In pursuit of gene variation of consequence to human health and disease

Yuhu Liang

Academic advisor: Anders Krogh
Department of Computer Science, University of Copenhagen

7th December, 2022

UNIVERSITY OF
COPENHAGEN

穷则独善其身，达则兼济天下

《孟子》

Preface

The thesis entitled "In pursuit of gene variation of consequence to human health and disease" is submitted to the PhD School of The Faculty of Science to meet the requirements of the obtaining a PhD degree at the University of Copenhagen.

The work presented in this thesis contains two projects, that I atttached in the Chapter 7. The first project was initiated from March 2019 primarily at the Section for Computational and RNA Biology, Department of Biology, Univeristy of Copenhagen, under the supervsion of professor Anders Krogh. The second one was carried out in the period between July 2021 and November 2022, at the Machine Learning section, Department of Computer Science and Center for Health Data Science (HeaDS), University of Copenhagen. The Ph.D. project was financially supported by the China Scholarship Council (201804910693) and the Graduate School of Sciences, University of Copenhagen.

The thesis is divided into two parts. Chapter 1 is an introduction for the research of human genome, transcriptome and neural network. I also briefly introduced the research aims and hypotheses, methods and summary of the studies which I worked in the last almost four years (chapter 2 to 5). The chapter 7 contains two papers, the first paper was supervised by Prof. Krogh with the help of Christian Grønbæk and Prof. Pieor Fariselli. The second paper described a generative decoder model, which was used on Cancer data research. I collaborated with Viktoria Schuster, Iñigo Prada-Luengo and Prof. Krogh with the big help from Thilde Terkelsen on this project.

Abstract

From the invention of Sanger sequencing, to the birth of current high-throughput and long-read methodologies, sequencing technology has become an vital tool for scientific research. Biologists released the first version of the human genome in 2001, and continued to refine it over the following years until the complete and final genome sequence was published in 2022. In parallel, the 1000 Genome project has revealed the extent of human genetic variation and polymorphisms, filling a gap in our knowledge about the diversity of the human mutational landscape. Transcriptome sequencing provides a means to study the changes in gene expression patterns and related signaling pathways affected by diseases and other biological processes. With the advancement of computer science, machine learning has been introduced into the field of biological and medical research. Using ML approaches scientists hope to find the biological signals and patterns hidden within massive datasets.

The first chapter of this thesis provides an overview of the human genome, transcriptome research and different machine learning algorithms, including their applications in biological and medical research.

The last chapter centers around two projects I worked on during my Ph.D. In the first project, simply called DNA prediction, we employed a Central model, a Markov model and a bi-directional Markov model to estimate the probability of the occurrence of four nucleotide types at a site based on its context sequence - the input for these models were the human reference genome. The results show that the base prediction of the human genome was above 50% on average, which should be compared to random guessing (25%). We applied the predicted results to SNP databases, and found that the alternative alleles showed higher probabilities than reference bases for somatic SNPs. In addition, we developed a substitution model to calculate the base mutability. Here, we found that the α matrix relies on a much smaller context sequences, and in the prediction results of the model with one base to each side, we found that cytosine (C) has a higher mutability to thymine (T) in CpG sites. Additionally, our substitution model fits the somatic mutations very well..
In the second project, we developed a generative nerual network consisting of decoder and a Gaussian mixture model - hence, we called it a deep generative decoder model. We applied the decoder model to the study of gene expression data. We used normal individual bulk RNA sequencing samples from the GTEx

database to train our model, and made a matrix to show how well the samples can be clustered together by tissue type and their distribution within different Gaussian components. We found that, except for three tissues with a small sample size, the majority of tissue types independently dominated a Gaussian component. Then, the cancer samples from the TCGA database were used to evaluate whether our trained model could generate new data points and match them to the correct Gaussian component of the corresponding tissue. Additionally, our sophisticated model can be used to predict the probability of genes being differentially expressed, by using the negative binomial distribution in our model, which can be used for N-of-1 research. Compared to DESeq2, a commonly used method to obtain differential expressed genes (DEGs), the number of DEGs provided by our model is much smaller. However, in the enrichment expected fraction analysis of driver genes and the analysis of subtype-specific related genes of breast cancer, our model shows a good performance.

Resumé (Dansk)

Fra opfindelsen af Sanger-sekventering, til fødslen af nuværende high-throughput -og long-read metoder, er sekventeringsteknologi blevet et vigtigt redskab indenfor den videnskabelige forskning. Biologer fremlagde den første version af det menneskelige genom i 2001, og fortsatte med at forfine dette i mange år efter, indtil den fuldstændige og endelige genomsekvens blev offentliggjort i 2022. Sideløbende har 1000 Genome-projektet afsløret omfanget af menneskelig genetisk variation og polymorfier, og udfyldt et hul i vores viden omkring variabiliteten af det menneskelige mutationslandskab. Transkriptomsekventering giver os mulighed for at studere ændringer i genekspressionsmønstre og relaterede enzym/protein signaler som påvirkes af sygdomme og andre biologiske processer. Med de store fremskridt inden for datalogi, er maskinlæring blevet en central del af den biologiske og medicinske forskning. Ved at bruge maskinlærings-metoder håber forskere at kunne udlede de biologiske mønstre, som ligger gemt i massive datasæt.

Det første kapitel i min Ph.d. afhandling giver et overblik over det humane genom, transkriptomforskning, samt forskellige maskinlæringsalgoritmer, herunder deres anvendelser indenfor biologisk og medicinsk forskning.

Det sidst kapitel centrerer sig omkring to projekter, jeg har arbejdet på under min Ph.d. I det første projekt, her kaldet DNA-forudsigelse, brugte vi en central model, en Markov-model, samt en tovejs Markov-model til at estimere sandsynligheden for forekomsten af hver af de fire nukleotide på en specifik position i genomet, baseret på kontekstsekvens - inputtet til disse modeller var det humane referencegenom. Resultaterne viste, at basisforudsigelsen af det humane genom var over 50% i gennemsnit, hvilket skal sammenholdes med et tilfældig gæt på 25%. Vi anvendte de forudsagte resultater med SNP-databasen og fastslog, at de alternative alleler viste højere sandsynlighed end referencebaser for somatiske SNP'er. Ydermere udviklede vi en substitutionsmodel til at beregne basismutabiliteten. Her fandt vi, at vores α-matrix er afhængig af en mindre kontekstsekvens, og i vores output fra modellen med en base til hver side, ser vi at cytosin (C) har en højere mutabilitet til thymin (T) i CpG-regioner. Ydermere passer vores substitutionsmodel godt på de somatiske mutationer.

I det andet projekt udviklede vi en generativ model, som består af et deck-oder neuralt netværk og en Gaussisk blandingsmodel, som vi kalder en dyb generativ dekodermodel. Vi anvendte dekodermodellen til at studere genekspressionsmønstre. Normale individuelle RNA-sekventeringsdata fra GTEx-databasen blev brugt til at træne vores model og vi generede et matrix for at vise, hvorledes data kan grupperes efter vævstype og deres fordeling inden for forskellige Gaussiske komponenter. Vi fandt, at bortset fra tre væv med kun få prøver, dominerede størstedelen af vævstyper hver især en af de Gaussiske komponenter. Dernæst blev kræftprøver fra TCGA-databasen brugt til at evaluere hvorvidt vores trænede model kunne generere nye datapunkter og matche disse med den korrekte Gaussiske komponent af det tilsvarende væv. Vores sofistikerede model kan bruges til at forudsige sandsynligheden for at et gen er differentielt udtrykt, ved at benytte den negative binomiale fordeling i vores model, som kan bruges til N-af-1 forskning. Sammenlignet med DESeq2, en populær og anvendt metode til at opnå differentielt udtrykte gener (DEG'er), er antallet af DEG'er givet fra vores model en del mindre. Imidlertid viser vores model en god præstation i den forventede berigelsesfraktionsanalyse af drivergener, samt i analysen af subtypespecifikke-relaterede brystkræft gener.

Acknowledgements

Around four years ago, I turned a new page of my life. I decided to do my Ph.D. study in Copenhagen, a beautiful and cozy city. Over the past four years, I met so many great people, who supported and helped me in life, work, etc.

First and foremost, I woule like to show my deep gratitude to my supervisor Prof. Anders Krogh, for giving me the opportunity to work on these interesting projects in your group and helping me a lot in my academic research. Thanks for always being available for scientific disscussion with me. During the time of working together with Anders, I not only learned something new, know more mathematics and start doing machine learning, but also saw how to do projects with critical and independent thinking from you. Moreover, Anders gave me a relaxing work environment, and you sent me to the international conferences and encouraged me to take the summer or winter school outside Denmark. All these made my Ph.D. study more fulfilling.

Secondly, I would like to thank the entire Krogh group people, including in Christia of course. You are the amazing people, I really appreciate that I had a super good time with all of you. Thanks to Henrike Zschach and Thilde Terkelsen, for creating a good atmosphere in both, work time and social events. Also, I am very grateful for your invaluable advice on some specific biological questions and my confusion about my Ph.D. studying. I sincerely thank Dr. Prada and Viktoria Schuster. Without your kind helping hand, it is hard to finish my project, successfully. Another thanks goes to the remaining HeaDS people, Chloé, Diana, Jennie, Alex, Tugce, Jonas and Conor. The center grew from nothing to all of you here. I am so happy that I could join this center. I truly value the time I spent working and hanging out together with you.

A special thanks goes to binf people. All of you gave me a homely and friendly environment in Bioinformatics center, which made me adapt to a new place quickly. By the way, you've changed me in some ways, especially from a non-drinker to one who is more comfortable with life in a danish way. Thanks for all those friday beers and parties.

The other special thanks should goes to Frederik Otzen Bagger. First of all, thank you for having me in your group during my external exchange period. Importantly, I benefited a lot from your insightful feedback when I participated in the generative model project for single cell research. In addition, it is worth mentioning that working in the hospital was such a special experience for me.

Then, I would like to thank all my Chinese friends, Jingwen Liu, Yufang Deng, Jiayi Yao, Wei Du, Haifeng Zheng, Zhengfu Zhao and Haotian Zheng, who were also studying in Copenhagen. We cooked Chinese food together, traveled together, partied together, drank together. All these are great memories for me.

Last but least, I am deeply grateful to my parents. During the COVID-19 time, I haven't been able to visit you for nearly 4 years. Thanks for your unconditional support, love and encouragement all the time.

Yuhu Liang, Copenhagen, Denmark, November 2022.

List of Abbreviations

DNA	deoxyribonucleic acid
RNA	ribonucleic acid
mRNA	messenger RNA
tRNA	transfer RNA
rRNA	ribosomal RNA
A	Adenine
T	Thymine
C	Cytosine
G	Guanine
U	Uracil
HGP	Human Genome Project
NGS	Next-Generation Sequencing
1KGP	1000 Genome Project
SMRT	Single Molecule Real-Time
DBG	de Brujin Graph
OLC	Overlap-Layout-Consensus
SNP	Single Nucleotide Polymorphism
DEGs	Differential Expressed Genes
CNN	Convolutional Neural Networks
DBN	Deep Belief Networks
RNN	Recurrent Neural Networks
LSTM	Long Short Term Memory
AE	Autoencoder
VAE	Variational Autoencoder
GAN	Generative Adversarial Network
SGD	Stochastic Gradient Descent
RMSprop	Root Mean Square Prop
DGM	Deep Generative Models
DGD	Deep Generative Decoder model
SNV	Single Nucleotide Variants
BM14	Bidirectional Markov model with 28 bases as context
GTEx	Genotype-tissue Expression
TCGA	The Cancer Genome Atlas dataset

Papers included in the thesis

The findings of the PhD project are reviewed and discussed in this thesis, and the following additional research papers:
† First author
* Corresponding author

1) Yuhu Liang†, Christian Grønbæk, Piero Fariselli, Anders Krogh*
Context dependency of nucleotide probabilities and variants in human DNA. BMC Genomics 23, 87 (2022).

2) Yuhu Liang†, Iñigo Prada-Luengo†, Viktoria Schuster†, Thilde Terkelsen, Anders Krogh*
A generative model of normal tissue gene expression enables differential expression in cancer with *one* sample. Manuscript in preparation

Papers not included in the thesis

† First author
* Corresponding author

1) Christian Grønbæk†, Yuhu Liang, Desmond Elliott, Anders Krogh*
Context dependent prediction in DNA sequence using neural networks. PeerJ 10, e13666.

Contents

1 Introduction

In the 19th century, deoxyribonucleic acid (DNA) was for the first time isolated by Friedrich Miescher, a doctor from Switzerland [1]. DNA was shown to carry genetic information via a Pneumococcus experiment in 1928 [2]. Around 20 years later, Alfred Hershey and Martha Chase identified the genetic function of DNA in 1952 [3]. One year after, J. D. Watson and F. H. C. Crick reported the molecular structure of DNA [4], which had a profound influence on the scientific research of the later generations. In 1957, Crick laid out the central dogma of molecular biology, from DNA - to RNA - to proteins, which suggested that genetic information only has one direction between DNA, RNA and proteins [5]. These and many more important discoveries have laid a solid foundation for biological studying. On the other hand, the development of computer technology and the breakthrough in sequencing technology in the recent 20 years provided us with the possibility of studying life on earth by using big data.

At the core of an organisms' genetic content are nucleic acids, bio-macromolecules located in cells. There are two types of nucleic acids, called DNA and ribonucleic acid (RNA). DNA is a long polymer made up of repeating units of four different nucleotides: adenine (A), thymine (T), cytosine (C) and guanine (G). Normally, in living creatures, DNA is composed of two helical strands and bound together tightly, according to base pairing rules (A = T, and C $\equiv$ G), with hydrogen bonds, so both chain of this double stranded DNA have exactly the same genetic information. CG base pairs are more stable than AT base pairs because of the extra hydrogen bond in CG pairs. Hence, the binding strength of double chains is associated with the proportion of CG content [6]. In eukaryotes, DNA is mainly stored in the nucleus of every cell, with a samll amount in the mitochondria or chloroplast. DNA and histones combine together to form a higher-order structure called chromosomes. The human genome is comprised by 46 chromosomes, including sex chromosome X and Y [7]. The entire DNA polymer may contain hundreds of millions of nucleotides. For instance, chromosome one of the human genome, the largest chromosome, has about 220 million base pairs [8]. There are around three

billion base pairs in total in the human genome. However, in many species only a small fraction of the genome encodes proteins. Only about 1.5% of the human genome are protein-coding exons [9].

As for RNA, it has the similar structure to DNA. Commonly, RNA is a single chain polymer and generally much shorter in length than DNA. Another primary difference is that the base complementary to Adenine (A) is Uracil (U) in RNA, while in DNA the nitrogenous base is Thymine (T) [10] . There are essentially three kinds of RNA: 1. messenger RNA (mRNA) is the template for protein synthesis that carries information from the DNA; 2. transfer RNA (tRNA) which transfer amino acid by recognizing the genetic codons; 3. ribosomal RNA (rRNA) also plays an important role in the process of protein synthesis. It is a part of the ribosome that is responsible for translation in cells.

As science evolves, like physic, computer science and chemistry etc., so does biologists' understanding of DNA, RNA and other biological mechanisms. Especially in recent years, scientists used varieties of biological data as research materials, and then generated the results through computer processing, including but not limited to sequence assembly, sequence alignment, prediction of protein structure and gene expression analysis [11]. Moreover, scientists are starting to use machine learning models that let computers learn and capture the features of biological and medical datasets, which can help us in disease research, drug discovery and more.

1.1 Human Genome

The human genome is constituted of 23 chromosome pairs, including one pair of sex chromosome, each of which has hundreds or thousands genes (Figure 1.1). A gross estimation result showed that we have approximately $3.72*10^{13}$ cells [12] in the human body. All cells share the same genomic information. Genes on every chromosome do not line up next to each other, there is an intermediate intergenic region which could be regulatory elements or non-coding DNA. For a long time, people divided DNA into coding and non-coding DNA. The non-coding DNA is also commonly called 'Junk DNA', for those segments cannot be transcribe into functional RNA molecules [13]. With the development of high-throughput sequencing technology however, people began

to systematically study the function of non-coding regions. For example, transcription factors can specifically recognize some non-coding DNA in the vicinity of genes and interact with them to activate or inhibit gene expression.

During the DNA replication process, DNA repair mechanisms accurately fix mistakes, which makes the human genome seem quite stable. However, uncorrected nucleotide pairs will become permanent mutation in the next cell division if these mismatched pairs remain after the DNA repair process [14]. In another case, when a retrovirus enters the cells its RNA is converted into a double-stranded DNA by reverse transcription. The reversed DNA can be integrated into the infected cell's genome [15]. In addition, under the conditions of ionizing and ultraviolet radiations mutations of sequence or structure of the DNA chain can be induced. These are some examples for causes of mutations. Mutations that occur in somatic cells could have a chance to be present in different tissues if they happened in the very early stage of the cell development process, however these mutations cannot be inherited by offspring. Mutations that occur in the chromosomes of germ cell can be passed on to offspring.

Differences in the arrangement of the four nucleotides of A, T, C and G in genome lead to the different species. Even a single nucleotide change can make the difference in phenotypic characteristics in a population. And the deletion or mutation of genetic information is also one of the sources of many diseases. Therefore, it was an important step to obtain the full sequence information of human genome. In 1984, the Human Genome Project (HGP) plan was proposed by US government, then the great and largest biological collaboration between six countries was started in 1990. The first draft complete sequence of the human genome was generated in 2001 [17]. The first complete human genome was sequenced with Sanger sequencing [18]. The accuracy of the Sanger sequencing method is up to 99.99%, however since the time and economic costs are high it is difficult to apply widely. In 2004 and 2006 (Figure 1.2) two next-generation sequencing (NGS) technologies were introduced: 454 Life Science (Roche) and Solexa 1G (Illumina) [19, 20]. The advanced technology of NGS brings us to the next chapter of sequencing. The final complete genome with gapless assembly was finished in early2022 [21]. After the complete human genome was released, the studies on human population genetics and comparative genomics helped biologist acquire insight on genetic diversity

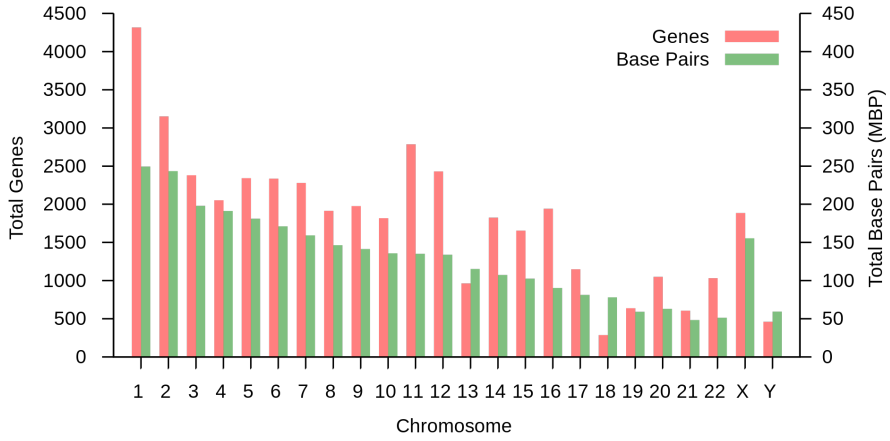

Figure 1.1. Statistics of the number of genes and base pairs on each human chromosome [16]

[22]. The international 1000 Genome Project, launched in 2008, aimed to create the most detailed map of human genetic variation across individuals, which can be used in biological and medical research [23]. The project declared more than 88 million variants. 84.7 million single nucleotide polymorphisms (SNPs), 3,600,000 insertions/deletions (indels, shorter than 50 nt.), and around 60,000 structural variants were found [24].

1.1.1 DNA Sequencing and Genome Assembly

In 1977 the first generation of DNA sequencing methods was invented by Sanger using the double-stranded termination method [25], and then the chemical degradation method was invented by Maxam and Gillbert [26]. The emergence of sequencing technology opened a new door in the field of biological research, which made it possible to decipher genes, genomes, transcriptome and proteome information. However, the low throughput was one of the fatal factors affecting its widespread application. NGS technology was developed in response to the increasing demand for sequencing throughput and time. NGS platforms can sequence millions of DNA segments at a time for a single individual via massively parallel sequencing method, which enables the sequencing of a whole genome within a short period of time [18].

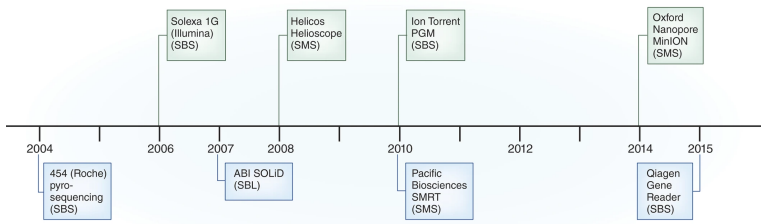

Figure 1.2. Timeline of the development of sequencing technology of each platforms. SBS: sequencing by synthesis; SMS: single molecule sequencing; SBL: sequencing by ligation. [20]

Here we focus on three established technology platforms [27]: 1) 454 method from Roche company. The basic principle is: one magnetic bead is used for one fragment of DNA, and then generate one read information for this fragment. When a dNTP is added to the template sequence, a pyrophosphoric acid will be removed to identify the base by detecting this pyrophosphoric acid. 2) ABI/SOLiD, ligation-sequencing, uses a mixture of single-chain fluorescent probes containing 8 bases instead of dNTP. These probes are paired with template DNA sequence according to the complementary base pairing rule. Every two bases determine a fluorescence signal, it is also called two probe-specific bases sequencing. 3) Illumina/Solexa method can sequence the DNA fragment along with synthesis. In this technology, modified DNA polymerase and dNTP with four kinds of fluorescent are added in the process of sequencing. It only allows a single base involved in each cycle because of the chemically cut 3'-hydroxyl-end of dNTP. And the nucleotide types can be read according to the fluorescence carried by dNTP.

The increase in sequencing throughput and the decrease in cost have led to the popularization of whole genome sequencing. NGS is getting more popular in the current research market, which is not only used in whole genome sequencing and transcriptome sequencing, and further used in population genomics, metagenomic sequencing, re-sequencing, cancer genome, genetic disease research and metabolic fields. Nonetheless, the weakness of NGS is obvious, the reads length around 250-300 bp for Illumina platform is too short [28]. Therefore, third generation sequencing technologies were published

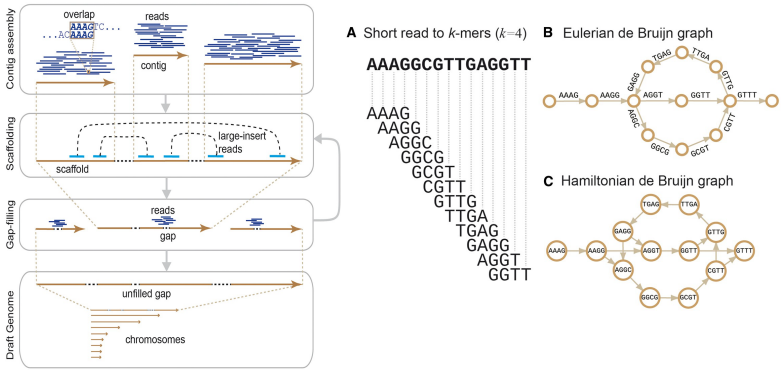

Figure 1.3. Illustration of genome assembly. A. The workflow of whole genome *de novo* assembly. B. The de Brujin graph approach. Figure modified from[31]

by PacBio (Single Molecule Real-Time, SMRT) and Oxford Nanopore Technologies. Different from the first two, the biggest highlight is that they use single molecular sequencing without doing PCR amplification. Ideally, the read length could be as long as we need [28]. The SMRT technology won't bring artificial mutations, no GC bias because we don't have to do the PCR. Secondly, the average sequencing length is around 10kb, the longest read can reach 54 kb. Another advantage is the accuracy rate of reads of up to 99% after self-correction if the sequencing depth above 10. However, the error rate for a single read is higher than the former two [29], and it is much more expensive. In Nanopore sequencing the reads are even longer, up to 150kb [28]. The other brilliant merit of Nanopore is that it can be used for RNA-seq sequencing directly, circumventing reverse transcription and PCR [30].

As we solve the sequencing problem, another important question is how do we assemble these reads in the right order. This drive the development of genome assembly software. Basically, the core algorithm for short reads assembly is using a de Brujin Graph (DBG) [32]. The software slices the sequencing reads into substrings of length k, which is called k-mer. These k-mers are used as nodes to build the DBG, from which then branches are

Assembler	Speed [a]	Memory efficiency [a]	N50 length [b]	Input data type	Assembly steps
Celera	+	+	+++	S,P,Li,L	C,S,G
ALLPATHS-LG	+	+	+++	P,Li (L[c])	E,C,S,G
ABySS	++	+++	++	S,P,Li	E,C,S
Velvet	++	++	+	S,P,Li	C,S
SPAdes	++	+++	++	P,Li	E,C,S
SOAPdenovo	+++	++	++	S,P,Li	C,S,G
SparseAssembler	++	+++	++	S,P,Li	C,S
SGA	+++	++	+	S,P,Li	E,C,S
MaSuRCA	+	+	+++	S,P,Li,L	C,S,G
Meraculous	++	++	++	P,Li	C,S,G
JR-Assembler	+	+	+++	S,P,Li	E,C,S,G

Note: +++: high; ++: medium; +:low.

In the 'Data Type' column, the symbols S, P, M and L refer to Single-end reads, Paired-end reads, Large-insert reads and Long reads, respectively.

In the 'Assembly steps' column, the symbols E, C, S and G refer to Error-correction, Contig assembly, Scaffolding and Gap-filling steps, respectively.

Figure 1.4. The list of genome assembly softwares for short reads. The + in the plot represents the evaluation in various aspects. Figure modified from[31]

remove if have low coverage or cannot be extended further. Thirdly, the DBG needs to be disassembled to get contigs, and scaffolds can be further obtained by mapping reads back to contigs. Pair-end reads will be used for closing gaps (Figure 1.3) [31]. The genome assembly softwares normally used in the data science market are shown in the Figure 1.4 [31].

Overlap-Layout-Consensus (OLC) is commonly used for 3rd generation sequencing assembly. In general, there are three steps for long reads assembly: 1) Pairwise alignment of all reads is conducted to find the overlap between segments. The overlap length of two reads needs to be higher than a chosen

threshold. 2) Obtain contigs by using the overlap information, and then generate scaffolds. 3) On the basis of reads quality, the sequence with the highest quality score is found in all of the contigs, which is called consensus. The final genome sequence can be obtained by using multi-sequence alignment for consensus. There are a few softwares that can be used for the 3rd generation sequencing assembly, including HGAP, Canu, FALCON, Flye and Miniasm[33].

1.1.2 Variants studies in Human Genome

Mutation and selection are essential and vital in the evolutionary process of species, which can increase the polymorphism of organisms to better adapt to environmental changes. Approximately, 60% mutations could have influence on proteins but won't cause harmful results for organisms [34]. These are called neutral mutations. The frequency of these mutations and selections fluctuates randomly [35, 36, 37]. However, the amount of mutations needs to be controlled since the stability of the genome is vital. There are also a certain percentage of harmful mutations which may affect the health or fertility of the organism if some essential genes are mutated. Thus, it is important to study the relationship between stability and plasticity of biological genome [38, 39]. One of the most common mutations is a SNP, which describes genomic locations where variation can occur [40].

In 2011, John C. Castle pointed that SNPs are more often detected in less conserved regions of the genomic sequence. He revealed that regulatory loci are highly conserved and have a low SNP rate. Furthermore, SNPs in coding regions have high conservation scores together with low SNP rates. It is worth mentioning that codon position three has the highest SNP rate and the lowest SNP rates was found at codon position two, which is consistent with the degeneracy of amino acids (Figure 1.5) [41].
Additionally, previous studies showed that the sequence context is one of the factors which influence mutation rates [42, 43, 44, 45]. The research was based on nucleotides neighboring SNPs, such as CpG dinucleotides. It was found that base C more often mutated to T in human genome. Few studies have used larger sequence contexts to study polymorphism rates. Varun Aggarwala and Benjamin F Voight made a model that expands the sequence context up to 7 bases, and this model could explain more than 81% of the variability in substitution probabilities [46]. Another study showed that motifs are associated with mutations as well. A novel motif associated with A to G mutations

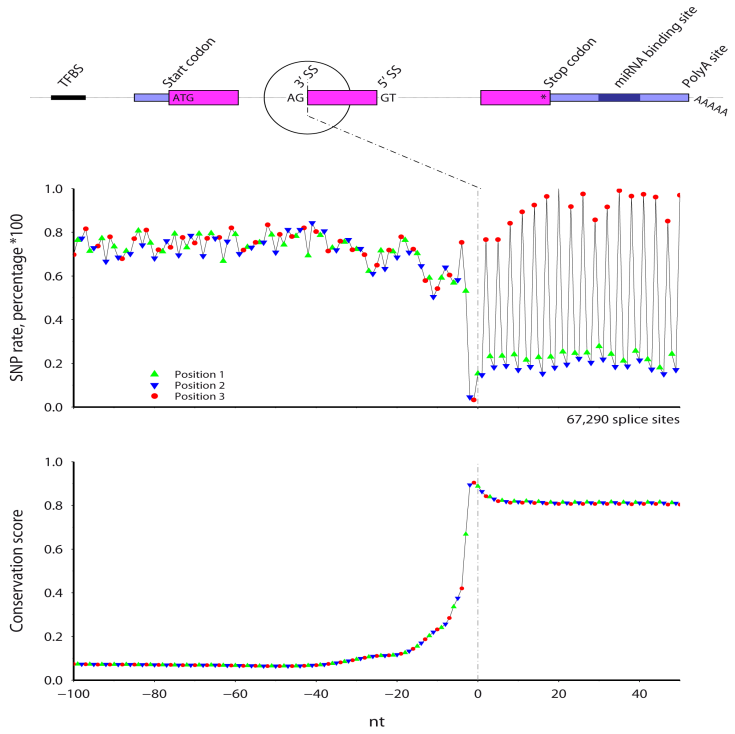

Figure 1.5. The SNP rate and conservation scores across 3' splice sites. The y-axis shows the SNP rate (top) and Conservation score from UCSC (bottom), x-axis shows the last 100 nt in introns and first 50 nt in coding regions[41]

was identified via analyzing the mutation process in the human germline and malignant melanoma [47]. Recently, we built a substitution model which was based on the context dependent nucleotide probabilities to calculate the mutability in human DNA. In our study, we found a significant C to T substitution probability in GpG contexts, ranging from 0.48 to 0.72 [39].

1.2 Human Transcriptome

Transcriptomics is based on RNA sequencing, which can be used for studying the gene expression and understanding the RNA regulations in cells. In 2005, Cheng J *etc.* indicated that around 5% of genomic sequences can be covered by detectable transfrags [48]. The annotation of the human genome was also limited, therefore the majority of the observed RNA fractions were not from known transcriptions [49]. Fortunately, rapid advances in sequencing technology have made up for that.

Unlike the genome, the transcriptome contains specific information about the time point and space in tissues [50]. Under different conditions such as environment and cell growth period the gene expression can be varied. As research has continued to evolve, the methods of transcriptomics have become increasingly diverse. For example, single cell sequencing is another great breakthrough technology, which can research the heterogeneity across cells [51, 52]. However, a challenge we faced was how to isolate high-quality single cells. In 2014 a research group from Peking University successfully prepared high quality single cell transcriptome sequencing samples[53]. In 2016 PL Ståhl and colleagues developed a high-resolution method to study which genes are active in a tissue, and this method was named spatial transcriptomics [54]. I contrast to classic transcriptomics, this method can provide positional information and quantitative gene expression values. On the other hand, the research on human disease based on transcriptomics has also made great progress. As mentioned above, gene mutations may accompany the development of diseases. There is much evidence that mutated genes can be used as a marker for disease diagnosis. A study published in Cancer Research found several dysregulated transcripts which occurred many times in multiple cancer types [55]. In another case, RNA-seq for help with the diagnosis of Mendelian genetic diseases was first published in 2017 [56].

1.2.1 RNA Sequenceing and Downstream Analysis

Once scientists were able to sequence the human genome, they turned their eyes to RNA sequencing. In the past decade, RNA-seq played an important role in differential gene expression analysis [57]. Nowadays, RNA-seq has been used in a lot of different aspects, including RNA structure, RNA translation, single cell studies, spatial transcriptomics and RNA-protein interaction [58]. Unlike DNA sequencing, transcripts are much more complex and one gene corresponds to more than one transcript because of alternative splicing [59].

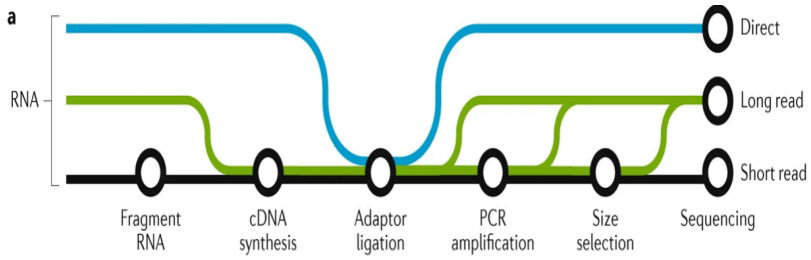

Figure 1.6. Library preparation schematic of different RNA sequencing methods. The black line shows the method for short reads RNA-seq; green for long reads RNA-seq and blue presents the direct RNA-seq methods for long reads RNA-seq. Figure modified from [58]

Secondly, RNA is not stable enough for sequencing directly, with the exception of Nanopore technology. Accordingly, we need to prepare a cDNA library before sequencing [60, 61].
In order to obtain a good quality cDNA library purification is a vital step after extracting RNAs since rRNA accounts for over 80% of the total RNA [62]. Expression of rRNA is stable in different tissues, which means rRNA can provide less useful information for our experiments. Thus, we purify RNA in many of the cases to improve the utilization of mRNA sequencing data [63]. The poly(A) method is primarily used in eukaryotes because of structural differences of the mRNA from prokaryotes. The eukaryotic mRNA has a 3' poly(A) tail which can be enriched for by Oligo(dT) magnetic beads [64, 65]. Another method is the removal of rRNA which is often used in prokaryotes RNA sequencing. The next step is to construct a cDNA library by reverse transcribing the RNA. Here, we can choose to reverse transcribe mRNA first and then fragment cDNA, or we can break the mRNA and then reverse transcribe the fragments to cDNA (Figure 1.6) [58].

In continuation of the above topic we are going to talk about the existing RNA sequencing platforms. Short-read sequencing technology from Illumina has been used to sequence more than 90%of the published data deposited in NCBI and other databases. This method is very robust, and multiple tests and comparisons showed a strong correlation between intra- and inter-platform

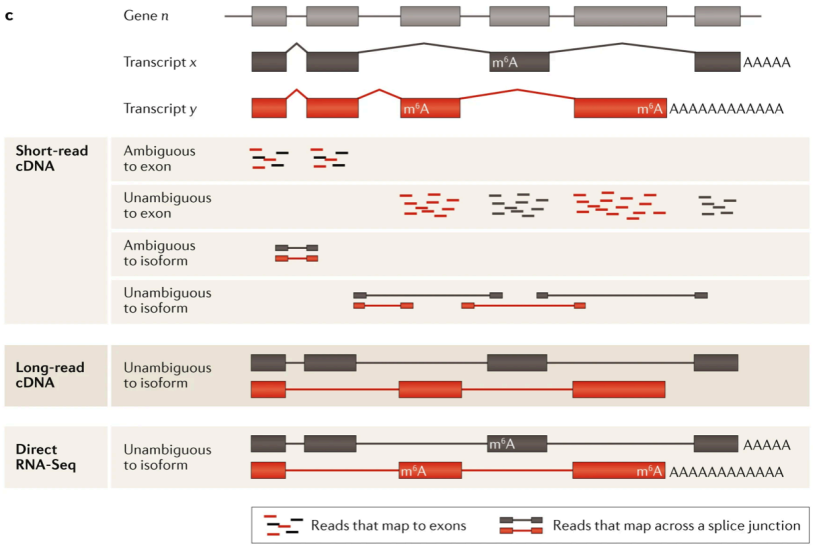

Figure 1.7. Comparison of three different RNA-seq methods. Figure modified from [58]

results of RNA-seq [66, 67]. However, this method can introduce some artificial biases and limit detection and quantitation of isoforms [68]. As for long-read RNA-seq, PacBio technology was able to detect full length transcript cDNA reads (~ 15kb) which helped with annotating novel transcripts [69]. In addition it effectively reduces the false positive rate of splice junction sites (Figure 1.7), while in short-read RNA-seq this can only be done by relying on the optimization of computing algorithms [70]. The nanopore method, long-read direct RNA-seq, has obvious advantages as we discussed in the human genome sequencing section. RNA base modifications have the potential to be detected by this method and it facilitates the discovery of updates in the field of epitranscriptomics [30, 71, 58].

Long-read sequencing has advantages for the read length and is suitable for studying the structural information of transcripts, like isoforms and splice-junctions. The short-read RNA-seq method is applicable for gene quantification and to study differentially expressed genes by comparing case and control.

At present, short-read RNA-seq is still the most commonly used technology for transcriptomics analysis. In order to productively utilize the biological data from RNA-seq a range of software has been developed, from *de novo* assembly or reads mapping to differential gene expression to pathway annotation in the end [72, 73]. The first step is to map the fastq data to a reference transcriptome or genome, after quality control using tools such as TopHat2 [74] or HISAT2 [75]. However, for some species, we don't have reference genomes with high quality annotation file. We then have to consider RNA-seq *de nove* assembly [76]. Trinity is the first transcriptome *de nove* assembly tool developed independently of genome assembly software [77]. There are some other tools such as StringTie [78] and SOAPdenovo-Trans [79]. The next important step is to quantify the abundance of transcripts. Previous studies have shown that the quantitative process can be hugely affected by the use of different quantification tools [80, 81, 82]. The five most commonly used tools are RSEM [83], CuffLinks [84], eXpress [85], Kallisto [86] and Saiffish [87]. RSEM and eXpress can quantify known genes, and RSEM uses an expectation maximization model to assign ambiguous reads. They show a high accuracy of computing along with a long calculating time. Kallisto and Saiffish are reference-free alignment softwares with shorter run time, however they can generate bias.
After the expression matrix has been obtained, a statistical model can be established for calculating the significantly differentially expressed transcripts. The four most commonly used methods for this step are DESeq2 [88], edegR [89], limma-voom [90] and EBseq [91]. Importantly, filtering and normalization should be done before this step.
Regarding the annotation, there are several R packages available such as enrichGO and enrichKEGG [92] for GO terms and KEGG [93] for pathway enrichment analysis based on the differentially expressed genes we have found. msigdbr is another R package which can be used for GTRD, TFT and Reactome annotation based on MSigDB database [94].

1.2.2 Gene Expression in Human Disease Research

Human disease is an abnormal life activity process caused by the disorder of autostable regulation under harmful effects. It may influence the function of some or all tissues, organs and systems of a human. Many diseases can cause mutations in genes or alter gene expression by turning transcription factors on or off. With the development of sequencing technology gene expression values became widely used in human disease research. For example, a study [95] used scRNA-seq data to infer regulatory networks, and a new correlation metric

was shown that can calculate the correlations among genes. The metric was applied on mouse organs [96], a mouse model of Alzheimer disease [97] and human pancreatic tissue of healthy people and type2 diabetes patients [98] for generating regulatory networks [95].

Diabetes is one of the major diseases affecting humans, with many complications. Insulin signaling is the center of metabolic control and it can prevent many chronic diseases including diabetes [99]. Research by Melissa L. Hancock *et al* has shown that the binding of insulin receptors to promoters is mediated by host cell factor 1 (a coregulator) and transcription factors. The lack of insulin increases the risk of impaired binding of insulin receptor on chromatin [100].
Gene expression data can also be sued together with DNA methylation data in disease research. Recently, a study by Palou *et al.* integrated these two data types to study cardiovascular disease [101].

Another serious human disease is cancer, recognized by the World Health Organization as the second leading cause of death worldwide. Cancer kills almost 10 million people every year. When oncogenes are activated cancer occurs in the human body. Altered cells can proliferate uncontrollably and eventually form tumors. Hence, understanding the pathogenesis of cancer and pathways involved in it is of great importance. The development of RNA-seq technology has opened up new horizons in cancer research. Breast cancer, one of the most common cancers, has been found to be associated with the Wnt signaling pathway. Typically, Wnt signalling is a conserved signaling pathway that plays a vital role in cell proliferation, differentiation and survival and calcium homeostasis [102, 103, 104, 105]. It is regulated by the β-catenin signaling pathway in the cell nucleus. TCF/LEF, a family of transcription factors, was reported to mediate β-catenin signaling. However, in 2018 a research group announced that they found Wnt/β - catenin pathway could regulate target genes independently by using whole transcriptome sequencing analysis[106].

The development of transcriptomics, including single-cell sequencing, long-read sequencing, time- and spatially-resolved transcriptomics among others, has helped scientists study and explain human diseases.
As we know, the change of gene expression values can directly reflect the impact of disease or other factors on the organism. Most of the current methods

use DEseq2 to calculate differentially expressed genes. Although the public database is growing rapidly, it is still difficult to find good controls in the process of disease. Therefore, in my second project, we developed a deep neural network model to learn the features of the healthy individuals' gene expression data across all tissues. We hope the model could help with future disease research.

1.3 Machine Learning

In 1959, the term machine learning was coined by Arthur Samuel [107]. It is a technique that expectes computer to learn knowledege like human beings, finding useful knowledge and capture information features from vast amounts of data. Compared to the human brain computers have a larger memory, faster calculation speed and more stable computing capabilities. In the past few decades, mathematicians and computer scientists have tried various methods to improve the learning capability of machines. It is hoped that the computer can handle complex scientific problems based on the excellent characteristics mentioned above. To put this concept in simply, machine learning is a process of using mathematical model with several different parameters to teach a computer to use known information and to optimize the model [108], then find the best solutions for the real problems. We should evaluate this optimized model if it can solve the similar questions, afterwards.

Briefly, the machine learning method constitutes three parts: 1) Input data, also referred to as the training set; 2) Modeling, the process by which a machine learns according to a given algorithm; 3) the Model, form an efficient model that captures the features behind the data with the optimal parameters. According to the different training methods of machine learning, it can be divided into three categories: supervised learning, unsupervised learning and reinforcement learning [109]. The fist two approaches are more often used in real-world problem-solving.

In supervised learning a mathematical model is trained on a dataset with desired output (label) and then evaluated with test data [110]. The goal of the model is to learn how to calculate the desired answer and get the correct results when it faces new input [111]. Supervised learning can be used for regression and classification tasks.

In unsupervised learning there is no label information on the training data. The computer learns, generalizes and summarizes the features of the input

dataset through the given mathematical model. In general, the interest is in the patterns discovered by the model. It can be used for classification and dimensionality reduction [112].

Machine learning is widely used to deal with a variety of real-world problems, including disease study, drug discovery, speech recognition and more [113].

1.3.1 Markov Model

A Markov model is a kind of statistical model which can be used for probability prediction [114]. It is a discrete time stochastic process with Markov property in mathematics. During the Markov process for a given present state it is irrelevant to predict the future state by using the past information. In other words, the future state only relies on the current state. Markov models are used as forecasting models in many areas, like price trends [115] and weather prediction [116].

In biological research, Markov models are often used to explain biological evolution. Choudhuri described that cis-regulatory elements prediction can be used by first order Markov model [117]. As previous studies revealed that the choice of bases is highly context dependent in human DNA [46, 38, 47]. The study of the first project in this thesis employed the Markov model to predict probabilities of DNA nucleotides. Compared with the Central model, which uses up to 7 bases in length to each side to predict the probability of the observed nucleotide, however, for the Markov model, the probability of the observed base is only related to its previous sequence. Thus, the first 14 bases sequence of this nucleotide can be used to make predictions. The number of free parameters is the same as the Central model. DNA has double strands, therefore, we predicted the probability of a base both from the forward and reverse side, where the model has 28 bases as context sequences [39].

1.3.2 Deep Neural Network

Deep neural networks are part of machine learning, inspired by the discovery of different activation states in cats' visual cells when they saw different objects [118, 119, 120], where many neurons were connected to each other [121]. Many such neural network models have been developed to deal with different

structures of data and complicated problems, including but not limited to Convolutional neural networks (CNN), Deep belief networks (DBN), Recurrent neural networks (RNN), Long Short term memory networks (LSTM), Autoencoders (AE), Variational autoencoder (VAE) and Generative adversarial networks (GAN). In this section, we mainly introduce CNN, RNN and LSTM, three neural network models commonly used in biomedical research. VAE and GAN models will go into the next section, which will introduce deep generative models in detail.

There are several important factors for machine learning with neural network models: 1) Data splitting - split the data into training set, testing set and validation set. The training set is used for the calculation of model parameters, and the validation set is used to evaluate the model performance and adjust the hyperparameters, as well as to check whether there is overfitting in the model training process. The testing set is used for the final model evaluation [122]. 2) hyperparameters - one or some external parameters in the model that need to be manually set, which cannot be trained through the training set [123]. 3) Activation functions - it is used to form the nonlinear layers in deep learning architecture and simulate the nonlinear transform from the input to the output by combining with other layers. It can help improve the robustness of the model, alleviate the problem of vanishing gradient and accelerate convergence of the model. Choosing a suitable activation function is closely related to if the model can effectively learn the training dataset [124]. Sigmoid, tanh, ReLU, leaky ReLU, Maxout, softplus and softmax are some often used activation functions. 4) Loss function - a mathematical function for evaluating the difference between a model's predicted value and the true value [125]. 5) Optimization method - an algorithm for training models with gradient descent in deep learning. Among them, the algorithms commonly used in biomedical research are Stochastic Gradient Descent (SGD) - each sample can be used to optimize the model, so the optimal θ can be calculated by randomly picking a part of the samples [126, 127, 128]; Momentum (Gradient descent with Momentum); root mean square prop (RMSprop); and Adam - an algorithm that combines the Momentum and RMSprop together [129].

Another indispensable part of machine learning is to construct an appropriate neural network model architecture. Firstly, CNN is a neural network model with many applications in image processing, speech and face recognition

etc. directions [130, 131]. It can effectively retain the features of the training set, and the complex problem is simplified by reducing the dimensionality of a large number of parameters to a small number of parameters. Generally, CNN is composed of Convolution layer, Pooling layer and Fully connected layer [127]. Since the CNN model is very good at processing image data, many biologists use this model to analyze medical image data. For instance, Kooi published a study using the CNN model to detect lesions with mammography [132]. Additionally, the CNN model can also be used for genome sequence target prediction and protein structure prediction [133, 134, 135, 136]

Most neural network models have corresponding inputs and outputs. However, for sequential data, such as text data and biological DNA sequence data, whose input is related to each other. It is necessary to develop a specialized neural network architecture, which is what we call RNN. RNN can bring the previous output results into the next hidden layer for training together. But the downside of this model is also obvious, that is, short-term memory matters much more than long-term memory. Therefore, it is limited in training on longer sequential data [137, 138, 139]. Because of the advantages of the RNN model, it was often employed to identify some specific input sequences on the genome, detect the binding sites of transcription factors and DNA methylation, and predict the secondary structure of proteins [140, 141, 142, 143].
LSTM is a derivative model of RNN, which optimizes the shortcomings of RNN's short-term memory impact. LSTM can capture important information and ignore less important content when it is used for longer sequential data training scenarios [144]. The application of this model is also very broad, such as speech recognition, images analysis, disease prediction and stock forecast [145, 146, 147, 148].

LSTM is also a good model for predicting the probability of DNA sequences. In our previous work, we developed various LSTM models to predict the probability of nucleotides in the human genome, with one of the best models showing an accuracy of nearly 54% [149]. Furthermore, Vinyals *etc* reported a possible combination of CNN and LSTM together for automatically generating image caption [150].

1.3.3 Deep Generative Models

Deep generative models (DGM) are a class of methods that combine generative models and deep learning. Since DGM has both the learning ability of deep neural networks and the prediction ability of probability models, DGM can be used to estimate the probability likelihood of each sample in the unsupervised training loop, and then generate new samples that conform to this distribution [151]. VAE and GAN are the two most popular methods.

GAN is a type of unsupervised algorithm, which is composed of a generator and discriminator. During the training, the discriminator can be used to automatically assess and continuously optimize the model, and the generator is used to fit the distribution of real data and generate extremely realistic new data [152]. Compared with other models, GAN can better simulate the distribution of data. Second, GNA is less restrictive on generator functions. And the Markov chain is not necessary for GAN. However, GAN has a serious model collapse issue, where generator generates a large amount of the same pattern of data, resulting in a lack of diversity of generated data [152, 153, 154]. Although GAN is difficult to train, much biomedical research is still based on this model. For instance, GAN was used to predict the molecular process of Alzheimer's disease by using RNA-seq data in Jinhee Park's group[155].

VAE is another type of DGM. The VAE model learns to capture data features through encoder, and converts them into a low-dimensional and easy-to-represent form in latent space, which can be decoded back to the original real data as losslessly as possible through the decoder [156]. Compared with GAN, VAE introduces latent variables, has a more complete mathematical theory, and it is easier to train the model. We can reduce the dimensionality of the latent space by methods like PCA, t-SNE, UMAP *etc.* or directly set the latent space to 2D, and visualize these data points in a 2D plot because the latent space has a good continuity during training [157, 156].
In biology, most models are not based on raw count data and require a few data-processing steps, but deep generative models can be well compatible with this [158]. Therefore, VAE is widely used in the analysis of both bulk-RNA-seq and single cell RNA-seq data [159, 160].

Last year, V. Schuster and A. Krogh published a method to train a de-

coder and presentation layer without an encoder. In this method, the decoder can be self-trained by learning the presentation layer of training samples and the weights of the decoder. In the comparison of the decoder and the autoencoder model in the image data, it is found that decoder can better learn low-dimensional data [161].
Based on this work, they developed a new model called the Deep Generative Decoder (DGD), which can be applied to single-cell gene expression data. This DGD model consists of three parts: Representation, Gaussian mixture model (GMM) and decoder. Among them, the representation has m-dimensions and is learned as trainable parameters and participates in the entire training process. As a kind of generative model, GMM is used to guide the data distribution of latent space in the DGD model. The GMM consists of K mixture components, each of which has a mean μ, and diagonal covariance $\sum$ and a mixture coefficient c. As for decoder, it can be any kind of neural network model [162, 163]. Since the DGD model can learn single-cell expression data very well. Therefore, we applied this model, in parallel, to the gene expression dataset of bulk-RNA seq data.

2 Aims and Hypotheses

This section is intended to summarize the research aims and underlying assumptions of my Ph.D. project. The primary goal is to study practical biological questions by designing mathematical models.

2.1 Aims

• Project 1
The main question we studied in this project was how to predict the probability of bases and their mutability in the human genome from the context. In detail, we aim to: 1) Implement and evaluate the predictions of the Central model, Markov Model and bidirectional Markov model for the bases at each position in the human reference genome. And compare the prediction accuracy between the models. 2) Analyze the predictive capability of bidirectional Markov model (BM14), with 28 bases as context, in different regions of the genome. 3) Investigate the prediction results of BM14 in variants. 4) Combine with the BM14 model's output, a substitution model is implemented to estimate the base substitution rate of the known variants.

• Project 2
The goal is to build a neural network model that can be used for human disease research. In this project, we aim to: 1) Train a deep generative decoder model with the GTEx dataset and assess whether it can cluster samples from the same original tissue. 2) Investigate whether the model can lead cancer samples to their corresponding normal tissue clusters. 3) Calculate differentially expressed genes and evaluate the potential application of the model in cancer research and N-of-1 sample research.

2.2 Hypotheses

• Project 1
Previous studies have shown that in the process of biological evolution, the

mutation and selection in chromosomal DNA is in a stable-variable balance. The probability of a nucleotide at a particular position also depends on its context sequences. Therefore, we hypothesize that each nucleotide in the human genome can be probabilistically predicted based on the information of its neighboring bases, and then calculate the mutation rate based on the prediction result and the information in the SNP database.

• Project 2
Gene expression data in disease tissue differs from corresponding normal tissue, which can be identified by differential expression genes. However, one of the limitations of such research is the lack of controls. We hypothesize that the feature information of different normal tissues can be learned by developing a deep generative decoder model. And when faced with disease samples, the model has the potential to find its nearest normal tissue and identify differential expression genes.

3 Methods and Dataset

This section summarizes the databases and datasets we used in my Ph.D. projects. As well as briefly introducing methods.

3.1 Database

• We downloaded the human reference genome, version GRCh38.p13, and its annotation bed files from NCBI and UCSC Table Browser, respectively.

• Different SNP datasets, including 1KPG variants data, ClinVar SNPs and Somatic SNPs were downloaded from 1KGP, NCBI and COSMIC, respectively.

• Raw count files of gene expression data from Genotype-tissue Expression normal individuals and The Cancer Genome Atlas dataset (TCGA) cancer samples were downloaded from Recount3 platform.

• Cancer driver genes were downloaded from DriverDB3 database, and breast PAM50 genes were obtained from R's build-in 'genefu' library.

3.2 Methods

• Project 1
The models are implemented in the C language, and count the context sequences for each nucleotide by a Burrows-Wheeler transform method. The details of the mathematics can be found in the attached paper 1. The software is available at GitHub: https://github.com/AndersKrogh/abwt/releases/tag/v1.2.1a

• Project 2
We developed a deep generative decoder model with an input layer-two hidden

layers-an output layer architecture. The input layer has 50 dimensions, and the two hidden layers have 500 and 8,000 hidden units, respectively. As for the output layer, its unit number is consistent with the number of genes in the training set. ReLU is used as the activation function in the model. The latent space is presented by a Gaussian mixture model, which consists of 45 mixture components. For each component, it has a 50-dimensional mean and diagonal covariance vector. The model was used to train on GTEx dataset across all tissues, and then we mainly studied Breast cancer. The details of the analysis work can be found in the second attachment.

4 Summary of the Projects

This section summarizes the mian results of the two projects included in the Ph.D. thesis.

4.1 Project 1

Context dependency of nucleotide probabilities and variants in human DNA

Status: Published - BMC Genomcics

In the DNA prediction project, we first implemented the Central model to predict the probabilities of nucleotides with a given context sequence size k. Here, the central model can be simply written as, for the observed base x_i at the position i on genome DNA, its calculated probability is

$$P(x_i|x_{i-k},\ldots,x_{i-1},x_{i+1},\ldots,x_{i+k}).$$

We find that as k increases, the prediction accuracy also increases. However, when $k = 7$, the model will have $3{*}4^{2k} = 0.8$ billion free parameters. This is the upper bound that our model can predict, because of the fixed size of the human reference genome. In order to avoid the problem of overfitting, we introduce the interpolation, that is, the predicted probabilities of the order k model are used to regularize the model of order $k + 1$. The average prediction accuracy of the Central model ($k = 7$)for the whole genome is around 49%.
Our best model, the bidirectional Markov model (BM14), achieves an average prediction accuracy of over 51% for $k = 14$. In a predictive analysis of different regions of the human genome, we found that the repetitive sequences have - as expected - a higher accuracy. Different repeat types have great differences as well, the simple repeat type is as high as 88%, but LINE1 is only about 63%. Among all regions, the least accurate prediction was in the coding region, at only 36%, but still higher than the random guessing (25%).

We analyzed the performance of the BM14 model in the SNPs database. We refer to the predicted probability of reference allele simply as Pref, and the predicted probability of alternative allele as Palt. In the analysis of the 1KGP dataset, we did a density plot to show Pref - Palt across all SNP sites, and we found that there is a peak on the right side close to 1. However, when we removed the SNPs with low allele frequency (rare SNPs), a peak appeared on the left side, which means Palt has a higher probability. The peak on the Pref side has decreased, and the density plot is gradually symmetrical. In the analysis of somatic SNPs, it was found that there is a clear trend of shifting to higher Palt direction.
Based on the results of the BM14 model, we developed a substitution model to estimate the SNPs' substitution rate. In our study we found that in the α matrix with $k = 1$ for CpG contexts, where C has a greater probability to mutate towards T, which ranges from 0.48 to 0.72. In non-CpG contexts, its maximum substitution rate is only 0.22.

4.2 Project 2

A generative model of normal tissue gene expression enables differential expression in cancer with one sample

Status: Manuscript in preparation

In this project, we used the gene expression counts data of healthy individuals from GTEx database to form a training set with 31 tissues, 17,072 samples, and 16,883 genes after removing low-expressed genes and non-protein-coding genes to train our deep generative decoder model. The latent space of the model is represented by a Gaussian Mixture model with 45 mixture components.
After 200 epochs of training, our model can cluster samples of the same tissue origin in a component. Visualizing the results, we found that almost all tissues independently dominate a component except for the three tissues of bladder, fallopian tube and cervix, which may be due to the small sample size. Uterus and ovary share a component nearly half-and-half, probably because the two tissues are too similar.

We then wrote a function that uses the trained model to generate new data points. The results showed that the majority of cancer samples we used could be correctly matched to components corresponding to their tissues. On this basis, we use the negative binomial distribution within the model to calculate the probability of being a differential expression gene.
After the model could predict differential genes, our first work was to evaluate the false positive rate of the obtained genes. We randomly selected a sample from the breast tissue in the GTEx testset, pretending to be a disease sample, and used the function to predict differentially expressed genes. At the same time, we used the same samples as case, and then used all breast tissue samples in the GTEx training set as controls to calculate differentially expressed genes by DESeq2. We repeated the experiment for 20 tims. The results indicated that our model gave nearly 0 genes (average 4.25 genes) when we set the absolute value of log2FoldChange > 1 and P-adjust value < 0.01. However, the average number of differentially expressed genes given by DESeq2 is over 75.
Following this, we assessed the predictive power of marker genes for breast cancer. We took two datasets, the breast cancer driver genes and PAM50 genes. We compared the predicted ability of our model and DESeq2 for marker genes in these two datasets by calculating enrichment scores. The results showed that our model has a higher enrichment ratio for marker genes than DESeq2, both in the multi-sample test and in the N-of-1 sample test for the four subtypes of breast cancer.
In addition, we selected 7 genes for N-of-1 sample research of the four subtypes of breast cancer, and we did this experiment 20 times for each subtype to show the robustness of our model. The results left a deep impression on us. For example, ERBB2 is a gene that is specifically up-regulated only in the HER-2 type, and our results showed that ERBB2 gene can be predicted in half of the samples in the HER-2 type. But in the other three subtypes, the gene was almost absent from the predicted significant gene list. Another example is EGFR, a gene that is specifically down-regulated only in type Luminal B. In our study, this gene could be found in almost all of the 20 experimentally predicted significant gene list in Luminal B, with negative log2FoldChange values of Luminal B type. In the other types, however, few can be found. Interestingly, MMP11 is an up-regulated gene in HER-2, Luminal A and Luminal B, but not in Basal-like type. Our results revealed that this gene can be found in the list of significantly up-regulated genes in all of the 20 experiments in the first three subtypes, while basal-like also can find this gene

13/20 times.

Furthermore, we expanded our study to 10 other cancers to evaluate our model comprehensively.

5 Conclusion and Perspectives

In this Ph.D. thesis, we introduce three different models that can predict nucleotide probability at that position based on its context sequences in human genomic DNA. And the models have been packaged into software open for use. The prediction results can help us better understand the structure of DNA. In the prediction of different regions of the genome, it is found that the prediction ability of the model for coding regions is limited. Probably because the genome needs to have the ability to encode different proteins, while the high predictive power means the low content information. Our further substitution model was found to fit the observed mutations well, especially somatic mutations. Most importantly, the α matrix can rely on smaller context sequences. In parallel, our model was employed to predict the nucleotide probabilities of *E. coli*, *A. thaliana*, *C. elegans* and *S. cerevisiae* genomes to assess the generalizability of our model. Since our model is limited by the number of free parameters. there is an upper bound to reliable prediction, we used LSTM, a deep neural network model, in parallel to predict base probabilities on the human reference genome. The predicted accuracy improved by about 2%.

Due to the development of sequencing technology, there has been an explosive growth of genome and transcriptome sequencing data. Among them, gene expression data is an intuitive representation of the regulation of an organism's genes, and it is also an often used data type for studying disease and other biomedical research. Therefore, in addition to developing simple models for research and applications on DNA to help us study mutations and genome structure. We find the deep neural network model, especially the deep generative model, has advantages in the application of gene expression data. In many cases of disease research, it is difficult to find suitable controls to screen genes that are specifically overexpressed or suppressed by disease. We found that it is possible to develop a deep generative decoder model that uses GTEx healthy individuals as a training dataset to learn the differences between human tissues and the intrinsic connections between genes from gene expression data to help biomedical research and future applications.

The results show that our model can cluster the samples of training set well in different Gaussian components, and can correctly match untrained cancer samples to components corresponding to their tissues.
DESeq2 is one of the standard tools for differential expression genes analysis. Although this method can handle single-sample research, there will still be artificial bias. Our model has better compatibility with single-sample data. In the comparative analysis with DESeq2, our model outperformed DESeq2 both in false positive rate and in enrichment analysis of cancer driver genes and PAM50 genes. Secondly, our model can also identify breast cancer subtype-specifically expressed genes when we did N-of-1 cancer research study.

Doctors also differ in how they treat and administer patients for different subtypes of the same cancer. Therefore, identifying a patient's cancer type is critical. Taking breast cancer as an example, it is thought to have four different subtypes: Basal-like, HER-2, Luminal A and Luminal B. However, some studies indicated that Luminal B may potentially continue to be divided into different subtypes. Based on the ideas mentioned above and the performance of our model in current studies, maybe we can try to use TCGA-tumor data to train the model to better distinguish different subtypes of cancer.

6 References

[1] Ralf Dahm. Discovering dna: Friedrich miescher and the early years of nucleic acid research. *Human genetics*, 122(6):565–581, 2008. 1

[2] Fred Griffith. The significance of pneumococcal types. *Epidemiology & Infection*, 27(2):113–159, 1928. 1

[3] Alfred D Hershey and Martha Chase. Independent functions of viral protein and nucleic acid in growth of bacteriophage. In *Die Entdeckung der Doppelhelix*, pages 121–139. Springer, 2017. 1

[4] James D Watson and Francis HC Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953. 1

[5] Matthew Cobb. 60 years ago, francis crick changed the logic of biology. *PLoS biology*, 15(9):e2003243, 2017. 1

[6] Tigran V Chalikian, Jens Völker, G Eric Plum, and Kenneth J Breslauer. A more unified picture for the thermodynamics of nucleic acid duplex melting: a characterization by calorimetric and volumetric techniques. *Proceedings of the National Academy of Sciences*, 96(14):7853–7858, 1999. 1

[7] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001. 1

[8] Marshal Mandelkern, John G Elias, Don Eden, and Donald M Crothers. The dimensions of dna in solution. *Journal of molecular biology*, 152(1):153–161, 1981. 1

[9] Tyra G Wolfsberg, Johanna McEntyre, and Gregory D Schuler. Guide to the draft human genome. *Nature*, 409(6822):824–826, 2001. 2

[10] L Stryer, JL Tymoczko, and JM Berg. Biochemistry 5th ed freeman. *WH and Company*, 41, 2002. 2

[11] Pierre Baldi and Søren Brunak. *Bioinformatics: the machine learning approach*. MIT press, 2001. 2

[12] Eva Bianconi, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, Maria Chiara Pelleri, Simone Tassani, Francesco Piva, et al. An estimation of the number of cells in the human body. *Annals of human biology*, 40(6):463–471, 2013. 2

[13] Alexander F Palazzo and T Ryan Gregory. The case for junk dna. *PLoS genetics*, 10(5):e1004351, 2014. 2

[14] Leslie Pray. Dna replication and causes of mutation. *Nature education*, 1(1):214, 2008. 3

[15] Jonathan P Stoye. Endogenous retroviruses: still active after all these years? *Current biology*, 11(22):R914–R916, 2001. 3

[16] Wikipedia. Human genome — Wikipedia, the free encyclopedia. 4

[17] US DOE Joint Genome Institute: Hawkins Trevor 4 Branscomb Elbert 4 Predki Paul 4 Richardson Paul 4 Wenning Sarah 4 Slezak Tom 4 Doggett Norman 4 Cheng Jan-Fang 4 Olsen Anne 4 Lucas Susan 4 Elkin Christopher 4 Uberbacher Edward 4 Frazier Marvin 4, RIKEN Genomic Sciences Center: Sakaki Yoshiyuki 9 Fujiyama Asao 9 Hattori Masahira 9 Yada Tetsushi 9 Toyoda Atsushi 9 Itoh Takehiko 9 Kawagoe Chiharu 9 Watanabe Hidemi 9 Totoki Yasushi 9 Taylor Todd 9, Genoscope, CNRS UMR-8030: Weissenbach Jean 10 Heilig Roland 10 Saurin William 10 Artiguenave Francois 10 Brottier Philippe 10 Bruls Thomas 10 Pelletier Eric 10 Robert Catherine 10 Wincker Patrick 10, Institute of Molecular Biotechnology: Rosenthal André 12 Platzer Matthias 12 Nyakatura Gerald 12 Taudien Stefan 12 Rump Andreas 12 Department of Genome Analysis, GTC Sequencing Center: Smith Douglas R. 11 Doucette-Stamm Lynn 11 Rubenfield Marc 11 Weinstock Keith 11 Lee Hong Mei 11 Dubois JoAnn 11, Beijing Genomics Institute/Human Genome Center: Yang Huanming 13 Yu Jun 13 Wang Jian 13 Huang Guyang 14 Gu Jun 15, et al. Initial sequencing and analysis of the human genome. *nature*, 409(6822):860–921, 2001. 3

[18] Ayman Grada and Kate Weinbrecht. Next-generation sequencing: methodology and application. *The Journal of investigative dermatology*, 133(8):e11, 2013. 3, 4

[19] Richard E Green, Johannes Krause, Susan E Ptak, Adrian W Briggs, Michael T Ronan, Jan F Simons, Lei Du, Michael Egholm, Jonathan M Rothberg, Maja Paunovic, et al. Analysis of one million base pairs of neanderthal dna. *Nature*, 444(7117):330–336, 2006. 3

[20] Elaine R Mardis. Dna sequencing technologies: 2006–2016. *Nature protocols*, 12(2):213–218, 2017. 3, 5

[21] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, et al. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022. 3

[22] Rasmus Nielsen. In search of rare human variants. *Nature*, 467(7319):1050–1051, 2010. 4

[23] 1000 Genomes Project Consortium et al. A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061, 2010. 4

[24] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015. 4

[25] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977. 4

[26] Allan M Maxam and Walter Gilbert. A new method for sequencing dna. *Proceedings of the National Academy of Sciences*, 74(2):560–564, 1977. 4

[27] Karl V Voelkerding, Shale A Dames, and Jacob D Durtschi. Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry*, 55(4):641–658, 2009. 5

[28] Christoph Bleidorn. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and biodiversity*, 14(1):1–8, 2016. 5, 6

[29] Pushpendra K Gupta. Single-molecule dna sequencing technologies for future genomics research. *Trends in biotechnology*, 26(11):602–611, 2008. 6

[30] Daniel R Garalde, Elizabeth A Snell, Daniel Jachimowicz, Botond Sipos, Joseph H Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, et al. Highly parallel direct rna sequencing on an array of nanopores. *Nature methods*, 15(3):201–206, 2018. 6, 12

[31] Jang-il Sohn and Jin-Wu Nam. The present and future of de novo whole-genome assembly. *Briefings in bioinformatics*, 19(1):23–40, 2018. 6, 7

[32] Phillip EC Compeau, Pavel A Pevzner, and Glenn Tesler. How to apply de bruijn graphs to genome assembly. *Nature biotechnology*, 29(11):987–991, 2011. 6

[33] Edward S Rice, Richard E Green, et al. New approaches for genome assembly and scaffolding. *Annu Rev Anim Biosci*, 7(1):17–40, 2019. 8

[34] Haiwei H Guo, Juno Choe, and Lawrence A Loeb. Protein tolerance to random amino acid change. *Proceedings of the National Academy of Sciences*, 101(25):9205–9210, 2004. 8

[35] Amir Yassin, Héloïse Bastide, Henry Chung, Michel Veuille, Jean R David, and John E Pool. Ancient balancing selection at tan underlies female colour dimorphism in drosophila erecta. *Nature communications*, 7(1):1–7, 2016. 8

[36] Nicolas Derome, Karine Métayer, Catherine Montchamp-Moreau, and Michel Veuille. Signature of selective sweep associated with the evolution of sex-ratio drive in drosophila simulans. *Genetics*, 166(3):1357–1366, 2004. 8

[37] Nicolas Derome, Emmanuelle Baudry, David Ogereau, Michel Veuille, and Catherine Montchamp-Moreau. Selective sweeps in a 2-locus model for sex-ratio meiotic drive in drosophila simulans. *Molecular Biology and Evolution*, 25(2):409–416, 2008. 8

[38] Ingo Schubert and Giang TH Vu. Genome stability and evolution: attempting a holistic view. *Trends in plant science*, 21(9):749–757, 2016. 8, 16

[39] Yuhu Liang, Christian Grønbæk, Piero Fariselli, and Anders Krogh. Context dependency of nucleotide probabilities and variants in human dna. *BMC genomics*, 23(1):1–15, 2022. 8, 9, 16

[40] Fred W Allendorf, Paul A Hohenlohe, and Gordon Luikart. Genomics and the future of conservation genetics. *Nature reviews genetics*, 11(10):697–709, 2010. 8

[41] John C Castle. Snps occur in regions with less genomic sequence conservation. *PLoS One*, 6(6):e20660, 2011. 8, 9

[42] David N Cooper and Hagop Youssoufian. The cpg dinucleotide and human genetic disease. *Human genetics*, 78(2):151–155, 1988. 8

[43] Samuel T Hess, Jonathan D Blake, and RD Blake. Wide variations in neighbor-dependent substitution rates. *Journal of molecular biology*, 236(4):1022–1033, 1994. 8

[44] Alan Hodgkinson and Adam Eyre-Walker. Variation in the mutation rate across mammalian genomes. *Nature reviews genetics*, 12(11):756–766, 2011. 8

[45] Jedidiah Carlson, Adam E Locke, Matthew Flickinger, Matthew Zawistowski, Shawn Levy, Richard M Myers, Michael Boehnke, Hyun Min Kang, Laura J Scott, Jun Z Li, et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nature communications*, 9(1):1–13, 2018. 8

[46] Varun Aggarwala and Benjamin F Voight. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature genetics*, 48(4):349–355, 2016. 8, 16

[47] Yicheng Zhu, Teresa Neeman, Von Bing Yap, and Gavin A Huttley. Statistical methods for identifying sequence motifs affecting point mutations. *Genetics*, 205(2):843–856, 2017. 9, 16

[48] Jill Cheng, Philipp Kapranov, Jorg Drenkow, Sujit Dike, Shane Brubaker, Sandeep Patel, Jeffrey Long, David Stern, Hari Tammana, Gregg Helt, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *science*, 308(5725):1149–1154, 2005. 10

[49] Martin C Frith, Michael Pheasant, and John S Mattick. The amazing complexity of the human transcriptome. *European journal of human genetics: EJHG*, 13(8):894–897, 2005. 10

[50] Mihaela Pertea. The human transcriptome: an unfinished story. *Genes*, 3(3):344–360, 2012. 10

[51] James Eberwine, Jai-Yoon Sul, Tamas Bartfai, and Junhyong Kim. The promise of single-cell sequencing. *Nature methods*, 11(1):25–27, 2014. 10

[52] Xiaoning Tang, Yongmei Huang, Jinli Lei, Hui Luo, and Xiao Zhu. The single-cell sequencing: new developments and medical applications. *Cell & bioscience*, 9(1):1–9, 2019. 10

[53] Aaron M Streets, Xiannian Zhang, Chen Cao, Yuhong Pang, Xinglong Wu, Liang Xiong, Lu Yang, Yusi Fu, Liang Zhao, Fuchou Tang, et al. Microfluidic single-cell whole-transcriptome sequencing. *Proceedings of the National Academy of Sciences*, 111(19):7048–7053, 2014. 10

[54] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016. 10

[55] Bogumil Kaczkowski, Yuji Tanaka, Hideya Kawaji, Albin Sandelin, Robin Andersson, Masayoshi Itoh, Timo Lassmann, Yoshihide Hayashizaki, Piero Carninci, and Alistair RR Forrest. Transcriptome analysis of recurrently deregulated genes across multiple cancers identifies new pan-cancer biomarkers. *Cancer research*, 76(2):216–226, 2016. 10

[56] Beryl B Cummings, Jamie L Marshall, Taru Tukiainen, Monkol Lek, Sandra Donkervoort, A Reghan Foley, Veronique Bolduc, Leigh B Waddell, Sarah A Sandaradura, Gina L O'Grady, et al. Improving genetic diagnosis in mendelian disease with transcriptome sequencing. *Science translational medicine*, 9(386):eaal5209, 2017. 10

[57] Scott J Emrich, W Brad Barbazuk, Li Li, and Patrick S Schnable. Gene discovery and annotation using lcm-454 transcriptome sequencing. *Genome research*, 17(1):69–73, 2007. 10

[58] Rory Stark, Marta Grzelak, and James Hadfield. Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019. 10, 11, 12

[59] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12):1413–1415, 2008. 10

[60] Fatih Ozsolak and Patrice M Milos. Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2):87–98, 2011. 11

[61] Radmila Hrdlickova, Masoud Toloue, and Bin Tian. Rna-seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, 8(1):e1364, 2017. 11

[62] Lu Wang, Yu Liang, Rongzhi Lin, Qiuchan Xiong, Peng Yu, Jieyi Ma, Maosheng Cheng, Hui Han, Xiaochen Wang, Ganping Wang, et al. Mettl5 mediated 18s rrna n6-methyladenosine (m6a) modification controls stem cell fate determination and neural function. *Genes & diseases*, 2020. 11

[63] Brian T Wilhelm, Samuel Marguerat, Ian Goodhead, and Jürg Bähler. Defining transcribed regions using rna-seq. *Nature protocols*, 5(2):255–266, 2010. 11

[64] Ryan D Morin, Matthew Bainbridge, Anthony Fejes, Martin Hirst, Martin Krzywinski, Trevor J Pugh, Helen McDonald, Richard Varhol, Steven JM Jones, and Marco A Marra. Profiling the hela s3 transcriptome using randomly primed cdna and massively parallel short-read sequencing. *Biotechniques*, 45(1):81–94, 2008. 11

[65] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008. 11

[66] A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature biotechnology*, 32(9):903–914, 2014. 12

[67] Sheng Li, Scott W Tighe, Charles M Nicolet, Deborah Grove, Shawn Levy, William Farmerie, Agnes Viale, Chris Wright, Peter A Schweitzer,

Yuan Gao, et al. Multi-platform assessment of transcriptome profiling using rna-seq in the abrf next-generation sequencing study. *Nature biotechnology*, 32(9):915–925, 2014. 12

[68] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, et al. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012. 12

[69] Maria Cartolano, Bruno Huettel, Benjamin Hartwig, Richard Reinhardt, and Korbinian Schneeberger. cdna library enrichment of full length transcripts for smrt long read sequencing. *PLoS One*, 11(6):e0157779, 2016. 12

[70] Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, Gunnar Rätsch, Nick Goldman, Tim J Hubbard, Jennifer Harrow, Roderic Guigó, et al. Systematic evaluation of spliced alignment programs for rna-seq data. *Nature methods*, 10(12):1185–1191, 2013. 12

[71] Piroon Jenjaroenpun, Thidathip Wongsurawat, Taylor D Wadley, Trudy M Wassenaar, Jun Liu, Qing Dai, Visanu Wanchai, Nisreen S Akel, Azemat Jamshidi-Parsian, Aime T Franco, et al. Decoding the epitranscriptional landscape from native rna sequences. *Nucleic acids research*, 49(2):e7–e7, 2021. 12

[72] Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, et al. De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. *Nature protocols*, 8(8):1494–1512, 2013. 13

[73] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):1–19, 2016. 13

[74] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):1–13, 2013. 13

[75] Daehwan Kim, Ben Langmead, and Steven L Salzberg. Hisat: a fast spliced aligner with low memory requirements. *Nature methods*, 12(4):357–360, 2015. 13

[76] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010. 13

[77] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Trinity: reconstructing a full-length transcriptome without a genome from rna-seq data. *Nature biotechnology*, 29(7):644, 2011. 13

[78] Mihaela Pertea, Daehwan Kim, Geo M Pertea, Jeffrey T Leek, and Steven L Salzberg. Transcript-level expression analysis of rna-seq experiments with hisat, stringtie and ballgown. *Nature protocols*, 11(9):1650–1667, 2016. 13

[79] Yinlong Xie, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, Guangzhu He, Shengchang Gu, Shengkang Li, et al. Soapdenovo-trans: de novo transcriptome assembly with short rna-seq reads. *Bioinformatics*, 30(12):1660–1666, 2014. 13

[80] Cheng Yang, Po-Yen Wu, Li Tong, John Phan, and May Wang. The impact of rna-seq aligners on gene expression estimation. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 462–471, 2015. 13

[81] Mingxiang Teng, Michael I Love, Carrie A Davis, Sarah Djebali, Alexander Dobin, Brenton R Graveley, Sheng Li, Christopher E Mason, Sara Olson, Dmitri Pervouchine, et al. A benchmark for rna-seq quantification pipelines. *Genome biology*, 17(1):1–12, 2016. 13

[82] Thomas P Quinn, Tamsyn M Crowley, and Mark F Richardson. Benchmarking differential expression analysis tools for rna-seq: normalization-based vs. log-ratio transformation-based methods. *BMC bioinformatics*, 19(1):1–15, 2018. 13

[83] Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):1–16, 2011. 13

[84] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562–578, 2012. 13

[85] Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature methods*, 10(1):71–73, 2013. 13

[86] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525–527, 2016. 13

[87] Rob Patro, Stephen M Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. *Nature biotechnology*, 32(5):462–464, 2014. 13

[88] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014. 13

[89] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010. 13

[90] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):1–17, 2014. 13

[91] Ning Leng, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart MG Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, 29(8):1035–1043, 2013. 13

[92] Tianzhi Wu, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Lang Zhou, Wenli Tang, LI Zhan, et al. clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141, 2021. 13

[93] Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, et al. Kegg for linking genomes to life and the environment. *Nucleic acids research*, 36(suppl_1):D480–D484, 2007. 13

[94] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011. 13

[95] Giovanni Iacono, Ramon Massoni-Badosa, and Holger Heyn. Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome biology*, 20(1):1–20, 2019. 13, 14

[96] Tabula Muris Consortium, Overall coordination Schaum Nicholas 1 Karkanias Jim 2 Neff Norma F. 2 May Andrew P. 2 Quake Stephen R. quake@ stanford. edu 2 3 f Wyss-Coray Tony twc@ stanford. edu 4 5 6 g Darmanis Spyros spyros. darmanis@ czbiohub. org 2 h, Logistical coordination Batson Joshua 2 Botvinnik Olga 2 Chen Michelle B. 3 Chen Steven 2 Green Foad 2 Jones Robert C. 3 Maynard Ashley 2 Penland Lolita 2 Pisco Angela Oliveira 2 Sit Rene V. 2 Stanley Geoffrey M. 3 Webber James T. 2 Zanini Fabio 3, and Computational data analysis Batson Joshua 2 Botvinnik Olga 2 Castro Paola 2 Croote Derek 3 Darmanis Spyros 2 DeRisi Joseph L. 2 27 Karkanias Jim 2 Pisco Angela Oliveira 2 Stanley Geoffrey M. 3 Webber James T. 2 Zanini Fabio 3. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, 2018. 14

[97] Hadas Keren-Shaul, Amit Spinrad, Assaf Weiner, Orit Matcovitch-Natan, Raz Dvir-Szternfeld, Tyler K Ulland, Eyal David, Kuti Baruch, David Lara-Astaiso, Beata Toth, et al. A unique microglia type associated with restricting development of alzheimer's disease. *Cell*, 169(7):1276–1290, 2017. 14

[98] Åsa Segerstolpe, Athanasia Palasantza, Pernilla Eliasson, Eva-Marie Andersson, Anne-Christine Andréasson, Xiaoyan Sun, Simone Picelli, Alan

Sabirsh, Maryam Clausen, Magnus K Bjursell, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell metabolism*, 24(4):593–607, 2016. 14

[99] Jérémie Boucher, André Kleinridders, and C Ronald Kahn. Insulin receptor signaling in normal and insulin-resistant states. *Cold Spring Harbor perspectives in biology*, 6(1):a009191, 2014. 14

[100] Melissa L Hancock, Rebecca C Meyer, Meeta Mistry, Radhika S Khetani, Alexandre Wagschal, Taehwan Shin, Shannan J Ho Sui, Anders M Näär, and John G Flanagan. Insulin receptor associates with promoters genome-wide and regulates gene expression. *Cell*, 177(3):722–736, 2019. 14

[101] Guillermo Palou-Márquez, Isaac Subirana, Lara Nonell, Alba Fernández-Sanlés, and Roberto Elosua. Dna methylation and gene expression integration in cardiovascular disease. *Clinical epigenetics*, 13(1):1–12, 2021. 14

[102] Catriona Y Logan and Roel Nusse. The wnt signaling pathway in development and disease. *Annu. Rev. Cell Dev. Biol.*, 20:781–810, 2004. 14

[103] Hans Clevers. Wnt/β-catenin signaling in development and disease. *Cell*, 127(3):469–480, 2006. 14

[104] Yuko Komiya and Raymond Habas. Wnt signal transduction pathways. *Organogenesis*, 4(2):68–75, 2008. 14

[105] Andy J Chien, William H Conrad, and Randall T Moon. A wnt survival guide: from flies to human disease. *Journal of Investigative Dermatology*, 129(7):1614–1627, 2009. 14

[106] Nikolaos Doumpas, Franziska Lampart, Mark D Robinson, Antonio Lentini, Colm E Nestor, Claudio Cantù, and Konrad Basler. Tcf/lef dependent and independent transcriptional regulation of wnt/β-catenin target genes. *The EMBO journal*, 38(2):e98873, 2019. 14

[107] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959. 15

[108] John R Koza, Forrest H Bennett, David Andre, and Martin A Keane. Automated design of both the topology and sizing of analog electrical circuits using genetic programming. In *Artificial intelligence in design'96*, pages 151–170. Springer, 1996. 15

[109] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006. 15

[110] Stuart J Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010. 15

[111] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018. 15

[112] V Golovko, O Ignatiuk, Yu Savitsky, T Laopoulos, A Sachenko, and L Grandinetti. Unsupervised learning for dimensionality reduction. In *Proc. of Second Int. ICSC Symposium on Engineering of Intelligent Systems EIS*, pages 140–144, 2000. 16

[113] Junyan Hu, Hanlin Niu, Joaquin Carrasco, Barry Lennox, and Farshad Arvin. Voronoi-based multi-robot autonomous exploration in unknown environments via deep reinforcement learning. *IEEE Transactions on Vehicular Technology*, 69(12):14413–14423, 2020. 16

[114] Paul A Gagniuc. *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons, 2017. 16

[115] Edmundo G de Souza e Silva, Luiz FL Legey, and Edmundo A de Souza e Silva. Forecasting oil price trends using wavelets and hidden markov models. *Energy Economics*, 32(6):1507–1519, 2010. 16

[116] Shriya Kaneriya, Sudeep Tanwar, Srushti Buddhadev, Jai Prakash Verma, Sudhanshu Tyagi, Neeraj Kumar, and Sudip Misra. A range-based approach for long-term forecast of weather using probabilistic markov model. In *2018 IEEE international conference on communications workshops (ICC workshops)*, pages 1–6. IEEE, 2018. 16

[117] Supratim Choudhuri. *Bioinformatics for beginners: genes, genomes, molecular evolution, databases and analytical tools*. Elsevier, 2014. 16

[118] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962. 16

[119] David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959. 16

[120] Chensi Cao, Feng Liu, Hai Tan, Deshou Song, Wenjie Shu, Weizhong Li, Yiming Zhou, Xiaochen Bo, and Zhi Xie. Deep learning and its applications in biomedicine. *Genomics, proteomics & bioinformatics*, 16(1):17–32, 2018. 16

[121] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982. 16

[122] Jason Brownlee. What is the difference between test and validation datasets. *Machine Learning Mastery*, 14, 2017. 17

[123] Kizito Nyuytiymbiy. Parameters and hyperparameters in machine learning and deep learning. *Medium. https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac (April 22, 2021)*, 2021. 17

[124] Jason Brownlee. How to choose an activation function for deep learning. *Machine Learning Mastery*, 2021. 17

[125] Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same? *Neural computation*, 16(5):1063–1076, 2004. 17

[126] Léon Bottou et al. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8):12, 1991. 17

[127] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 17, 18

[128] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010. 17

[129] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 17

[130] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3642–3649. IEEE, 2012. 18

[131] Masakazu Matsugu, Katsuhiko Mori, Yusuke Mitari, and Yuji Kaneda. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks*, 16(5-6):555–559, 2003. 18

[132] Thijs Kooi, Geert Litjens, Bram Van Ginneken, Albert Gubern-Mérida, Clara I Sánchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Medical image analysis*, 35:303–312, 2017. 18

[133] Zeming Lin, Jack Lanchantin, and Yanjun Qi. Must-cnn: a multilayer shift-and-stitch deep convolutional architecture for sequence-based protein structure prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. 18

[134] Sheng Wang, Shunyan Weng, Jianzhu Ma, and Qingming Tang. Deepcnf-d: predicting protein order/disorder regions by weighted deep convolutional neural fields. *International journal of molecular sciences*, 16(8):17315–17330, 2015. 18

[135] Jiyun Zhou, Qin Lu, Ruifeng Xu, Lin Gui, and Hongpeng Wang. Cnnsite: Prediction of dna-binding residues in proteins using convolutional neural network with sequence features. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 78–85. IEEE, 2016. 18

[136] Guishan Zhang, Tian Zeng, Zhiming Dai, and Xianhua Dai. Prediction of crispr/cas9 single guide rna cleavage efficiency and specificity by attention-based convolutional neural networks. *Computational and structural biotechnology journal*, 19:1445–1457, 2021. 18

[137] Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5:64–67, 2001. 18

[138] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010. 18

[139] Ahmed Tealab. Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal*, 3(2):334–340, 2018. 18

[140] Md-Nafiz Hamid and Iddo Friedberg. Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. *Bioinformatics*, 35(12):2009–2016, 2019. 18

[141] Mohanad A Deif, AA Solyman, Mehrdad Ahmadi Kamarposhti, Shahab S Band, and Rania E Hammam. A deep bidirectional recurrent neural network for identification of sars-cov-2 from viral genome sequences. *Math. Biosci. Eng*, 18(6):8933–8950, 2021. 18

[142] Hongfei Li, Yue Gong, Yifeng Liu, Hao Lin, and Guohua Wang. Detection of transcription factors binding to methylated dna by deep recurrent neural network. *Briefings in Bioinformatics*, 23(1):bbab533, 2022. 18

[143] Gianluca Pollastri, Darisz Przybylski, Burkhard Rost, and Pierre Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Structure, Function, and Bioinformatics*, 47(2):228–235, 2002. 18

[144] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 18

[145] Hasim Sak, Andrew W Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. 2014. 18

[146] R Rajeev, J Abdul Samath, and NK Karthikeyan. An intelligent recurrent neural network with long short-term memory (lstm) based batch normalization for medical image denoising. *Journal of medical systems*, 43(8):1–10, 2019. 18

[147] Yangseon Kim, Jae-Hwan Roh, and Ha Young Kim. Early forecasting of rice blast disease using long short-term memory recurrent neural networks. *Sustainability*, 10(1):34, 2017. 18

[148] Jiayu Qiu, Bin Wang, and Changjun Zhou. Forecasting stock prices with long-short term memory neural network based on attention mechanism. *PloS one*, 15(1):e0227222, 2020. 18

[149] Christian Grønbæk, Yuhu Liang, Desmond Elliott, and Anders Krogh. Context dependent prediction in dna sequence using neural networks. *PeerJ*, 10:e13666, 2022. 18

[150] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 18

[151] Lars Ruthotto and Eldad Haber. An introduction to deep generative modeling. *GAMM-Mitteilungen*, 44(2):e202100008, 2021. 19

[152] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016. 19

[153] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4502–4511, 2019. 19

[154] Vandana Kushwaha, GC Nandi, et al. Study of prevention of mode collapse in generative adversarial network (gan). In *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*, pages 1–6. IEEE, 2020. 19

[155] Jinhee Park, Hyerin Kim, Jaekwang Kim, and Mookyung Cheon. A practical application of generative adversarial networks for rna-seq analysis to predict the molecular progress of alzheimer's disease. *PLoS computational biology*, 16(7):e1008099, 2020. 19

[156] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. 19

[157] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 19

[158] Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sønderby, Tune H Pers, and Ole Winther. scvae: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415–4422, 2020. 19

[159] Gregory P Way, Michael Zietz, Vincent Rubinetti, Daniel S Himmelstein, and Casey S Greene. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. *Genome biology*, 21(1):1–27, 2020. 19

[160] Gaoyang Li, Shaliu Fu, Shuguang Wang, Chenyu Zhu, Bin Duan, Chen Tang, Xiaohan Chen, Guohui Chuai, Ping Wang, and Qi Liu. A deep generative model for multi-view profiling of single-cell rna-seq and atac-seq data. *Genome biology*, 23(1):1–23, 2022. 19

[161] Viktoria Schuster and Anders Krogh. A manifold learning perspective on representation learning: Learning decoder and representations without an encoder. *Entropy*, 23(11):1403, 2021. 20

[162] Viktoria Schuster and Anders Krogh. The deep generative decoder: Using map estimates of representations. *arXiv preprint arXiv:2110.06672*, 2021. 20

[163] Viktoria Schuster and Anders Krogh. The deep generative decoder: Map estimation of representations improves modeling of single-cell rna data. *arXiv preprint arXiv:2110.06672v2*, 2021. 20

7 List of Research Publications

7.1 Context dependency of nucleotide probabilities and variants in human DNA

Liang *et al. BMC Genomics* (2022) 23:87
https://doi.org/10.1186/s12864-021-08246-1

BMC Genomics

RESEARCH ARTICLE Open Access

Context dependency of nucleotide probabilities and variants in human DNA

Yuhu Liang[1,5], Christian Grønbæk[2], Piero Fariselli[3] and Anders Krogh[1,4,5*]

Abstract

Background: Genomic DNA has been shaped by mutational processes through evolution. The cellular machinery for error correction and repair has left its marks in the nucleotide composition along with structural and functional constraints. Therefore, the probability of observing a base in a certain position in the human genome is highly context-dependent.

Results: Here we develop context-dependent nucleotide models. We first investigate models of nucleotides conditioned on sequence context. We develop a bidirectional Markov model that use an average of the probability from a Markov model applied to both strands of the sequence and thus depends on up to 14 bases to each side of the nucleotide. We show how the genome predictability varies across different types of genomic regions. Surprisingly, this model can predict a base from its context with an average of more than 50% accuracy. For somatic variants we show a tendency towards higher probability for the variant base than for the reference base. Inspired by DNA substitution models, we develop a model of mutability that estimates a mutation matrix (called the alpha matrix) on top of the nucleotide distribution. The alpha matrix can be estimated from a much smaller context than the nucleotide model, but the final model will still depend on the full context of the nucleotide model. With the bidirectional Markov model of order 14 and an alpha matrix dependent on just one base to each side, we obtain a model that compares well with a model of mutability that estimates mutation probabilities directly conditioned on three nucleotides to each side. For somatic variants in particular, our model fits better than the simpler model. Interestingly, the model is not very sensitive to the size of the context for the alpha matrix.

Conclusions: Our study found strong context dependencies of nucleotides in the human genome. The best model uses a context of 14 nucleotides to each side. Based on these models, a substitution model was constructed that separates into the context model and a matrix dependent on a small context. The model fit somatic variants particularly well.

Keywords: DNA context, Markov model, DNA substitution model

Background

The evolution of species can be followed in chromosomal DNA, which has undergone mutations and selection, and mutational processes have been essential for the development of life on earth. On the other hand mutations need to be controlled, because if an essential gene is mutated it may result in severe disease or loss of viability. This balance between plasticity and stability is important for sustaining stable life forms [1]. The question we ask in this study is, how this balance is reflected in the local sequence properties of human DNA and how the sequence context affects mutations. More precisely, we consider models of mutability that depend on the sequence context of e.g. k bases on each side of the position in question.

It is well known that the sequence context influences mutational processes. For instance, the mutation of C to T is much more common in CpG dinucleotides than in other

*Correspondence: akrogh@di.ku.dk
[1] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
[5] Part of this work was carried out at Department of Biology, University of Copenhagen, Copenhagen, Denmark
Full list of author information is available at the end of the article

contexts in the human genome [2, 3], and previous studies have reported that the immediate neighbouring bases (up to a 7 base context) influence mutation rates [4–7]. Another study showed point mutations can be affected by sequence motifs [8]. The cellular machinery includes components for maintaining genome integrity, such as DNA repair mechanisms, which result in mutational biases [9, 10] and other processes may lead to other biases. These mechanisms together govern the intrinsic mutability. Following [11], we use the term mutability rather than mutation rate, because we are not considering the detailed evolutionary process and there is no time in our models, although the same ideas are easily applicable to estimation of context sensitive mutation rates.

Models of mutability can be estimated from observed variants by simply estimating the probability of a mutation given a context. However, such models are estimated from fairly small and biased sets of variants without utilizing the mutability foot-print in the genome. Here we propose to split the context dependent mutability into a nucleotide distribution and a variant part. The nucleotide distribution can be estimated from the whole genome and the variant part from variants, thereby allowing the two parts to have different context sizes. Due to the size of the human genome, the context dependent nucleotide distribution can be estimated from a much larger context than the variant part. The variant part can depend on a *smaller* context and can thus be estimated from a small number of variants.

In the first part of the paper, we focus on estimation of the probability of observing a base in the genome, given a context. One measure to quantify the context sensitivity is predictability. In a random sequence of nucleotides with no context sensitivity, we would only be able to predict a given base with an accuracy of 25% (random guessing), so this is the lower boundary of predictability. However, due to the mutational biasses discussed above and the repetitive nature of genomes, we would expect that a genome is more predictable than a random sequence. We show that a human genomic base can be predicted with an average of 51% using our most sophisticated model.

In the second part of the paper, we estimate a mutability model based on the context dependent nucleotide distribution found. For a fixed context dependent nucleotide distribution model, we show that the mutability is not very sensitive to the context size of the variant part. We compare to a simple mutability model conditioned on a 7 base context as in [5] and show that they differ between different types of mutations.

Knowledge of the background probability is important for a lot of models and the models described in this work can form a basis for other modelling efforts in the future. It has been shown, for instance, that a high-order Markov model can improve motif discovery over a simple background model [12]. Similarly our models of mutability can be useful in future studies of mutations in disease, where the mutability can be used to e.g. identify unexpected mutations.

Results

Context modeling of the human genome

In our first model, the Central model, (Fig. 1), we simply estimate the conditional probability of a nucleotide given k bases to each side. For base x_i at a genomic position i these probabilities are written as

$$P(x_i|x_{i-k},\ldots,x_{i-1},x_{i+1},\ldots,x_{i+k}).$$

They are estimated from the genomic frequencies of the 4 possible $(2k+1)$-mers of the given context. A $k=3$ model corresponds to a neighbourhood of 7 as used in [5], and we use this model as our baseline. Since we are estimating frequencies from all positions on both strands, they are automatically strand symmetric.

One can use other values of k as long as a model can be reliably estimated. As the 4 probabilities sum to one, there are $3*4^{2k}$ free parameters in the model, so the $k=3$ model has around 12,000 free parameters, which can easily be estimated from the 6 billion sites of the two strands of the human genome. A $k=7$ model has approximately 0.8 billion free parameters, and is thus the upper limit of what we can hope to reliably estimate for a genome like the human. Even with $k=7$ there are many contexts that occur only once or very rarely. To avoid over-fitting, we have used an interpolated Central model in which a model of order k is used to regularize a model of order $k+1$ and so on (see Methods). For our second model, we have used a central model with $k=7$ and interpolated from $k=4$.

A Markov model of order k yields probabilities of the four bases conditional on the k previous bases. A Markov model also can be used to estimate from both strands, as above, which means that for base i, it can give two different probabilities: $P(x_i|x_{i-1},\ldots,x_{i-k})$ on the direct strand and $P(\hat{x}_i|\hat{x}_{i+1},\ldots,\hat{x}_{i+k})$ on the opposite strand, where $\hat{x}_i$ means the complementary base to base x_i. Note that these models are estimated from both strands as the central models, which means that a model estimated using a 5' context is identical to the complementary of a model estimated using a 3' context and therefore, without loss of generality, we always assume 5' models.

Our third model is a bidirectional Markov model (Fig. 1) of order $k=14$, interpolated from $k=8$. It is called bidirectional, because we use the *average* between the probability of x_i from one strand and the probability of $\hat{x}_i$ from the opposite strand as explained above. Note that this model with $k=14$ has the same number of free parameters $(3*10^{14})$ as the central model with $k=7$ described above, because both use 14 bases as context. However, the bidirectional Markov model actually uses a

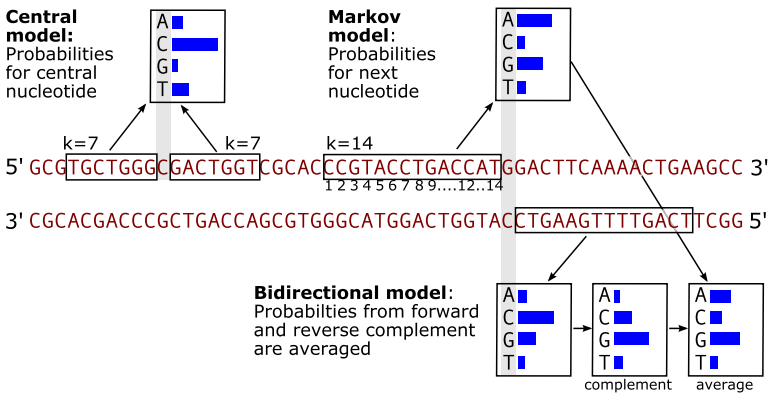

Fig. 1 Illustration of the three models used. A DNA sequence is shown with the complement sequence below. The blue histograms illustrate nucleotide probabilities. The central model with $k = 7$ (upper left) predicts the base in the middle from the adjacent nucleotides in the boxes to the left and right. For this illustration, C has highest probability, which happens to coincide with the correct nucleotide at the position. The Markov model (top right) of order $k = 14$ predicts a nucleotide from the previous 14. In this example A has highest probability although G is the actual reference probability. The bidirectional model (bottom right) use the same model on the reverse complement strand. In this example C has the highest probability, which coincides with the complement base at the position. The probabilities are translated to the direct strand and averaged with the forward model

context of 28 bases for prediction, because of the averaging over the two directions. This model is called BM14 in the following.

We have developed a program written in C that implements these different models. Instead of saving counts for each context, it dynamically calculates the count based on a Burrows-Wheeler encoded genome [13] to save memory. The performance of our models can be evaluated by the accuracy, which is the fraction of positions, where the most probable base given the context equals the actual base in the reference genome. The accuracy on the human genome is shown in Fig. 2 for the different models mentioned above (Supplementary Table S1, S2).

For the baseline model there is a strong correlation between the GC content and the accuracy on each chromosome. In Supplementary Table S3, we show GC content [14] with the accuracy and find a Pearson correlation of 0.90 for the baseline model with the lowest accuracy of around 38% for Chromosome 2–6 that has GC content of 38–40% and the highest accuracy of around 42% for chromosome 19, which has the highest GC content of 48%. For the $k = 7$ central model and BM14, the picture is less clear. Although they have correlations of 0.70 and 0.53 with GC content, the two chromosomes with the best prediction accuracy are chromosome 19 (GC 48%) and chromosome Y (GC 40%) at opposite ends of the GC scale.

For estimating the performance shown in Fig. 2, we have used leave-one-out cross-validation at the nucleotide level. It means that when estimating the probabilities for a given site in the genome, that site is excluded in the counts for model estimation. Because the k-mers overlap, one may argue that it is not proper cross-validation, but more fulfilling a minimum requirement that the site itself should not be used for estimating the model. Therefore we have also done a chromosome-based cross-validation for comparison and calculated the overall accuracies for each chromosome using a model estimated from the *other* chromosomes. The difference between nucleotide-based and chromosome-based cross validation is only 0.5 percentage points (p.p.) on average, but for the Y chromosome, it is more than 3 p.p. (Supplementary Table S1, S2 and Supplementary Fig. S1). Chromosome Y is known to differ from other chromosomes by being more heterochromatic and contain mostly repetitive regions [15], and therefore the model performs poorly on this chromosome when estimated only from other chromosomes.

With interpolation it is in principle possible to go beyond $k = 14$, because for contexts with zero counts, the probabilities are equal to a lower order estimate, so it should adapt without over-fitting. We have not explored higher k so much, but in Supplementary Fig. S2, we have run the bi-directional Markov model from $k = 10$ to $k = 20$ for different values of the interpolation constant described in Methods. The figure shows results for chromosome 20 and the model estimated from all the other chromosomes. Up to $k \simeq 14$ the models steeply improve and are almost insensitive to the interpolation constant. Above $k = 14$ we still see a monotonous improvement that seems to level off at around 52% for the best model. Chromosome 20 was chosen for this experiment, because

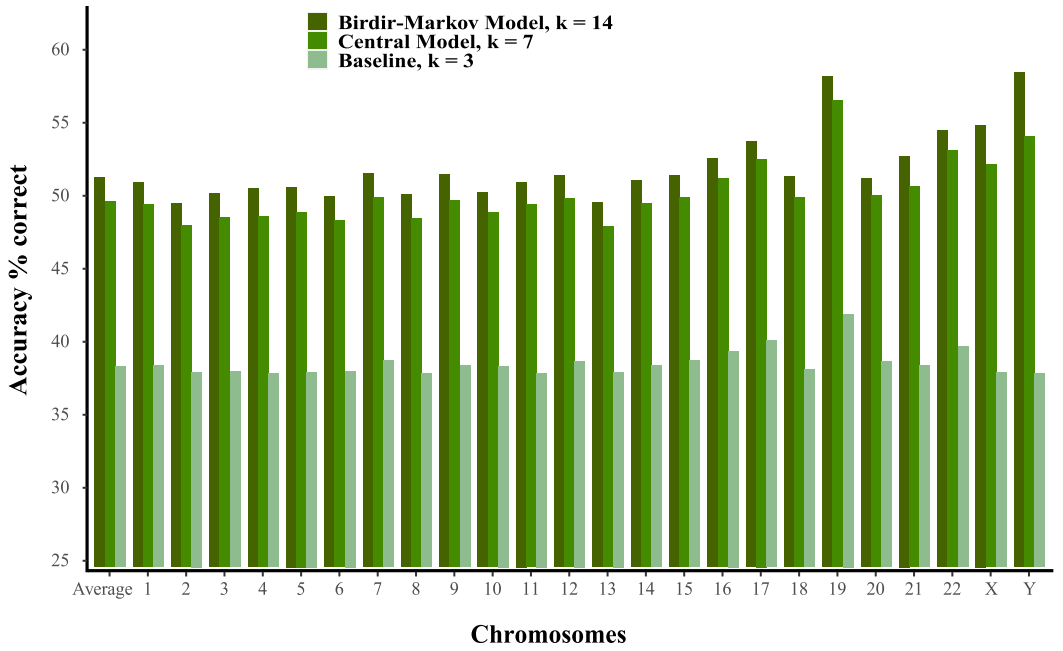

Fig. 2 Prediction accuracy for the three models. Baseline with $k = 3$, Central model with $k = 7$ and the bidirectional Markov model with $k = 14$ (BM14). The bar-plot shows accuracy for each chromosome and average accuracy on the whole genome. Results using nucleotide-based cross-validation

it is small and has a prediction accuracy similar to the average for the BM14 model. It clearly shows that interpolation improves the model although not by a great deal for $k < 14$. Importantly, interpolation at any strength ensures that zero counts do not occur, which would otherwise result in undefined probabilities.

The predictive performance of BM14 on different regions in the human genome is shown in Fig. 3. As expected, the model predicts repetitive sequences very well with an overall accuracy of 64%, but there are quite large differences between different types of repeats. The most common type of repeat in the human genome, the ALU sequences, is 87% correctly predicted, whereas LINE1 for instance is only at 63% (Supplementary Table S4). These differences are most likely due to differences in conservation of the different types of repeats.

The probability of the nucleotide in the reference genome given its context varies throughout the genome. The density of this probability, which we call the reference probability, is shown for different genomic regions in Fig. 4. For each feature except for CDS there are two peaks of which one is due to repeats. However, in positions where the reference probability is above 0.4, repeats account for a large proportion compared to other features. (Supplementary Table S5).

To further elucidate the predictability across different regions, we show in Fig. 5 the reference probabilities across human 3' and 5' splice sites that averaged over all introns annotated in Chromosome 1 (Chr1). The probability shows a large jump from a level of almost random prediction ($\sim$ 0.28) in the coding region to a fairly high value ($\sim$ 0.36) in the intron. The conservation plot in the same figure presents an opposite trend.

To test whether the model can be improved for non-repeat regions, we estimated a restricted model from everything *outside* coding regions and repeats. There is little difference between the restricted model and the full one in terms of prediction accuracy or reference probability as seen in (Supplementary Fig. S3) and we did not analyze this model further.

We briefly examined the performance of a bidirectional Markov model on some other species. Because of the smaller genome sizes, we used an interpolated bidirectional Markov model of order $k = 10$ in this analysis. The density plot of the reference probabilities (Supplementary Fig. S4A) shows that a single main peak occurs for human and *E.coli* genomes. *A. thaliana*, *C. elegans* and *S. cerevisiae* have two peaks. The peak towards low probability is enriched in coding sequence as can be seen from Supplementary Fig. S4B, where the density is plotted separately

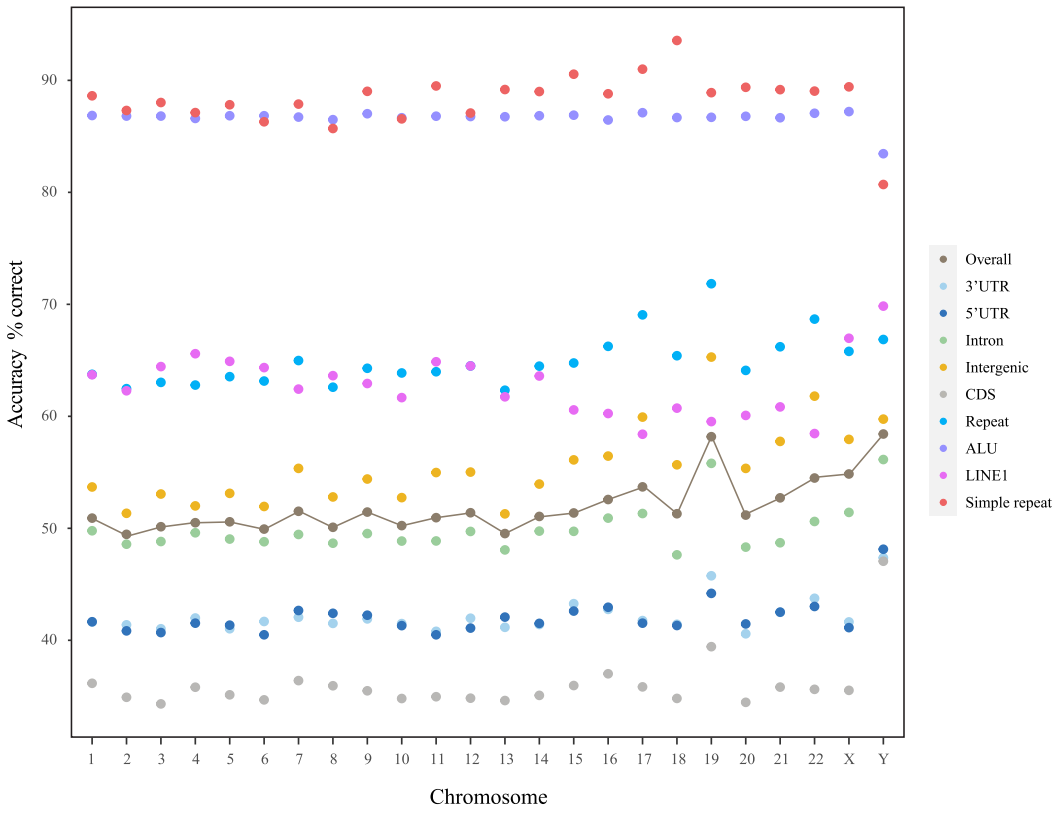

Fig. 3 Prediction accuracies for BM14 in different regions across all chromosomes. The accuracy for different features on the chromosome 1 to Y, is indicated by colored dots. The line shows the overall accuracy for each chromosome

for CDS regions and other regions. In positions where the reference probability is above $\sim$ 0.55, the density of human is higher than that of other species, which is most likely caused by repeats in human genome.

In the other eukaryotic genomes the prediction accuracy of the models were 45% for *C. elegans*, 40% for *A. thaliana*, and 38% for *S. cerevisiae*.

Variants

We next evaluated BM14 on variant datasets. We assume that our models are valid for all genomes, and variants found in population studies, such as the 1000 Genomes Project (1KGP) [16], should be predicted with the same accuracy as the corresponding positions in the reference genome. We identified $\sim$ 73 million bi-allelic single nucleotide polymorphisms (SNPs) in the 1KGP. The probability of the reference (Pref) was plotted against the probability of the alternative (Palt) shown in Fig. 6 for the $k = 7$ central model and BM14. The latter shows a larger concentration of sites in the middle of the plot. Note the unexpected asymmetry between the corners at Pref$\simeq$ 1 and Palt$\simeq$ 1 for both models.

This asymmetry is also reflected in the fact that the reference allele had the highest probability in 38.82% of cases and the alternative allele in only 24.20% for BM14. The density plot of Pref-Palt in Fig. 7A also shows a peak near 1 when all SNPs are used. However, when rare SNPs are ignored, the right peak decreases in size and a peak in the left side of the plot appears and the density becomes symmetric when only including SNPs with allele frequency above 20%. The far majority of SNPs with a reference probability higher than 0.875 in the 1KGP dataset belong to repeats.

We also compared Pref and Palt for different types of single nucleotide variants (SNVs) in coding (Fig. 7B) and non-coding regions (Supplementary Fig. S5). Clin-

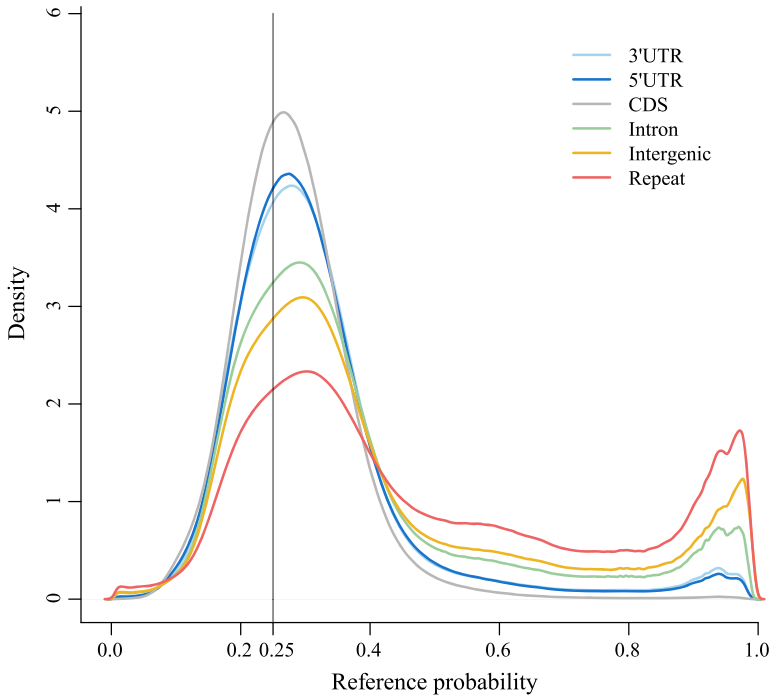

Fig. 4 Density profile of reference probabilities in different genomic regions obtained with BM14

ically relevant mutations from the Clinvar database are almost indistinguishable from 1KGP in coding regions and indeed a Kolmogorov–Smirnov (KS) test gives a p-value of 0.18 showing an insignificant difference (see Supplementary Table S6). On the contrary, somatic mutations have a clear tendency to mutate towards a more probable base (Palt > Pref) supported by a $p < 10^{-15}$ in the KS test. In non-coding regions, the somatic mutations are also shifted towards a higher probability for the alternative and have the same peak at high reference probability as 1KGP.

To see if there is a difference between damaging and benign SNPs, we show the same densities for Polyphen2 predictions [17] on Chr1 in Fig. 7C. On Chr1 there is a total of 32,841 SNPs classified as benign and 15,299 SNPs classified as damaging. There is a small, but significant (KS test ($p < 10^{-15}$, see Supplementary Table S6)), shift of the damaging SNPs towards higher probability of the alternative allele. We saw that for only 21% of damaging SNPs the reference allele had the highest probability whereas for 29% the alternative allele had the highest probability. For benign SNPs, these numbers are 26.5% and 24%. This difference is highly significant (Chi-squared test $p \simeq 10^{-9}$, see Supplementary Table S7).

Context-dependent models of substitutions

It is possible to estimate context dependent models of single nucleotide substitutions from a set of known variants. Since SNV sampling is very biased and variants are not fully observed, the context size needs to be much smaller than for the nucleotide distribution models described above. In the previously mentioned work [5] a seven nucleotide context is used. Here we want to explore the possibility of using our genome models to obtain models of substitutions. The rationale is that to maintain the context dependent nucleotide probabilities, they must be reflected in the mutability.

We assume the genome has reached approximate equilibrium. To keep this state, the mutability towards a nucleotide should be higher, the higher the probability of that nucleotide is in the given context. Therefore we set the probability of a mutation from a to b to be proportional to the probability of nucleotide b (in that context) with a constant that depends on the nucleotides and which can also depend on the context. This model is inspired by the general time-reversible stationary Markov model [18, 19], in which the off-diagonal rates are $\mu_{ab} = \alpha_{ab}\pi_b$ with symmetric α_{ab} for nucleotides $a \neq b$ and the equilibrium distribution $P(a) = \pi_a$. The mathematical theory does not

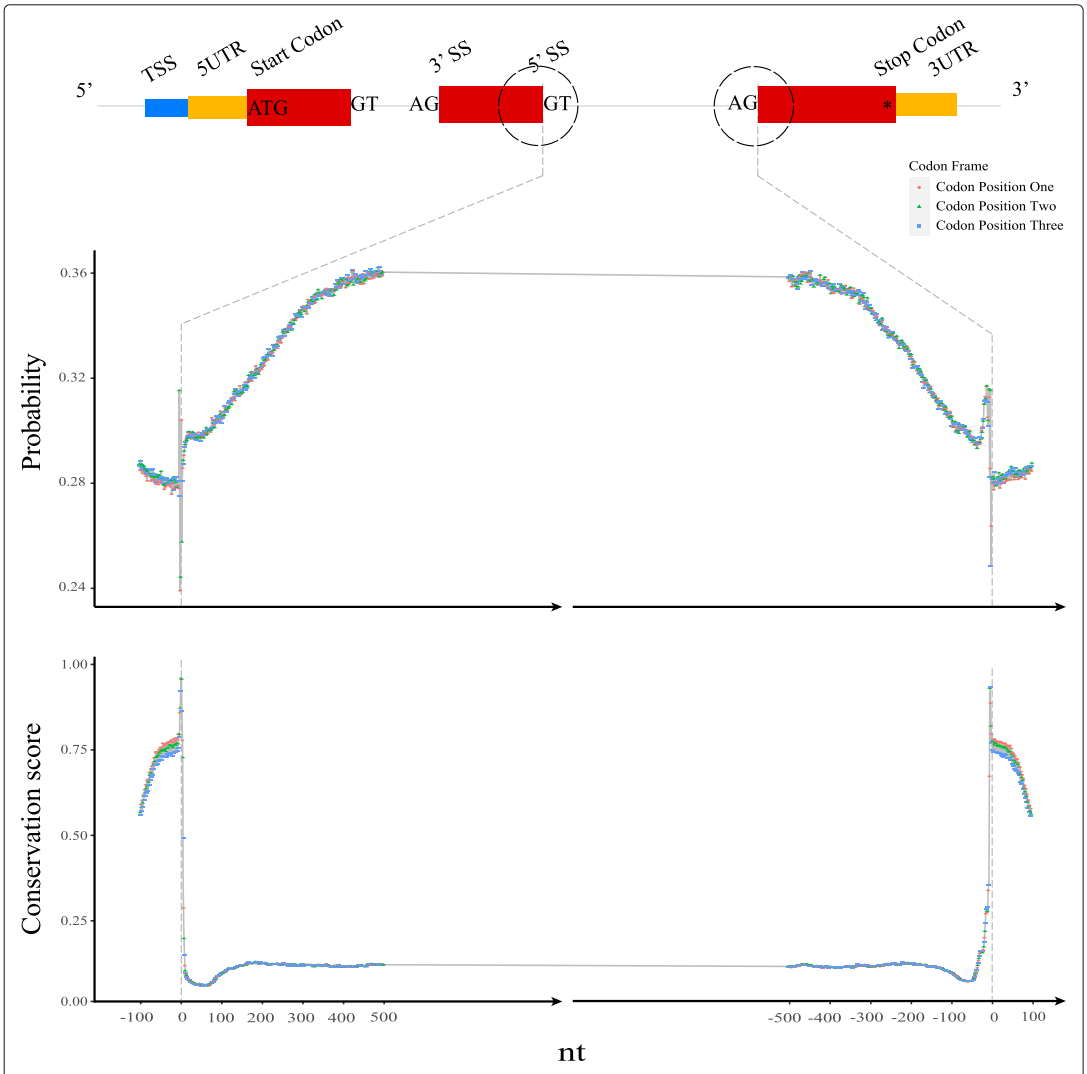

Fig. 5 Probabilities (top) and conservation score (bottom) of reference bases across 3′ and 5′ splice sites. The probabilities of the reference bases by BM14 were averaged for each position for the first/last 100 nt in coding sequence and 500 nt in introns. The conservation score is PhastCons100Way from the UCSC browser

apply directly here, because reversibility is too restrictive, so we do not require the α matrix to be symmetric, but we can still estimate an α matrix that best fits a set of variants. For lack of a better term, we call α the "alpha matrix".

Whereas the nucleotide distribution can be estimated from the whole genome using large contexts, the αs must be estimated from observed mutations. We hypothesize that the αs are less context dependent, and thus can be estimated from a smaller context than the nucleotide distributions. Details of the estimation procedure is described in Methods.

We estimated αs from all chromosomes except Chr1 for symmetrical contexts of size 0, 3, 5, and 7 ($k = 0$, 1, 2, and 3) using SNPs from the 1KGP and the BM14 model for the nucleotide distribution. The alpha matrix is shown in Table 1 (left) for $k = 0$. Notice that it is essentially strand-symmetric, but not symmetric in normal matrix-sense, so it violates reversibility. Similarly, we estimated a simple

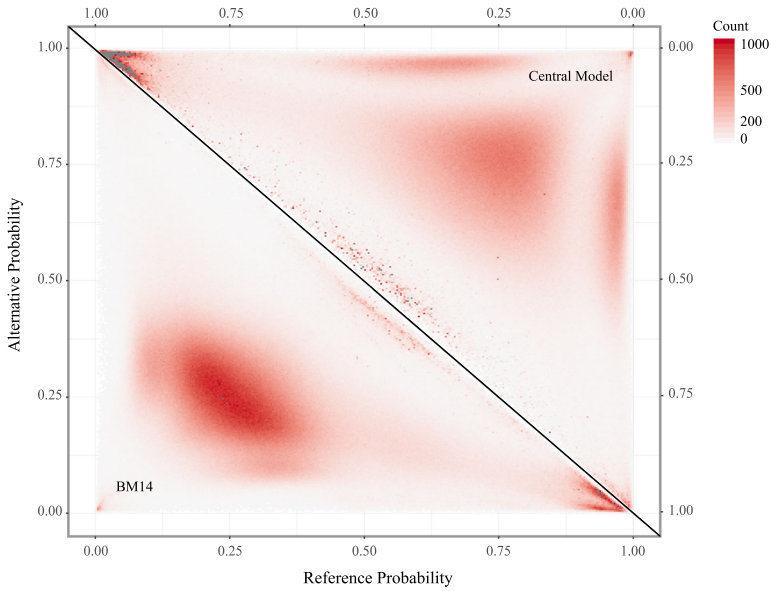

Fig. 6 Triangle plot for probabilities of Ref-Alt alleles. Probabilities of reference and alternative alleles were estimated by the $k = 7$ central model (upper right triangle) and the $k = 14$ bidirectional Markov model (BM14, lower left triangle) on SNPs from the 1000 Genomes Project

conditional model with a 7-mer context ($k = 3$) from the same data, which is called the simple model in the following. The simple model is similar to one of the models in [5], but the variants used for estimation are slightly different. The models were then applied to Chr1 where we calculated the probability of a mutation given the context for all positions with an observed SNP. The total fraction of sites with probability above 0.25 is very small for all models, see Fig. 8A. In Fig. 8B the fraction of sites with a certain mutability that has an observed SNP is plotted against mutability for some of the models. Ideally these should be linear, but we see a significant deviation from linear for the simple model and for the α models with $k > 0$. The models with $k = 1$–3 behave almost the same, and up to a substitution probability of ~ 0.25 they are very close to the simple model.

Above a mutability of 0.25, our models with $k > 0$ deviate significantly from the diagonal line. It turns out that

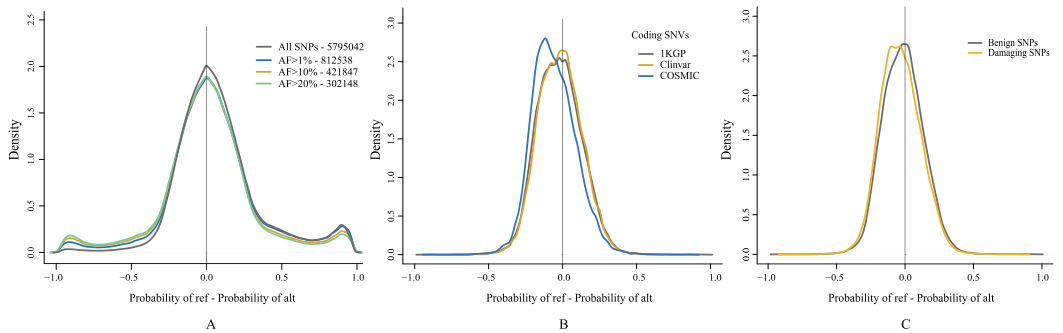

Fig. 7 Density profiles of Pref - Palt for SNPs on Chromosome 1. **A** SNPs from 1KGP. The different lines represent SNPs with allele frequencies greater than 0, 0.01, 0.1 and 0.2, respectively. SNP counts are shown in the legend after the dash. **B** Density profiles show variants of ClinVar, somatic mutations (COSMIC) and 1KGP database in coding regions. **C** Densities of damaging and benign variants predicted by Polyphen-2 based on HumanVar database and annotated on 1KGP database by ANNOVAR software

Table 1 α matrixes for $k = 0$ and $k = 1$ estimated by substitution model

	A	C	G	T
a α matrix, $k = 0$				
A	–	0.019	0.066	0.012
C	0.025	–	0.034	0.096
G	0.096	0.033	–	0.025
T	0.012	0.065	0.019	–
b α matrix, $k = 1$, CG sites only				
ACG	0.041	–	0.041	0.717
CCG	0.035	–	0.066	0.555
GCG	0.062	–	0.035	0.566
TCG	0.043	–	0.048	0.483

a: The α matrix for $k = 0$ estimated from all chromosomes except Chr1. **b**: The part of the α matrix for $k = 1$ corresponding to contexts with CG preceded by one base, so they correspond to mutations of C in these contexts

these rare reference genome sites with high substitution probability are mainly CpG sites. The alpha matrix for $k = 1$ is shown in Table 1 for the CG contexts, where it is evident that the C to T values are very large, ranging from 0.48 to 0.72, which should be compared to the largest α of 0.22 that is not a CG context, see (Supplementary Table S8). For contexts where the T has high probability according to the nucleotide distribution, the substitution probabilities will become large, because it is the product of α and the nucleotide probability. It suggests – as expected – that these substitutions are very likely at unselected positions.

We applied the model also to SNVs from Clinvar and COSMIC as shown in Fig. 8C for $k = 1$ and for the simple model. The number of variants with mutability values above 0.3 for the $k = 1$ model is relatively small. For Clinvar only 296 SNVs out of 42000 have a mutability larger than 0.3 and for COSMIC this number is 2760 out of 120000. It means that the data are noisy as seen in Fig. 8C, but it is evident that the somatic SNVs from COSMIC follow the model more closely than germline SNPs in this domain.

Discussion

We developed context dependent models of the nucleotide distribution in the human genome. The most advanced one, a bi-directional Markov model with a context of 14 nucleotides to each side, can predict a nucleotide with 51% accuracy. We use interpolation from lower orders, so it is in principle possible to go above $k = 14$, but we saw that this did not change the model very much, and the predictability of just above 50% is close to an upper limit for this type of model.

In this work our objective has been to apply simple interpretable models to the problem. Previous studies have applied neural networks to the human genome by sequence context to obtain DNA representations for other tasks. This has been used for prediction of the effect of non-coding variants [20] and the regulatory code of the accessible genome [21], for instance. The DNAbert model [22] is more related to the present work. It is a transformer neural network, which in the pre-training is trained to predict k-mers (k=3-6) from the surrounding sequence context. However, the focus is on using it for other prediction tasks, and direct comparison to our models is not possible. We have used neural networks ourselves for the same task for prediction of bases from

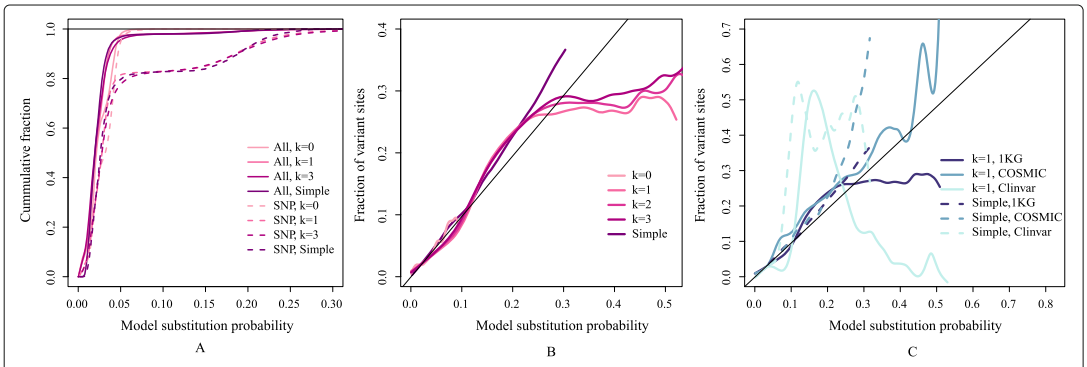

Fig. 8 Substitution model. Model substitution probabilities shown for the models with context-insensitive α (k=0), the ones with α depending on 1, 2, and 3 bases to each side (k=1, 2, 3), and the simple model conditioned on the 3 bases to each side. The model substitution probability for a site is the sum of the probabilities for the three possible substitutions. **A** The cumulative distribution of model substitution probabilities for all sites (solid lines) and for SNPs (dashed) on Chr1 shown for the five models. Note that for all models there are very few sites with substitution probability above 0.3. **B** The fraction of sites on Chr1 with an observed variant in the 1000 Genomes project (1KGP) plotted against p. The y values are SNP counts in small probability intervals (10^{-4}) divided by total counts. The curves are smoothed with splines. Estimates are noisy for larger probabilities due to low counts. **C** As **B** for SNPs in 1KGP, Clinvar and COSMIC for the $k = 1$ model and simple only. For latter two, counts are scaled so they sum to the number of SNPs in the 1KGP set for Chr1. For high mutability values there are few SNPs, so the curves are very noisy especially for Clinvar

the context [23]. Using a larger context in the neural network leads to marginally better prediction accuracy, but more importantly differences in performance depending on context.

The high predictability of our model is, to a large extent, due to repeats. It is interesting that approximately half the human genome is said to be repetitive [24], which superficially coincides with the predictability, but an exact definition of repetitive regions is a challenge and some report a higher repetitive fraction (see e.g. [25]). For *A. thaliana* and *C. elegans* the predicability was 40% and 45%, respectively, and they both have 12-13% repeats [26], and although the model was of lower order, it suggests that predictability could be used as a measure of the repetitiveness of a genome. This, however, would require more extensive analyses.

Not surprisingly, the predictability is highly dependent on the type of the genomic region. Coding regions can be predicted with only 36% accuracy, whereas Alu repeat regions are at 87% and simple repeats even higher (Fig. 3). When looking more closely at splice sites we see – as expected – a negative correlation between conservation and the probability of the reference base (Fig. 5), although such a correlation is weak, when looked at genome wide due to the lack of conservation of repeats. There are also differences between chromosomes, where especially the Y chromosome and Chr19 stand out with higher predictability than others, which is likely due to their high repeat content.

The model was applied to the genomes of *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Escherichia coli*, and *Saccharomyces cerevisiae*. Due to the smaller genome sizes a bidirectional Markov model with $k = 10$ was used. The large differences between species observed is an indication of quite different composition of genomes. Interestingly some species have two peaks in the density of the reference probability, which is partly explained by differences between coding regions and non-coding.

We compared the probability of the reference allele to the alternative allele on single nucleotide variants from the 1000 Genomes Project. There is a peak with SNPs that have a reference probability close to one, which skews the distribution away from symmetry (Fig. 7A). Almost all SNPs in this peak (with reference probabilities over 0.875) fall in repeat regions and one possibility is that some of them are mapping artefacts. They also have relatively low allele frequencies, and when considering only SNPs with high allele frequency, the plot becomes symmetric. Therefore, another factor that may explain the asymmetry is that the reference genome, which is not a genome of a single individual, contains very few rare alleles.

The difference between the probability of the reference allele and the alternative allele for coding SNVs in the 1000 Genomes Project was compared to SNVs from somatic mutations and clinically relevant SNPs from Clinvar (Fig. 7B). Here we see a statistically significant shift of somatic SNVs towards higher probability for the alternative allele, which suggest that somatic mutations tend to favor more probable bases. Similarly, we see a significant difference between damaging and benign SNPs (as classified by ANNOVAR) as seen in Fig. 7C. Surprisingly, the damaging SNPs seem to have a higher probability according to our model than benign ones.

The sequence models presented here estimate distributions of the bases for a given context and reflect inherent properties of the cellular machinery responsible for replication, error correction, and so on, as well as the physical properties of DNA, such as curvature and bendability. A mutation that moves a base closer to this distribution is likely to be more probable than one that moves it away, at least if selection is ignored. To explore this, we have derived a model that takes the context dependent nucleotide distribution into account.

In our model, we are assuming that the variation of a site in the human DNA can be described by a context sensitive continuous Markov model with a rate matrix that is a product between the nucleotide distribution and an "alpha matrix". The alpha matrix can be estimated from known variants and it can depend on a smaller context than the model for the nucleotide distribution and can be estimated from a relatively small number of SNVs. It means that our model for mutability have a very large context due to the context dependent nucleotide distribution even if the alpha matrix uses a smaller context.

The model does not depend strongly on the context size for the alpha matrix for contexts of the two neighbours or larger ($k \geq 1$). Our models behave very similarly to a simple mutability model, which is estimated from SNPs alone and a context of three nucleotides to each side except in a regime of very high mutability (Fig. 8B). Our models seem to over-estimate the SNP mutability from 1KGP when the values are larger than about 0.25. However, this is not the case for somatic mutations, and the mutations seem to be well-described by these models (Fig. 8C).

The model is inspired by the general time-reversible model from evolutionary theory, which has six free parameters corresponding to a symmetric alpha matrix, and with rates depending on the equilibrium distribution. However, although time-reversibility would be desirable, it is not likely that the context dependent nucleotide distribution we estimate is an equilibrium distribution for the entire genome. In fact, when inspecting the estimated alpha matrix for zero context (Table 1) and a context of one nucleotide to each side (Supplementary Table S8), it is evident that it is not symmetric. For the latter there are very large deviations from symmetry for contexts with NCG, where N can be any base. In these contexts, α_{CT} is

consistently 10-20 times larger than α_{TC} corresponding to a strong tendency to mutate from CG to TG.

Even if the α matrix depends on a small context, the substitution still depends on the full context of the nucleotide distribution. This construction is very attractive, because substitution models estimated from variants alone need to have small contexts due to the limited number of variants and the strong sampling biases.

Conclusions

There are strong context dependencies of nucleotides in genomes. We have shown how one can estimate a model of the nucleotide probabilities depending on contexts up to 14 nucleotides to each side. Building on these models, it was shown how it is possible to make models of mutations that combine the context dependent nucleotide probabilities with a mutation matrix, called the alpha matrix, to give mutation probabilities ("mutabilities") that depend on the same large context. It was shown that these models fit observed mutations very well and especially somatic ones. Importantly, the alpha matrix can depend on a much smaller context of just one to three bases to each side and does not depend strongly on this parameter.

These models can form the basis for a better understanding of human mutations and we believe it will be possible to use them in a wide range of applications from GWAS studies to analysis of somatic mutations.

Methods

Conditional probability models for the central base

The base at position i (chromosome, coordinate) in the reference genome is called x_i and the symmetric sequence context around it is called

$$s_i(k) = x_{i-k}, x_{i-k+1}, \ldots, x_{i-1}, x_{i+1}, x_{i+2}, \ldots, x_{i+k}. \quad (1)$$

If it is clear from the context which k, we call it s_i to ease notation. To estimate the conditional probability of base b at position i, we use the counts $n(b|s_i)$ of the occurrences in the same context throughout the reference genome (on both strands):

$$P(b|s_i) = \frac{n(b|s_i) - \delta_{b,x_i}}{N(s_i) - 1}, \quad (2)$$

where

$$N(s_i) = \sum_b n(b|s_i).$$

We use the Kronecker δ_{b,x_i}, which is 1 if $x_i = b$ and otherwise 0, to ensure that we only count *other* contexts, when estimating probabilities at position i. This is leave-one-out cross-validation and is discussed further below.

For large contexts, the counts become small and thus the probabilities cannot be reliably estimated. To interpolate between different orders of the model, we use regularization by pseudo-counts obtained from the $k-1$ model. Specifically, for order k, we define pseudo-counts

$$r(b|s_i(k)) = \gamma P(b|s_i(k-1)),$$

where γ is the strength of pseudo-counts. Now the model of order k is estimated as before, but using the actual counts plus pseudo-counts,

$$P(b|s_i(k)) = \frac{n(b|s_i(k)) - \delta_{b,x_i} + r(b|s_i(k))}{N(s_i(k)) - 1 + \gamma}.$$

The advantage of pseudo-counts is that they have minor influence, when there is plenty of data (actual counts are high), but have strong effect at low counts. With $k = 4$ counts are on average $6 * 10^9/4^9 \simeq 23000$, so we assume that psudo-counts are not needed. Therefore, our interpolated model starts with unregularized estimates for $k = 4$, and then use the pseudo-counts iteratively for $k = 5$ to $k = 7$ for the interpolated model. We used a strength of $\gamma = 100$ for the pseudo-counts (a few experiments showed that the model is relatively robust to changes in γ, see below).

Markov models

In a Markov model of order k, the probability of a base is conditioned on the k previous bases. If we redefine the k-context in (1) to be the k previous bases,

$$s_i(k) = x_{i-k}, x_{i-k+1}, \ldots, x_{i-1},$$

we can use exactly the same formulation as above. In this case however, the context size is not $2k$ letters as above, but only k letters. Therefore, one can estimate Markov models up to sizes around $k = 14$ for the human genome, and we used a model interpolated from $k = 8$ to $k = 14$ analogously to the central interpolated model described above.

Due to the interpolation, larger k are possible, and we performed a small experiment with k ranging from 10 to 20 and with four different values of the interpolation constant γ resulting in Supplementary Fig. S2. These tests were done only on chromosome 20 with a model estimated from all chromosomes *except* 20. Although small gains can be obtained with larger k values and different γ, we decided to stick to our initial choice of $k = 14$ and $\gamma = 100$.

Estimating a "forward" Markov model from both strands of the human genome will automatically make it strand-symmetric. For a given position in the genome, the model can therefore give two sets of base probabilities: one for the forward strand and one for the reverse strand. Our final Markov probabilities are the average between the two as described in the main text and referred to as bidirectional.

Cross-validation

Our way of estimating the conditional probability of seeing one of the four bases given the surrounding context can be seen as a leave-one-out procedure. In particular, the estimate depends on the reference base at the considered position as well as the context. To obtain an estimate that is independent of the reference base at the position, a natural way to proceed is to consider the average of the four base-dependent estimates over all occurrences of the given context. This average turns out to be equal to the estimate that includes all positions. To see this, average (2) over all sites (skipping the k dependence for clarity) gives the probability of a base b:

$$\bar{P}(b|s) = \frac{1}{N(s)} \sum_{b'} n(b'|s) \frac{n(b|s) - \delta_{b,b'}}{N(s) - 1}.$$

Here the base we are summing over is called b' to distinguish it from the base b in question. Since $\sum_s n(b'|s)\delta_{b,b'} = n(b|s)$, we get

$$\bar{P}(b|s) = \frac{1}{N(s)(N(s)-1)}(N(s)n(b|s) - n(b|s)) = \frac{n(b|s)}{N(s)}.$$

We also assessed our models by cross-validation by chromosomes. One chromosome was used as test data, and the remaining chromosomes as training data. We repeated this step 24 times to calculate the fraction correct predictions for each chromosome.

Substitution models

A simple model estimates mutability as the fraction of all sites with context $\hat{s}$ having a specific mutation. More specifically,

$$P_{\text{Simple}}(a \to b|\hat{s}) = \frac{n(a \to b|\hat{s})}{n(a|\hat{s})}. \tag{3}$$

Here $n(a \to b|\hat{s})$ is the number of observed mutations $a \to b$ in context $\hat{s}$ and $n(a|\hat{s})$ is the number of times we see reference base a in context $\hat{s}$ (as above). We use $\hat{s}$ to indicate that the context may be different from the context s for the genome model above. We have used this model with a symmetric context of three bases to each side, which we call the simple model.

We will now derive a continuous time Markov model with context dependent substitution rates $\mu_{ab|s}$ that takes the nucleotide distribution into account. We also assume a constant evolutionary time, which is infinitesimally small compared to the rates, so we can approximate the substitution probability by the first-order term in the Taylor expansion of an exponential

$$P(a \to b|s) \simeq \delta_{a,b} + \mu_{ab|s},$$

where time is set to 1. The diagonal rates are $-\sum_{b \neq a} \mu_{ab|s}$, so in the following we will not write the diagonal terms. For a stationary, reversible Markov model with $P(a|s)$ as equilibrium probabilities the rates can be written as

$$P(a \to b|s) \simeq \mu_{ab|s} = \alpha_{ab|s} P(b|s) \quad (a \neq b).$$

with a symmetric matrix α_{ab}. This is the general time-reversible six-parameter model (see e.g. [19]). Inspired by this model, we assume that mutability is given by the same equation, but without requiring that the nucleotide distribution is the equilibrium distribution and without requiring that α is symmetric.

The above expression factorizes the rates into the nucleotide distribution and the α-term that encapsulates the mutations. Now we assume the αs depend on a *smaller* context $\hat{s}$ than the context s for the genome model $P(a|s)$, so the above can be written as

$$P(a \to b|s) \simeq \mu_{ab|s} = \alpha_{ab|\hat{s}} P(b|s) \quad (a \neq b) \tag{4}$$

In analogy with (3), $P(a \to b|s) = n(a \to b|s)/n(a|s)$ with s instead of $\hat{s}$, so combining with the above

$$n(a \to b|s) \simeq n(a|s)\alpha_{ab|\hat{s}} P(b|s) \quad (a \neq b)$$

To estimate the αs we sum over all contexts that contains $\hat{s}$, which we write as $s|\hat{s} \subseteq s$, so

$$n(a \to b|\hat{s}) = \sum_{s|\hat{s} \subseteq s} n(a \to b|s) \simeq \alpha_{ab|\hat{s}} \sum_{s|\hat{s} \subseteq s} n(a|s) P(b|s)$$

The last sum depends only on the nucleotide distribution. It can be rewritten as a sum over all positions in the genome, where the reference base, r_i, equals a and where the context is $\hat{s}$. We call this term $Z_{ab|\hat{s}}$,

$$Z_{ab|\hat{s}} = \frac{1}{n(a|\hat{s})} \sum_{s|\hat{s} \subseteq s} n(a|s) P(b|s) = \frac{1}{n(a|\hat{s})} \sum_{i|r_i = a \wedge \hat{s} \subseteq s_i} P(b|s_i),$$

For convenience, it is normalized by $n(a|\hat{s})$, so it is the average probability of base b over all positions with reference base a and context $\hat{s}$. As an estimate of α we then have

$$\alpha_{ab|\hat{s}} = \frac{1}{Z_{ab|\hat{s}}} \frac{n(a \to b|\hat{s})}{n(a|\hat{s})} = \frac{P_{\text{Simple}}(a \to b|\hat{s})}{Z_{ab|\hat{s}}}$$

Note that we can rewrite the original probability (4) in terms of the simple model as

$$P(a \to b|s) \simeq \frac{P(b|s)}{Z_{ab|\hat{s}}} P_{\text{Simple}}(a \to b|\hat{s})$$

for $\hat{s} \subseteq s$. The factor is 1 when $\hat{s} = s$, so the models are identical as they should be when they use the same context. The equation directly shows how the wider context from the genome model can modulate the simpler estimate. If the probability of base b in context s is larger than the mean $Z_{ab|\hat{s}}$, the mutability becomes larger than in the simple model, and if it is smaller, the mutability becomes smaller.

The first order approximation assumes the rates are small. When calculating the total mutability of a site, we therefore use the approximation $1 - P(a \rightarrow a|s) \simeq 1 - e^{\mu_{aa|s}}$. For small α's it makes little difference whether it is the exponentiated form or not.

Data

The human reference genome, GRCh38.p13, was downloaded from NCBI (released March 2019 by Genome Reference Consortium). We considered only primary assemblies of chromosomes 1 to 22 and X, Y. Genomic annotation bed files were downloaded from UCSC Table Browser. These are 3'-UTR, 5'-UTR, CDS, Introns, Genes, and Repeats. Conservation scores file (PhastCons100way) was downloaded from the UCSC as well.

Variants were downloaded from the 1000 Genomes project (released March 2019, phased 20190312_biallelic_SNV_and_INDEL) in VCF format. The INDELs were filtered from 1KGP dataset.

ClinVar (clinvar_20200310.vcf) [27, 28] and somatic mutations (CosmicCodingMuts.vcf and CosmicNonCodingVariants.vcf) [29] data were obtained from NCBI and COSMIC, respectively.

The genomes and GFF files of *Arabidopsis thaliana* (TAIR10.1), *Caenorhabditis elegans* (WBcel235), *Escherichia coli* (str. K-12 substr. MG1655), *Saccharomyces cerevisiae* (R64) were downloaded from NCBI.

Data analysis

Model implementation Counting of k-mers and estimation of probabilities is implemented in the C programming language. The program counts the contexts for each site using a Burrows-Wheeler transform (BWT) [30] rather than storing the k-mers, because it is much more efficient for the interpolated models. The program is called predictDNA and relies on an index built with the program makeabwt.

One program, called makeabwt, is used for construction of an index from a fasta file containing the genome sequences. If there are multiple sequences, they are concatenated with termination symbols in between and the suffixes are sorted. The BWT is constructed from the sorted suffixes and saved. An FM index [31] is constructed to ease the search of the BWT. To limit memory usage, the values are stored in first-level checkpoints for every 2^16 positions as long integers (8 byte) and for every 256 positions the difference from the nearest first-level checkpoint is stored as a short integer (two bytes). We used an index containing both the forward and reverse complements strands of the genome.

Another program, called predictDNA, use the index to look up k-mers. This is done using the standard backward search of the BWT/FM-index [31]. The size of the resulting suffix interval equals the number of the k-mers in the genome and these are used for calculating the conditional probabilities.

The advantage of using a BWT is that the index can be used with any k and thus facilitates the interpolated models. An naive approach using table-lookup would require a new table for each value of k and a table of $4^{15} \simeq 10^9$ integers for $k = 14$, which corresponds to 4GB of memory and this would become 16GB for $k = 15$, etc. The index used for this work use around 8GB of memory.

Model Performance We calculated the probabilities of the four bases for every position in the human genome using the software predictDNA we developed. We tested different k's, but used the same interpolation constant, $\gamma = 100$, for all models. We counted the correct sites for which the reference alleles gave the highest probabilities of the four bases, to calculate the fraction correct for each chromosome.

Furthermore, we overlapped the bed files with models' outputs via bedtools [32, 33] to get the feature-specific fraction correct and predicted probabilities. These were used to obtain the performance of our models for different regions of human genome.

Based on CDS bed file and human genome fasta file, we calculated average probabilities for the positions around the human 3' and 5' splice sites. We included 500 nucleotides beforer and 100 after the 3' splice site and, similarly, 500 before and 100 after the 5' splice. Besides, we extracted the conservation scores of PhastCons100Way for the same regions [34]. Those results were shown in Fig. 5.

SNP Variants Analysis We kept only single nucleotide bi-allelic variants in 1KGP, ClinVar and COSMIC databases for the following analysis, and we filtered INDELs. Based on central model and BM14 results, reference and alternative allele probabilities for each SNP sties in these three databases were extracted. The triangle plots (Fig. 6) were made by using reference probabilities against alternative probabilities of all SNPs in 1KGP database.

In order to understand the possible asymmetry shown by the cluster of many sites in the corners of the triangle plot, we separated SNPs with allele frequency greater than 0, 0.01, 0.1 and 0.2. To present the different types of SNPs in coding and non-coding parts, we did the density plots also by using Pref minus Palt for SNPs in 1KGP, ClinVar and COSMIC databases. Additionally, we used ANNOVAR software [35] to annotate benign and damaging SNPs on 1KGP, which were predicted by PolyPhen2 [17]. These are sites associated with single genetic disease.

We developed the subsitution model to estimate the mutability of SNVs as described above. We estimated the α matrix for $k = 0, 1, 2, 3$ for all SNPs 1KGP outside of

Chr1. The model was applied to chromosome 1, where we calculated the probability of a mutation from the BM14 and the alpha matrices. These were compared to observed SNVs in 1KP, ClinVar, and COSMIC on Chr1.

Test Bi-directional Markov Model on Other Species The bi-directional Markov model with was tested on the chosen species and also human genome. We used $k = 10$, $\gamma = 100$, and interpolated from $k = 6$, instead of using the same parameters as BM14, that is because of the smaller genome size of these species. The densities of the reference base probabilities were plotted (Supplementary Fig. S4A). We separated the CDS and non-coding regions of A.*thaliana*, C. *elegans* and S. *cerevisiae* according to the GFF files and made a density plot to show the distributions of CDS and non-coding of these three species.

Software

Our software is open source and available at GiHub: https://github.com/AndersKrogh/abwt/releases/tag/v1.2.1a. We wrote several scripts in Perl and Python for data analysis and these are all available in the GitHub release. The usage of these scripts is described in README files. All the figures made in R and this code is also available.

Abbreviations

BM14: Bidirectional Markov model with 14 bases as context; p.p.: percentage points; CDS: Coding Sequence; Chr: Chromosome; Pref: Probability of reference; Palt: Probability of alternative; 1KGP: 1000 Genomes Project; SNP: Sigle nucleotide polymorphism; SNV: Single nucleotide variants; BWT: Burrows-Wheeler transform

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-08246-1.

Additional file 1: Supplementary tables: Table S1, S2, S3, S4, S5, S6, S7, S8.
Additional file 2: Supplementary figures: Figure S1, S2, S3, S4, S5.

Acknowledgements

We thank Hanne Munkholm for her big help and support with compute servers.

Authors' contributions

AK and PF initiated the project. YL and AK performed most analyses and drafted the paper with assistance from CG and PF. All authors participated in revision and approved the final version.

Funding

YL acknowledges China Scholarship Council (Grant 201804910693) for Ph.D. financial support. AK and PF acknowledge visiting fellowship support from the Italian Ministry for Education, University and Research for the programme "Dipartimenti di Eccellenza 20182022D15D18000410001" delivered to University of Torino.
The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript

Availability of data and materials

All data used in this study are publicly available. All data can be downloaded from NCBI, UCSC, 1KGP and COSMIC database as we mentioned in our methods.
The links to the genomes of the species we used:
Homosapiens (https://www.ncbi.nlm.nih.gov/genome/?term=GRCh38.p13),
Arabidopsisthaliana (https://www.ncbi.nlm.nih.gov/genome/?term=TAIR10.1),
Caenorhabditiselegans (https://www.ncbi.nlm.nih.gov/genome/?term=WBcel235),
Escherichiacoli (https://www.ncbi.nlm.nih.gov/genome/?term=Escherichia+coli),
Saccharomycescerevisiae (https://www.ncbi.nlm.nih.gov/genome/?term=Saccharomyces+cerevisiae) The CDS, Introns, 3'-UTR, 5'-UTR, Genes, Repeats and Conservation score are download from UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables)
1000 Genomes Project (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/)
clinvar_20200310 was used for Clinical SNPs analysis (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/archive_2.0/2020/)
Coding and non-coding mutations of COSMIC (https://cancer.sanger.ac.uk/cosmic/download)

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark. [2]Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark. [3]Department of Medical Sciences, University of Torino, Torino, Italy. [4]Center for Health Data Science, University of Copenhagen, Copenhagen, Denmark. [5]Part of this work was carried out at Department of Biology, University of Copenhagen, Copenhagen, Denmark.

Received: 31 August 2021 Accepted: 10 December 2021
Published online: 31 January 2022

References

1. Schubert I, Vu GT. Genome stability and evolution: attempting a holistic view. Trends Plant Sci. 2016;21:749–57.
2. Cooper DN, Youssoufian H. The CpG dinucleotide and human genetic disease. Hum Genet. 1988;78:151–5.
3. Hess ST, Blake JD, Blake RD. Wide variations in neighbor-dependent substitution rates. J Mol Biol. 1994;236:1022–33.
4. Krawczak M, Ball EV, Cooper DN. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet. 1998;63:474–88.
5. Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet. 2016;48:349–55.
6. Carlson J, Locke AE, Flickinger M, Zawistowski M, Levy S, Myers RM, Boehnke M, Kang HM, Scott LJ, Li JZ, et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. Nat Commun. 2018;1:1–13.
7. Forsdyke DR. Complementary oligonucleotides rendered discordant by single base mutations may drive speciation. Biol Theory. 2021;27:1–5.
8. Zhu Y, Neeman T, Yap VB, Huttley GA. Statistical methods for identifying sequence motifs affecting point mutations. Genetics. 2017;205:843–56.
9. Lind PA, Andersson DI. Whole-genome mutational biases in bacteria. Proc Natl Acad Sci. 2008;105:17878–83.
10. Pearson CE, Edamura KN, Cleary JD. Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet. 2005;6:729–42.
11. Zavolan M, Kepler TB. Statistical inference of sequence-dependent mutation rates. Curr Opin Genet Dev. 2001;11:612–5.

12. Thijs G, Lescot M, Marchal K, Rombauts S, B. DM, Rouze P, Moreau Y. A higher order background model improves the detection of regulatory elements by Gibbs sampling. Bioinformatics. 2001;17:1113–22.
13. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics. 2009;25:1754–60.
14. Piovesan A, Pelleri MC, Antonaros F, Strippoli P, Caracausi M, Vitale L. On the length, weight and gc content of the human genome. BMC Res Notes. 2019;12(1):1–7.
15. Bachtrog D, Charlesworth B. Towards a complete sequence of the human Y chromosome. Genome Biol. 2001;2:1016–1.
16. Consortium TGP. A global reference for human genetic variation. Nature. 2015;526:68–74.
17. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
18. Tavaré S. Some probabilistic and statistical problems in the analysis of dna sequences. Lect Math Life Sci. 1986;17:57–86.
19. Felsenstein J, Felenstein J. Inferring Phylogenies, vol 2. Sunderland: Sinauer Associates; 2004.
20. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods. 2015;12(10):931–4.
21. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016;26:990–9.
22. Ji Y, Zhou Z, Liu H, Davuluri RV. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. Bioinformatics. 2021;37(15):2112–20.
23. Grønbæk C, Liang Y, Elliott D, Krogh A. Prediction of DNA from context using neural networks. bioRxiv. 2021.
24. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
25. de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS genetics. 2011;7:1002384.
26. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. Unknown Month 2013. http://www.repeatmasker.org.
27. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46:1062–7.
28. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;4:980–5.
29. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al. Cosmic: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2019;47:941–47.
30. Burrows M, Wheeler DJ. A block-sorting lossless data compression algorithm. Technical report. 1994.
31. Ferragina P, Manzini G. Opportunistic data structures with applications. In: Proceedings 41st Annual Symposium on Foundations of Computer Science. IEEE; 2000. p. 390–8.
32. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
33. Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. Curr Protoc Bioinforma. 2014;47:11–2.
34. Castle JC. SNPs occur in regions with less genomic sequence conservation. PLoS ONE. 2011;6:20660.
35. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:164.

Publisher's Note

Table S1: The predicted accuracy and cross validation of baseline model, k = 3.

Central Model, order k=3 (alpha = 100)			
Chromosome	Number of sites		Fraction correct
	total	correct	
1	230480065	88449502	0.383762
2	240548028	91245767	0.379325
3	198099967	75176460	0.379487
4	189752565	71817909	0.378482
5	181265162	68770233	0.379390
6	170078432	64622953	0.379960
7	158970017	61563427	0.387264
8	144768070	54795829	0.378508
9	121790280	46736090	0.383742
10	133262493	51086053	0.383349
11	134533646	50943862	0.378670
12	133137642	51492163	0.386759
13	97982998	37165783	0.379308
14	90568005	34730400	0.383473
15	84641219	32750038	0.386928
16	81805817	32208050	0.393713
17	82919922	33234391	0.400801
18	80089284	30501920	0.380849
19	58440710	24486635	0.419000
20	63943723	24697782	0.386242
21	40088313	15374295	0.383511
22	39159489	15541934	0.396888
X	154892834	58708281	0.379025
Y	26414686	9994145	0.378356
Total	2937633367	1126093902	0.383334

Central Model, order k=3 (alpha = 100)			
Chromosome	Number of sites		Cross Validation
	total	correct	
1	230480065	89886242	0.389996
2	240548028	92736305	0.385521
3	198099967	76409704	0.385713
4	189752565	72982568	0.384620
5	181265162	69910974	0.385683
6	170078432	65676753	0.386150
7	158970017	62561156	0.393541
8	144768070	55693874	0.384711
9	121790280	47496328	0.389985
10	133262493	51903471	0.389483
11	134533646	51754529	0.384696
12	133137642	52311190	0.392911
13	97982998	37790335	0.385683
14	90568005	35297775	0.389738
15	84641219	33277492	0.393159
16	81805817	32700600	0.399734
17	82919922	33730878	0.406789
18	80089284	31000166	0.387070
19	58440710	24842890	0.425096
20	63943723	25084465	0.392290
21	40088313	15621913	0.389687
22	39159489	15784648	0.403086
X	154892834	59585778	0.384690
Y	26414686	10148446	0.384197
Total	2937633367	1144188480	0.389493

Table S2: The predicted accuracy for Central model (k = 7) and Bidir-Markov model (k = 14).

Central Model, order k=4-7 (alpha = 100)			
Chromosome	Number of sites		Fraction correct
	total	correct	
1	230478925	113814952	0.493819
2	240547772	115376423	0.479640
3	198099743	96126911	0.485245
4	189752429	92180027	0.485791
5	181264874	88608732	0.488836
6	170078312	82169320	0.483126
7	158969865	79257155	0.498567
8	144767982	70168345	0.484695
9	121789920	60514603	0.496877
10	133261869	65113913	0.488616
11	134533518	66481601	0.494164
12	133137410	66291292	0.497916
13	97982830	46914046	0.478799
14	90567813	44809395	0.494761
15	84641083	42237607	0.499020
16	81805649	41867514	0.511792
17	82919546	43550084	0.525209
18	80088914	39937510	0.498665
19	58440646	33047717	0.565492
20	63943011	31995875	0.500381
21	40087905	20299423	0.506373
22	39159105	20800461	0.531178
X	154892583	80792188	0.521601
Y	26414219	14284130	0.540774
Total	2937625923	1456639224	0.495856

Bidir, interpol. Markov chain, order k=8-14 (alpha = 100)				
Chromosome	Number of sites		Fraction correct	Cross Validation
	total	correct		
1	230477064	117325638	0.509056	0.503261
2	240547335	118970085	0.494581	0.491506
3	198099351	99318219	0.501356	0.498554
4	189752191	95820651	0.504978	0.502674
5	181264370	91682968	0.505797	0.502634
6	170078102	84914118	0.499265	0.496488
7	158969599	81908691	0.515248	0.509518
8	144767828	72514932	0.500905	0.495853
9	121789290	62659881	0.514494	0.506384
10	133260788	66945484	0.502364	0.497800
11	134533294	68544761	0.509500	0.503384
12	133137004	68419682	0.513904	0.510150
13	97982536	48527934	0.495271	0.493617
14	90567477	46246593	0.510631	0.509077
15	84640845	43483462	0.513741	0.503680
16	81805355	43006907	0.525722	0.516664
17	82918900	44528474	0.537012	0.524948
18	80088309	41085342	0.513000	0.505543
19	58440534	34002677	0.581834	0.573294
20	63941765	32723225	0.511766	0.506975
21	40087191	21132708	0.527169	0.519941
22	39158433	21336374	0.544873	0.537980
X	154892149	84956352	0.548487	0.540570
Y	26413407	15429883	0.584169	0.551892
Total	2937613117	1505485041	0.512486	0.507172

The cross validation result of Bidir-Markov model shows on the right table.

Table S3: Spearman correlation of predicted accuracies and GC% for each Chromosome

Chromosome	GC%	Spearman correlation of Accuracy - GC% Fraction correct Baseline	Central Model (k = 7)	BM14
1	41.72	0.383762	0.493819	0.509056
2	40.23	0.379325	0.479640	0.494581
3	39.67	0.379487	0.485245	0.501356
4	38.24	0.378482	0.485791	0.504978
5	39.51	0.379390	0.488836	0.505797
6	39.61	0.379960	0.483126	0.499265
7	40.70	0.387264	0.498567	0.515248
8	40.16	0.378508	0.484695	0.500905
9	41.28	0.383742	0.496877	0.514494
10	41.54	0.383349	0.488616	0.502364
11	41.54	0.378670	0.494164	0.509500
12	40.77	0.386759	0.497916	0.513904
13	38.55	0.379308	0.478799	0.495271
14	40.83	0.383473	0.494761	0.510631
15	42.03	0.386928	0.499020	0.513741
16	44.58	0.393713	0.511792	0.525722
17	45.32	0.400801	0.525209	0.537012
18	39.78	0.380849	0.498665	0.513000
19	47.94	0.419000	0.565492	0.581834
20	43.80	0.386242	0.500381	0.511766
21	40.94	0.383511	0.506373	0.527169
22	47.00	0.396888	0.531178	0.544873
X	39.53	0.379025	0.521601	0.548487
Y	40.03	0.378356	0.540774	0.584169
Spearman correlation		0.784518	0.579691	0.476625
Pearson correlation		0.897925	0.706432	0.532642

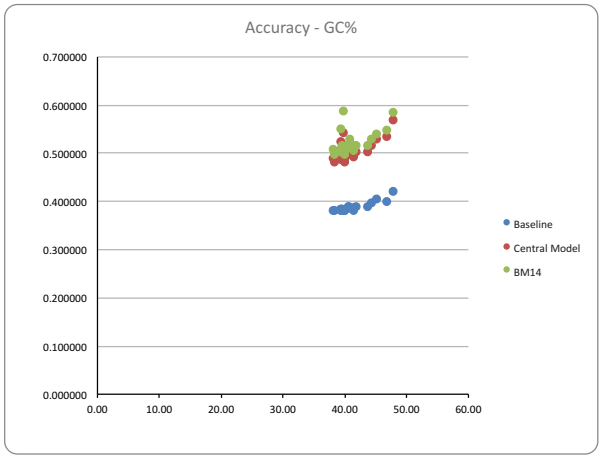

Table S4: The predicted accuracy in different genome regions based on Bidir-Markov model, $k = 14$.

Chromosome	Bidir, interpol. Markov chain, order k=8-14 (alpha = 100) Fraction correct																			
		CDs		Intergenic			Intronic			UTR-3			UTR-5			Repeat			Different Repeat Classes		
	total	correct	Fraction correct	total	correct	Fraction correct	total	correct	Fraction correct	total	correct	Fraction correct	total	correct	Fraction correct	total	correct	Fraction correct	ALU	LINE1	Simple repeat
1	3613817	1306656	**0.361572**	84301159	45258497	**0.536867**	139092980	69243240	**0.497820**	8008924	3333987	**0.416284**	4631839	1929146	**0.416497**	119056918	75897946	**0.637493**	0.8686	0.6371	0.8862
2	2623685	916030	**0.349139**	92297532	47385879	**0.513404**	143292330	69613502	**0.485815**	6270743	2595099	**0.413842**	3682967	1504167	**0.408412**	118241969	73866541	**0.624707**	0.8681	0.6228	0.8731
3	2034864	698330	**0.343183**	70371457	37342432	**0.530647**	123895792	60483421	**0.488180**	5051341	2072638	**0.410314**	3032862	1233938	**0.406856**	101617514	64044795	**0.630254**	0.8681	0.6444	0.8802
4	1421634	509130	**0.358130**	82785978	43051670	**0.520036**	103853008	51512448	**0.496013**	3737146	1569378	**0.419940**	2131915	885324	**0.415272**	98046953	61559347	**0.627856**	0.8660	0.6559	0.8712
5	1628337	572110	**0.351346**	76147646	40451992	**0.531231**	101798788	49926430	**0.490442**	4336308	1779034	**0.410265**	2503457	1035243	**0.413525**	93200114	59217538	**0.635381**	0.8684	0.6491	0.8782
6	1797767	623554	**0.346849**	72091517	37454842	**0.519546**	94269687	45999894	**0.487961**	4205028	1752497	**0.416762**	2548187	1031866	**0.404941**	85025516	53696837	**0.631538**	0.8684	0.6435	0.8630
7	1702420	619646	**0.363980**	61087944	33809137	**0.553450**	94380108	46671822	**0.494509**	4104292	1726447	**0.420644**	2448377	1044583	**0.426643**	81490138	52951978	**0.649796**	0.8672	0.6243	0.8788
8	1209831	434776	**0.359369**	54914112	28994546	**0.527998**	87157868	42420888	**0.486713**	3226788	1339345	**0.415071**	2179053	924137	**0.424100**	74215234	46456104	**0.625965**	0.8649	0.6363	0.8570
9	1442642	512044	**0.354935**	53624114	29173800	**0.544043**	65209939	32297536	**0.495285**	3188567	1336289	**0.419088**	1851956	782286	**0.422411**	62883721	40430890	**0.642947**	0.8702	0.6293	0.8902
10	1382518	481170	**0.348039**	54382720	28679922	**0.527372**	75910976	37087826	**0.488570**	3250963	1348743	**0.414875**	1869909	772470	**0.413106**	66408069	42415549	**0.638711**	0.8666	0.6167	0.8656
11	2109046	737280	**0.349580**	51966102	28570396	**0.549789**	78820526	38515885	**0.488653**	4575885	1866999	**0.408008**	2890929	1170482	**0.404881**	70699091	45234293	**0.639814**	0.8680	0.6487	0.8950
12	1853897	645757	**0.348324**	48616223	26746498	**0.550156**	80883600	40217253	**0.497224**	4947205	2076431	**0.419718**	2723412	1118963	**0.410868**	70435780	45434167	**0.645044**	0.8677	0.6450	0.8708
13	646040	223661	**0.346203**	48685421	24972698	**0.512940**	47612421	22887443	**0.480703**	1890628	778232	**0.411626**	1083361	455775	**0.420705**	48147855	30011123	**0.623312**	0.8675	0.6174	0.8918
14	1170110	410479	**0.350804**	33761575	18213986	**0.539489**	54550219	27136881	**0.497466**	2997736	1241380	**0.414106**	1699243	705406	**0.415130**	46756190	30150597	**0.644847**	0.8684	0.6360	0.8900
15	1232151	443061	**0.359583**	24801621	13915857	**0.561087**	57380388	28533387	**0.497267**	3465872	1499541	**0.432659**	2031152	865463	**0.426095**	43445719	28137423	**0.647645**	0.8689	0.6057	0.9055
16	1521767	563169	**0.370076**	28945476	16338960	**0.564474**	50314246	25617435	**0.509149**	3724555	1592604	**0.427596**	2304213	989672	**0.429505**	42267029	28000470	**0.662466**	0.8646	0.6025	0.8880
17	2055458	736647	**0.358386**	27345217	16387229	**0.599272**	52259106	26817776	**0.513169**	4286474	1789941	**0.417579**	2835063	1177311	**0.415268**	43046304	29725718	**0.690552**	0.8711	0.5840	0.9100
18	563581	196155	**0.348051**	37831200	21059274	**0.556664**	40880833	19469918	**0.476260**	1916887	794178	**0.414306**	1009571	417117	**0.413163**	40760659	26660499	**0.654074**	0.8668	0.6073	0.9356
19	2367720	933529	**0.394273**	19219508	12548353	**0.652897**	35946019	20058817	**0.558026**	4117806	1884150	**0.457562**	2480838	1096186	**0.441861**	34840214	25030442	**0.718435**	0.8670	0.5953	0.8890
20	846489	291718	**0.344621**	28244601	15634161	**0.553527**	33950098	16405392	**0.483221**	1835311	744707	**0.405766**	1124155	466065	**0.414591**	34153300	21891472	**0.640977**	0.8679	0.6008	0.8938
21	355609	127395	**0.358245**	18459731	10661460	**0.577552**	20803643	10132437	**0.487051**	1085540	461393	**0.425035**	754291	320653	**0.425105**	20673206	13687135	**0.662071**	0.8666	0.6084	0.8917
22	756270	269400	**0.356222**	14867058	9188157	**0.618021**	22602536	11438817	**0.506086**	2011024	879824	**0.437500**	1182773	508826	**0.430198**	20907244	14360035	**0.686845**	0.8706	0.5846	0.8904
X	1347799	478847	**0.355281**	86789543	50283538	**0.579373**	65143216	33499885	**0.514250**	2737928	1140080	**0.416402**	1563584	643129	**0.411317**	96223499	63318321	**0.658034**	0.8721	0.6697	0.8942
Y	95364	44874	**0.470555**	17955819	10728499	**0.597494**	7947104	4461566	**0.561408**	382290	181100	**0.473724**	325239	156545	**0.481323**	16524846	11048610	**0.668606**	0.8345	0.6984	0.8070
Total average	35778816	12775418	**0.357067**	1189493274	646851783	**0.543804**	1677955431	830449899	**0.494918**	85355241	35784017	**0.419236**	50888346	21234753	**0.417281**	1529067082	983227830	**0.643025**	0.8666	0.6276	0.8832

Table S5: The predicted average probabilities in different genome regions based on Bidir-Markov model, $k = 14$.

Bidir, interpol. Markov chain, order k=8-14 (alpha = 100)						
	Average Probability (Ref)					
Chromosome	CDs	Intergenic	Intronic	UTR-3	UTR-5	Repeat
1	0.284398	0.403085	0.375882	0.316466	0.312159	0.475164
2	0.279192	0.383054	0.366286	0.314754	0.307805	0.461253
3	0.278367	0.396859	0.368700	0.313361	0.308536	0.466471
4	0.283072	0.385895	0.371263	0.314933	0.308952	0.461346
5	0.279506	0.396156	0.370164	0.311306	0.308938	0.470572
6	0.278677	0.387985	0.368093	0.315637	0.303838	0.466445
7	0.282557	0.415837	0.372873	0.316082	0.313052	0.485514
8	0.281892	0.393981	0.366793	0.313037	0.313177	0.463238
9	0.279441	0.404889	0.372060	0.312249	0.307835	0.477780
10	0.278189	0.396600	0.369809	0.313084	0.307314	0.475369
11	0.278581	0.416209	0.369960	0.311594	0.307849	0.478778
12	0.279281	0.414982	0.376775	0.321811	0.313894	0.482580
13	0.278639	0.382248	0.361672	0.310374	0.310420	0.459961
14	0.279622	0.406438	0.375560	0.317286	0.313113	0.480891
15	0.281748	0.424731	0.375462	0.324673	0.318428	0.486987
16	0.285458	0.427076	0.386496	0.323519	0.318401	0.501964
17	0.280311	0.468020	0.392803	0.319234	0.313653	0.534678
18	0.279760	0.433472	0.359645	0.317651	0.310302	0.501284
19	0.293708	0.505405	0.425240	0.341372	0.326675	0.554286
20	0.276961	0.421177	0.367999	0.310108	0.309141	0.484946
21	0.279895	0.428169	0.361199	0.315645	0.311104	0.490754
22	0.279257	0.465401	0.385303	0.328594	0.319071	0.525004
X	0.280497	0.427725	0.383470	0.312335	0.306479	0.482864
Y	0.307826	0.418311	0.395924	0.318556	0.317295	0.472119
Total average	0.281951	0.416821	0.375810	0.317236	0.311976	0.485010

Table S6: Kolmogorov–Smirnov test of "Probability of Ref - Probability of Alt" distributions

Two sample Kolmogorov-Smirnov test		
Data:	1KGP	ClinVar
D = 0.00702:	p-value = 0.177	
alternative hypothesis: two-sided		

Two sample Kolmogorov-Smirnov test		
Data:	1KGP	COSMIC
D = 0.10173	p-value < 2.2e-16	
alternative hypothesis: two-sided		

Two sample Kolmogorov-Smirnov test		
Data:	ClinVar	COSMIC
D = 0.10225	p-value < 2.2e-16	
alternative hypothesis: two-sided		

Two sample Kolmogorov-Smirnov test		
Data:	Benign SNPs	Damaging SNPs
D = 0.06165	p-value < 2.2e-16	
alternative hypothesis: two-sided		

Table S7: Chi-Squared of Damaging and Benign SNPs

Counts	Ref-Highest	Alt-Highest	Rest SNPs
Damaging	3233	4449	7617
Benign	7864	8704	16273

Pearson's Chi-squared test

Expected	Ref-Highest	Alt-Highest	Rest SNPs
Damaging	3526.7	4180.1	7592.3
Benign	7570.3	8972.9	16297.7

X-squared = 61.325, df = 2, p-value=4.824e-14

Comparing only Alt highest to the rest

Counts	Alt-Highest	Rest SNPs
Damaging	4449	10850
Benign	8704	24137

Pearson's Chi-squared test

Expected	Alt-Highest	Rest SNPs
Damaging	3526.7	11772.3
Benign	7570.3	25270.7

X-squared = 34.772, df = 1, p-value=3.707e-09

Table S8: The α matrix for $k = 1$ estimated from all chromosomes except Chr1.

α matrix, $k = 1$				
	A	C	G	T
AAA		0.023	0.035	0.010
ACA	0.021		0.028	0.087
AGA	0.039	0.043		0.027
ATA	0.013	0.126	0.015	
AAC		0.019	0.060	0.012
ACC	0.049		0.028	0.087
AGC	0.069	0.027		0.019
ATC	0.025	0.062	0.013	
AAG		0.036	0.040	0.008
ACG	0.041		0.041	0.717
AGG	0.061	0.061		0.015
ATG	0.017	0.223	0.020	
AAT		0.016	0.095	0.012
ACT	0.018		0.037	0.064
AGT	0.064	0.037		0.018
ATT	0.012	0.094	0.016	
CAA		0.020	0.072	0.008
CCA	0.017		0.035	0.058
CGA	0.487	0.048		0.043
CTA	0.008	0.050	0.023	
CAC		0.020	0.084	0.011
CCC	0.025		0.059	0.067
CGC	0.570	0.035		0.063
CTC	0.010	0.032	0.026	
CAG		0.027	0.080	0.009
CCG	0.035		0.066	0.555
CGG	0.549	0.066		0.035
CTG	0.009	0.080	0.027	
CAT		0.020	0.224	0.017
CCT	0.015		0.061	0.061
CGT	0.712	0.041		0.041
CTT	0.008	0.039	0.036	

α matrix, $k = 1$				
	A	C	G	T
GAA		0.014	0.033	0.009
GCA	0.027		0.019	0.068
GGA	0.059	0.028		0.028
GTA	0.009	0.066	0.013	
GAC		0.012	0.045	0.017
GCC	0.037		0.026	0.084
GGC	0.085	0.026		0.038
GTC	0.017	0.045	0.012	
GAG		0.026	0.032	0.010
GCG	0.062		0.035	0.566
GGG	0.067	0.058		0.025
GTG	0.011	0.083	0.020	
GAT		0.013	0.062	0.025
GCT	0.019		0.027	0.069
GGT	0.087	0.028		0.049
GTT	0.012	0.059	0.018	
TAA		0.016	0.046	0.015
TCA	0.018		0.024	0.049
TGA	0.049	0.024		0.018
TTA	0.015	0.046	0.016	
TAC		0.013	0.067	0.009
TCC	0.028		0.029	0.060
TGC	0.068	0.019		0.027
TTC	0.010	0.033	0.014	
TAG		0.023	0.050	0.008
TCG	0.043		0.048	0.483
TGG	0.058	0.035		0.017
TTG	0.008	0.071	0.020	
TAT		0.015	0.128	0.013
TCT	0.027		0.043	0.039
TGT	0.087	0.028		0.021
TTT	0.010	0.035	0.023	

Supplementary Figures for

Context dependency of nucleotide probabilities and variants in human DNA

Yuhu Liang[1,†], Christian Grønbæk[2,†], Piero Fariselli[3] and Anders Krogh[1,4,*,†]

[1]Department of Computer Science, University of Copenhagen, Denmark
[2]Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Denmark
[3]Department of Medical Sciences, University of Torino, Italy
[4]Center for Health Data Science, University of Copenhagen, Denmark
[†]Part of this work was carried out at Department of Biology, University of Copenhagen, Denmark

Corresponding author: Anders Krogh, email: akrogh@di.ku.dk

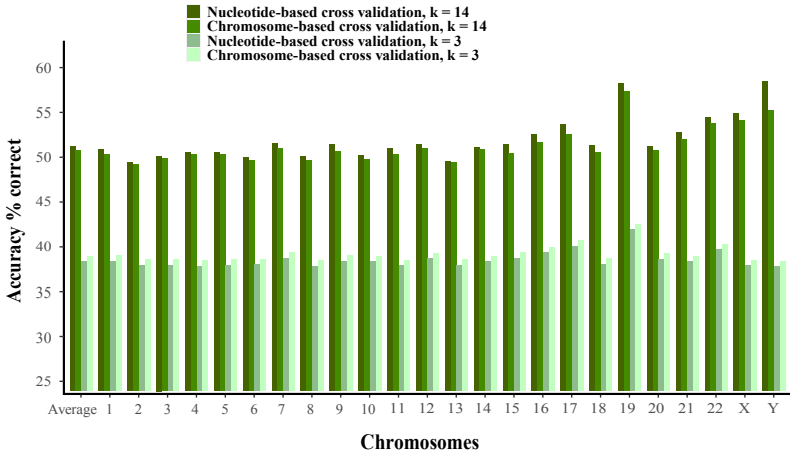

Figure S1: Chromosome based cross validations for baseline model and Bidir-Markov model. For each chromosome, the overall prediction accuracy is calculated for a model is estimated from the other chromosomes (chromosome-based cross validation). The overall average is weighted by chromosome sizes. These are compared to the nucleotide-based cross validation accuracies used in Figure 1.

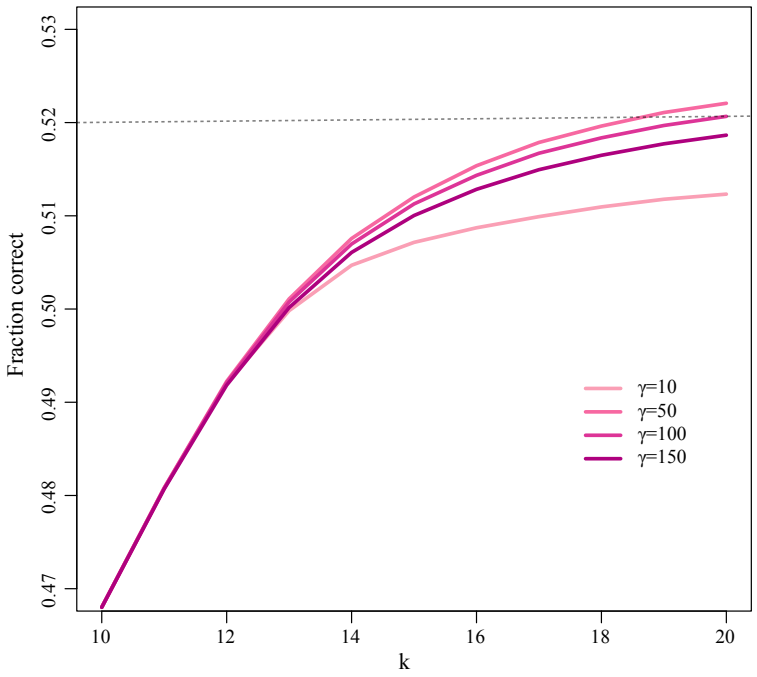

Figure S2: Accuracy of the bi-directional k-th order Markov model for different strengths of regularization, γ. Results are shown only for Chromosome 20 with the model estimated from all the other chromosomes.

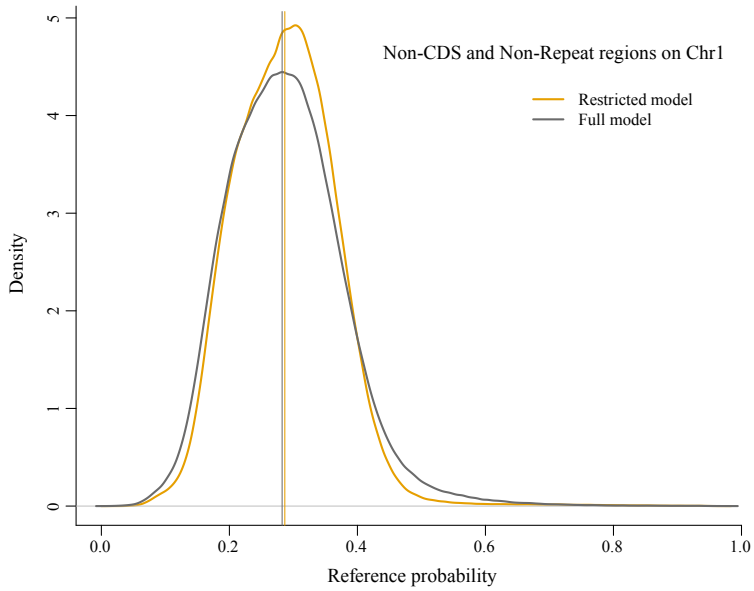

Figure S3: Comparison of restricted and full model based on density profile of reference probabilities. Density profile of the reference probabilities for the full model was shown as a dark grey line and the other for a model estimated on non-repeat and non-coding regions on Chromosome 1. The yellow and gray vertical lines represent the median probabilities of restricted model and full model, which are 0.286578 and 0.282368, respectively.

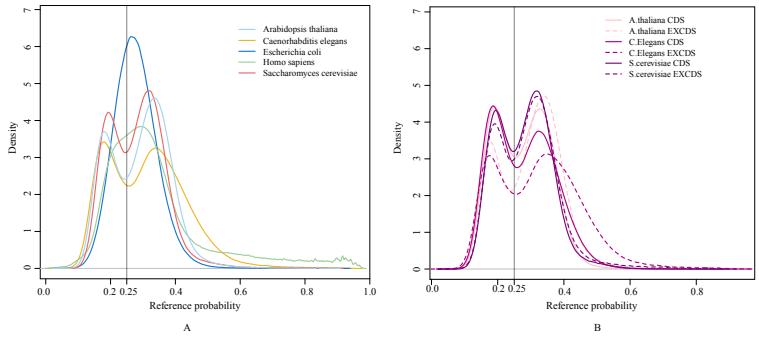

Figure S4: Density profile of reference probabilities of different species. A. Those species were estimated via 10-k context bidirectional Markov model, γ = 100 interpolated from 6. B. Density plots of CDS regions and non-CDS for the species, which have two peaks in Figure S4A.

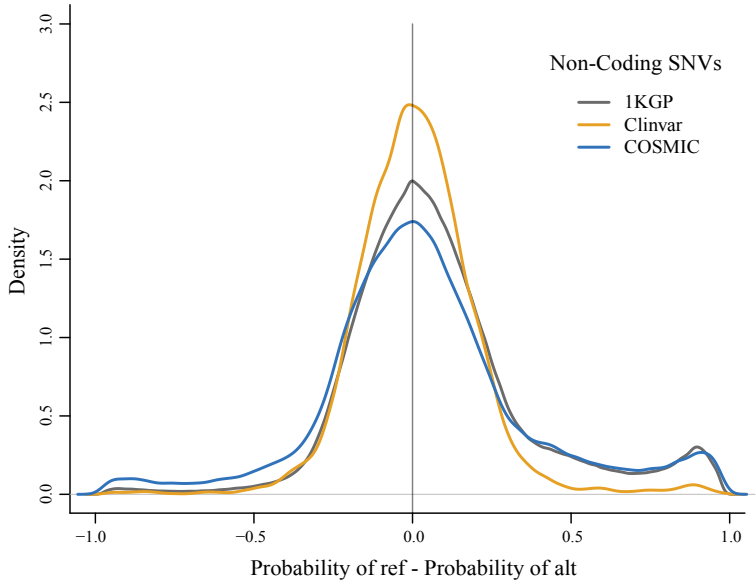

Figure S5: Density profiles of Pref - Palt for SNPs on Chromosome 1. Density profiles show ClinVar, somatic mutations (COSMIC) and 1KGP SNPs in Non-Coding regions, respectively.

7.2 A generative model of normal tissue gene expression enables differential expression in cancer with one sample

A generative model of normal tissue gene expression enables differential expression in cancer with *one* sample

Yuhu Liang[1*], Inigo Prada Luengo[1*], Viktoria Schuster[2*], Thilde Terkelsen[2], Anders Krogh[1,2]

1) Department of Computer Science, University of Copenhagen, Denmark
2) Center for Health Data Science, University of Copenhagen, Denmark

*) equal contributions

Abstract

Differential gene expression analysis in bulk RNA sequencing data between disease and control is challenging due to lack of good controls and the heterogeneous nature of the samples. Here we present a deep generative model that frees us of the need for controls. The model is trained on RNA-Seq data from normal tissue and tested on cancer samples. For most cancer samples, the model infers representations in clusters of normal tissues identical to the cancer origin. The overall probability of a cancer sample, which is given by the model, is lower than that of an adjacent normal sample. This indicates that the model can distinguish cancer from normal samples. From the model we can derive a p-value for each gene in a sample. In a detailed analysis of breast cancer, we show that when comparing *a single cancer sample* to the model *without a paired control,* the significant genes are enriched in known cancer driver genes and marker genes for cancer subtypes. This enrichment is much greater than in standard DESeq2 analysis *with* paired control samples. In a control experiment comparing normal vs. normal the model barely finds any false positives, whereas the standard comparison using DESeq2 results in hundreds of false positives.

Introduction

Cellular function varies with cell type and environment. Differences in cell function can largely be characterized by gene expression profiles, and analysis of gene expression data has become a standard for studying differences in cells and tissues, in part driven by advances in next-generation RNA sequencing (RNA-Seq) technologies. In many diseases, such as cancer, gene expression patterns deviate strongly from those observed in normal states. By identifying differentially expressed genes (DEGs), we can pinpoint genes involved in the onset or progression of disease, which could present biomarkers or potential drug targets in personalized treatment (Burska et al. 2014; Kamel and Al-Amodi 2017).
Despite the great potential of differential expression analysis, the methods employed for this type of analysis are often found to have low reproducibility and return thousands of

significant DEGs (Cui et al. 2021), making a clinical interpretation challenging. This is to a large extent due to the lack of good controls, which is a common problem in the study of diseases. In cancer studies, controls are most often tissue samples from healthy individuals (potentially matched on available clinical parameters) or, alternatively, normal adjacent tissues (NATs) from the cancer patients themselves. The benefit of the latter is a reduction of person-specific biological variance. However, NATs from cancer patients have been shown to display field cancerization (Aran et al. 2017), meaning that these samples are not truly normal. Inversely, normal samples from other individuals will display genetic heterogeneity and are thus not suitable for direct comparison with classical methods, especially at low sample numbers (Li et al. 2022; Vihinen 2022). Lastly, a general problem in relation to bulk sequencing data is that samples differ in cell type composition. This problem may in part be alleviated by taking this into account using weighted averages of the closest normal samples (Rapin et al. 2014; Vivian et al. 2020).

The most generally applied method for differential expression analysis relies on count statistics using negative binomial (NB) distributions to account for over-dispersion (Love, Huber, and Anders 2014). Neural networks and Machine learning in general have been increasingly applied to the field of transcriptomics in the last two decades. Applications range from quality control using simple regression and mixture models (McDermaid et al. 2018) over identifying DEGs and biomarkers using random forests (Abbas and El-Manzalawy 2020) or convolutional neural networks (Kakati et al. 2022) to digital pathology (Schmauch et al. 2020) using multi-layer perceptrons. Other approaches have been suggested to learn biologically meaningful representations from gene expression data (Altman et al. 2019) present a variational autoencoder (Kingma and Welling 2013) trained on cancer transcriptomes with the potential to predict therapeutic responses. Another generative neural network, SOPHIE (Lee et al. 2022) identifies cancer-specific genes from a collection of normal and cancer datasets. So far, available methods seem to either be limited to requiring paired or manually curated controls or having to be trained on very specific datasets including cancer samples.

In this work, we present a model of gene expression in normal tissue which addresses the problem of finding good controls and enables differential expression analysis in cancer using only a single sample. The model is an extension of the Deep Generative Decoder (DGD) (Schuster and Krogh 2021, 2022) a generative neural network which learns a probabilistic low-dimensional representation of the data.

Our model is trained on the Genotype-Tissue Expression (GTEx) data (Lonsdale et al. 2013) with around 20,000 bulk samples from 31 different human tissues and 948 individuals. Briefly, the model learns parameters with two goals. Firstly, the neural network parameters in the decoder are learned to best describe all data in a low-dimensional space, the representation. Secondly, the model learns a most probable representation for each sample and returns a NB distribution over count values for each gene. For samples that are not from normal tissue, such as cancer samples, we can infer a nearest representation in the model to use as the control.This inferred control is informed by the whole training data instead of a limited and biased set of control samples. We therefore expect it to yield a less noisy control and more precise judgment of differential expression. In order to test this hypothesis, we apply the model to cancer samples from the Cancer Genome Atlas (TCGA) program

(Cancer Genome Atlas Research Network 2008). From the NB distribution over gene counts, we can derive a p-value for each gene in a sample and can thus identify a set of significant DEGs. We focus on the analysis of breast cancer (BC) and calculate enrichment of known cancer driver genes and subtype marker genes among the significant genes. This is compared to a standard case-control analysis using DESeq2.
In conclusion, we find that our model of normal gene expression drastically improves differential expression analysis by yielding fewer false positives and extending the analysis to single cancer samples (N-of-one) *without controls*. We believe that this method can have a significant impact on the utility of gene expression analysis, target identification and therefore personalized treatment.

Results

The goal of our method is to construct a deep generative model that learns how genes are expressed across human tissues. Our model, a deep generative decoder (DGD) (Schuster and Krogh 2022) learns a low dimensional *representation* for every sample. The representation (or latent) space has a dimension of 50 and representations are distributed according to a mixture of Gaussians with 45 components, closely matching the number of tissues covered in our data (Fig1A). A decoder neural network with several hidden layers maps the latent space to sample space, resulting in a negative binomial distribution for each gene (Fig. 1B). We infer the parameters of the representations, Gaussian mixture and decoder by training our model on a random subset containing 90% of GTEx data (17072 samples), while leaving the remaining 10% (1903 samples) as a test set (Supplementary Table S1).

A generative model of gene expression for bulk samples

After training, we first decided to evaluate whether the latent space of the DGD model distinguishes different tissues. We first performed principal component analysis (PCA) of the latent space (Fig. 2A), finding that the DGD is able to find well-separated representations. Our learned Gaussian mixture model (GMM) over the latent space adds structure to representations and ideally, each mixture component should gather samples that originate from the same tissue. To test this, we assign each sample to the GMM component with highest probability and evaluate how samples are distributed across components (Fig. 2B). The matrix shows that almost all GMM components are assigned to samples dominated by only one tissue. Correspondingly, most tissues are represented by a few GMM components – for most tissues only one. Interestingly, we observe that some tissues with known biological substructure are divided across several GMM components. For example, the DGD splits the subtypes of the brain in 8 components; the esophagus and adipose in 3; and colon into 2. In summary, these results show that the DGD model learns how genes are expressed across human tissues, while being able to find separable low dimensional tissue representations.

Finding closest-normal comparison sets for cancer samples

Next, we evaluated whether our model of normal gene expression could find meaningful representations for cancer samples. To do so, we used our model (trained on GTEx) to find representations for tumor samples (Supplementary Table S2) in The Cancer Genome Atlas dataset (TCGA) by maximizing the probability of a representation for a cancer sample, while leaving the decoder neural network and GMM parameters fixed (Fig. 3A). We interpret the representation as the closest-normal sample to the tumor. To start with, we evaluated the ability of our model to detect out-of-distribution examples (i.e. anomalous expression profiles) by calculating the probability of each sample, given our model (Figure 3C). We observe that TCGA-cancer samples generally have a much lower probability than GTEx samples while TCGA-normal samples have intermediate probabilities. Afterwards, we assessed if our model matches tumor samples to their healthy counterparts (Fig3B). We find that our model closely matches most tumors to their healthy normal and 11 out of 14 tissues have a classification percentage higher than 80%. The three tissues with low classification accuracies are bladder, stomach and esophagus.

Detecting cancer differentially expressed genes without controls

We extended the DGD to detect differentially expressed genes. Our model performs a two-tailed negative binomial test for the distribution generated from the closest-normal representation (material and methods).To test the DGD, we focused on breast cancer, as both TCGA and GTEx contain a large number of samples (Fig. 4A). However, we perform our experiments in an N-of-1 fashion to resemble common clinical settings. As a comparison, we benchmark the DGD against DEseq2 (Love, Huber, and Anders 2014), a widely used statistical method to detect differentially expressed genes.

We first assess the specificity in a normal vs. normal analysis, using healthy breast tissue from GTEx. It is assumed that there should be no or very few DEGs when comparing normal samples and therefore the number of DEGs functions as a proxy to specificity. We randomly selected 1 sample from the breast test set (42 samples) and compared against the 440 control samples from the training set using DEseq2, and we repeated this process 20 times. We used the same random samples to compare the DGD model. The number of significant genes of these are shown in Fig. 4B for varying p-adjusted values. We did the same analysis using five randomly selected samples from the training set as the comparison set (full line and box plot in Fig. 4B). Ideally, this analysis should give no significant genes, but DEseq2 found many genes differentially expressed, calling the 179.05 and 76.15 genes for the 1 versus 5 and 1 versus all comparisons (p-adjusted < 0.01 & log fold-change >1). The DGD, on the contrary, found almost no false positives with e.g. an average of 4.25 for 1 vs model.

To compare the sensitivity of the DGD and DEseq2 we analyzed their ability to correctly identify marker genes known to be differentially expressed in breast cancer. Two sets of BC-related genes were curated for the purpose (I) driver genes from the DriverDBv3 database (Liu et al. 2020), and (II) the PAM50 (Parker et al. 2009) set of subtype-specific BC genes (Materials and Methods). As a metric, we calculated a gene enrichment score, which

is the fraction of genes among the significant ones divided by the expected by random chance.

We evaluated enrichment scores across PAM50 breast cancer subtypes, namely, basal-like, HER2, luminal A and luminal B. For the purpose of this analysis we applied clinical filters to ensure greater homogeneity of samples (Supplementary Table S3). We performed experiments similar to those described above, randomly selecting one sample (20 repetitions) and comparing it to the rest of the samples and the model for DEseq2 and DGD, respectively. The DGD obtained higher enrichment scores than DEseq2 for all subtypes in regards to both driver genes (DriverDBv3) and PAM50 genes (Fig. 4C). DGD obtained an average enrichment score of 3.46, and attained a particularly high score for the luminal A subtype in the PAM50 marker set. In comparison, the DEseq2 average score was 1.71, without being high for any particular subtype (note that a score of 1 means no enrichment). In summary, the results highlight how the DGD maintains a very high specificity without sacrificing sensitivity.

Next, we selected a subset of PAM50 genes to evaluate whether the DGD captures the expression differences between subtypes. Briefly, we used three criteria for inclusion of genes in our downstream analysis: the gene was well-studied in breast cancer (Wirapati et al. 2008), had a significant p-value in at least 10 replicates and had different expression patterns across subtypes (Supplementary Table S4). Altogether, our filtering led to 28 across the four subtypes. For each gene we evaluated whether the DGD detected differential expression (i.e. a significant p-value) as well as the expression trend (i.e. upregulation or downregulation). The DGD correctly determined 21 out of the 28 gene expression patterns (Fig 4. D-G), obtaining similar performances across the subtypes. The best case was the basal subtype (6 out 7 correctly determined), while the performance was identical for the rest of the subtypes (5 out 7). Summarizing across genes, *Errbb2*, *Esr1* and *Pgr* were correctly called in all subtypes and *Mlph* and *Mmp11* were correct in 3 out of 4. DGD only had bad performances for *Egfr* and *Tmem45*, which were wrongly called in 2 out 4 and 3 out 4 subtypes, respectively. In short, these results show that the DGD is able to detect gene expression patterns which are specific to breast cancer subtypes.

As a final analysis, we extended our model to the rest of TCGA cancers in order to evaluate if the DGD could be applied to other cancer types. As above, we performed 1 versus model experiments using 20 repetitions and we evaluated the enrichment score of the DriverDB gene set for each cancer type. The DGD found more cancer marker genes than expected by chance for all cancer types (mean enrichment score 2.33, mean range 1.27 - 4.00). Specifically, the scores were higher for kidney renal clear & papillary cell carcinomas (mean enrichment of 3.65 and 3.18, respectively) and thyroid carcinoma (mean enrichment 4.00). Collectively, we here show that DGD is able to find marker genes across various cancer subtypes.

Discussion

A lot of attention is currently directed towards single-cell RNA sequencing due its potential higher resolution and recent advances in its scalability. However, bulk sequencing is still the work-horse for clinical use. Differential expression analysis between disease and normal has

relied on solid statistical methods, but are challenged by the difficulty in obtaining suitable control samples and large enough sample sizes. Here, we introduce a method that requires no biological replicates and matched controls.. The model we present learns the gene expression of normal tissue samples and generates a normal sample closest to the disease sample at hand.

In order to assess the model's capability to find meaningful representations and to generalize unseen data, we evaluated the model on data held out during training. The model clusters representations well in a tissue-dependent manner and typically assigns only one to a few Gaussian mixture components to each tissue type. Interestingly, we see that some of the complex tissues are spread over several components. Most notably, brain tissue is spread over eight components, whereas some other tissues are spread over two to three components. We also see some tissue types mixed together. For some, this is not surprising as several tissues are underrepresented in the data. Examples are the component that comprises samples from uterus, ovary, fallopian tube and cervix, and the two components in which small intestine and colon are mixed. These mixed clusterings additionally make sense as their tissues stem from larger systems (female reproductive system and digestive tract), which also explains some other mixed components such as one modeling colon and stomach. Given that the representations are found in a completely unsupervised fashion, we find the clustering to be remarkably interpretable.

Besides providing well clustered representations, the model of normal gene expression can replace control samples in the differential gene analysis of a disease. This is achieved by finding the closest normal representation in the model. The generated sample of this representation is in turn used to compute a probability distribution for the expression counts of each gene. As a sanity check, we calculated total probabilities of GTEx test, TCGA normal and TCGA cancer samples and find that the probabilities derived from the model of normal are highest for GTEx test data and decrease strongly with TCGA cancer as expected. Probabilities for TCGA normal samples lie between the two, which is consistent with previous findings that suggest that adjacent normal tissue carry traits of cancer (Aran et al. 2017) The separation we find seems to be much clearer than previously reported (Vivian et al. 2020), although a direct comparison is difficult. When analyzing the quality of the integration of cancer samples into latent space, we observe that most representations of cancer samples are consistent with the tissue of origin. For most cancers, more than 80% of samples end up in the expected tissue. These results are again consistent with (Vivian et al. 2020).

It is difficult to assess the performance of differential gene expression analysis without knowing the ground truth. We have compared our approach to a standard analysis in two ways using breast cancer as a case. Firstly, we assessed the specificity of our model and DESeq2 by comparing normal vs normal as a negative control. While DESeq2 yields large numbers of significant genes (1.75% of all genes, given log2FoldChange >1; P-adj < 0.05), here interpreted as false positives, our model reports only 0.03% of the genes to be differentially expressed under the same threshold. Secondly, we calculate the enrichment of relevant known cancer genes among the significant genes derived from differential expression analysis of breast cancer samples. In this positive control, we used a set of known breast cancer driver genes and the PAM50 set of genes breast cancer subtypes. In the general breast cancer case, we see consistently higher enrichment when using the DGD

compared to DESeq2. This is also true, in most part, for the cancer subtypes. However, there are some outliers with respect to the expected driver genes. For instance TMEM45B should be upregulated in the HER2 subtype, but is downregulated on average according to our model. Same for MLPH gene, which should be downregulated in Luminal B, but our model doesn't detect this gene as significant in Luminal B subtype in any of the 20 experiments (Tishchenko et al. 2016). Yet, we do not expect a perfect concordance between the PAM50 panel and tumor samples as there are many patient-specific factors that can affect the expression of a gene. Another possible reason is that Luminal B is in some ways more like Luminal A, and in some ways more like HER-2 (Yersal and Barutca 2014). Altogether, our evaluation on breast cancer shows that, in this case, the DGD returns much fewer false positives and a higher proportion of truly relevant genes compared to DESeq2.

The model introduced and discussed here presents an important step towards the use of bulk gene expression analysis for precision medicine. Given the fact that the DGD does not require paired control samples, the potential impact of the model for differential expression analysis with *single disease samples* is immense. Because of the great performance on even single samples, we see an especially high potential in application to rare diseases. . Additionally, the results of differential expression analysis will contain much fewer false positives, which will enable us to find genes that are truly involved in disease and thus increase the possibility to find druggable targets and understand the disease on an individual level as it has not been possible before with bulk data.

Methods

The model

Architecture and hyperparameters

The full model consists of the learned representation, a GMM as the parametrized distribution over latent space and a decoder as presented in (Schuster and Krogh 2022). Each representation of a sample receives a 50-dimensional vector initialized with zero. The architecture of the decoder consists of an 50-dimensional input layer which is fed with the representations, two hidden layers and an output layer with its units corresponding to the number of genes in the data. The two hidden layers are of size 500 and 8000, respectively and are immediately followed by ReLU activation (Fukushima 1975; "Rectified Linear Units Improve Restricted Boltzmann Machines" n.d.). The output layer's values are transformed into expression counts by the Negative Binomial layer (NB layer). The gene-specific dispersion parameters are initialized with 2. Unlike in the scDGD from (Schuster and Krogh 2022), the decoder outputs are passed through ReLU activation and scaled with the sample gene expression mean in order to achieve a predicted gene expression value. The GMM consists of 45 mixture components. The priors are a mollified Uniform with spread 7 and sharpness 10 for the means, a Gaussian with mean 1 (corresponds to a standard deviation of 0.1) and standard deviation 1 for the negative logarithmic diagonal covariance, and a Dirichlet with alpha 5.

Training

The model is trained for 200 epochs with a batch size of 256. The optimizer of choice is Adam (Kingma and Ba 2014) without weight decay and betas 0.5 and 0.9. Because the

representations are updated every epoch, decoder, representation and GMM have their own optimizer instances with learning rates 1e-4, 1e-2 and 1e-2, respectively.

Evaluation
Representations for single test samples are learned as described in (Schuster and Krogh 2022). For each new datapoint, new representations are initialized from the component means. This results in 45 representations per sample. These are trained on the frozen model for 10 epochs, after which the best representations per sample are selected and trained for another 50 epochs.

In order to learn a single representation for multiple samples, we assume that samples *x* are conditionally independent:

$$P(z|x_1,\ x_2 \ldots\ x_i) = P(z|x_1)P(z|x_2)\ldots P(z|x_i)$$

We can therefore simply obtain a single representation by using the summed negative log-probability masses (the losses) of all samples of interest.

Differential expression
We extended the DGD to find differentially expressed genes for tumor samples. Learning new representations for a set of tumor samples is performed as described above. The learned representations are the closest-normal for each tumor. In our setup, we want to test if counts for a given gene are significantly different between the tumor and the normal tissue output of the neural network. Let $NB(m_i, r_i)$ be the re-scaled negative binomial distribution for gene i, and x_i be the actual count in the tumor sample. We define the null hypothesis H_0: $m_i = x_i$. In other words, the mean of the negative binomial distribution and the tumor expression count are equal. The p-value for the probability of x_i originating from the negative binomial distribution can be calculated by summing all K counts with an occurrence probability lower than that for x_i:

$$\text{p-value} = \sum_{k=0}^{K} P(K = k|NB(m_i, r_i))\ .\ I\{P(K = k|\ NB(m_i, r_i))\} \leq P(K = x_i|NB(m_i, r_i))$$

The above expression yields an exact p-value for the negative binomial distributions. However, it requires summing over all read counts across genes. For the sake of efficiency, we therefore obtain an asymptotic p-value by summing over an evenly spaced grid of 10^4 in the domain of K.

Data

Data collection and processing

The raw gene count expression data from the Genotype-Tissue Expression (GTEx) and the Cancer Genome Atlas (TCGA) were downloaded from the Recount3 database (https://rna.recount.bio/),using the built-in R packages. Additionally the sample metadata files were acquired through the Recount3 platform (Wilks et al. 2021).
At the time of download (09th-Feb-2022) there were 31 different tissue types in GTEx and one NA Study category, with a total of 19,214 individual samples (Supplementary Table S1). 133 samples from the NA Study class were removed from the dataset as they had no tissue information, and the drop duplication function was used when we trained our model. We employed the filterByExpr (Filter Genes By Expression Level) (Chen, Lun, and Smyth 2016), an R function, to get rid of the low expressed genes by using the default parameters. For our analysis we only retained protein coding genes based on the annotation file namely 'GTEx gene', that was downloaded from UCSC Table Browser. After filtering (Chen, Lun, and Smyth 2016)) the GTEx dataset contained a total of 18,975 samples and 16,883 annotated protein coding genes.

We matched the genes from the filtered GTEx set with those in the TCGA , and separated the TCGA samples into a Normal Adjacent set and Tumor set, in accordance with the metadata file.

TCGA tissue selection

To evaluate whether our model could learn and generate new data points that can correctly match to the corresponding tissue of the GTEx dataset. We selected 12 different tissues under three conditions: 1) The TCGA tissues must correspond to GTEx dataset. 2) Have at least 10 adjacent normal samples from each cancer-type (Zeng et al. 2019). 3) Include the Adrenal and Brain tumor samples although they don't have adjacent normal samples, because we would like to compare our result with John Vivian's work (Vivian et al. 2020). There are 6111 TCGA tumor samples and 624 TCGA adjacent normal samples. (Supplementary Table S2)

TCGA breast cancer subset

To obtain a homogenous breast cancer (BC) dataset for testing we curated the TCGA BC samples to only include primary tumors from women between 40-70 years of age. We excluded samples with low tumor cell percentage (defined as < 50%) or a high level of necrosis (defined as > 5 %), in addition to samples from patients with known metastasis, stage iv or stage x tumors, or prior cancer diagnosis. For a full list of selection criteria and columns from metadata used for curation see Supplementary Table S4. The number of available BC samples for analysis was reduced from 1256 to 395.

Cancer Driver Genes and PAM50 genes

We downloaded a list of cancer driver genes for each cancer type from DriverDBv3 (Liu et al. 2020).
The PAM50 gene set used for BC subtype classification (Basal, Luminal A, Luminal B and Her2-enriched) (Parker et al. 2009) was downloaded though the R-package *genefu (Gendoo et al. 2016)*. As our model testing and comparison with DEseq2 pertained to the expression profile of BC subtype tissue vs normal tissue (i.e. not between subtypes), we filtered the PAM50 dataset to only include the genes which were specific to a single subtype, and/or which could distinguish

BC subtype(s) from normal tissue. We noted the expected directionality of each of the PAM50 genes (up or down regulated) for a contrast. The selection of genes were based in part on literature (Coleman and Anders 2017; Vaca-Paniagua et al. 2015; Hu et al. 2013) and in part on the robust normalized PAM50 scores (Gendoo et al. 2016). The PAM50 geneset was reduced to 34 genes.

Analysis

Evaluation of tissue specificity

Clustering performance of the model according to tissue type was evaluated based on the GMM probability densities for each sample's representation. For this purpose, all GTEx training samples are assigned the GMM component that achieves the highest probability density for their inferred representations. We calculate the percentage of each tissue per component as the number of samples of a given tissue clustered in a given component divided by the total number of samples assigned to this component.

$$percentage = \frac{\#\ of\ tissue{-}specific\ samples\ clustered\ in\ component}{\#\ of\ total\ samples\ in\ component}\ x\ 100$$

Matching TCGA to normal tissues

We use the TCGA data described in the Data section in order to evaluate the mapping of unseen, out-of-distribution data onto the latent space. New representations for all 6111 TCGA tumor samples are learned with the DGD trained on GTEx data. The resulting GMM probability densities for the TCGA representations are used to evaluate how well new samples are matched to the correct tissues of the training representation. We define the "correct" tissue as the tissue that best represents a given GMM component. Our evaluation metric is the percentage of TCGA samples of a given tissue matched to the corresponding GMM component with respect to the total number of TCGA samples for this tissue.

$$\%\ TCGA\ in\ correct\ clusters = \frac{\#\ of\ TCGA{-}tumor\ samples\ in\ "correct"\ component}{\#\ of\ TCGA{-}tumor\ samples\ of\ \ tissue\ \ in\ total}\ x\ 100$$

Bladder samples are evaluated differently due to the lack of a bladder-specific component in the normal model. Instead, we evaluate a correct match as TCGA and GTEx bladder samples assigned to the same component(s).

Comparing GTEx and TCGA gene expression predictions

The predicted gene expression of the model is given as the mean of a NB distribution, which is the product of the NN output and the mean expression of the sample. We calculate the negative log-probability mass (the reconstruction loss) of each sample across all 16,883 genes. We do this for three datasets: GTEx test, TCGA-Adjacent normal and TCGA-Tumor. For this comparison, we use subsets containing 10 tissues, namely Adrenal, Brain, Breast, Colon, Kidney, Liver, Lung, Prostate, Stomach and Thyroid. For the analysis of each tissue, we randomly select 100 samples from each set if the dataset has more than 100 samples, otherwise we take all samples from the set. We apply this analysis for a pan-tissue comparison based on 8 tissues because Adrenal and Brain are missing in the TCGA-Adjacent subset. For a fair comparison, we ensure equal numbers of samples for a given tissue across the three datasets. If all datasets have more than 20 samples for a given tissue, we randomly select 20 samples from each subset for that tissue. Otherwise, we choose the lowest number of samples available for a tissue and subsample the other datasets down to that number. As an example, there are only 7 kidney samples in the GTEx test set. We thus select 7 samples from each TCGA dataset. The sample numbers of TCGA tissues are shown in supplementary Table S1,S2.

Differential expression analysis in TCGA Breast Cancer

The cancer samples provide a unique opportunity to evaluate the capability of our model to perform Differential expression analysis (DEA) due to known cancer driver genes. DEA performed by our model is compared to DESeq2 and the resulting sets of DEGs are analyzed with respect to their enrichment in cancer driver genes.

For a general comparison, we perform 30 multi-sample experiments using 5 random breast cancer samples (cases) from the population of 40-50 year-old caucasian females. This leaves us with 166 samples. Genes that result in absolut log2-fold changes greater than 1 and adjusted P-values below 0.01 are accepted as differentially expressed. The enrichment score is then given as the normalized number of DEGs belonging to the group of breast cancer driver genes or PAM50 genes, respectively.

$$ES = \frac{enriched\ cancer\ marker\ genes * 16{,}883\ genes}{significant\ genes * cancer\ marker\ genes}$$

We perform a comparable DEA with DESeq2 using 5 random GTEx samples (control) under the same conditions (40-50 year-old females).

We also perform single-sample analyses of the four available breast cancer subtypes: Basal-like (84 samples), HER-2 (37 samples), Luminal A (176 samples) and Luminal B (84 samples), both for our model and DEseq2. We randomly choose one sample from each of the four subtypes as a case sample, and use all GTEx breast tissue samples (40-70 year-old females, 143 samples in total) as controls in the DEseq2 method. The experiment is repeated 20 times for each subtype.

False positive analysis

In order to assess the quality of the model's DEA, we perform an experiment to quantify its false positive rate. We therefore select a random GTEx breast sample from the test set (42 samples) as a false case sample. We perform DEA with both our model and DESeq2 to arrive at false positive DEGs (absolute log2-fold change greater than 1) for a range of adjusted P-values ranging from 0.01 to 0.1. We perform this 20 times and report the resulting DEGs as false positives. As controls for DESeq2, we choose 5 controls, which are randomly selected from the GTEx training set (440 samples). We also perform the analysis for DESeq2 using all breast samples, mentioned above, as controls.

Enrichment analysis of Cancer Driver Genes for multiple cancer types

Eleven different cancer types are involved in this analysis including Breast cancer, which are Adrenocortical Carcinoma, Bladder Urothelial Carcinoma, Brain lower Grade Glioma, Breast Carcinoma, Colon Adenocarcinoma, Kidney cancer (Kidney Chromophobe, Kidney Renal Clear Cell Carcinoma, Kidney Renal Papillary Cell Carcinoma), Liver Hepatocellular Carcinoma, Lung Adenocarcinoma, Prostate Adenocarcinoma, Stomach Adenocarcinoma and Thyroid Carcinoma. We perform 20 single-sample experiments for each of the cancer types. For each cancer type, its respective cancer driver gene list was used in the enrichment score calculation.

Acknowledgements

We would like to acknowledge discussions with Jonas Sibbesen and we thank the Department of Computer Science and the Bioinformatic Center for providing the computing server.

Funding

AK is funded by two grants from the Novo Nordisk Foundation: Center for Basic Machine Learning Research in Life Science (NNF20OC0062606) and Quantum for Life (NNF20OC0059939). AK and IPL receive funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017549, 'Genomed4all'. YL is supported by the China Scholarship Council (Grant 201804910693).

References

Abbas, Mostafa, and Yasser El-Manzalawy. 2020. "Machine Learning Based Refined Differential Gene Expression Analysis of Pediatric Sepsis." *BMC Medical Genomics* 13 (1): 122.

Altman, Russ B., A. Keith Dunker, Lawrence Hunter, Marylyn D. Ritchie, Tiffany A. Murray, and Teri E. Klein. 2019. *Biocomputing 2020 - Proceedings Of The Pacific Symposium*. World Scientific.

Aran, Dvir, Roman Camarda, Justin Odegaard, Hyojung Paik, Boris Oskotsky, Gregor Krings, Andrei Goga, Marina Sirota, and Atul J. Butte. 2017. "Comprehensive Analysis of Normal Adjacent to Tumor Transcriptomes." *Nature Communications*. https://doi.org/10.1038/s41467-017-01027-z.

Burska, A. N., K. Roget, M. Blits, L. Soto Gomez, F. van de Loo, L. D. Hazelwood, C. L. Verweij, et al. 2014. "Gene Expression Analysis in RA: Towards Personalized Medicine." *The Pharmacogenomics Journal* 14 (2): 93–106.

Cancer Genome Atlas Research Network. 2008. "Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways." *Nature* 455 (7216): 1061–68.

Chen, Yunshun, Aaron T. L. Lun, and Gordon K. Smyth. 2016. "From Reads to Genes to Pathways: Differential Expression Analysis of RNA-Seq Experiments Using Rsubread and the edgeR Quasi-Likelihood Pipeline." *F1000Research* 5 (June): 1438.

Coleman, William B., and Carey K. Anders. 2017. "Discerning Clinical Responses in Breast Cancer Based On Molecular Signatures." *The American Journal of Pathology* 187 (10): 2199–2207.

Cui, Weitong, Huaru Xue, Lei Wei, Jinghua Jin, Xuewen Tian, and Qinglu Wang. 2021. "High Heterogeneity Undermines Generalization of Differential Expression Results in RNA-Seq Analysis." *Human Genomics* 15 (1): 7.

Fukushima, K. 1975. "Cognitron: A Self-Organizing Multilayered Neural Network." *Biological Cybernetics* 20 (3-4): 121–36.

Gendoo, Deena M. A., Natchar Ratanasirigulchai, Markus S. Schröder, Laia Paré, Joel S. Parker, Aleix Prat, and Benjamin Haibe-Kains. 2016. "Genefu: An R/Bioconductor Package for Computation of Gene Expression-Based Signatures in Breast Cancer." *Bioinformatics* 32 (7): 1097–99.

Hu, Ying, Ling Bai, Thomas Geiger, Natalie Goldberger, Renard C. Walker, Jeffery E. Green,

Lalage M. Wakefield, and Kent W. Hunter. 2013. “Genetic Background May Contribute to PAM50 Gene Expression Breast Cancer Subtype Assignments.” *PloS One* 8 (8): e72287.
Kakati, Tulika, Dhruba K. Bhattacharyya, Jugal K. Kalita, and Trina M. Norden-Krichmar. 2022. “DEGnext: Classification of Differentially Expressed Genes from RNA-Seq Data Using a Convolutional Neural Network with Transfer Learning.” *BMC Bioinformatics*. https://doi.org/10.1186/s12859-021-04527-4.
Kamel, Hala Fawzy Mohamed, and Hiba Saeed A. Bagader Al-Amodi. 2017. “Exploitation of Gene Expression and Cancer Biomarkers in Paving the Path to Era of Personalized Medicine.” *Genomics, Proteomics & Bioinformatics* 15 (4): 220–35.
Kingma, Diederik P., and Jimmy Ba. 2014. “Adam: A Method for Stochastic Optimization.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1412.6980.
Kingma, Diederik P., and Max Welling. 2013. “Auto-Encoding Variational Bayes.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1312.6114v10.
Lee, Alexandra J., Dallas L. Mould, Jake Crawford, Dongbo Hu, Rani K. Powers, Georgia Doing, James C. Costello, Deborah A. Hogan, and Casey S. Greene. 2022. “SOPHIE: Generative Neural Networks Separate Common and Specific Transcriptional Responses.” *Genomics, Proteomics & Bioinformatics*, October. https://doi.org/10.1016/j.gpb.2022.09.011.
Li, Dongmei, Martin S. Zand, Timothy D. Dye, Maciej L. Goniewicz, Irfan Rahman, and Zidian Xie. 2022. “An Evaluation of RNA-Seq Differential Analysis Methods.” *PloS One* 17 (9): e0264246.
Liu, Shu-Hsuan, Pei-Chun Shen, Chen-Yang Chen, An-Ni Hsu, Yi-Chun Cho, Yo-Liang Lai, Fang-Hsin Chen, et al. 2020. “DriverDBv3: A Multi-Omics Database for Cancer Driver Gene Research.” *Nucleic Acids Research* 48 (D1): D863–70.
Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. “The Genotype-Tissue Expression (GTEx) Project.” *Nature Genetics* 45 (6): 580–85.
Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.
McDermaid, Adam, Xin Chen, Yiran Zhang, Cankun Wang, Shaopeng Gu, Juan Xie, and Qin Ma. 2018. “A New Machine Learning-Based Framework for Mapping Uncertainty Analysis in RNA-Seq Read Alignment and Gene Expression Estimation.” *Frontiers in Genetics* 9 (August): 313.
Parker, Joel S., Michael Mullins, Maggie C. U. Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, et al. 2009. “Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes.” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 27 (8): 1160–67.
Rapin, Nicolas, Frederik Otzen Bagger, Johan Jendholm, Helena Mora-Jensen, Anders Krogh, Alexander Kohlmann, Christian Thiede, et al. 2014. “Comparing Cancer vs Normal Gene Expression Profiles Identifies New Disease Entities and Common Transcriptional Programs in AML Patients.” *Blood* 123 (6): 894–904.
“Rectified Linear Units Improve Restricted Boltzmann Machines.” n.d. Accessed December 5, 2022. https://openreview.net/forum?id=rkb15iZdZB.
Schmauch, Benoît, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, et al. 2020. “A Deep Learning Model to Predict RNA-Seq Expression of Tumours from Whole Slide Images.” *Nature Communications* 11 (1): 3877.
Schuster, Viktoria, and Anders Krogh. 2021. “A Manifold Learning Perspective on Representation Learning: Learning Decoder and Representations without an Encoder.” *Entropy* 23 (11). https://doi.org/10.3390/e23111403.
———. 2022. “The Deep Generative Decoder: MAP Estimation of Representations Improves Modeling of Single-Cell RNA Data.” *arXiv*, November. https://doi.org/10.48550/arXiv.2110.06672.

Tishchenko, Inna, Heloisa Helena Milioli, Carlos Riveros, and Pablo Moscato. 2016. “Extensive Transcriptomic and Genomic Analysis Provides New Insights about Luminal Breast Cancers.” *PloS One* 11 (6): e0158259.

Vaca-Paniagua, Felipe, Rosa María Alvarez-Gomez, Hector Aquiles Maldonado-Martínez, Carlos Pérez-Plasencia, Veronica Fragoso-Ontiveros, Federico Lasa-Gonsebatt, Luis Alonso Herrera, et al. 2015. “Revealing the Molecular Portrait of Triple Negative Breast Tumors in an Understudied Population through Omics Analysis of Formalin-Fixed and Paraffin-Embedded Tissues.” *PLOS ONE*. https://doi.org/10.1371/journal.pone.0126762.

Vihinen, Mauno. 2022. “Individual Genetic Heterogeneity.” https://doi.org/10.3390/genes13091626.

Vivian, John, Jordan M. Eizenga, Holly C. Beale, Olena M. Vaske, and Benedict Paten. 2020. “Bayesian Framework for Detecting Gene Expression Outliers in Individual Samples.” *JCO Clinical Cancer Informatics* 4 (February): 160–70.

Wilks, Christopher, Shijie C. Zheng, Feng Yong Chen, Rone Charles, Brad Solomon, Jonathan P. Ling, Eddie Luidy Imada, et al. 2021. “recount3: Summaries and Queries for Large-Scale RNA-Seq Expression and Splicing.” *Genome Biology* 22 (1): 323.

Wirapati, Pratyaksha, Christos Sotiriou, Susanne Kunkel, Pierre Farmer, Sylvain Pradervand, Benjamin Haibe-Kains, Christine Desmedt, et al. 2008. “Meta-Analysis of Gene Expression Profiles in Breast Cancer: Toward a Unified Understanding of Breast Cancer Subtyping and Prognosis Signatures.” *Breast Cancer Research*. https://doi.org/10.1186/bcr2124.

Yersal, Ozlem, and Sabri Barutca. 2014. “Biological Subtypes of Breast Cancer: Prognostic and Therapeutic Implications.” *World Journal of Clinical Oncology* 5 (3): 412–24.

Zeng, William Z. D., Benjamin S. Glicksberg, Yangyan Li, and Bin Chen. 2019. “Selecting Precise Reference Normal Tissue Samples for Cancer Research Using a Deep Learning Approach.” *BMC Medical Genomics* 12 (Suppl 1): 21.

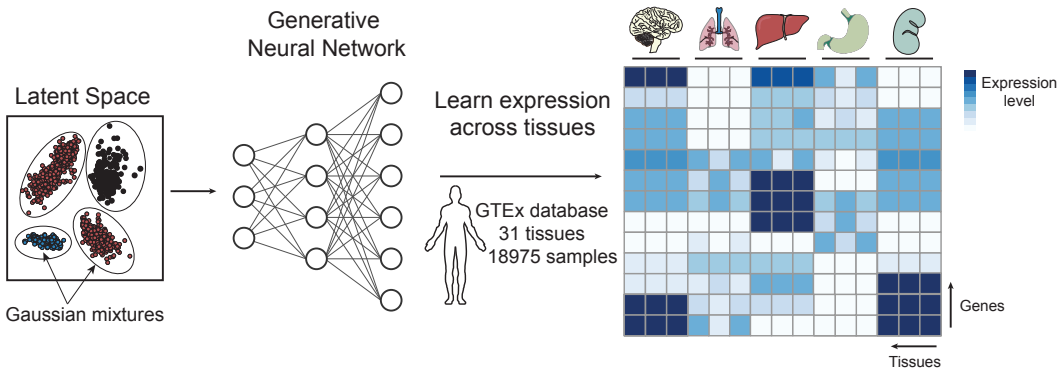

Figure1: The bulk deep generative decoder.The latent space is parameterized with a Gaussian Mixture Model (left). A generative neural network trained on GTEx maps the latent space to the data space. Altogether, the model learns the gene expression distribution across bulk tissues, as illustrated in the heatmap (right).

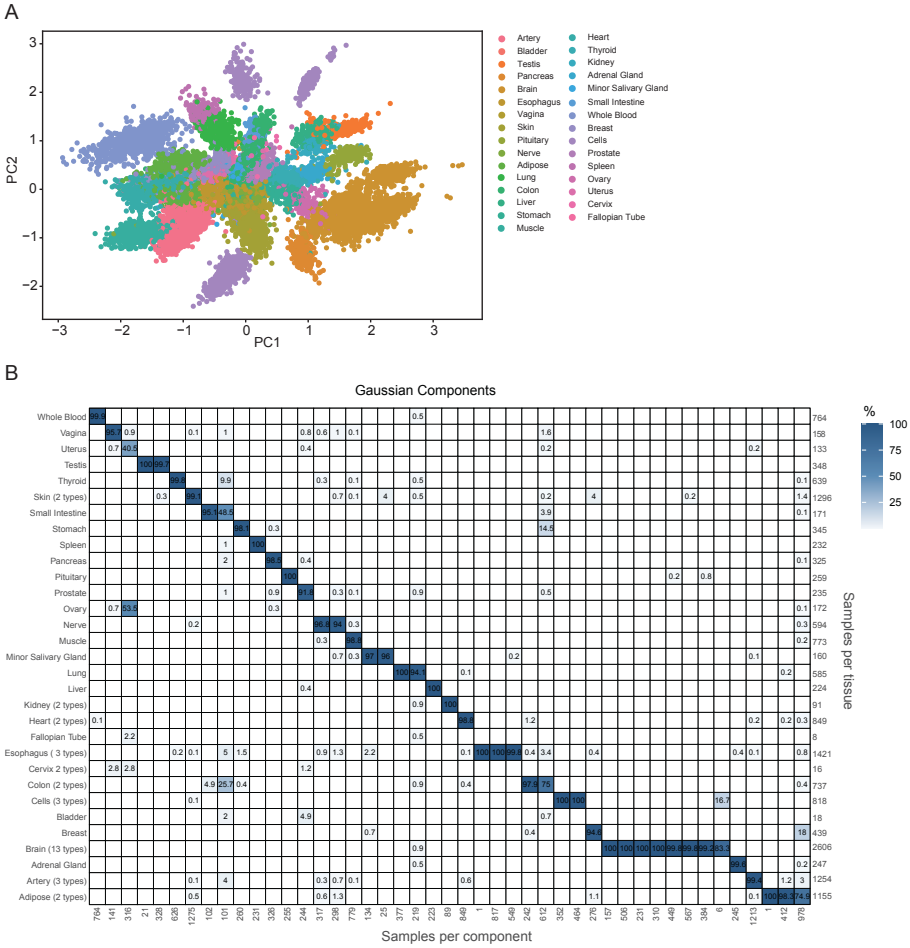

Figure2: Representations for normal tissue. A. The PCA plot shows the clustering of 31 tissues in the latent space reduced from 50 dimensions to 2. Each dot corresponds to a training sample. B. The matrix plot on the right shows the percentage of train samples from each tissue (y-axis) in each Gaussian mixture component (x-axis). The tissue types and component numbers are sorted for an optimal diagonal view. The number of samples assigned to each component are given by the numbers below the x-axis.

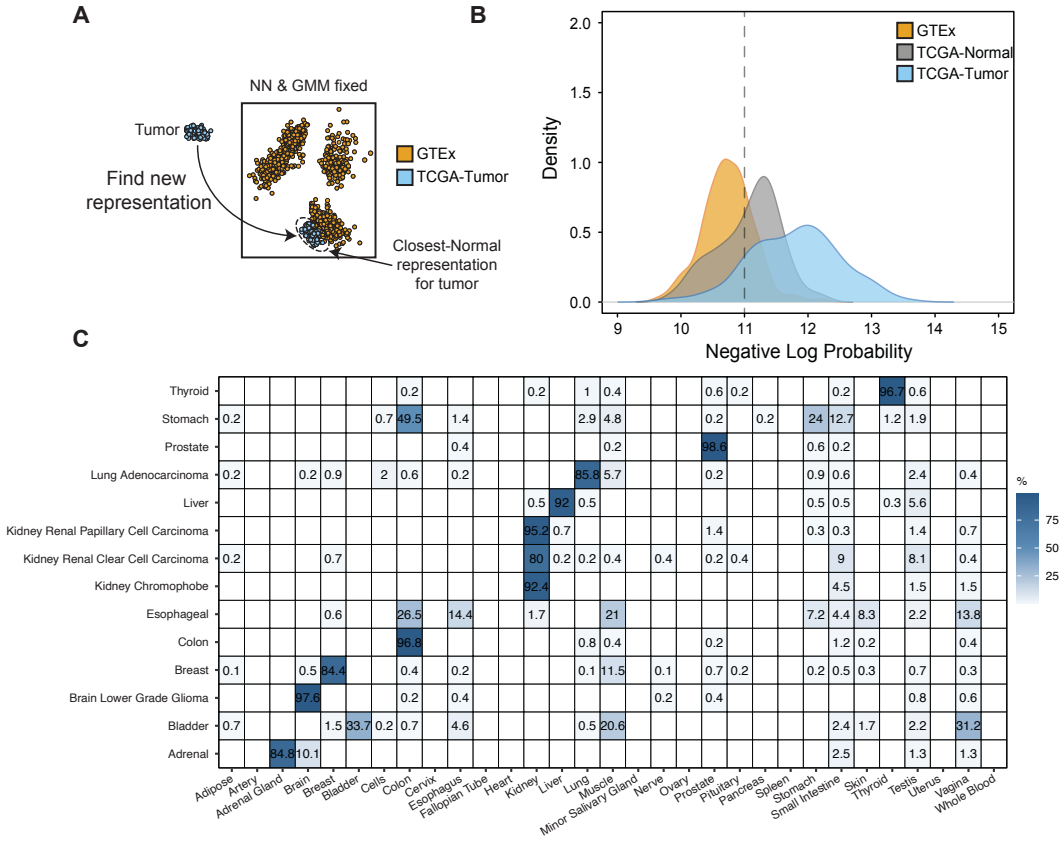

Figure3: Closest normal representation for cancers. A. Schematic representation for finding the closest-normal representations for TCGA tumors. We find the representation by maximizing the probability of the representation, while leaving the GMM and neural network fixed. B. Negative log probability distributions for GTEx samples (test samples), TCGA-normal and TCGA tumor samples. C. Matrix showing the percentage of TCGA tumors assigned to GMM components. The rows represent TCGA tumor samples and the columns show tissue-specific GMM components.

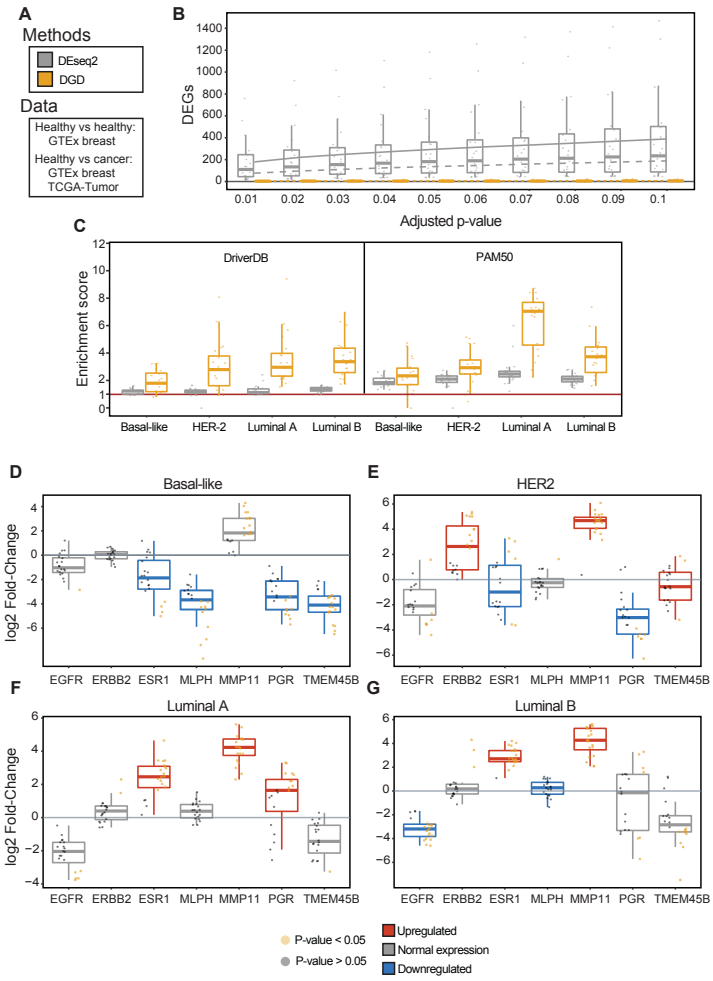

Figure4: Differential analysis of Breast cancer and its subtypes. A. Schematic overview of the samples used in our experimental set-up. B. Control experiment comparing normal test samples against the model (yellow) and against controls using DESeq2 (gray). 20 random test samples were chosen (shown with dots) and summarized by the boxplots for different cut-off on p-adjusted (x-axis). For DESeq2 the whole training set of breast tissue from GTEx was used as the control.The mean is shown by the dashed gray line. We also tested DEseq2 with a random subset of 5 controls and showed the mean in the solid gray line. C. Enrichment score across breast cancer subtypes for driver genes and PAM50 genes between our model and DESeq2. 20 random cancer samples were selected and compared to the model (yellow) and GTEx samples of breast (gray, see Methods for sample selection). The enrichment of cancer driver genes and PAM50 among the significant genes (p-adj<0.01) is shown for each subtype of breast cancer. C. Comparison of enrichment score for Breast cancer driver genes and PAM50 genes for DESeq2 and DGD. The enrichment scores were calculated based on the set differentially expressed genes obtained by each method. D. 1 versus model (DGD) and 1 versus GTEx train (143 samples) breast cancer subtype specific enrichment scores. D-G Breast cancer specific differential expression analysis on a subset of 7 marker genes, using 20 repetitions. The box-plots are colored based on whether the gene is known to be differentially expressed in a cancer subtype. The dots are colored based on the p-value obtained by DGD in each replication experiment.

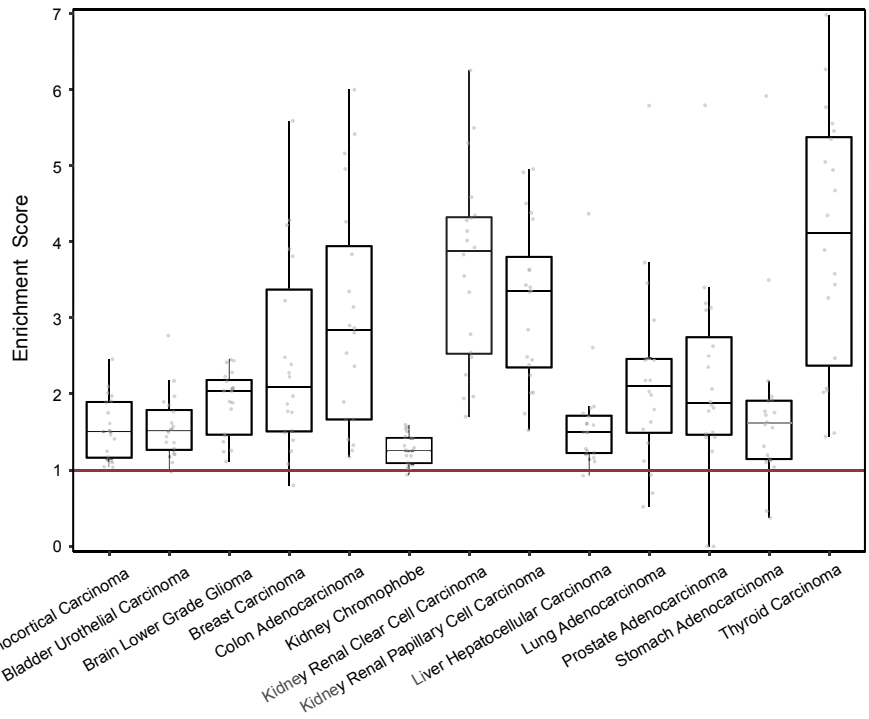

Figure5: The N-of-1 sample research of enrichment analysis for Cancer driver genes of different cancer types. Single sample was randomly selected from each TCGA-tumor type to calculate the differentially expressed genes by using our model. It was repeated 20 times for each cancer.

Table S1 GTEx samples							
Tissue Site Detail field	**Number of Sample per tissue**	**Number of Sample in training set**	**Number of Sample in test set**	**Tissue Site Detail field**	**Number of Sample per tissue**	**Number of Sample in training set**	**Number of Sample in test set**
Adipose Tissue	1293	1155	138	Ovary	195	176	19
Adrenal Gland	274	249	25	Pancreas	360	327	33
Artery	1398	1259	139	Pituitary	301	265	36
Bladder	21	18	3	Prostate	263	237	26
Brain	2931	2625	306	Minor Salivary Gland	178	160	18
Breast	482	440	42	Muscle	881	775	106
cell	920	826	94	Skin	1420	1302	118
Cervis Uteri	19	16	3	Small Intestine	193	171	22
Colon	822	741	81	Spleen	255	232	23
Esophagus	1577	1429	148	Stomach	384	349	35
Fallopian Tube	9	8	1	Testis	410	349	61
Heart	942	852	90	Thyroid	706	639	67
Kidney	98	91	7	Uterus	159	138	21
Liver	251	226	25	Vagina	173	160	13
Lung	655	589	66	whole blood	852	772	80
Nerve	659	597	62	Study NA	133	-	-

Table S2 TCGA tumor samples and TCGA adjacent sampes			
Tumor Tissue	**Number of Sample**	**Adjacent Normal Tissue**	**Number of Sample**
Adrenocortical Carcinoma	79	Adrenal	0
Bladder Urothelial Carcinoma	414	Bladder	19
Brain Lower Grade Glioma	532	Brain	0
Breast Invasive Carcinoma	1142	Breast	114
Colon Adenocarcinoma	505	Colon	41
Esophageal Carcinoma	185	Esophageal	13
Kidney Chromophobe	66	Kidney	25
Kidney Renal Clear Cell Carcinoma	546	Kidney Renal Clear Cell	72
Kidney Renal Papillary Cell Carcinoma	291	Kidney Renal Papillary Cell	32
Liver Hepatocellular Carcinoma	374	Liver	50
Lung Adenocarcinoma	542	Lung	110
Prostate Adenocarcinoma	506	Prostate	52
Stomach Adenocarcinoma	416	Stomach	37
Thyroid Carcinoma	513	Thyroid	59

Table S3 Clincial Filters for Breast Cancer Subtypes			
Column Name	**Include**	**Exclude**	**Note**
gdc_cases.project.primary_site	Breast		
gdc_experimental_strategy	RNA-Seq		
gdc_cases.demographic.gender	female		
gdc_cases.diagnoses.tumor_stage		stage iv & stage x	
gdc_cases.samples.sample_type	Primary Tumor		
cgc_case_pathologic_stage		Stage IV & Stage X	
cgc_sample_sample_type	Primary Tumor		
cgc_case_age_at_diagnosis	40-70		Match with GTEX
cgc_case_new_tumor_event_after_initial_treatment	NO | NA		
cgc_case_prior_diagnosis		Yes	
cgc_case_gender	FEMALE		
cgc_slide_percent_tumor_nuclei	50		Minimum
cgc_slide_percent_necrosis	5		Maximum
cgc_slide_percent_tumor_cells	50		Minimum
cgc_slide_percent_neutrophil_infiltration	5		Maximum
xml_history_of_neoadjuvant_treatment		Yes	
xml_distant_metastasis_present_ind2	NO | NA		
xml_first_nonlymph_node_metastasis_anatomic_sites	NA		
gdc_cases.demographic.race	*white*		*Only if in GTEX*
cgc_case_race	*WHITE*		*Only if in GTEX*
cgc_drug_therapy_pharmaceutical_therapy_type			*We would like to set this to NA, BUT if I do this we lose most samples as most patients got treatment. So, for now I did Chemo + NA.*
xml_radiation_therapy			*Same as above*

Table S4 The number of times genes were identified as significant in each of the four subtypes and the direction of their expression								
Gene	**Basal**	**Her2**	**LumA**	**LumB**	**Basal**	**Her2**	**LumA**	**LumB**
ANLN	up	up						
BCL2				down				4
BIRC5		up		up				
CCNB1				up				
CDC6		up		up		3		2
CDH3			down	down				
CEP55	up	up		up				
EGFR				down				16
ERBB2		up				11		
ESR1	down	down	up	up	4	6	16	19
FGFR4	down			down	1			5
FOXA1	down							
FOXC1		down		down		6		7
GPR160	down	up	up	up	2	9	5	11
GRB7		up				8		
KNTC2	up			up				
KRT14	down	down	down	down		2		1
KRT17		down		down				
KRT5		down		down		3		4
MELK	up			up				
MIA		down	down	down		4	1	5
MLPH	down			down	10			
MMP11		up	up	up		19	20	20
MYBL2	up			up	1			1
MYC		down				4		
NAT1			up					
PGR	down	down	up		9	7	10	
PHGDH			down	down			6	9
RRM2		up				1		
SFRP1		down	down	down		9	5	16
SLC39A(	down	down	up	up	1		13	12
TMEM45	down	up			16	2		
TYMS	up			up	1			1
UBE2T	up	up		up	1			3
The threshold to identify significant gens:								
log2FoldChange > 1 or < -1								
P-adjust value < 0.05								