result, this dataset is actually biased by an overrepresentation of disease-related mutations. If a set of data from random sampling in the general population is used, the test statistics might be significant.

Figure 3-10: Distribution of sample variance at 20 bp resolution of human mitochondrial genome. (End of Chapter 3)
Chapter 4

DNA-based information analysis of RNA-seq data

The results in this chapter have been included in a published paper first-authored by this dissertation author [74]. Used under permission from BMC Genomics (see Appendix).

Project background

In 2009, I joined the research group led by Professor Hong Ma. The Ma group at that time was collaborating with another research group led by Professor Gong Chen, and the joint project between them was to use NGS technologies to seek new insights into early brain development. What they did initially was using Illumina RNA-seq platform to sequence the transcriptomes of embryonic day 18 and postnatal day 7 mouse brain cortices. A paper has been published [72] prior to my contribution. That paper has described a large number of differentially expressed genes between these two stages and many new splicing variants of known genes. Due to the limitations of available tools and methods at the time of that publication, the expression patterns and associated information were not yet analyzed. With my experiences in analyzing NGS data, and then the newly available tool TopHat [73], I started to look into these transcriptomes, especially activities originated from unannotated regions of the mouse genome.
Introduction

The C-value paradox and the extra information provided by transcriptome

As shown in table 1-1, the genome size of a bacterium is about ten times larger than that of a virus. Similarly, the genome size of a unicellular eukaryotic organism is about ten times larger than that of a bacterium. And again, genome size is about ten times bigger in a multicellular eukaryotic organism than in a unicellular eukaryote. A larger genome has the potential to harbor more information. This observation agrees with the concepts introduced in information theory: the larger genome has more information, and thus could sustain a more complex living system.

However, it is well known that the total gene number is similar among multicellular eukaryotes. In addition, multicellular eukaryotic genome size (C-value, as coined by Hewson Swift in 1950 [230]) does not correlate well with the organism complexity, although multicellular eukaryotes differ greatly in terms of development, physiology and behavior. This phenomenon is called the C-value paradox, introduced by Thomas in 1971 [231]. This appears to disagree with the concepts outlined in information theory. However, the splicing mechanism can produce multiple transcripts of different functions and expression profiles from the same genomic sequence. What’s more, different DNA binding, RNA binding and RNA editing proteins provide additional control to gene expression variability. The above processes probably contribute to the additional information required for more complex living systems while still maintain a relative stable genome size. A dynamic transcriptome is capable of conveying additional information beyond the genome.
The transcriptome and its regulation contribute significantly to eukaryotic diversity. The Functional Annotation of the Transcriptome of Mammalian Genome (FANTOM) projects (FANTOM 1-4) have demonstrated the complexity of transcriptomes in several aspects, including non-coding RNAs [232], antisense transcription [232, 233], regulated retrotransposon expression [234], and alternative promoter usage, splicing and polyadenylation [235]. All these studies have pointed out the extra information conveyed through a dynamic and complex transcriptome.

**Further transcriptome complexity revealed by RNA-seq**

Recent high-throughput RNA-seq [70] technologies have provided unprecedented capability to analyze cellular, tissue-specific, or organismal gene activities across a broad spectrum. They also revealed the transcriptomic complexity during cell differentiation [236, 237] and organ development [72]. Furthermore, individuals of the same species have transcriptomic differences such as expression variation among humans [238]. Another level of transcriptomic complexity has been revealed by extensive analysis of novel splicing variants from known exons [72, 236-239]. In addition, thousands of transcripts from previously unannotated (non-exonic) genomic regions have been reported [236-241]; they are either named TUF (Transcripts of Unknown Function) [242] or unannotated TAR (Transcriptionally Active Region) [243]. Some of the unannotated TARs are large intergenic noncoding RNAs that function in embryonic stem cell pluripotency and cell proliferation [244, 245], while most unannotated TARs have no known function.

It has been reported that undifferentiated human stem cells have elevated expression of unannotated TARs compared with differentiated neural progenitor cells [236]. Our recent study has also detected additional transcripts from intergenic regions and introns in mouse embryonic
and neonatal brain cortex [72]. Mammalian neural development is a complex process involving cell division, differentiation, migration, axon guidance, synaptogenesis, and synaptic plasticity. The characterization of stage specific unannotated TARs during early brain development could provide clues regarding the roles these unannotated TARs might play in determining neural fate and in regulating neuronal functions.

Assessing extra information conveyed by neurodevelopmental transcriptomes

To further investigate the transcriptome dynamics and to better understand the possible roles of unannotated TARs in early neural development, we have analyzed the RNA-seq datasets from embryonic and postnatal mouse brain cortex that we generated recently [72], as well as seven additional RNA-seq datasets covering both neural and non-neural tissues (Mortazavi et al. 2008; Wu et al. 2010). These nine transcriptome datasets include data from human embryonic stem cell (hESC) and its subsequently differentiated forms (N1, early initiation; N2, neural progenitor; and N3, early glial-like cell) [236], embryonic day 18 (E18) and postnatal day 7 (P7) mouse brain cortices [72], and adult mouse brain (AMB), liver (AML), and muscle (AMM) [71].

Through a systematic analysis of these nine datasets, we found several unique characteristics of the transcriptomes in early neural development. We found that, although the genome was not as pervasively transcribed as previously reported [246], most of the genomic regions at 1 Mb resolution had detectable RNA-seq signals. We also found that the transcriptomes from neural tissues possessed several genome-wide characteristics resembling those of stem cells. Interestingly, the E18 cortex shows the highest level of unannotated transcript expression comparing to P7 and adult brains. Furthermore, the intronic unannotated expressions are associated with GO terms for neurogenesis, neural signaling and negative regulation.
Importantly, few of the unannotated TARs in E18 and P7 cortices are connected with known transcripts, suggesting potential novel functions of these TARs during brain development.

**Analyses and Results**

**Mapping gigabytes of RNA-seq data**

To examine the genomic distribution of transcriptomic reads, we mapped all RNA-seq data by the TopHat software [73], which was designed to map RNA-seq data with moderate IT (information technology) infrastructure. Embryonic, neonatal and adult mouse data were mapped onto the mouse reference genome (UCSC mm9, NCBI build 37) as described in the Methods section. For comparison, adult mouse liver and muscle RNA-seq data were analyzed in the same manner. Human ESC and its differentiation data were mapped onto the human reference genome (UCSC hg19, NCBI build 37). Only uniquely matched reads were further analyzed. Because the data amounts available for the downstream analysis varied for different tissues (Table 4-1), the read count data were first normalized against the available dataset size, measured in base pairs, for each tissue. To accommodate differences in the details of library preparation and sequencing procedures, we adjusted for the sequencing quality in all data sets according to the quality computation method of the Illumina 1.3 pipeline. E18 and P7 RNA-seq data had the largest percentage of mappable reads in this group, approximately 60%, while other data sets had about 30% mappable data. The AMB RNA-seq data had the largest quantity, with more than 1,000 million mapped base pairs.
Table 4-1: RNA-seq mapping result using TopHat.

<table>
<thead>
<tr>
<th>Sequencing Platform</th>
<th>Illumina</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accession Number</td>
<td></td>
</tr>
<tr>
<td>SRR037165 to SRR037170</td>
<td>SRR001356 to SRR001358</td>
</tr>
<tr>
<td>SRR037193 to SRR037198</td>
<td>SRR001357 to SRR006488</td>
</tr>
<tr>
<td>SRR037205 to SRR037226</td>
<td>SRR006490 and SRR006491</td>
</tr>
<tr>
<td>Read Type</td>
<td></td>
</tr>
<tr>
<td>35bp PE and 35bp SE</td>
<td>36bp PE and 36bp SE</td>
</tr>
<tr>
<td>33bp SE</td>
<td></td>
</tr>
<tr>
<td>Original Read Count</td>
<td></td>
</tr>
<tr>
<td>14.4M PE and 4.4M SE</td>
<td>15.8M PE and 7.0M SE</td>
</tr>
<tr>
<td>15.6M PE and 11.6M SE</td>
<td>22.4M PE and 3.0M PE</td>
</tr>
<tr>
<td>19.6M PE and 3.0M PE</td>
<td>10.0M PE and 3.0M SE</td>
</tr>
<tr>
<td>22.4M PE and 3.0M PE</td>
<td>10.4M PE and 3.6M SE</td>
</tr>
<tr>
<td>1,162</td>
<td>1,352</td>
</tr>
<tr>
<td>1,774</td>
<td>1,819</td>
</tr>
<tr>
<td>853</td>
<td>857</td>
</tr>
<tr>
<td>853</td>
<td>2,936</td>
</tr>
<tr>
<td>2,503</td>
<td>1,977</td>
</tr>
<tr>
<td>Original RNA-seq data size (Mbp)</td>
<td></td>
</tr>
<tr>
<td>401</td>
<td>458</td>
</tr>
<tr>
<td>628</td>
<td>594</td>
</tr>
<tr>
<td>594</td>
<td>491</td>
</tr>
<tr>
<td>628</td>
<td>525</td>
</tr>
<tr>
<td>594</td>
<td>1,086</td>
</tr>
<tr>
<td>525</td>
<td>683</td>
</tr>
<tr>
<td>571</td>
<td>571</td>
</tr>
<tr>
<td>Mapped by Tophat* (Mbp)</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>34</td>
</tr>
<tr>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>58</td>
<td>61</td>
</tr>
<tr>
<td>61</td>
<td>37</td>
</tr>
<tr>
<td>37</td>
<td>27</td>
</tr>
<tr>
<td>29</td>
<td></td>
</tr>
</tbody>
</table>

* Allowing 1 hit by TopHat

1. Data from Wu et al.
2. Data from Mortazavi et al.

Medium and high resolution genome-wide transcription activity profile

Despite a dataset size up to 1,000 Mbp (AMB), only about 2% of the mouse genome had been mapped with RNA-seq reads, unlike a previous report that suggested a more pervasively transcribed genome from detailed analyses of ~1% of the human genome [246]. A recent study using single- and paired-end RNA-seq and tiling arrays also concluded that the genome is not as pervasively transcribed as previously reported [240]. Nevertheless, when the mouse genome was divided into 1 million base-pair long intervals, we observed that 85% of the intervals had 100 or
more detected RNA-seq reads, as shown in Figure 4-1 (Each 1Mb interval is represented by a horizontal box, with color ranging from black to green. A box with black color means that there was no exon in this given interval. The brighter the green color, the higher the exon content (percentage) this given interval possesses. The scale at the bottom of the map illustrates color representation of the exon percentage of the region. Red square at the left side of the chromosome represents the detected number of RNA-seq reads for the given region in E18 stage, while the blue square at the right side represents the detected number of RNA-seq reads in P7 stage. Expression level is data size normalized. Scale for expression level is at the bottom of the map.).

However, the majority of chromosome regions of the 100 kb size had no detected transcription, as illustrated for X-chromosome in Figure 4-2 (All symbols represent the same information as in Figure 4-1, except each horizontal box now represents only a 100 Kb genome region.). Whereas many regions showed transcription at both the E18 and P7 stages, some regions were specifically active in one of these stages, suggesting that a selective set of genes are turned on and off from embryonic to neonatal brain stages, consistent with the previous finding of preferential expression of several thousand genes at either one of these stages [72]. Besides heterochromatin regions (centromeres and telomeres), some large genomic regions with very low annotated gene content had no detectable reads, including several regions in Chromosome 7 (Figure 4-1) and the region in chromosome X from 26,000,000 to 32,000,000 bp (Figure 4-2).
Figure 4-1: Genome-wide expression map at 1 Mb resolution for E18 and P7.

See text for detailed explanation.
Figure 4-2: 100 Kb resolution expression map of chromosome X.

See text for detailed explanation.

After further dividing the genome into 10 kb intervals, most of the 10 kb intervals had no detected transcription (Figure 4-3A). Most of the reads in highly expressed intervals (> 1,000
RNA-seq reads) were mapped to known exons. Among intervals with 1,000 to 10,000 detected reads, more intervals had intergenic transcripts than intronic transcripts, with very few intervals having all three types (exonic, intronic and intergenic) of transcripts. For intervals with 100 or more RNA-seq reads, there were 3 times more intervals expressing intronic signal in E18 stage than in P7 stage. The E18 stage also had slightly more intervals with intergenic transcripts than the P7 stage, although the numbers of intervals with exonic transcripts were similar between the two stages (Figure 4-3B).

Figure 4-3: Genomic expression analysis in 1 Kb interval.

A. Number of intervals in different categories of RNA-seq read coverage. RNA-seq reads were categorized as: exonic, which were reads mapped to known exon, intronic, which were reads mapped to known intron, and intergenic, which were reads mapped to known intergenic region. There were more highly expressed intervals with exonic expression than intronic or intergenic ones. B. Venn diagram illustration of the above result. Only intervals containing more than 100 reads were considered, as indicated by the dotted line in A.
We found that expression level was positively correlated with the exon contents of the given interval. The higher exon percentage the interval had, the higher the number of detected RNA-seq reads in the interval. At 1 Mb interval size, the Pearson’s Correlation Coefficient between exonic read number and exonic content percentage was 0.60; in contrast, the Pearson’s Correlation Coefficient between intronic read number and exonic content percentage was only 0.18. The Pearson’s Correlation Coefficient between intergenic read number and exonic content percentage was also quite low, at 0.26. At 100 kb interval size, the correlation decreased but the general trend was maintained. While the exon-rich regions had more reads for exons than introns or intergenic regions, the exon-sparse regions had similar numbers of intronic and intergenic reads as the exon-rich region. In addition, at 1 Mb resolution, the exonic expression level had modestly positive correlation with both intronic \( (R = 0.43) \) and intergenic expression level \( (R = 0.37) \) in the same region. However, at 100 kb resolution, the aforementioned correlation became very weak.

**Genome-wide comparison revealed similarity between neural and stem cell transcriptomes**

To understand the neural transcriptome characteristics, we compared mouse cortical RNA-seq data at E18 and P7 stages with available adult mouse brain, liver, and muscle RNA-seq data [71], as well as RNA-seq data from human embryonic stem cells (hESC) and neural cells (N1, N2, N3) immediately differentiated from hESC [236]. We first analyzed transcriptome properties at the chromosome level, using a method slightly modified from Mortazavi et al. [71] as detailed in the Methods section, and labeled as RPKM* (similar to RPKM; formula (40)).

In addition to the above-mentioned mapping of the 5 mouse datasets (E18 and P7 cortices, and adult brain, liver and muscle), we also mapped the 4 human RNA-seq datasets
Based on the 85% identity calculated from coding regions between mouse and human genome previously [62], there would be on average 5 mismatches per 35 bp RNA-seq read length. We found that the threshold of 2 mismatches per 35 bp read achieved the best balance between specificity and sensitivity for this cross-species mapping. Increasing the number of allowed mismatches resulted in fewer uniquely and correctly mapped reads, while decreasing this number resulted in fewer total mappable reads. With the threshold of allowing maximum 2 mismatches for RNA-seq mapping, this would mean little cross-species mappable reads if the differences between coding regions of human and mouse were distributed evenly. Surprisingly, on average 6% of the total human RNA-seq reads could be mapped to the mouse genome, or 11.5% of the reads mappable to the human genome. The majority (80%) of the reads mapped to the mouse genome were also mapped to the human genome.
Figure 4-4: Chromosomal expression level between stages.

Measured in RPKM* as described in formula (40).

- **A.** RNA-seq reads of all neural samples, along with hESC reads, mapped onto mouse reference genome.
- **B.** Comparison between adult mouse brain, liver and muscle.
- **C.** RNA-seq reads of four human samples mapped onto human reference genome.
- **D.** Standard deviation (StdDev) of chromosomal expression level across datasets.
- **E.** Mitochondrial expression level across datasets.

### Table 4-2: Chromosomal expression across stages (read count)

<table>
<thead>
<tr>
<th></th>
<th>hESC</th>
<th>N1</th>
<th>N2</th>
<th>N3</th>
<th>E18</th>
<th>P7</th>
<th>AMB</th>
<th>AMM</th>
<th>AML</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>92476</td>
<td>89415</td>
<td>121003</td>
<td>119972</td>
<td>576642</td>
<td>617325</td>
<td>1664591</td>
<td>836121</td>
<td>1101579</td>
</tr>
<tr>
<td>chr2</td>
<td>116351</td>
<td>133228</td>
<td>195010</td>
<td>231845</td>
<td>869368</td>
<td>977243</td>
<td>2739929</td>
<td>1190394</td>
<td>1066153</td>
</tr>
<tr>
<td>chr3</td>
<td>68408</td>
<td>76658</td>
<td>92907</td>
<td>99243</td>
<td>472070</td>
<td>486040</td>
<td>1082967</td>
<td>597877</td>
<td>1091763</td>
</tr>
<tr>
<td>chr4</td>
<td>97551</td>
<td>107289</td>
<td>123988</td>
<td>137954</td>
<td>620423</td>
<td>635012</td>
<td>1724011</td>
<td>681761</td>
<td>1070897</td>
</tr>
<tr>
<td>chr5</td>
<td>76508</td>
<td>80364</td>
<td>83364</td>
<td>94414</td>
<td>676641</td>
<td>740624</td>
<td>1914836</td>
<td>503478</td>
<td>2708046</td>
</tr>
<tr>
<td>chr6</td>
<td>81063</td>
<td>78476</td>
<td>110257</td>
<td>102125</td>
<td>513719</td>
<td>538667</td>
<td>1331478</td>
<td>496872</td>
<td>926322</td>
</tr>
<tr>
<td>chr7</td>
<td>102920</td>
<td>115680</td>
<td>123946</td>
<td>144323</td>
<td>726801</td>
<td>928153</td>
<td>3047195</td>
<td>3394482</td>
<td>1550147</td>
</tr>
<tr>
<td>chr8</td>
<td>58980</td>
<td>61532</td>
<td>74119</td>
<td>79362</td>
<td>542771</td>
<td>626383</td>
<td>1802582</td>
<td>1503853</td>
<td>837867</td>
</tr>
<tr>
<td>chr9</td>
<td>77712</td>
<td>93932</td>
<td>102339</td>
<td>110617</td>
<td>574405</td>
<td>649769</td>
<td>1667401</td>
<td>874972</td>
<td>1260666</td>
</tr>
<tr>
<td>chr10</td>
<td>67252</td>
<td>73139</td>
<td>77072</td>
<td>89781</td>
<td>514335</td>
<td>565546</td>
<td>1533719</td>
<td>574747</td>
<td>755932</td>
</tr>
<tr>
<td>chr11</td>
<td>152849</td>
<td>166420</td>
<td>196167</td>
<td>215689</td>
<td>945122</td>
<td>1068548</td>
<td>2946991</td>
<td>1417059</td>
<td>1176264</td>
</tr>
<tr>
<td>chr12</td>
<td>51725</td>
<td>54751</td>
<td>71727</td>
<td>76497</td>
<td>466206</td>
<td>484446</td>
<td>968689</td>
<td>342598</td>
<td>826123</td>
</tr>
<tr>
<td>chr13</td>
<td>42278</td>
<td>47373</td>
<td>51176</td>
<td>63538</td>
<td>374994</td>
<td>351659</td>
<td>744911</td>
<td>313915</td>
<td>395946</td>
</tr>
<tr>
<td>chr14</td>
<td>50762</td>
<td>49722</td>
<td>56582</td>
<td>60894</td>
<td>335300</td>
<td>343972</td>
<td>917395</td>
<td>406472</td>
<td>631873</td>
</tr>
<tr>
<td>chr15</td>
<td>69589</td>
<td>81617</td>
<td>79826</td>
<td>88766</td>
<td>454657</td>
<td>546428</td>
<td>1489998</td>
<td>697665</td>
<td>749796</td>
</tr>
<tr>
<td>chr16</td>
<td>40058</td>
<td>39991</td>
<td>47270</td>
<td>60747</td>
<td>339079</td>
<td>369717</td>
<td>826537</td>
<td>276625</td>
<td>646429</td>
</tr>
<tr>
<td>chr17</td>
<td>63137</td>
<td>72609</td>
<td>82793</td>
<td>84524</td>
<td>475870</td>
<td>498847</td>
<td>1311107</td>
<td>413794</td>
<td>839180</td>
</tr>
<tr>
<td>chr18</td>
<td>51396</td>
<td>50382</td>
<td>58234</td>
<td>65976</td>
<td>331399</td>
<td>337969</td>
<td>855284</td>
<td>265056</td>
<td>516904</td>
</tr>
<tr>
<td>chr19</td>
<td>49964</td>
<td>55859</td>
<td>67622</td>
<td>67584</td>
<td>335801</td>
<td>408372</td>
<td>1174643</td>
<td>750516</td>
<td>899498</td>
</tr>
<tr>
<td>chrX</td>
<td>65669</td>
<td>63436</td>
<td>98576</td>
<td>86681</td>
<td>332539</td>
<td>358502</td>
<td>809519</td>
<td>223777</td>
<td>286129</td>
</tr>
<tr>
<td>chrY</td>
<td>621</td>
<td>372</td>
<td>593</td>
<td>779</td>
<td>1008</td>
<td>645</td>
<td>1033</td>
<td>1697</td>
<td>2438</td>
</tr>
</tbody>
</table>

### Table 4-3: Principal inertias (eigenvalues) of Correspondence Analysis

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.043309</td>
<td>0.019397</td>
<td>0.001323</td>
<td>0.000912</td>
<td>0.000177</td>
</tr>
<tr>
<td>Percentage</td>
<td>66.34%</td>
<td>29.71%</td>
<td>2.03%</td>
<td>1.4%</td>
<td>0.27%</td>
</tr>
<tr>
<td>Accumulated Percentage</td>
<td>66.3%</td>
<td>96.1%</td>
<td>98.1%</td>
<td>99.5%</td>
<td>99.7%</td>
</tr>
</tbody>
</table>
Figure 4-5: Sample/Stage clustering based on Correspondence Analysis of chromosomal expression profiles.

Similar stages were closer to each other than dissimilar ones. Analysis done using ca package in R.

We then compared the chromosomal expression profile between mouse (E18, P7, AMB, AMM and AML) and human (hESC, N1, N2 and N3) samples. Despite the fact that smaller fractions of the human reads were mapped onto the mouse reference genome, largely due to DNA sequence dissimilarity between these two species, both mouse and human neural data sets were highly similar in terms of expression level relative to total mapped reads at the scale of individual chromosomes (Figure 4-4A). Furthermore, all neural tissues had similar profiles to that of stem cells, but quite different from those of adult liver and muscle, particularly for some chromosomes,
such as 5 and 7 (Figure 4-4B). To summarize the above information, we applied Correspondence Analysis on the chromosomal expression profile. The chromosomal expression distribution among all stages, using mouse genome as reference, was first measured in read counts (Table 4-2). Correspondence Analysis (CA) was performed using the ca package in R [247]. The first two dimensions resulted from the CA could explain 96% of the differences in the original 9×21 dimensions, which indicated that the first two dimensions were representative (Table 4-3). The result was plotted using these two dimensions (Figure 4-5). E18 and P7 were clustered together with hESC, N1, N2 and N3. AMB was also very close to the aforementioned cluster. AML and AMM were significantly further away than AMB.

As an alternative way to compare the mouse and human data, we also mapped the human data onto the human reference genome (hg19; [18]), for comparative analysis using previously identified syntenic/orthologous genomic regions between mouse and human [18, 62]. These studies defined 217 conserved syntenic blocks between the human and mouse genomes. Chromosomal expression profiles in early-differentiated human neural cells were very similar to that of human embryonic stem cells (Figure 4-4C). Even with different chromosome numbers and organizations, neural chromosomal expression profiles were also very similar between human neural cells and mouse neural tissue samples between syntenic/orthologous genomic regions (Figure 4-4B, C). For example, the most highly expressed chromosome in mouse was chromosome 11, whose human counterpart is chromosome 17, which was the second most highly expressed chromosome in human. The most highly expressed chromosome in human was chromosome 19, whose mouse counterparts are distributed on chromosomes 7, 8 and 19, among which chromosomes 7 and 19 were also highly expressed in neural tissues.
To assess the variation in expression levels between chromosomes for different tissues/organs, we calculated the standard deviation for the distribution of individual chromosome expression level, or RPKM* values, for each mapped transcriptome dataset (Figure 4-4D). We found that the standard deviation for E18 was the lowest among mouse samples, while the standard deviation for the stem cells was the lowest in human samples (Figure 4-4D). These results indicated that the mouse E18 brain cortex and human embryonic stem cells use chromosomes more evenly than other organs/tissues.

It is well known that the brain has a very high metabolic rate, consuming a significant amount of energy while lacking substantial energy reserve tissues. Thus normal brain functions depend on mitochondria as the crucial energy provider. To examine the mitochondrial genome expression-level changes across different developmental stages, we plotted the normalized mitochondrial expression level, measured in RPKM*, across all nine datasets and normalized against the dataset size. We found that in human datasets, compared with stem cells, differentiated neural cells had a higher level of mitochondrial expression (Figure 4-4E), increasing from the N1 to N2 stages, then maintaining a similarly high level at the N3 stage. Similarly in the mouse brain, mitochondrial expression progressively increased from the E18 embryonic stage to the P7 neonatal stage, then to the adult stage. The adult mouse brain had a similar level of mitochondrial expression to that of the adult liver, while the adult mouse muscle had the highest level of mitochondrial expression among all analyzed organs/tissues, consistent with high energy demand for muscle contraction.

To assess the similarity between neural and stem cell transcriptomes further, we compared the transcriptomes between human and mouse using only 1-to-1 orthologous gene pairs between these species. A total of 12168 orthologous gene pairs were identified using the MGI
orthology database. The expression level of each gene was measured in RPKM, with the modification that detected base pairs from exons were used instead of read number in the RPKM formula to accommodate read length differences between datasets. Again the expression level for a given gene at a given stage was normalized against the RNA-seq dataset size. We then added 1 to the calculated expression level value for each gene, to ensure valid logarithm transformation. The calculated value was then log₂ transformed. We first calculated the gene expression correlation between hESC and the rest of the samples. As expected, the correlation was high between hESC and the cells derived from hESC. Among the rest, E18 had the highest correlation with hESC (R = 0.61), and P7 had the second highest (R = 0.57). The data were then analyzed in MeV (MultiExperiment Viewer) from the TM4 suite [248]. After using different clustering methods (Hierarchical clustering and K-means), we found that neural datasets from human and mouse were consistently grouped together with the stem cell dataset, separate from the liver and muscle datasets (Figure 4-6A). This suggested that the neural and stem cell transcriptomes were globally more similar in terms of orthologous gene expression than they were to liver and muscle transcriptomes.
Figure 4-6: Sample/Stage clustering based on Orthologous gene expression pattern.

A. Clustering of nine RNA-seq datasets using orthologous genes between human and mouse. Center panel: expression level heat map overview of 12168 orthologous genes. The sample tree on top is derived from Pearson’s Correlation (PC) distance matrix, while the one at the bottom is derived from Euclidean Distance (ED) matrix. B. Scatter plot of orthologous gene expression level between selected stages. Genes without detectable expression were not included.

Figure 4-7: Selected gene expression level at different stages.

A. Expression level of pluripotency-related genes across stages. Expression level was measured in both raw RNA-seq data size (bp) and RPKM methods. B. RT-PCR results showing the expression of Sox2, Myc and Klf4, with Sox11 as the positive control.
We then further analyzed the correlation of the expression in two different tissues/stages among co-expressed genes between the tissues/stages (Figure 4-6B). Among all mouse samples, although E18 was the one with the highest correlation with hESC, E18 was still more similar to mouse neural transcriptomes in terms of expression level correlation. In particular, E18 and P7 transcriptomes were much more correlated with each other than with hESC, suggesting that the similarity between E18 or P7 cortex and hESC is relatively limited.

We also analyzed the genes associated with the pluripotency of stem cells. Sox2, Myc, Oct4 and Klf4, which are four genes that have been found in multiple studies to convert human and mouse somatic cells to induced pluripotent stem (iPS) cells [249-253]. Similarly, Oct4 and Sox2, plus two other factors, Nanog and LIN28A, were also able to induce iPS cells from human somatic cells [254]. We found that Sox2, Myc and Klf4 were detectable in all neural samples we analyzed (Figure 4-7A). Specifically, the E18 stage had the highest Sox2 and Myc expression among all mouse samples. However, Nanog and LIN28A were not detected at either the E18 or P7 stage. RT-PCR experiment was also carried out with primers specifically targeting Sox2, Myc and Klf4. The results supported the expression of Sox2, Myc and Klf4 in E18 and P7 cortices (Figure 4-7B). Due to the presence of multiple homologues of Oct4 (also called Pou5f1) in the mouse genome, neither RNA-seq nor RT-PCR could identify specific expression for Oct4/Pou5f1. Interestingly, Sox11, which encodes a transcription factor and was previously reported to be expressed in glial cells [255], was highly expressed in the E18 cortex and significantly down-regulated in the P7 cortex (Figure 4-7A & B). Because cortical neurons are mainly generated at the late embryonic stage whereas glial cells are mainly generated in the postnatal stage, the high level expression of Sox11 in the E18 cortex suggests an additional role during early brain development besides its proposed function in glial cells.
Expression pattern of selected neurodevelopmental disorders associated genes

Figure 4-8: Expression pattern of genes associated with neurodevelopmental disorders.

**A.** Expression of selected GABAA receptor genes. **B.** Expression of selected Autism spectrum disorders associated genes.
We further analyzed the expression of genes potentially associated with neurodevelopmental disorders in both neural tissues and stem cells. Autism spectrum disorders (ASD), together with schizophrenia and mental retardation, are typically characterized as neurodevelopmental disorders. Genome-wide association studies (GWAS) have identified many genes related to ASD. Among these genes, many have been found to relate to the GABAergic neurotransmission system. Here, we analyzed the expression of 20 genes encoding different GABAA receptor subunits and 25 genes that have been proposed to be associated with ASD [256, 257] (Figure 4-8A and B). Although GABA receptor genes in general showed low expression levels in non-neural tissue and stem cells, the gene for GABA\textsubscript{A}R alpha5 (GABRA5) subunit showed a very high level of expression in hESCs, which lacks a GABAergic system, suggesting a novel function of this gene in embryonic stem cells. Its expression was not detectable after initial differentiation (N1) and then observed again after further differentiating into neural cells (N2). It also showed an increase from E18 to P7, but a reduction from P7 to adult brain, consistent with more restricted localization in the adult brain [258]. In addition, GABRQ and GABRP also showed modest expression in hESCs. Furthermore, the gene encoding the GABA\textsubscript{A}R delta (GABRD) subunit showed the highest expression level among all GABA\textsubscript{A}R subunit genes in the adult brain. Since delta is specifically localized at extrasynaptic sites and mediates tonic inhibition rather than normal fast inhibition, this result emphasizes the importance of tonic inhibition in regulating adult brain activity. Finally, genes for GABA\textsubscript{B} receptor (GABBR1 and GABBR2) subunits showed significant differential expression during brain development, with GABBR1 dominant from E18 through adult brain while GABBR2 only expressed highly in the adult brain.

Among genes associated with ASD, \textit{RGS4, DTNBP1, NLGN2, STX7, MECP2, ARVCF}, and \textit{PPP3CC} all showed high-level expression from the embryonic to adult brain. One important
finding was that while both *NLGN1* and *NLGN2* showed high-level expression at the E18 to P7 stages, consisting with their synaptogenic functions, *NLGN1* expression was significantly reduced in the adult brain, suggesting that the relevant function might be fulfilled by other cell adhesion molecules. This is also consistent with the current understanding that many cell adhesion molecules can trigger glutamatergic synapse formation as NLGN1 does, but only NLGN2 is capable of inducing GABAergic synaptogenesis [259, 260]. One surprising finding is that DISC1, a well-studied gene associated with schizophrenia [261], showed very low expression at the E18/P7 stages and still low in adult brain. However, DISC1 was modestly expressed in hESCs and the expression decreased after neural differentiation, suggesting that DISC1 might play an important role in stem cell functions.

**High level of unannotated expression at E18 and P7 stages**

To obtain an accurate set of unannotated TARs, we first subtracted the TARs overlapping with annotated exons, tRNAs or rRNAs. We then removed TARs overlapping with repeats. We also excluded TARs within 2 kb vicinity of the first and last exon to avoid promoter, TSS and TTS associated transcription activity, which have been previously studied [262-264]. We call the set of unannotated TARs generated from the aforementioned procedure the filtered unannotated TARs, and they include both intronic TARs and intergenic TARs. We found that the E18 stage had the highest percentages of both intronic and intergenic reads, at 5% and 3% of the E18 total data, respectively (Figure 4-9A). P7 stage had the second highest percentages, with about 1.1 % and 2.6% for introns and intergenic regions, respectively. In comparison, almost all the other datasets had less than 1% of the data matching intronic or intergenic regions. Among human
samples, surprisingly, the stem cell stage had the lowest percentage of unannotated TARs, while the neural progenitor cell N2 stage had the highest. This result is different from a previous report [236], which did not use methods that rigorously removed repeats as in this study.
Figure 4-9: Unannotated transcript expression.

A. Percentage of unannotated transcriptionally active regions (TARs) data of each dataset.
Column in red color represents intronic TAR, and column in blue color represents intergenic TAR. X-axis represents different datasets, while Y-axis represents the percentage. B to E: An intronic TAR found located inside mouse PCBP2. B: RNA-seq reads mapped by TopHat to this location. Each black bar represents a RNA-seq read. Because the coverage on the 3’ end (right side) of the gene is too high, only part of the reads mapped at this region is shown. Each blue horizontal line connecting two bars indicates a splice-junction-spanning read (read split by splicing activity). Red lines highlighted such reads spanning the intronic TAR and the exon on its left. C: conservation score from 30-Way multi-species genome alignment. Notice the region with high conservation score co-localized with the intronic TAR. D: current PCBP2 gene model, with green box representing known exons. E: amino acid alignment of the ORF of intronic TAR and its human and dog homologues. Gray column plot below: each column represents the similarity of the amino acids in that column. F: RT-PCR validation for intronic TARs. For each intronic TAR inferred from RNA-seq mapping result, primers were picked to validate the connection between the intronic TAR and the flanking exon (upstream/downstream), between the flanking exons, as well as the expression of the intronic TAR.

**Concordant expression changes between intronic TARs and flanking exons**

To investigate the relationship between an intronic TAR and its flanking exons, we examined their respective expression levels (as measured in RPKM; [71]). There were 488 genes with intronic TARs at either the E18 or P7 stage. We found a strong positive correlation between the intronic expression and the flanking exonic expression. From E18 to P7, we found 436 genes with decreased intronic expression levels; among these 436 genes, 242 (56%) also had decreased exonic expression (Table 4-4). Even more strikingly, 52 genes had increased intronic expression level from E18 to P7 (Table 4-4), with 43 (83%) having increased exonic expression levels. The correlation was highly significant, with a p-value of 0.0001 from Fisher’s Exact Test for the association of the exonic and intronic expression levels.
Table 4-4: Intronic and exonic expression level changes for genes with intronic TARs (from E18 to P7).

<table>
<thead>
<tr>
<th>Number of Genes With</th>
<th>Also With Increased Intrinsic Expression</th>
<th>Also With Decreased Intrinsic Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased Exonic Expression</td>
<td>43</td>
<td>194</td>
</tr>
<tr>
<td>Decreased Exonic Expression</td>
<td>9</td>
<td>242</td>
</tr>
</tbody>
</table>

**Few unannotated TARs connected with known exons**

The strong concordant correlation between the previously unannotated intronic TARs and flanking exons suggested that the intronic TAR and its flanking exon might be parts of the same RNA transcript. To test this hypothesis, we focused on the E18 and P7 datasets, which had the largest percentage of filtered unannotated TARs. A paired-end read with one end located in the unannotated TAR and the other in a known exon would be strong evidence that this intronic TAR and the known exon are parts of the same mRNA. However, it is in principle possible that the mapping positions could be erroneous. In addition, the existing mathematical and statistical models for determining the connection between TARs [237] are designed for RNA-seq data from cDNAs generated with random primers. They are not applicable to poly-dT primed data, which have a 3’ bias. So we first devised a model suitable for both priming techniques (Methods; formula (41)), which reports the presence of the physical connection between expressed TARs...
and known exons. Using known adjacent exons and single exon genes (SGEs) with detected reads as positive and negative controls in a simulation, the formula had success rates of 93% and 100%, respectively.

Using formula 41, we found that only a very small percentage of the unannotated intronic TARs were connected with known exons (Table 4-5) as supported by the RNA-seq reads. Although a large fraction (70% for E18, 60% for P7) of the unannotated TARs was found to be connected to other regions mapped elsewhere, these mapped regions were very short, fragmented and unannotated regions with low read coverage. Surprisingly, some unannotated TARs were connected to read ends that mapped to multiple chromosomes, possibly due to mapping to repetitive sequences, erroneous mapping or possible cross-chromosome splicing (trans-splicing), a rare phenomenon that was previously observed [265].

Table 4-5: Physical connection between unannotated TARs and known exons.

<table>
<thead>
<tr>
<th>Connected with Known Exons</th>
<th>E18 Brain Cortex</th>
<th>P7 Brain Cortex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic TAR</td>
<td>Standalone</td>
<td>Intrinsic TAR</td>
</tr>
<tr>
<td>19</td>
<td>117</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Intergenic TAR</td>
<td>Non-Standalone</td>
<td>Intergenic TAR</td>
</tr>
<tr>
<td>2</td>
<td>404</td>
<td>41</td>
</tr>
<tr>
<td>2</td>
<td>71</td>
<td>46</td>
</tr>
<tr>
<td>13</td>
<td>123</td>
<td>46</td>
</tr>
<tr>
<td>Other (Non-Standalone</td>
<td>Non-Standalone</td>
<td>Other (Non-Standalone Multi-Chromosome)</td>
</tr>
<tr>
<td>14</td>
<td>404</td>
<td>14</td>
</tr>
<tr>
<td>8</td>
<td>71</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>123</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>46</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>554</td>
<td>168</td>
</tr>
<tr>
<td></td>
<td>126</td>
<td>94</td>
</tr>
</tbody>
</table>
Few newly discovered intronic TARs were present in known mRNA and EST databases

To test whether there is other evidence for the intronic TARs, we searched the data in NCBI’s cDNA/mRNA and EST databases. Among 554 intronic TARs detected at E18 stage, 176 (32%) had no matches in NCBI databases. Similarly, among 168 intronic TARs at P7 stage, 49 (29%) had no matches in NCBI databases. Therefore, our results provide the first evidence for these TARs being expressed. Among the matching NCBI database entries, 11 (2%) of the 378 for the E18 stage and 7 (4%) of the 119 for the P7 stage were from the same stages, but none of them was from the brain cortex.

We then examined the splicing pattern of the mRNA and EST records matched to our detected intronic TARs and found two classes of intronic TARs: (1) with records suggesting that the TARs were standalone, without connection to known exons; (2) with some records suggesting that the TARs were standalone while other records suggesting that they were connected to known exons. 304 out of 378 (80%) intronic TARs at E18 and 75 out of 119 (63%) intronic TARs at P7 belonged to the first class. For the second class of intronic TARs, on average, the ratios for records supporting standalone transcripts to those for connections to known exons were 4.2 and 2.8 for the E18 and P7 stages, respectively. Taking together, the comparison with NCBI’s cDNA/mRNA and EST databases strongly suggested that most of our detected intronic TARs were not connected with known exons and thus were novel transcripts.
### Table 4-6: E18 intronic TARs with hits in miRbase

<table>
<thead>
<tr>
<th>Query</th>
<th>Query Length</th>
<th>Subject</th>
<th>Subject Length</th>
<th>BitScore</th>
<th>E-value</th>
<th>Alignment Length</th>
<th>Matched</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1-33852746-33857360</td>
<td>4615</td>
<td>mmu-mir-1935</td>
<td>60</td>
<td>77.8</td>
<td>1.66E-06</td>
<td>55</td>
<td>51</td>
<td>0</td>
</tr>
<tr>
<td>chr1-56989661-56990743</td>
<td>1083</td>
<td>mmu-mir-1935</td>
<td>60</td>
<td>85.7</td>
<td>8.28E-08</td>
<td>51</td>
<td>49</td>
<td>0</td>
</tr>
<tr>
<td>chr2-6719696-6721827</td>
<td>2132</td>
<td>mmu-mir-1935</td>
<td>60</td>
<td>79.8</td>
<td>6.12E-07</td>
<td>52</td>
<td>49</td>
<td>0</td>
</tr>
<tr>
<td>chr3-158158463-158160069</td>
<td>1607</td>
<td>mmu-mir-1935</td>
<td>60</td>
<td>71.9</td>
<td>6.78E-06</td>
<td>60</td>
<td>54</td>
<td>0</td>
</tr>
<tr>
<td>chr4-45834328-45836371</td>
<td>2044</td>
<td>mmu-mir-1935</td>
<td>60</td>
<td>93.7</td>
<td>5.60E-09</td>
<td>55</td>
<td>53</td>
<td>0</td>
</tr>
<tr>
<td>chr11-104173757-104179094</td>
<td>5338</td>
<td>mmu-mir-1935</td>
<td>60</td>
<td>95.6</td>
<td>1.65E-08</td>
<td>60</td>
<td>57</td>
<td>0</td>
</tr>
<tr>
<td>chr11-57103743-57105116</td>
<td>1374</td>
<td>mmu-mir-1935</td>
<td>60</td>
<td>69.9</td>
<td>6.14E-06</td>
<td>55</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>chr12-118489259-118490932</td>
<td>1674</td>
<td>mmu-mir-153-2</td>
<td>87</td>
<td>157</td>
<td>9.24E-17</td>
<td>87</td>
<td>85</td>
<td>0</td>
</tr>
<tr>
<td>chr12-6E308091-6E309299</td>
<td>1209</td>
<td>mmu-mir-1935</td>
<td>60</td>
<td>85.7</td>
<td>8.28E-08</td>
<td>55</td>
<td>52</td>
<td>0</td>
</tr>
<tr>
<td>chr14-55697945-55699546</td>
<td>1602</td>
<td>mmu-mir-1935</td>
<td>60</td>
<td>69.9</td>
<td>6.14E-06</td>
<td>55</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>chr19-43008931-43008045</td>
<td>1115</td>
<td>mmu-mir-1935</td>
<td>60</td>
<td>87.7</td>
<td>6.09E-08</td>
<td>60</td>
<td>56</td>
<td>0</td>
</tr>
<tr>
<td>chr19-16227992-16228149</td>
<td>358</td>
<td>mmu-mir-1935</td>
<td>60</td>
<td>73.8</td>
<td>1.66E-06</td>
<td>53</td>
<td>49</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 4-7: E18 intronic TARs with hits in lncRNAdb

<table>
<thead>
<tr>
<th>Query</th>
<th>Query Length</th>
<th>Subject</th>
<th>Subject Length</th>
<th>BitScore</th>
<th>E-value</th>
<th>Alignment Length</th>
<th>Matched</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr6-49205737-49207540</td>
<td>1894</td>
<td>B2 SINE RNA</td>
<td>177</td>
<td>311</td>
<td>1.22E-37</td>
<td>177</td>
<td>172</td>
<td>0</td>
</tr>
<tr>
<td>chr7-75179481-75180102</td>
<td>1522</td>
<td>B2 SINE RNA</td>
<td>177</td>
<td>165</td>
<td>1.56E-18</td>
<td>142</td>
<td>128</td>
<td>2</td>
</tr>
<tr>
<td>chr11-104173757-104179094</td>
<td>5338</td>
<td>B2 SINE RNA</td>
<td>177</td>
<td>301</td>
<td>7.33E-36</td>
<td>176</td>
<td>170</td>
<td>0</td>
</tr>
<tr>
<td>chr11-24004772-24006928</td>
<td>2157</td>
<td>B2 SINE RNA</td>
<td>177</td>
<td>272</td>
<td>1.98E-32</td>
<td>177</td>
<td>168</td>
<td>1</td>
</tr>
<tr>
<td>chr17-5104261-5111031</td>
<td>6771</td>
<td>B2 SINE RNA</td>
<td>177</td>
<td>103</td>
<td>5.58E-10</td>
<td>104</td>
<td>91</td>
<td>0</td>
</tr>
<tr>
<td>chr19-40419656-40421427</td>
<td>1772</td>
<td>B2 SINE RNA</td>
<td>177</td>
<td>311</td>
<td>1.22E-37</td>
<td>177</td>
<td>172</td>
<td>0</td>
</tr>
</tbody>
</table>
Few intronic TARs present in miRbase and lncRNAdb

We then compared our intronic TARs in miRNA database miRbase [266] and long non-coding RNA database lncRNAdb [267]. Although we found no significant hits in these two databases for any intronic TARs observed at P7 stage, we did find 12 and 6 hits for intronic TARs at E18 stage in miRbase and lncRNAdb, respectively (Table 4-6 and 4-7). However, all 6 intronic TARs with hits in lncRNAdb were mapped to the same lncRNA, B2 SINE RNA, which was from a SINE repeat element. In addition, 11 of the 12 intronic TARs having hits in miRbase mapped to the same miR-1935 miRNA, and the remaining one mapped to miR-153-2. Otherwise, we did not detect significant hits for other types of RNAs.

Cases of high sequence conservation and coding potential of intronic TARs

To obtain clues about possible function of the intronic TARs using sequence similarity to other mammalian genes, we investigated whether unannotated TARs corresponded to any highly conserved region using the 30-Way Multiz Alignment & Conservation track in UCSC Genome Browser [268]. We found that there were 554 and 168 unannotated TARs at E18 and P7 stages, respectively; among these, 67 in E18 and 21 in P7 matched regions of highly conserved sequences. For example, a TAR on chromosome 15 (102324092-102324772) was localized to an intron of the mouse PCBP2 gene encoding the major cellular poly(rC)-binding protein [269]. In addition, there were RNA-seq reads spanning this intronic TAR and the upstream exon (Figure 4-9B, red reads), indicating that this previously unannotated TAR was spliced with a known exon. Moreover, this TAR had a significant overlap with a highly conserved region located in the 3’ most intron, which was identified by mammalian conservation study using 30-Way Multiz
Alignment & Conservation track data (Figure 4-9C and D) [268]. An Open Reading Frame (ORF) was also predicted inside this TAR and was conserved among the PCBP2 genes of human and dog (99% similar in amino acid sequences; Figure 4-9E), but not opossum. PCR and ABI 3730 resequencing results further verified that this TAR is indeed part of an mRNA (Figure 4-9F) with a connection between this TAR and the upstream exon, consistent with RNA-seq results. However, PCR product between this intronic TAR and the downstream exon was not detected, in agreement with the RNA-seq results. This TAR was very likely to represent an alternative 3’ UTR with a potential coding region.

In addition, an intronic TAR with an ORF inside the ATP2B1 gene located on chromosome 10 (98481907-98482067) shares 99.3% identity to the 20th exon of human ATP2B1 isoform a (ATP2B1a). Human ATP2B1 has two splicing variants: ATB2B1a and ATB2B1b, which differ in the usage of the 20th exon. Previous studies showed that ATP2B1a has a specific expression at synapses whereas ATB2B1b is expressed in most tissues [269, 270]. Thus this TAR is likely to encode a neuron-specific exon of the mouse ATP2B1 gene. We also found another expressed region on chromosome 7 (112781296-112781396) that shares 87.5% identity with a part of the second 3’ UTR exon of the human Trim3 gene. Trim3 (or BERP) is expressed in the brain and encodes a RING finger protein that regulates GABAR cell surface expression [271]. Another intronic TAR located in the NRXN1 gene on chromosome 17 (90854147-90854636) has the potential for coding Neurexin 1, a neuronal cell adhesion molecule interacting with neuroligins to promote synapse formation and maturation [272]. The ORFs in these intronic TARs were highly similar to parts of human ATP2B1, BERP and NRXN1 genes, respectively. A number of other intronic TARs, such as those in CHD3, TSC22, and SRCAP, were either similar to known human exons or supported by mouse gene predictions and mRNA and/or EST data in the NCBI database.
Three other intronic TARs were located, respectively, in the Zeb2 gene on chromosome 2 (44953049-44955802), the Ntrk3 gene on chromosome 7 (85484006-85485464), and the Odz2 gene on chromosome 11 (36491704-36492013), within introns that are more than 10 kb long. These TARs did not match mRNA or EST records in the NCBI database, nor were they similar to protein sequences in the NCBI database. Nevertheless, these three TARs were conserved in rat, human, dog and opossum genomes, matching annotated introns in the orthologous genes in human and rat. Our RNA-seq data did not detect physical connection between the TARs and known exons; the lack of connection between the TARs and the flanking exons were further supported by the observations that PCR was successful when both primers were located inside a particular intronic TAR, but not able to generate products when a primer in the intronic TAR region was combined with another primer in one of the flanking exons (Figure 4-9F). As a control, the correct PCR product was obtained using primers matching the two flanking exons of the given intronic TAR (Figure 4-9F). Therefore, these three intronic TARs were most likely standalone transcripts that were not linked with the flanking exons.

**Association of intronic TARs in neurodevelopmental processes**

Although we have found that few of these intronic TARs were physically connected with the exons of the corresponding genes in the same mRNA, the fact the same genomic regions encode both the transcripts with exons and the intronic TARs suggests some association between these intronic TARs and the exonic genes. To further examine the functional implication of the intronic TARs, it is informative to study the corresponding genes.
Figure 4-10: Neural signaling GO terms significantly overrepresented by genes with intronic TARs.

The coloring of the node box indicates the statistical support for the overrepresentation. The color scale at the bottom of the figure illustrates the different levels of statistical support. The node box with white color means that that node was not statistically overrepresented. However these are included to illustrate the relationship between statistically significant nodes.
Figure 4-11: Neurogenesis GO terms significantly overrepresented by genes with intronic TARs.

The coloring of the node box indicates the statistical support for the overrepresentation. The color scale at the bottom of the figure illustrates the different levels of statistical support. The node box with white color means that that node was not statistically overrepresented. However these are included to illustrate the relationship between statistically significant nodes.
Figure 4-12: Regulation GO terms significantly overrepresented by genes with intronic TARs.

The coloring of the node box indicates the statistical support for the overrepresentation. The color scale at the bottom of the figure illustrates the different levels of statistical support. The node box with white color means that that node was not statistically overrepresented. However these are included to illustrate the relationship between statistically significant nodes.
Table 4-8: Exonic and intronic RNA-seq read number comparison between E18 and P7 for exon guidance related genes.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Number of detected reads in E18</th>
<th>Number of detected reads in P7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>From Exon</td>
<td>From Intron</td>
</tr>
<tr>
<td>Slit3</td>
<td>43</td>
<td>1012</td>
</tr>
<tr>
<td>Robo2</td>
<td>2123</td>
<td>4900</td>
</tr>
<tr>
<td>Robo1</td>
<td>1569</td>
<td>1872</td>
</tr>
<tr>
<td>Nrp1</td>
<td>991</td>
<td>1688</td>
</tr>
<tr>
<td>Nrcam</td>
<td>775</td>
<td>1150</td>
</tr>
<tr>
<td>Klf7</td>
<td>1032</td>
<td>826</td>
</tr>
<tr>
<td>Gli3</td>
<td>295</td>
<td>133</td>
</tr>
<tr>
<td>Ephb2</td>
<td>697</td>
<td>845</td>
</tr>
<tr>
<td>Ephb1</td>
<td>2006</td>
<td>1773</td>
</tr>
<tr>
<td>Dcc</td>
<td>1294</td>
<td>6509</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Read Ratio (Intron/Exon)</th>
<th>3.860</th>
<th>1.802</th>
</tr>
</thead>
</table>

| Intrinsic Read Ratio (E18/P7) | 3.67 |
| Exonic Read Ratio (E18/P7)   | 1.61 |

| Myc  | 358 | 13  | 229 | 8   |
| Actb | 8847| 144 | 8236| 503 |
| Tuba1a| 4029| 63  | 3724| 214 |

<table>
<thead>
<tr>
<th>Read Ratio (Intron/Exon)</th>
<th>0.023</th>
<th>0.051</th>
</tr>
</thead>
</table>

| Intrinsic Read Ratio (E18/P7) | 0.30 |
| Exonic Read Ratio (E18/P7)    | 1.09 |

To study the nature of proteins encoded by the genes with intronic TARs, we analyzed their enrichment in Biological Process Gene Ontology (GO) using agriGO [273]. The GO annotations for all expressed genes at either stage were used as a reference for comparison to determine possible enrichment of specific GO categories. A total of 316 unique genes contained
filtered intronic TARs among a total of 10657 genes with detected reads at E18, while only 119 genes contained filtered intronic TARs among 10901 genes expressed at P7. E18 data had 59 statistically overrepresented GO terms, but P7 data only had 4 statistically overrepresented GO terms, 3 of which are shared between these stages. The 60 overrepresented GO terms could be mapped onto only 3 major branches: neural signaling (Figure 4-10), neurogenesis (Figure 4-11) and regulation (Figure 4-12). The GO terms of these branches are closely related with neural developmental events occurring at E18 stages.

For neural signaling related GO terms at E18, two subgroups form largely parallel interactions: the first subgroup mainly functions in regulating neural system process, while the second subgroup carries out signal transmission (Figure 4-10). The terminal node was the regulation of synaptic transmission, which combines the aforementioned two functions. Among the nodes with strong statistical support were “transmission of nerve impulse”, “neurological system process”, and “synaptic transmission”. Another interesting aspect is that the regulatory relations in the first subgroup were positive, while they were negative in the second subgroup. For neural signaling related GO terms at P7, however, only one function group, similar to the aforementioned 2nd subgroup, was identified. It had 3 nodes and the regulatory relations between the nodes were negative (Figure 4-10).

Nearly half (28, all at E18) of the enriched GO terms were for neurogenesis, with extensive interconnections between nodes and no obvious functional subgroups (Figure 4-11). The functions of these nodes include many aspects of neural development, such as cell morphogenesis, neurogenesis and neuron differentiation, eventually diverging into two termini: (1) axongenesis and axon guidance and (2) dendrite morphogenesis. It is also striking that all regulatory relationships were negative.
As shown in Figure 4-11, the third major group of overrepresented GO terms at E18 was for regulation. They were mainly about negative regulations, consistent with the negative regulatory relations identified between the majority of the nodes for neural signaling and neurogenesis groups. Specifically, strong statistical support was found for negative regulation of “metabolic process”, of “gene expression” and of “biosynthetic process”, ending in that of RNA polymerase II-dependent transcription. For P7, only one node received strong statistical support: regulation of molecular function (Figure 4-12).

Because axon guidance is critical for correct formation of the neural circuit during neural development, we then further analyzed the axon guidance node at E18. A total of 10 genes with filtered intronic TARs were assigned to this node by GO. According to the KEGG Pathway database [274], among the 10 genes, Robo1, Robo2, Nrp1, Dcc, Ephb1 and Ephb2 were all receptors involved in axon guidance pathway. They were all involved in the regulation of the cytoskeleton dynamics and axon repulsion activity. The average ratio of intronic read number to exonic read number for these 10 genes was 3.9 at E18, but was 1.8 at P7 (Table 4-8). As a comparison, we also examined three well-known genes: Myc, β-actin (Actb) and tubulin (Tuba1a), and found that their expression levels (exonic reads) were very similar between E18 and P7. However, the average ratios of intronic read number to exonic read number for these three genes were 0.023 and 0.051 at E18 and P7, respectively. These ratios were less than 1/100 (E18) or 1/30 (P7) of those for the 10 axon guidance associated genes. From a different perspective, for the 10 axon guidance associated genes, the average exonic reads ratio (E18/P7) was 1.61, suggesting a slight reduction in expression. In contrast, while the average intronic reads ratio (E18/P7) was 3.67, representing a much bigger reduction of the intronic transcripts. As a reference, the 3 house keeping genes had an average ratio of 1.09 for E18/P7 exonic reads,
whereas the intronic read numbers for the 3 genes were too low to compare accurately. Therefore, the 10 axon guidance-related genes had significantly more intronic reads than exonic reads (P-value = 0.001, Chi-square with Yates correction) and more so at E18 than P7 (P-value = 0.001), suggesting a possible role of these intronic transcripts in modulating axon guidance at E18 cortex.

**Materials and Methods**

**RNA-seq mapping**

RNA-seq data for hESC, N1, N2 and N3 were obtained from NCBI Sequence Read Archive SRP002079. RNA-seq data for adult mouse brain, liver and muscle tissues were obtained from NCBI Sequence Read Archive SRA001030. RNA-seq data for mouse embryonic day 18 and postnatal day 7 brain cortices were the same as described previously [72]. Its NCBI Sequence Read Archive accession number is SRP007262. The protocol for dissection of the mouse cortex was approved by IACUC committee of Pennsylvania State University and in accordance with the US Federal guidelines. All quality scores were then transformed into FASTQ ASCII code by original quality score plus 64. TopHat was selected for mapping these RNA-seq data. SRP002079 data were mapped onto human genome (UCSC hg19, NCBI Build 37), and the rest were mapped onto mouse genome (UCSC mm9, NCBI Build 37), both with the following parameters: --solexa-qual, -g 1. The same parameters were used when we mapped all nine datasets onto mouse reference genome. Although TopHat was instructed to report only unique hit (-g 1), it sometimes could not fully suppress multiple hits (personal communication with Cole Trapnell, TopHat author, on Feb 22\textsuperscript{nd} 2011). Results were then further screened against RepeatMasker [275] database of the corresponding species to further eliminate possible ambiguous hits.
Data normalization

Normalization was done according to previously published RPKM method [71] with the following adjustment: length normalization was done against chromosome size $L_C$ when we were studying the chromosomal expression level distribution. Also, unlike the original RPKM concept, detected base pair size $C_{bp}$ was used instead of read numbers to accommodate different read-lengths from different RNA-seq datasets (33bp, 35bp and 36bp). For the same reason, total mapped base pairs $N_{bp}$ was used in normalization against data size. And thus the RPKM* label was used here to distinguish these differences:

$$ RPKM^* = \frac{C_{bp} \times 10^{12}}{N_{bp} \times L_C} $$

Data size normalization was done against mappable data size instead of original data size generated from sequencer. This was to accommodate systematic sequence quality and mapping percentage differences from different datasets.

Unannotated transcriptionally active region (TAR) calling

After the RNA-seq data were mapped to the target genome, regions with continuous read coverage that were within close proximity to each other were then chained together, thus forming the transcriptionally active regions (TARs). Only TARs longer than 100 bp and with more than 5X coverage were considered. These TARs were then compared with UCSC Known Gene [276]. TARs that did not overlap with UCSC Known Gene annotation features were then compared with known tRNA annotation [277] and custom-complied rRNA annotation. To further eliminate possible false-positives from repeat, TARs that were not included in any of the above annotations were then mapped back to genome with BLAST [9]. All regions with significant hits elsewhere in
the genome were discarded. The remaining unannotated TARs were then filtered by their distance to their nearest exons. All unannotated TARs that were too close to known exons or genes were discarded as these may originated from previously reported small exon variations [278].

**RT-PCR experiment for intronic TARs**

Sample preparation and RNA extraction were done according to the procedures described previously [72]. For the RT-PCR experiment, total RNA was isolated from mice E18 and P7 cortical tissues by using Ambion RNAqueous-Midi Total RNA Isolation Kit (Catalog#1911). One microgram of RNA was reverse transcribed into cDNA by using Biolabs DyNAmo cDNA Synthesis Kit (Catalog# F-470L).

To validate the expression of several pluripotency-related genes: approximately 1/20 of the first strand cDNAs was used as a template for PCR with gene-specific primers. PCR was carried out for 25 cycles of 94°C for 20s, 54°C for 30s, and 74°C for 40s. 10ul of PCR products was separated on 0.8% (w/v) agarose gels containing ethidium bromide and visualized by UV light. A secondary PCR was performed for P7 with same primers by using 1ul first round PCR products, for 32 cycles.

To validate the expression of specific intronic TARs, primer sequences were chosen within the intronic TAR, between the intronic TAR and the upstream exon, between the intronic TAR and the downstream exon, and between the upstream exon and the down stream exon. RT-PCR was carried out using the cDNAs as template with Taq polymerase for 22 cycles (add the temperature info). PCR product was sequenced at the Genomics Core Facility at Penn State using an ABI 3730 machine.
Modeling connection between unannotated TARs and known transcripts

The general question can be abstracted of how to determine whether a given detected transcriptionally active region (TAR) was on the same RNA with other exons/transcripts using paired-end information, i.e., there was a physical connection between the given TAR and another exon. If a given unannotated TAR is long and has many internal RNA-seq reads, its number of paired-end reads with one end located at a known exon should mean something different that a given unannotated TAR which is short and with comparably less internal RNA-seq reads. We thus propose that the support for the aforementioned physical connection between the unannotated TAR and a known exon should be evaluated as a function of the length of the given unannotated TAR, the coverage (number of internal RNA-seq reads) of the unannotated TAR and the number of paired-end RNA-seq reads linking this unannotated TAR and another known exon. RNA-seq mappers also tend to have a lower mapping capability if they need to map a partial read at the end of an exon or TAR.

We first assumed that the paired-end read distribution inside a given TAR were either in a uniform distribution (in the case of using random primer), or in a skewed distribution (in the case of using ploy-T primer). Under this assumption, if the given TAR was part of a larger transcript, the number of paired-end reads at its end(s) should be similar to the average number of paired-end reads over the entire TAR. The following formula was used to calculate the estimated number of paired-end reads \((Ne)\) at only one end of the given TAR, on the assumption that this TAR was part of a larger transcript:

\[
Ne = \frac{Ni \times Lc}{2 \times Lr} - \frac{T \times Ni \times (Ls - M)}{Lr}
\]  

(41)
Ni, two times the total number of paired-end reads with both ends located inside the given region (to reflect each end in a pair).

Ls, read length for one end of a paired-end read.

Lc, clone length of a paired-end read, which is $2 \times Ls$ plus the insert size.

Lr, length of the given TAR.

M, maximum number of allowed mismatches of the mapping algorithm.

T, a correction factor. Splice junction spanning reads will have two partial matches to two discrete genomic regions. This value represents the success rate of the algorithm in mapping partial reads to the end of a given region, normally between 0 and 1. 0 means the algorithm cannot map partial reads to the end of a given region. 1 means the algorithm can map 100% of the partial reads to the end of a given region. Given the fact TopHat is designed to do RNA-seq mapping, the T value we picked was 0.99.

If data indicated that there was a significant amount of links from the given TAR to both upstream and downstream exons, the Ne should be doubled since there should be reads covering both ends of the given TAR. If the size of the given region Lr was smaller than that of the clone length Lc, then by theory all paired-end reads from this TAR should be reads linking this given TAR with other region(s).

If the actual number of paired-end reads connecting a given TAR with other regions was significantly less than (in this case we used 20%) Ne, the given TAR is thought to be a standalone transcript. Otherwise, this region was inferred as non-standalone, which means some level of splicing activity. More specifically, if a significant portion of the aforementioned paired-end reads had the other ends located in annotated exon(s), this given TAR was thought to be part of a
known transcript. However, if the aforementioned reads were connecting more than one chromosome, then this TAR was thought to be multi-chromosome linked.

Testing model effectiveness

To test the effectiveness of the proposed formula (formula (41)) in determining the connection between a given unannotated TAR and known exon(s), we must have positive controls that are known to be detectable in our dataset and are also known to be on the same RNAs with known exons. An expressed exon from a multi-exon gene would meet this requirement and should be able to serve as our positive control. To ensure that these exons were truly connected to known exons by RNA-seq reads, these exons were selected by hand through manual inspection of the RNA-seq mapping results using Integrative Genomics Viewer (IGV, http://www.broadinstitute.org/igv). For each selected exon, we made sure that there were multiple reads spanning the selected exon with at least one other known exon. To ensure a true representation of the genome wide situation from our test data, these selected exons were picked from different chromosomes, with different RNA-seq read coverages, different locations within a given gene and different relative distances to the 3’ end (Table 4-9). The proposed formula (formula (41)) was able to identify such exons as being physically connected with known exons with a success rate of 93%.

We also did the negative control test to determine the effectiveness of the proposed formula in determining that a given transcript has no physical connection with any known exon(s). Single exon genes (SEGs) in mouse genome were selected as the negative controls since they are known to be a standalone transcript. We first identified a list of SEGs which had RNA-seq reads in our dataset (Table 4-10). The proposed formula was able to identify these selected
SEGs as not being physically connected with any known exon(s) at a success rate of 100% (14% of the selected SEGs were determined, however, as multi-chromosome linked).

**Gene Ontology (GO) analysis**

Reference mouse GO annotation was obtained from the Jackson Laboratory’s MGI site (http://www.informatics.jax.org). Expressed genes were inferred from RNA-seq mapping results mapped to UCSC Known Gene. Expressed genes were then compared with reference mouse GO annotation. Identifier conversion between the UCSC Known Gene and the GO annotation was done using in-house script. Among all GO terms, only Biological Process GO terms were analyzed. We first calculated the number of genes mapped to a given GO term. For a gene with multiple GO terms, all terms were considered because one gene may be involved in multiple biological processes. If one GO term node was counted, all its parental nodes were excluded. Four sets of GO annotation were produced using the aforementioned procedure: all expressed genes in E18, all expressed genes in P7, genes with intronic TAR(s) in E18 and genes with intronic TAR(s) in P7.

For a given stage, GO annotation for the entire transcriptome and GO annotation for only genes containing intronic TARs were compared using agriGO server [273]. The statistical significance was determined by Fisher’s Exact Test, with Bonferroni Correction. The p-value threshold was preset at 0.05 and only GO terms with more than 5 hits were reported.
Table 4-9: Manually validated expressed exons physically connected with other exons

<table>
<thead>
<tr>
<th>Chromosome Name</th>
<th>Start Position</th>
<th>Stop Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr14</td>
<td>79688556</td>
<td>79688927</td>
</tr>
<tr>
<td>chr14</td>
<td>79690050</td>
<td>79690115</td>
</tr>
<tr>
<td>chr14</td>
<td>79691477</td>
<td>79691586</td>
</tr>
<tr>
<td>chr7</td>
<td>25207317</td>
<td>25207797</td>
</tr>
<tr>
<td>chr7</td>
<td>25206137</td>
<td>25206284</td>
</tr>
<tr>
<td>chr7</td>
<td>25209830</td>
<td>25209961</td>
</tr>
<tr>
<td>chr2</td>
<td>130533586</td>
<td>130535067</td>
</tr>
<tr>
<td>chr16</td>
<td>66664625</td>
<td>66664753</td>
</tr>
<tr>
<td>chr16</td>
<td>66661317</td>
<td>66663459</td>
</tr>
<tr>
<td>chr16</td>
<td>7353023</td>
<td>7353902</td>
</tr>
<tr>
<td>chr5</td>
<td>36200474</td>
<td>36200614</td>
</tr>
<tr>
<td>chr5</td>
<td>36192164</td>
<td>36192845</td>
</tr>
<tr>
<td>chr12</td>
<td>81268532</td>
<td>81269459</td>
</tr>
<tr>
<td>chr12</td>
<td>81269844</td>
<td>81270003</td>
</tr>
<tr>
<td>chr12</td>
<td>81272714</td>
<td>81272863</td>
</tr>
<tr>
<td>chr7</td>
<td>29681107</td>
<td>29681254</td>
</tr>
<tr>
<td>chr7</td>
<td>29681893</td>
<td>29682074</td>
</tr>
<tr>
<td>chr7</td>
<td>29677979</td>
<td>29679406</td>
</tr>
<tr>
<td>chr7</td>
<td>29679519</td>
<td>29679679</td>
</tr>
<tr>
<td>chr7</td>
<td>29683087</td>
<td>29683270</td>
</tr>
<tr>
<td>chr5</td>
<td>121975057</td>
<td>121976421</td>
</tr>
<tr>
<td>chr5</td>
<td>121975057</td>
<td>121975250</td>
</tr>
<tr>
<td>chr5</td>
<td>121975670</td>
<td>121975779</td>
</tr>
<tr>
<td>chr5</td>
<td>121976746</td>
<td>121976871</td>
</tr>
<tr>
<td>chr5</td>
<td>121978875</td>
<td>121979001</td>
</tr>
<tr>
<td>chr1</td>
<td>158738249</td>
<td>158739325</td>
</tr>
<tr>
<td>chr8</td>
<td>125411214</td>
<td>125411526</td>
</tr>
<tr>
<td>chr6</td>
<td>28376068</td>
<td>28376500</td>
</tr>
<tr>
<td>chr6</td>
<td>28375737</td>
<td>28375866</td>
</tr>
<tr>
<td>chr6</td>
<td>28374721</td>
<td>28374832</td>
</tr>
</tbody>
</table>

Table 4-10: Manually validated expressed single exon genes (SGEs)

<table>
<thead>
<tr>
<th>Chromosome Name</th>
<th>Start Position</th>
<th>Stop Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr12</td>
<td>17178905</td>
<td>17183331</td>
</tr>
<tr>
<td>chr12</td>
<td>28020080</td>
<td>28027577</td>
</tr>
<tr>
<td>chr5</td>
<td>35621214</td>
<td>35624412</td>
</tr>
<tr>
<td>chr15</td>
<td>84383827</td>
<td>84388253</td>
</tr>
<tr>
<td>chr2</td>
<td>158736503</td>
<td>158739633</td>
</tr>
<tr>
<td>chr7</td>
<td>69493162</td>
<td>69494813</td>
</tr>
<tr>
<td>chrX</td>
<td>119508169</td>
<td>119510994</td>
</tr>
<tr>
<td>chrX</td>
<td>149932774</td>
<td>149936101</td>
</tr>
<tr>
<td>chr3</td>
<td>61168428</td>
<td>61172625</td>
</tr>
<tr>
<td>chr14</td>
<td>109309205</td>
<td>109313456</td>
</tr>
<tr>
<td>chr12</td>
<td>8504564</td>
<td>8506791</td>
</tr>
<tr>
<td>chr4</td>
<td>88364195</td>
<td>88368412</td>
</tr>
<tr>
<td>chr8</td>
<td>73222847</td>
<td>73224515</td>
</tr>
<tr>
<td>chr9</td>
<td>123168174</td>
<td>123169907</td>
</tr>
</tbody>
</table>
Project summary and discussion

Similarity between neural and stem cell transcriptomes

In this study, we have investigated global characteristics of embryonic and neonatal neural transcriptomes, and compared with transcriptomes of the adult brain and embryonic stem cells. We found that embryonic and neonatal brain cortex transcriptomes correspond to most genomic regions at large scale of megabase intervals, but are unevenly distributed with positive correlation to exon density. In addition, neural transcriptomes are similar to that of embryonic stem cells, more than those of liver and muscle, in several features including chromosome level expression, and expression pattern of orthologous genes. Also, the E18 brain cortex transcriptome and hESC transcriptome showed relatively even chromosomal distribution and had lower mitochondrial expression.

Other than these global similarities, we noted another shared characteristic between neural expressed genes and genes important for pluripotent stem cells. Specifically, three genes, Sox2, Myc and Klf4 were detectable in all six neural samples, with high levels in E18 or P7 transcriptomes. The expression of these genes suggests that neural cells might need fewer factors to be converted to stem cells. Indeed Kim et al found that only two factors (Oct4 with either Klf4 or c-Myc), instead of four, were needed to revert neural stem cell to iPS cells [279]. Therefore, the similarity in transcriptome, including the expression of specific genes, such as Sox2, Myc and Klf4, between neural cells and stem cells suggests that neural cells might retain certain stem cell properties and have greater potential to be reprogrammed to be pluripotent.
Intronic TARs as regulators in early brain development

The mapping results of our transcriptome datasets revealed significant levels of intronic reads. We found that only a small portion of the intronic transcripts that we detected was on the same RNAs with any known exons. Recently, Klevebring et al. have reported that about 50% of the intronic expression was actually from the antisense strand [239], different from the sense exon-containing mRNAs of the same gene. Thus the intronic TARs detected here share some characteristic with the antisense transcripts, although our data lacked the strand information. Our finding that the level of intronic transcript is positively correlated with that of flanking exons is consistent with previous studies that antisense transcription may have both concordant and discordant regulation relative to the adjacent exons [233]. Furthermore, Faghihi et al. also reported regulation involving antisense RNA not mediated by the conventional RNA interference pathway [280], indicating additional mechanisms are important. Our results that the E18 brain cortex has significantly higher levels of intronic transcripts than other tissues/organs strongly suggest that such non-coding transcripts play important roles in regulating gene expression during embryonic brain development.

We also found that the mouse E18 embryonic brain had a concordant relation between intronic transcript and flanking exonic expression. This is unlike previous studies showing preferential localization of antisense transcripts in the upstream and downstream regions of the gene [281-283]. Our data have further indicated that the E18 embryonic brain showed enrichment of genes with intronic TARs in GO categories that are closely associated with neural functions. The E18 cortical neurons are actively engaged in neurogenesis, including axongenesis and synaptogenesis. For the significant GO terms associated with neurogenesis, all the regulatory relations between nodes were negative (Figure 4-11). Moreover, an entire group of significant
nodes was about negative regulation (Figure 4-12). However, at P7, intronic TARs were no longer associated with either neurogenesis or negative regulation. These findings suggest the involvement of intronic TARs in stage-specific regulation of neural developmental.

Recently, a subset of long ncRNAs was found to have an enhancer-like function [284]. Our data also indicated a correlation between the change in intronic transcript expression level and the change in the expression level of the corresponding gene. For example, the increased intronic expression is correlated with increased exonic expression for 10 axon guidance associated genes, whereas such correlation was not found for a control set of 3 housekeeping genes. The positive correlation in expression change between intronic TARs and the flanking exons further supports the idea that they have regulatory interactions, although it is formally possible that the intronic transcripts have functions unrelated to the genes represented by the flanking exons.

Our transcriptome analyses have revealed possible important mechanisms of gene function and regulation in developing brain, and uncovered a strong similarity between neural tissues and stem cells in terms of global transcriptional activities. These results provided novel insights regarding neural developmental gene functions that can be further investigated using molecular genetic, biochemical and electrophysiological experiments.

**Time-sensitive sampling and its application in revealing time-dependent complexity**

Current transcriptome collection methods all depend on RNA extraction and reverse-transcription. The initial sampling normally pools multiple cells or even tissues together. This approach not only homogenizes across different cells of different types and locations, but also
homogenizes different cells at different time points and developmental stages. We could call the former spatial homogenization and the latter temporal homogenization. The spatial and temporal homogenization makes the sampled transcriptome a mixture. This homogenization could be avoided by only sampling one single cell. The recent development of single cell sequencing technology has made this possible [285]. However, inside a living cell, the speed of RNA synthesis (RNA polymerase II) was estimated to be 33 nucleotides per second [286]. Thousands of genes are being transcribed to RNAs and possibly similar amounts of RNAs are being degraded at any moment inside a single cell. We could only assume that the sampled single cell is at a stable state, but far more likely the sample is temporally homogenized.

The entire human genome could only encode maximally 6.4 gigabits of information. In reality, there are numerous regions inside the genome without known activities or functions, for example, heterochromatin regions. It is estimated that only about 1.5% to 2% of the genome is protein coding. That gives us about 10 to 12 megabits of information content. It only equals less than 1.5 megabytes of disk storage space, or less than half of a typical 5-minute song in mp3 compression format. However, considering the number of surface molecules a cell has, the different signaling pathways in a cell and the possible types of cell-cell interactions, without mentioning the numerous functions a cell could perform, the information complexity of the simplest human cell should be much larger than the above number (10 to 12 megabits).

As mentioned at the beginning of this chapter, the complexity of the transcriptome augmented the genome complexity. In addition, the transcriptome complexity is time dependent. Given a transcript $X$, its function $F(X)$ could be conditioned upon time $t$. This means that the function of $F(X|t_1)$ might be different from $F(X|t_2)$. If we use $T_t$ to denote the transcriptome at time $t$, similarly there could be $F(X|T_{t_1})$ and $F(X|T_{t_2})$. If $T_{t_1}$ and $T_{t_2}$ are different, $F(X|T_{t_1})$ and
$F(X|T_{t_2})$ are possibly quite different given the mechanisms of molecular regulation inside a cell, although the transcript $X$ is the same. For example, in the classic ABC flower development model, class B genes together with class A genes will induce the development of petals, but class B genes together with class C genes will induce the development of stamens [287]. The complexity of the transcriptome is thus the sum of different transcriptomic complexities at different time points, or the integral as the time is continuous. This means the complexity is time-dependent. The stage specific transcriptome analysis and the comparison of such analyses across many stages, like in this study, help to understand this time-dependent complexity in living organisms and their processes.

(End of Chapter 4)
Chapter 5

Short-take: DNA-based protein sequence retrieval in unannotated genomes

The results in this chapter have been included in a paper to be submitted for publication.

Project background

In late 2009, then graduate student Xiaofan Zhou (received PhD in 2011) in my research group led by Professor Ma was trying to identify genes that are orthologous or strictly single-copy among all major organismal lineages. Genes that are single-copy in most species are crucial to determine correct organismal phylogenies. However, his efforts were hindered by the fact that most newly sequenced genomes have no annotation information or the released annotation information contain too many errors, as explained in Chapter 1.

Introduction

Challenges of genome annotation in the era of ultra high-throughput sequencing

According to the statistics from Genome Online Database [288] released on February 2010, there were 1196 finished genomes before that date. There were additional 5201 ongoing genome projects. In addition, the Genome 10K Project has also planed to sequence 10000 animal genomes in the near future. The explosively increasing number of finished genomes and ongoing genome projects is largely credited to the rapid technological advancement in DNA sequencing
technologies. As of late 2011, the data throughput had reached 55Gb per day (Illumina HiSeq2000) while the cost had reduced to below 10 Mb per dollar. Genomes that could only be completed through international collaborations and the involvement of thousands of researchers in the past are now feasible for a single research lab with just a few members. Genome sequences are now being produced at an unprecedentedly fast speed. However, the speed of genome annotation has not been significantly improved during the same time.

Genomes generated in the past decade are exclusively based on shotgun sequencing principle, which is sequencing through random sampling and parallelization. This is unlikely to change in the immediate future, as the latest DNA sequencing technologies are all based on the same principle to achieve higher data yield. These shotgun sequencing fragments, or reads, were assembled by automated genome assemblers. The assembled genomes were then annotated through automated genome annotators. Some widely used algorithms and tools for genome assembly and annotation have been introduced in Chapter 1. Due to the large amount of sequencing data, the possibility of manual curation of genome assembly and annotation is limited. Sequencing errors could contribute to assembly errors, and then could contribute to annotation errors. This cascading effect causes the accumulation of errors in the final annotated genome. In addition, existing profiles of annotation tools that were parameterized based on only a few well-studied model organisms might not perform well on the newly sequenced species. All these have detrimental effects on the quality of the annotation.

**Method of retrieving a certain gene or gene family based on homology in genomes**

Biologists who are familiar with certain genes or gene families might not be able to use many of the newly completed genome sequences because of either lack of genome annotation or
errors in annotation. In this regard, the problem of interest is that given a set of proteins, presumably all belonging to a single family, how to retrieve their close relative(s) from an unannotated genome. The traditional approach to this problem is to first annotate the given genome, and then through profile matching, find out the close relative(s) among all the newly annotated genes.

However, a full genome annotation is not cost effective if the biologists are only interested in only a few proteins. This is because, first, genome annotation normally requires fine-tuning the gene prediction model used for the given genome. Without transcriptome data or a large set of previously studied genes from the given species, the annotation process is actually biased by the available knowledge, which is normally not from the same species or even close to the given species. Second, for any moderately sized genome, the annotation process can take days if not weeks. Third, the errors accumulated from genome sequencing and assembly processes will have detrimental effects on the correctness of the annotation.

Given the above, it is desirable to only and directly find out the close relative(s) of the given proteins. This involves a direct search for the homologues of a given gene or gene family inside the genome. The simplest solution is to search for homologous regions inside the genome using local alignment tools, for example BLAST [9]. However, as we have discussed in Chapter 1, the boundaries of exons are not precisely determined in this manner and small exons are easily missed. Tools that are capable of rapid searching for homologous regions and at the same time accurately determining exon boundaries are needed for this purpose.


**Current homologue retrieving tools and their limitations**

Several tools have been specifically developed for this purpose. Among them, TARGeT [124] and FGF [122] are web-based tools, while GFscan [119] and GeneWise [121] are local command-line tools. GFscan is among the first tools to address this question directly. However, it is not widely used anymore mainly due to lack of support. Although introduced at the same time period, GeneWise has a more sophisticated and generalizable gene prediction model. GeneWise has better support and execution as well, so it is still widely used today. TARGeT uses BLAST [8, 9] directly as the main homology search method. But it has a simpler *ad hoc* algorithm for gene prediction. FGF also uses BLAST for homology search but it directly depends on GeneWise for gene prediction.

TARGeT and FGF only provides web-based access. Only selected genomes are available through their web interfaces. These two interfaces are not designed for high-throughput large-scale analysis, which normally involves thousands of families, with some of the families containing hundreds of members. In comparison, GeneWise is local and can be integrated into pipelines for large-scale analysis. However, GeneWise hasn’t implemented any mechanisms optimized for fast large-scale search. This makes GeneWise time and memory intensive for genome wide search.

Due to the limitations of the currently available tools, we built our own program called Phoenix. It can directly retrieve homologues based on input from a genomic sequence. It employed the same principle and two-step process as in FGF, which combines proven tools for both homology search and gene prediction.
Analyses and Results

Phoenix algorithm for homologue retrieval from genomic DNA sequence

First we assume that there are \( K \) genes in the given family, and \( k \) is one of the genes. In total there will be \( N \) new members of this given family in the given unannotated genome, and \( n \) is one of the putative members, i.e., a putative gene. \( S \) is the score for a significant match between a stretch of symbols (amino acids) from \( k \) and a stretch of symbols from \( n \). Obviously, \( S \) can be found out by TBLASTN search. By definition [9], the bitScore of a high-scoring segment pair (HSP) will be \( S \) and \( S \) is additive [289, 290]. Each \( k \) could give multiple \( S \)s for a single \( n \), represented by \( S^i_k \). \( S^i_k \) could arise from different parts or domains of \( k \). Although each part of a single protein contributes to a given putative gene equally, independent of its conservation, the conserved regions from the given family would contribute much more \( S \)s to a true putative gene than unconserved regions. This ensures proper weighting assigned to the conserved regions, which in turn preserves the given family’s signature. A stretch of DNA sequence’s potential for being a new member could then be measured. Collective support from the entire given gene family for \( n \), represented by \( P_n \), can be written as:

\[
P_n = \sum_k \sum_i S^i_k \tag{42}
\]

Thus \( P_n \) is a measure of the potential of the putative gene \( n \) (Figure 5-1a).

Several nearby \( S \) could be an indication of a putative gene. However, deducing the boundary of each putative gene from \( S \) is problematic, because intron length variation is quite significant both within a species and between different species. Extremely long introns among
higher eukaryotes also complicate the issue. Parameterization after the intron length distribution from known data would decrease the applicability of the model for other species, especially for newly sequenced species with intron length distribution greatly deviating from known ones. Since gene size is of several magnitudes smaller than genome size, and GeneWise operates comparatively fast at the gene-size level, it is thus reasonable to generously include regions flanking the actual gene, on the order of gene-size. GeneWise then could locate the precise exon boundaries. So we only need to select a distance that is large enough to break far away HSPs into different putative genes. An upper bound of the exon-exon distance distribution would be the obvious choice, however, the 95\textsuperscript{th} percentile would suffice to include most of the situation without adding unnecessarily large flanking regions and costly GeneWise running time. For example, the 95\textsuperscript{th} percentile of the exon-exon distance distribution in mouse is 19360 bp, while the 98\textsuperscript{th} and 99\textsuperscript{th} percentiles are 42407 and 62100 bp, respectively.

Due to large genome size and presence of remote homology, a putative gene with small $P_n$ is not necessarily a homologue to the given family, or even not necessarily a gene. Such small $P_n$ may come from only a few short HSPs or several HSPs with very small bitScores. Given that the genome remains the same for each $P_n$, these $P_n$ s could be compared directly. We first need to differentiate the high $P_n$ s from the low ones. The $P_n$ s are first sorted in descending order. A curve is then fitted to the distribution profile. If the inflection point of the fitted curve is present, the first (largest) inflection point will serve as the separation point (threshold $T$) between high and low $P_n$ s. If there is no inflection point for the fitted line, the following heuristic is applied: first we calculate the average for the top $b$ number of $P_n$ s, then a percentage is taken from the average. Any $P_n$ s larger than this percentage will be considered as high $P_n$ s, while anything smaller than this would be low ones. Thus the threshold $T$ in this case can be written as:
\[
T = \frac{a}{b} \sum_{i=1}^{4} P_i (P_i > P_{i-1} > P_{i+1})
\]

So \( a \) represents the percentage factor, and \( b \) stands for the number of the highest \( P_i \) value examined (Figure 5-1b). With the above method, the separation point between high and low \( P_i \) s is input-data dependent. This is very important for the convergence analysis that is discussed later.

In order to use GeneWise to predict a putative gene, a reference protein or a profile HMM needs to be provided. Although a profile HMM can be easily generated from the input gene family, using a protein from the given family that is the closest to the putative gene in question is more appropriate for this problem. For instance, the proteins in the given family might come from only a few well-studied species, and these species might not be a good representation of all available species. Proper weighting of the proteins in the given family may give a more even distribution, but selecting the one closest to the putative genes is more desirable. To select the closest reference protein is quite easy given the above procedure that is used to find out these putative genes. For each given protein, its contribution to each putative gene can be calculated by simply summing up all the bitScores of the HSPs that it provided to that putative gene. The putative gene with the highest sum would be the closest one to this given protein, and thus this given protein would be the reference for that putative gene (Figure 5-1c).

**Obtaining converging results using Phoenix**

Once the new homologues of the given gene or gene family are retrieved, the collection of this gene family is updated with the new members. It is thus logical to search for potential new members using the updated collection of sequences. This could be repeated until no further new
members can be retrieved, which means the input and output converge (convergence result). This means the search has reached a local maximum.

The dynamically determined threshold $T$ described previously ensures that the results won’t grow indefinitely. With our testing data, the convergence result could be achieved usually within three iterations.

Figure 5-1: Phoenix program algorithm.

Rapid and accurate JmjC domain family homologue retrieval by Phoenix
In order to assess the performance of this algorithm and the corresponding Phoenix software, we tested it on a set of manually curated Jumonji C (JmjC) domain-containing proteins. JmjC domain-containing proteins are important histone demethylases that were recently discovered to be able to reverse histone methylation in mammalian cell lines [291]. This family of genes is conserved from yeast to human [291]. Zhou and Ma found that JmjC domain was present in the ancestors of all major eukaryotic groups and experienced dynamic gene duplication and loss events along different lineages [292]. The JmjC domain-containing proteins are ideal candidate for testing Phoenix’s performance because of their variable conservation and copy numbers in different eukaryotes. Using the 58 JmjC domain-containing genes from the study by Zhou and Ma as reference, the JmjC domain-containing proteins were manually selected from ten representative eukaryotic genomes, ranging from animal to plant, from unicellular to multicellular. The ten genomes are from Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens, Oryza sativa, Physcomitrella patens, Populus trichocarpa, Saccharomyces cerevisiae, Schizosaccharomyces pombe and Selaginella moellendorffii. Since the JmjC genes were only recently discovered, all existing annotations for these ten genomes were not considered. This manual curation process retrieved 189 potential members of the JmjC family from these ten genomes. The same 58 reference JmjC genes and ten genomes were used to test Phoenix program. Phoenix was able to retrieve 188 members. After careful comparison, it was found that Phoenix missed only 2 members previously manually curated. Interestingly, the manual curation process also missed one member that was picked up by Phoenix software (Table 5-1). The success rate for Phoenix is 99%, only slightly lower than manual curation (99.5%). However, the manual curation process took weeks of time, but Phoenix completed within 50 minutes.
We also took a random subset of the 58 reference JmjC proteins (22 sequences, 38% of the original size) as the input for Phoenix. Relying only on 38% of the original input, this software was still able to retrieve 109 potential members (58% of the original) from the same ten genomes. Close homologues of the input proteins were all retrieved successfully as expected. Remote homologues could not be retrieved because of divergence. Moderately or closely related subfamilies without representation in the input can be retrieved in part or in full depending on their phylogenetic distances from the input sequences. The amino acid sequences recovered by neighboring sub-family members were almost the same as those recovered by the same sub-family members in the previous test using all 58 reference proteins.

Table 5-1: Performance of JmjC family homologue retrieval in 10 genomes by Phoenix.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Recovered</th>
<th>Missed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>189</td>
<td>1</td>
</tr>
<tr>
<td>Phoenix</td>
<td>188</td>
<td>2</td>
</tr>
</tbody>
</table>

Complex evolution history of elongation factor 2 gene family revealed using Phoenix

Orthologous genes or near-orthologous single-copy genes are crucial in reconstructing organismal phylogeny. These genes are commonly regarded as universal marker genes. However, a few of the widely used marker genes, for example, eEF1α and α-tubulin, were recently found to have a complex and non-orthologous evolution history [293, 294]. Phoenix program could be
used to search for gene homology and further to check for the orthology and copy numbers. Although sometimes pseudogenes will be recovered as well, these pseudogenes still reflect the history of duplication events. As an essential component of the eukaryotic protein synthesis machinery, \( eEF2 \) gene is a member of GTP-binding translation elongation factor family [295]. In our analysis of the \( eEF2 \) gene family members retrieved by Phoenix, we also observed a non-orthologous evolution pattern in five protist genomes (Figure 5-2). The genomes used here are from *Naegleria gruberi*, *Dictyostelium purpureum*, *Acanthamoeba castellanii*, *Tetrahymena thermophila* and *Paramecium tetraurelia*. Among them, genomes of *Paramecium tetraurelia* and *Acanthamoeba castellanii* are only partially assembled. Some of the \( eEF2 \) genes from *N. gruberi* are grouped together with *D. purpureum*, separated from the clade containing the rest of the *N. gruberi* \( eEF2 \) genes. The latter clade diverged from all the other \( eEF2 \) genes studied here. Similar situations could be observed in *D. purpureum* and *A. castellanii* \( eEF2 \) genes.
Materials and Methods

Genome sequence retrieval

The genomes of *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* were from UCSC Genome Browser database [268]. The genome of *Arabidopsis thaliana* was from The Arabidopsis Information Resource (TAIR) database [296]. The genomes of *Oryza sativa*, *Physcomitrella patens*, *Populus trichocarpa*, *Saccharomyces cerevisiae* and *Selaginella moellendorffii* were from PlantGDB.
The genome of *Schizosaccharomyces pombe* was from the Welcome Trust Sanger Institute. The genomes of *Naegleria gruberi*, *Dictyostelium purpureum* and *Tetrahymena thermophila* were from the Department of Energy Joint Genome Institute (JGI). The partially assembled genome of *Paramecium tetraurelia* was from ParameciumDB [298]. The partially assembled genome of *Acanthamoeba castellanii* was from Baylor College of Medicine.

**Phylogeny reconstruction**

Multiple sequence alignment was done using MUSCLE [30], which is introduced in Chapter 1. Conserved blocks from the resulting alignment were then extracted by Gblocks [299]. Maximum likelihood tree was built based on manually adjusted multiple alignment of the conserved blocks using RAxML [300], with the option GAMMALG and 100 bootstrap resampling.

**Prerequisites for running Phoenix program**

The Phoenix software is written in Java. A Java Runtime Environment (JRE), version 1.5 or later is required. BLAST [9] and GeneWise [121] need to be installed prior to running Phoenix since they are used to perform sequence search and gene prediction functions internally in Phoenix. The protein sequences of the gene family in question should be input in FASTA format. The genome sequences should also be in FASTA format. PhoenixPrep, a small java program distributed along with Phoenix, is needed to prepare the genome sequences and format them into NCBI blastall recognizable databases.
Phoenix program instructions and examples

The parameters for Phoenix could be provided either from command line or through a configuration file. A sample configuration file is provided in Table 5-3. If using a configuration file, the program could be invoked through a very simple command line argument (Table 5-4).

Table 5-3: Sample configuration file for running Phoenix

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INPUT_GENE FAM</td>
<td>/path/to/FILE/contain/query/protein/fasta</td>
</tr>
<tr>
<td>SUFFIX ID</td>
<td>IDofYourChoice</td>
</tr>
<tr>
<td>WORK_DIR</td>
<td>/path/to/DIR/for/output</td>
</tr>
<tr>
<td>BLAST_DIR</td>
<td>/path/to/DIR/contain/ncbi/blastall/app</td>
</tr>
<tr>
<td>NUMPROC</td>
<td>1</td>
</tr>
<tr>
<td>DATABASE_DIR</td>
<td>/path/to/DIR/contain/formatted/database</td>
</tr>
<tr>
<td>WISECONFIG_DIR</td>
<td>/path/to/DIR/contain/gene wise2/config/files</td>
</tr>
<tr>
<td>WISEAPP_DIR</td>
<td>/path/to/DIR/contain/gene wise2/app</td>
</tr>
<tr>
<td>WISE_PSEUDO</td>
<td>T</td>
</tr>
<tr>
<td>EXPECT VAL</td>
<td>1e-5</td>
</tr>
<tr>
<td>MIN_GENEDIST</td>
<td>20000</td>
</tr>
<tr>
<td>PERC_CUTOFF</td>
<td>50</td>
</tr>
<tr>
<td>TOP PERC IN</td>
<td>25</td>
</tr>
<tr>
<td>LENG_CUTOFF</td>
<td>50</td>
</tr>
<tr>
<td>PGENE_EXT</td>
<td>10000</td>
</tr>
</tbody>
</table>

Table 5-4: Phoenix command line argument with a configuration file

```
java -jar phoenix.jar -c configurationFileLocation
```

The details of these parameters are provided in the user’s manual distributed along with the program.

Using configuration file could simplify input for running Phoenix repeatedly or incorporating Phoenix into an analysis pipeline. However, all the parameters could be provided through command line arguments instead. The command line arguments that Phoenix recognizes are list in Table 5-5. The relation between configuration file parameters and command line
arguments are given in Table 4-6. However, if a configuration file is provided, the configuration file parameter setting takes precedence over command line arguments.

Table 5-5: Command line arguments of Phoenix

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-i</td>
<td>Path to a file which contains the interested gene family's protein sequences in FASTA format</td>
</tr>
<tr>
<td>-s</td>
<td>An ID defined by user to provide additional information in the names of the putative genes when they are produced, example: YS, [default value: YS]</td>
</tr>
<tr>
<td>-w</td>
<td>Path to a directory where the output files will be placed</td>
</tr>
<tr>
<td>-c</td>
<td>Path to a file which contains the controlling parameters of this program in format VAR_NAME=VALUE</td>
</tr>
<tr>
<td>-b</td>
<td>Path to a directory which contains the NCBI blastall executable</td>
</tr>
<tr>
<td>-d</td>
<td>Path to a directory which contains the interested/to-be-searched databases formatted by NCBI formatdb program</td>
</tr>
<tr>
<td>-a</td>
<td>Number of processors the NCBI blastall executable will use</td>
</tr>
<tr>
<td>-f</td>
<td>Path to a directory which contains the control files of the EMBL-EBI, i.e. $WISECONFIGDIR mentioned in the genewise/Wise2 manual</td>
</tr>
<tr>
<td>-g</td>
<td>Path to a directory which contains the genewise executable</td>
</tr>
<tr>
<td>-P</td>
<td>Boolean value to instruct genewise program to operate under -pseudo mode, acceptable value: T (in -pseudo mode) or F (not in), [default value: T]</td>
</tr>
<tr>
<td>-e</td>
<td>A Scientific Notation expect value threshold used by blastall/tblastn program, example: 1e-5, [default value: 1e-5]</td>
</tr>
<tr>
<td>-m</td>
<td>An Integer value used by this program to determine whether two neighboring HSPs are from the same gene, example: 20000, [default value: 20000]</td>
</tr>
<tr>
<td>-p</td>
<td>An Integer percentage threshold used by this program to filter out non-significant HSPs reported by tblastn, example: 50, [default value: 50]</td>
</tr>
<tr>
<td>-t</td>
<td>An Integer percentage value to instruct what percentage of the top BitScoreSums should be included, example: 25, [default value: 25]</td>
</tr>
<tr>
<td>-l</td>
<td>An Integer percentage threshold used by this program to filter out too short predicted proteins by genewise, example: 50, [default value: 50]</td>
</tr>
<tr>
<td>-x</td>
<td>An Integer value used by this program to do end extension on constructed putative genes, example: 10000, [default value: 10000]</td>
</tr>
</tbody>
</table>

Table 5-6: Relation between configuration file parameters and command line arguments
Phoenix output files and formats

The output file name is in the format of `SpeciesName.GeneFamilyName.ys.faa`, one file for each gene family per genome. These files are in amino acid FASTA format, and the first line of each entry is in such format:

```
>SpeciesName|SuffixID|GeneNumber
```

The `SpeciesName` is the one provided by the `-n` argument to PhoenixPrep, the `SuffixID` is the `-s` or `SUFFIX_ID` argument to Phoenix, and the `GeneNumber` is a unique number created by Phoenix for each putative gene predicted for the same species at the same run.

Phoenix program availability

The executables and instructions could be downloaded from

(http://www.synblex.com/phoenix/).
Project summary and discussion

Limitations

As demonstrated, gene or gene family homologues could be extracted from genomes without annotation or even from only partially assembled genomes. However, in the latter case, we have to assume that the gaps or break points between contigs do not overlap with regions where the genes of interest are located. A necessary but not sufficient condition is that the genomes have been assembled into contigs of length larger than that of average genes. This rules out the possibility of directly operating Phoenix on EST or RNA-seq data. This also means that genome assembly errors and sequencing errors would still influence the homologue retrieval process, as in any other genome annotation methods. Same as other homologue retrieval tools, the input also has a great impact on the behavior of Phoenix. A set of sequences of high similarity would retrieve sequences that are equally similar. A set of input of low similarity would retrieve sequences that are as divergent. A carefully prepared input dataset would have the best result.

Previously, the parallelization of genome annotation was done through genome segmentation and subsequent parallel annotation on different genome segments. With Phoenix, it is possible to achieve parallelization through parallel annotation of different gene families. Manually curated gene families by biologists who have extensive knowledge specifically on these gene families would be far superior models than those generated by automated methods without manual inspection. These manually curated families then could be used to annotate new genomes by Phoenix. A centralized server could be established to be a repository of these curated gene families and researchers could freely contribute their knowledge to this repository. However, as mentioned in Chapter 1, it is estimated that only 50% of the genes in a genome could be
identified by homology with known genes [93], so this approach would miss many lineage-specific genes. The only solution to the discovery of lineage-specific genes might be extensive transcriptome sampling.

**Boundary ambiguities between families**

The scientific definition of a gene family (or protein family) is based on descendence from a common ancestor. This means shared protein 3D structures, functions, as well as sequences in all descendant genes from the same ancestor. However, it is not possible to go back in time to find the ancestral sequence. In reality, we could only rely on the information from currently available proteins. Not every known protein has a verified 3D structure or a well-documented functional annotation. However, each protein has its amino acid sequence. Most of the time, the definition of a gene or protein family is entirely built upon sequence similarity. This is essentially a clustering problem in information science. As introduced in Chapter 1, the similarity between sequences is related to evolutionary distance.

The significance of sequence similarity could be measured using formulas introduced in Chapter 1. However, the cutoff between significance and insignificance is arbitrary, as different researchers might perceive different cutoffs as the appropriate one for their different studies. BLAST is the commonly used search tool and many researchers use E-value < $10^{-10}$ as the cutoff for significance, even though the E-value itself is context dependent and varies from study to study. In contrast, the bitScore calculated by BLAST is context independent. This is the reason that the similarity measurement used in here is based on bitScore instead of E-value.
Because of the ambiguity of the boundary conditions, the concept gene (or protein) family is ambiguous. The cutoff condition directly influences the cluster algorithm’s behavior. If a less stringent cutoff is provided, many sequences with lower similarities are clustered together. This is essentially how super families are identified. At a higher stringency, fewer numbers of sequences with higher similarities are clustered together. Several clusters formed with a higher cutoff could be all included in the same super family. By increasing the cutoff, a family could be divided into different sub-families. As a result, the behavior of the Phoenix software heavily depends on input sequences’ similarity. For example, if the input sequences were poorly defined and low in similarity, the output would be more likely to include irrelevant sequences. A tightly defined input has less entropy and thus more information. A loosely defined input has more entropy and thus less information.

**Differences between protein family and domain family in theory and in practice**

Most proteins contain multiple domains or motifs. For example, a transcription factor is expected to have at least one DNA binding domain and one activation domain to function as a transcription factor. They often contain other regulatory domains that could be controlled by other factors inside the cell.

However, protein families are often defined by a single domain or only a few domains. For example, the widely used Pfam database is based on known domains or motifs [211, 301]. As a result, a single protein could be assigned to different families based on its different domains. In practice, if a full-length protein is used for protein family search and it has multiple domains, the signal of its family signature domain could be diluted by other domains. Formula 42 and 43 in
this chapter take this into consideration and essentially reward domains supported by multiple members from the input. On one extreme, if only one single protein sequence is used in input, this mechanism could not work. On the other extreme, if a loosely defined gene family were provided, the signal of their family signature domains would also be weak.

**Model complexity and applicability**

Most current protein family classification algorithms depend on the profile HMM concept introduced in Chapter 1. A profile HMM allows insertion and deletion, and also different emission probabilities for different conserved sites. However, this method does not allow changes in the order of the domains. This is because the profile HMM structure illustrated in Figure 5-3 could only jump forward between states [20]. However, it is likely that domains could be reshuffled during evolution. Therefore, the Phoenix algorithm does not include a profile HMM. As a result, homologues with domains swapped in position could also be identified if a proper input is given.

![Figure 5-3: An example profile HMM model structure.](image-url)
Phoenix has implemented a very straightforward algorithm. Commonly used methods involving probabilistic modeling and parameterization based on training data were not considered. Similarly, the GeneWise program operated by Phoenix performs the exon boundary prediction with the simplest model and the least assumptions. This is because the currently well-studied model organisms only represent a small percentage of biodiversity. As more genomes from increasingly diversified evolutionary backgrounds become available, probabilistic models trained by current model organism genomes might not be able to perform correctly on the newly sequenced ones. Incorrect models could easily miss genes or exons. Gene prediction based on protein sequence conservation might be the only reliable method for retrieving protein homologues from highly diversified genomes.

(End of Chapter 5)
Chapter 6

Discussion

Studies in this dissertation included evolutionary event reconstruction, genealogy inference and identification, transcriptional activity monitoring and gene identification. Underpinning all these studies is DNA sequence analysis. Driven by the continued development of DNA sequencing technology, DNA sequence analysis is adopted in more and more fields as an integrated and reliable method and shows greater potential in both basic and applied researches. Before discussing its potentials and advantages, it is necessary to first consider the intrinsic limitations of DNA sequence analysis.

Intrinsic limitations of DNA sequence analysis

The foundation of any DNA sequence analysis is sequence alignment. The algorithms introduced by Smith and Waterman (SW) [21] for local alignment, as well as Needleman and Wunsch (NW) [7] for global alignment, are still the solution for inferring the optimal sequence alignment. However, due to their high computational cost, heuristic algorithms that could potentially miss the optimal alignment, but are much faster, are overwhelmingly more widely used in life sciences related fields. Although in most cases these heuristic methods produce the same results as the full SW or NW method, the possibility of missing the optimal solution is always present for any heuristic methods.
Usually, the most pronounced effects of different alignment algorithms and scoring schemes in DNA sequence analysis are related with inferring insertions and deletions. Sometimes, alignments are built with too many inappropriately introduced indels. Nucleotides in-between short indels are aligned together and then interpreted as mismatches. This situation leads to the overestimation of substitution events and subsequently, the overestimation of genetic distance between sequences. This is especially pronounced at regions that contain repeats. Some sequencing technologies may make this situation worse by determining the length of the repeats at a higher error rate. For example, sequencing-by-synthesis sequencers that do not use dye-terminators usually have high error rate in determining the correct number of nucleotides for homopolymer tracts. As a result, the number of SNPs estimated by mapping data from these sequencers is usually inflated because of incorrectly determined homopolymers followed by incorrectly introduced indels.

Generally, repeats are challenging to analyze. Limited by the maximum read length of the available sequencing technologies, most of the long repeats that cannot be sequenced as a single read could not be analyzed using the methods introduced in this dissertation. This is because these repetitive reads could not be reliably assembled nor be mapped, simply because there are not enough information to operate on due to their repetitive nature. Notable examples are the centromeres, telomeres and heterochromatin regions in human genomes. These regions are excluded in the finished human reference genome due to the aforementioned technical difficulties. Even shorter repeats that can be characterized with the current sequencing technologies usually complicate the DNA sequence analysis processes. For example, in Chapter 4, when the reads from repeats were vigorously removed based on known repeat databases and genomic repeat features, embryonic day 18 transcriptome had the highest percentage of previously unannotated transcription activities. However, when these repeats were not removed
as thoroughly, stem cell stage transcriptome had the highest percentage of previously unannotated transcription activities. Differences in just one single step of the analysis leads to two drastically different conclusions.

Regardless of the algorithms used, any alignment result is dictated by the scoring scheme. Any scoring scheme is derived from previously known alignment that is deemed to be correct. The decision of whether or not applying a certain scoring scheme to a specific problem is usually arbitrary. This is because it is impractical, if not impossible, to estimate a custom scoring scheme for each alignment. As a result, it is necessary to keep in mind the algorithms and scoring schemes used in generating a given alignment, and their impact on this given alignment and any results derived from this alignment. Furthermore, it is possible to use an expectation–maximization (EM) algorithm to co-maximize the likelihood or probability of both the scoring scheme and the alignment result. However, as with any EM algorithm, there is no guarantee that the final maximum is the global one instead of a local one. The initial parameters to start the EM process will have a huge impact on the final result, but these parameters are again from previous known estimates that are not necessarily correct in a new problem.

More generally, parameterization of any models used for analyzing DNA sequence has its own scope of usability or applicability. Parameters estimated from known data are not necessarily appropriate for the data to be studied. When the applicability of known parameters is expected to be questionable, it is better to choose models that are least affected by differences in parameters. On the other hand, when the nature of the known data is expected to be similar to that of the new data, models that best reflect the characteristics of the known pattern, normally with more parameters, should be used. In the latter case, the necessary training dataset size, the
number of parameters to estimate and the desirable variance of these estimated parameters are the factors to consider in choosing the ideal model.

For the above reason, the model chosen for developing Phoenix program is very simple and has very few parameters. This is because that Phoenix is designed for genomes with very different and possibly unknown patterns. On the other hand, the HapSearch system heavily depends on the parameters estimated from the 7975 currently available human mitochondrial genome sequences. This is because that it is designed only for human mitochondrial genomes, which are the same as in the training data.

The more extreme, but sometimes less obvious factors than inappropriate parameters are the inappropriate models. Any mathematical or statistical model used in DNA sequence analysis is based on a biological, evolutionary or molecular process. No matter how abstracted, simplified or sophisticated a model is, the applicability of this model is determined by the problem on hand. The assumptions or conditions of the model have to fit the real problem. For example, most of today’s phylogenetic reconstruction methods treat all leaf nodes as present-day samples. These methods estimate the genetic distance, evolutionary rate and phylogenetic relationship based on the assumption that the leaf nodes are of the same time period. This assumption is perfectly valid for most of the phylogenetic applications, since all sequences are taken from current living specimens and compared across species. However, when an ancient DNA sample is included, this assumption becomes erroneous. Obviously an ancient DNA is from a different time period from modern DNAs, and is probably close in terms of time to the parental nodes of some of the modern DNAs. Treating all leaf nodes as if they were from the same generation would be incorrect in this case. The polar bear phylogeny reconstruction study introduced in Chapter 2 is a perfect example, which used a 100,000-year old polar bear sample. Another example is the
validity of HMMs discussed in Chapter 5. If domain order is allowed to change, commonly used profile HMMs are inappropriate for describing protein families as these profile HMMs do not allow such domain order changes.

Another common fallacy is that the phylogenetic methods could be directly applied to population studies. Other than the parameters for interspecific processes are different from those in intraspecific processes, as introduced in Chapter 3, the evolutionary models should be considered differently as well. Similar to treating modern day samples and ancient samples as from the same time period, by directly applying some of the phylogenetic methods to population data, one actually treats all individuals in the population as from the same generation. There might not be too many observable effects when the time scale is greater, however, as the scale becomes finer, the effects will become more pronounced. In order to correctly apply some phylogenetic methods to population studies, some special designs in sampling are needed. For example, in Chapter 3, in order to use the tree method to estimate substitutions, only one sample from each country was used. This is because the generational differences between individuals from different countries are negligible for the given problem.

In summary, the design, parameterization and selection of DNA sequence analysis methods require prior knowledge. Therefore, the correctness of a given method is largely determined by the similarity of the known problem where the prior knowledge is derived and the new problem where the method will be applied. It is fair to assume that we have only encounter a fraction of all possible living organisms, cellular processes and genome organizations. As a result, it is always helpful to understand both the conditions of the new problem and the limitations of the available methods. After all, all methods will produce some results, but the end users without
the proper understanding of the used methods cannot distinguish the results from the correct methods from those from the incorrect methods.

Intrinsic advantages of DNA sequence analysis

The most obvious property of DNA sequence analysis is that it is an analysis process based on DNA sequences. In most cases, the analysis process is sequencing technology agnostic, organism agnostic and problem agnostic. This means that the same set of analysis methods could be used to solve drastically different problems. For example, all the studies included in this dissertation rely on the same DNA alignment principle, though different algorithmic implementations. Once DNA sequence is determined, the entire DNA analysis process is performed in silico. Digitized DNA information in essence is the same as any other type of digital information. This means that the same analytic methods developed in information sciences could be directly transferred into life sciences to analyze DNA sequence information. For example, as mentioned in Chapter 1, based on the information theory, the first life forms on this planet should be in RNA molecules but not protein molecules, simply because the differences in information content, as demonstrated by Yockey in 2005 [75]. In addition, decades of information technology development provide a mature platform for accelerated development of in silico DNA sequence analysis.

In additional, digitized DNA information is easy to store, index, search, retrieve and distribute, just as any other digital information. This means that genetic resources, in the form of DNA sequences, and information could be stored in centralized databases that can be publicly
accessed. This convenient way of distributing information greatly promoted the application of DNA information in almost all life sciences related fields. This is also a two-way traffic: more and more DNA information from new researches is deposited into these centralized databases. From this perspective, the combined efforts of the entire life sciences field are analogous to a giant open source software community. With sufficient information for a sequence from various biological studies, potential functions and other useful information could be predicted for similar sequences that are new. For example, as introduced in Chapter 1, new genes and genomes could be annotated simply by comparing the DNA sequences with reference sequences or patterns in the database with known functions.

Another feature to consider is that the DNA molecule is relatively stable, and can still be suitable for sequencing long after being separated from the containing living system. Certain conditions will even prolong the usability of DNA further, for example, low temperature and low humidity. Certain features of the DNA molecule also help. For example, a circular DNA molecule has no free ends, and thus is less susceptible to degradation. The above features allow the uncoupling of the maintenance of the living system and the analysis of the system. This provides great flexibility in experiment design, and also new opportunities in research. For example, in Chapter 2, the jawbone discovered from the Svalbard Islands is from a bear died 100,000 years ago. However, the DNA is still retrievable. With proper methods, the addition of this 100,000 years old DNA sequence depicted a more detailed evolutionary process than using modern DNA alone. An added benefit of using ancient DNA with correct time information, for example, geological dating in this case, is that the evolutionary rate could be estimated instead of predefined. This is possible because of the added time information from ancient DNA sample.
Also because of the ability to study DNA information without maintaining the carrier, microbes can now be characterized without being cultured. This is extremely helpful in detecting pathogens and environmental microbes that are difficult to culture. Even more helpful is that previously unknown microbes’ DNA information could also be captured simply through DNA sequencing. This, combined with reference information in centralized databases, depicted a far more detailed genetic picture of any environment (metagenomics). For example, it is estimated that 90% of human microbiome consists of currently unknown microbes, and among the 10% known, 90% is not cultivable. DNA information of the human microbiome becomes the key to understanding their influence in human health [302].

Last but not least, the evolutionary rate variation among different sites broadens the application scope of the DNA sequence analysis on a single DNA sequence. The differences in conservation are present globally along the entire genome, and also present locally between different parts of a single gene. Highly conserved regions could be used as markers for genus level identification. Moderately conserved regions could be used as markers for species level identification. Rapidly evolving regions could be used as markers for individual identification. For example, bacterial 23S rRNA has been widely used as markers for prokaryotic identification purposes [303]. Human mitochondrial D-loop region has a long history of being sequenced for individual identification purposes, although in Chapter 3 it is proven that the human mitochondrial genome alone, without mentioning only the D-loop part, does not possess enough diversity to be used as individual identification markers for sufficiently large populations. Such a broad spectrum of resolution suitable for various problems is located on the same DNA molecule and can be retrieved at once by DNA sequencing. The DNA sequence analysis then operates on this single information source with different methods to provide different information to different problems, all without any further traditional bench-top experiments.
In summary, DNA sequence analysis provides improved efficiency, broader applicability and greater flexibility. With the continued decrease in DNA sequencing cost and increase in data output, DNA sequence analysis shows greater potentials in both traditional life sciences related fields and new interdisciplinary fields, especially in real world applications.

**Potentials of DNA sequence analysis**

Currently, the most hotly discussed potential of the DNA sequence analysis application is personalized medicine. As discussed in Chapter 3, each individual human genome harbors at least 3 million unique SNPs from the rest of the population. Different individuals with different genetic backgrounds may response to the same medical treatment differently. As a result, if the genomic DNA information of a patient is available, healthcare providers can tailor the treatment specifically for him or her if the effect of a certain treatment on a certain genotype is known. More generally, if the genomic DNA information is available for an individual, his or her predispositions to various illnesses can be determined if sufficient biomedical research information regarding this has been accumulated. All the above could greatly improve the welfare of an entire society and all its members.

As of March 2012, the price of generating enough reads to map an entire human genome is still well above 1000.00 USD. Other than the cost for these raw sequencing reads, the cost of genome assembly or genome mapping, and the cost for digital information storage, analysis and transmission should also be factored in the cost of generating a usable genome for a single individual. The combined cost is still prohibiting for the general public. Above all, currently, the
single most significant obstacle in the aforementioned application of personalized medicine is the lack of known knowledge regarding the effect of a certain treatment on a certain genotype. There is still not enough biomedical information to fully take advantage of a patient’s genomic information, without even considering the cost of generating a usable genome. Many diseases that are currently known to be strongly associated with certain genotypes have no effective treatment. Other diseases that have relative weak association with certain genotypes do not warrant the need of genome sequencing. However, it is certain that the information regarding the effect of genotypic differences on medical treatment responses will increase with time, especially with the current trend of dramatically decreasing sequencing cost and the wide spread adoption of genomic analysis methods in life sciences related fields.

Even without going genome wide, applications of DNA sequence analysis in local regions of the genome can be beneficial. Gene-specific or locus-specific DNA sequence data could be easily obtained using targeted sequencing techniques. Instead of spreading the reads randomly along the entire genome, targeted sequencing focuses on selected regions in the genome using target-specific amplicons. This approach offers substantial cost reduction for achieving sufficient read coverage at these selected locations. This, combined with known information at these locations, realizes the principle of personalized medicine on a smaller scale with drastically increased efficiency.

Another clinical application possibility is the non-invasive genetic testing. Genetic traits, including blood type and MHC (or HLA) type, are determined primarily be genomic DNA sequences. Current methods for determining blood type of MHC type requires blood sampling, usually in the form of venipuncture. However, any nucleated human cell contains the same set of genomic DNA. It is thus possible to determine the blood type or MHC type by sequencing the
corresponding genomic regions from non-blood samples, for example, saliva or hair root. This will bypass the invasive method of blood drawing. However, the exact genotypic to phenotypic relationships are not clear yet for both blood and MHC type. In principle, such relationships could be modeled using the same methods introduced in Chapter 1 and Chapter 3. Again, the accumulation of sufficient clinical data is still the key to this type of application.

Aside from human DNA, the environmental DNA composition of our immediate surroundings could provide information helpful in improving human healthcare. As mentioned earlier, it is now possible to characterize the composition of the microbes in an environmental sample through DNA sequence analysis. With sufficient fast sequencing and analysis turnaround, it is possible to monitor the airborne pathogens in populated areas, for example, schools, hospitals and shopping centers. This will give the exact status of a potential epidemic, assuming the percentage of the pathogen is an indicator of the severity of the resulting epidemic. With sufficient airborne pathogen monitoring data, it is also quite possible to predict the future trend of certain epidemics, and suggest prevention methods and appropriate actions.

The aforementioned possible future applications are just a small part of the full potential of DNA sequence analysis. The above examples are all centered on human healthcare applications. However, all living cells from bacteria to mammals contain DNA. The same set of methods could be applied to other applications in fields other than human healthcare, for example, crop breeding and environmental monitoring. In summary, one should be cautious, in light of the limitations that have been discussed, yet firmly optimistic, because of the advantages, about the future applications of DNA sequence analysis.

(End of Chapter 6)
Appendix

Permissions for published results

PNAS author rights and permissions pertinent to this document


Author rights in general:

As a PNAS author, you and your employing institution or company retain extensive rights for use of your materials and intellectual property. You retain these rights and permissions without having to obtain explicit permission from PNAS, provided that you cite the original source:

The right to post a PDF of your article on your Web site or that of your employer's institution (provided that the institution is nonprofit).

The right to make electronic or hard copies of articles for your personal use, including classroom use, or for the personal use of colleagues, provided those copies are not for sale and are not distributed in a systematic way outside of your employing institution.

The right to post and update a preprint version of your article on a public electronic server such as the Web. See the information on electronic preprints below.

The right to permit others to use your original figures or tables published in PNAS for noncommercial and educational use (i.e., in a review article, in a book that is not for sale), provided that the original source is cited. Third parties need not ask PNAS for permission to use figures and tables for such use.
The right, after publication in PNAS, to use all or part of your article in a printed compilation of your own works, such as collected writings or lecture notes.

If your article is a "work for hire" made within the scope of your employment, your employer may use all or part of the information in your article for intracompany use.

The right to include your article in your thesis or dissertation.

The right to present all or part of your paper at a meeting or conference, including ones that are webcast, and to give copies of your paper to meeting attendees before or after publication in PNAS. For interactions with the media prior to publication, see the PNAS Policy on Media Coverage.

The right to publish a new or extended version of your paper provided that it is sufficiently different to be considered a new work.

The right to expand your article into book-length form for publication.

The right to reuse your original figures and tables in your future works.

Patent and trademark rights or rights to any process or procedure described in your article.

FAQ: Can I use my article in my dissertation or thesis without asking permission?
Yes, provided that you cite the original source and, for articles published in Volumes 90–105 (1993–2008), copyright notice.

Nature author rights and permissions pertinent to this document

[Retrieved from Nature (http://www.nature.com/reprints/permission-requests.html) on Nov 15th 2011]
Since 2003, ownership of copyright in original research articles remains with the Authors*, and provided that, when reproducing the Contribution or extracts from it, the Authors acknowledge first and reference publication in the Journal, the Authors retain the following non-exclusive rights:

To reproduce the Contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s).

They and any academic institution where they work at the time may reproduce the Contribution for the purpose of course teaching.

To reuse figures or tables created by them and contained in the Contribution in other works created by them.

To post a copy of the Contribution as accepted for publication after peer review (in Word or Tex format) on the Author's own web site, or the Author's institutional repository, or the Author's funding body's archive, six months after publication of the printed or online edition of the Journal, provided that they also link to the Journal article on NPG's web site (eg through the DOI).

NPG encourages the self-archiving of the accepted version of your manuscript in your funding agency's or institution's repository, six months after publication. This policy complements the recently announced policies of the US National Institutes of Health, Wellcome Trust and other research funding bodies around the world. NPG recognizes the efforts of funding bodies to increase access to the research they fund, and we strongly encourage authors to participate in such efforts.
Authors wishing to use the published version of their article for promotional use or on a web site must request in the normal way.

* Commissioned material is still subject to copyright transfer conditions

**BMC Genomics** author rights and permissions pertinent to this document

[Retrieved from BioMed Central (http://www.biomedcentral.com/info/authors/license) on Nov 15th 2011]

---

**BioMed Central copyright and license agreement**

In submitting a research article ('article') to any of the journals published by BioMed Central Ltd ('BioMed Central') I certify that:

1. I am authorized by my co-authors to enter into these arrangements.

2. I warrant, on behalf of myself and my co-authors, that:

   the article is original, has not been formally published in any other peer-reviewed journal, is not under consideration by any other journal and does not infringe any existing copyright or any other third party rights;

   I am/we are the sole author(s) of the article and have full authority to enter into this agreement and in granting rights to BioMed Central are not in breach of any other obligation. If the law requires that the article be published in the public domain, I/we will notify BioMed Central at the time of submission upon which clauses 3 through 6 inclusive do not apply;

   the article contains nothing that is unlawful, libellous, or which would, if published, constitute a breach of contract or of confidence or of commitment given to secrecy;

   I/we have taken due care to ensure the integrity of the article. To my/our - and currently accepted scientific - knowledge all statements contained in it purporting to be facts are true and
any formula or instruction contained in the article will not, if followed accurately, cause any injury, illness or damage to the user.

And I agree to the following license agreement:

BioMed Central Open Access license agreement

Brief summary of the agreement

Anyone is free:

to copy, distribute, and display the work;

to make derivative works;

to make commercial use of the work;

Under the following conditions: Attribution

the original author must be given credit;

for any reuse or distribution, it must be made clear to others what the license terms of this work are;

any of these conditions can be waived if the authors gives permission.

Statutory fair use and other rights are in no way affected by the above.

Full BioMed Central Open Access license agreement could be accessed from

(http://www.biomedcentral.com/info/authors/license)

(Identical to the 'Creative Commons Attribution License')

Pennsylvania State University Thesis Office policies pertinent to this document

[E-mail communication between this author and Penn State Thesis Office]

from Pauletta Leathers
to    Yazhou Sun  
date    Fri, Apr 1, 2011 at 9:20 AM  
subject  Re: Questions about including projects done in previous lab in PhD thesis  
mailed-by    psu.edu  

Yazhou  

I believe you are correct.  

Pauletta  

At 06:37 PM 3/31/2011, Yazhou Sun wrote:  
> Dear Pauletta,  
>  
> Thanks for the explanation. Since I will be rewriting in my thesis the  
> parts about the projects I've done in my two published papers, adding  
> some other analysis and removing some irrelevant analysis, I think I  
> won't run into the copyright problem you pointed out. But I am very  
> glad to know that the content of my thesis is largely solely  
> determined by my advisor and my committee. I think I'll talk with them  
> about putting these two projects into the content of my thesis. If  
> they agree, then I can use these two projects.  
>  
> Is my understanding correct? Thank you very much.  
>  
> Best  
> Yazhou Sun  
>  
> On Thu, Mar 31, 2011 at 3:16 PM, Pauletta Leathers wrote:  
> >> Yazhou  
> >>  
> >> The Graduate School policy on previously published work is as follows:  
> >>  
> >> Work by the thesis author which is published prior to thesis submission (or  
> >> is shortly expected to be published) may be accepted as part of the thesis,  
> >> provided the committee approves the work and the published material was  
> >> written specifically to fulfill thesis requirements. If you submit  
> >> previously published copyrighted work and you are not the copyright holder,  
> >> a letter of permission from the copyright holder must accompany the thesis.  
> >>  
> >> This policy may be found on page 15 of the current Thesis Guide. As a  
> >> general rule, if you published your work in a journal, you most likely  
> >> signed over copyright at the time of publication, so you will need written  
> >> (letter or email) permission from the journal to use it in your  
> >> dissertation. Also, keep in mind our policy about multiple-authored work:  
> >>  
> >> Your department may permit you to submit multiple-authored work as thesis  
> >> material if you are first author of the work. Your contributions must be  
> >> clearly and fully indicated in a preface to the thesis.
The Graduate School has no policy with regard to the academic relationship between you and your previous advisor/lab. I would recommend that you consult with Dr. Suzanne Adair, Assistant Dean of the Graduate School, if you anticipate any problems.

The Graduate School has no reason to object to your use of unpublished work. The content of your dissertation is, for the most part, a matter for your advisor and committee.

I hope this has been helpful. Please feel free to contact me again if I can be of assistance.

Pauletta Leathers
Editorial Assistant, The Graduate School

At 03:09 PM 3/30/2011, Yazhou Sun wrote:

Dear Thesis Office,

My name is Yazhou Sun, PhD student in Genetics program. Currently my advisor (Prof. Hong Ma, Department of Biology) and me are thinking about my graduation. I have some questions about what can be included in my thesis:

Here's my situation: I have done my first three-year research during my study here in another lab (also in Penn State, but from a different department from my current advisor's). Then I switched to the lab of my current advisor.

My first question: two of my projects done in my previous lab have been published, which means I have coauthored two papers with my previous advisor for each of these two projects. Is it possible that I can include these two studies I have done in my final PhD thesis. My current lab and my previous lab have no connection or collaboration but the projects I did in my first lab and the projects I did in my current lab can be unified in my PhD thesis. And I intend to only include the parts I did from these two projects in my previous lab?

What's the school policy about this situation? Is this allowed or prohibited? My previous advisor is not on my current thesis committee and the communication between these two lab are not well.

My second question: two of my projects done in my previous lab have not been published and very likely will never be published. Can I include the parts I've already done for these two unpublished projects?

I am more interested in including the parts from the two published
> paper, but also curious about school policy about the unpublished
> (likely never-publishable) results the students have done?
> Thank you very much.
> Sincerely yours,
> Yazhou Sun

Pauletta Leathers
Thesis Office (gradthesis@psu.edu)
The Pennsylvania State University
115 Kern Building
University Park, PA 16802
phone: 814/865-5448
fax: 814/863-4627
References:


VITA

Yazhou Sun

Yazhou Sun was born in Taian, Shandong province, P.R.China in 1983. Yazhou Sun studied in the University of Science and Technology of China from 2002 to 2006. His college major was bioscience but his curriculum also included many subjects in computer and information sciences. He started his PhD study in the Intercollege Graduate Program in Genetics at Pennsylvania State University immediately after his college graduation. His graduate studies focused on DNA sequence analysis, including new applications with high-throughput sequencing and the development of new analysis methods.