



TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

JAKE LIN

INTEGRATION AND ANALYSIS OF VIROME NGS DATA

Master of Science thesis

Examiner: prof. Olli Yli-Harja
Examiner and topic approved by the
Faculty Council of the Faculty of
Computing and Electrical Engineer-
ing on 10th April 2015

ABSTRACT

JAKE LIN: TUT Thesis Template

Tampere University of technology

Master of Science Thesis, 50 pages, 5 Appendix pages

July 2015

Master's Degree Programme in Science and Bioengineering

Major: Information Technology for Health and Biology

Examiner: Professor Olli Yli-Harja

Keywords: NGS analysis, diabetes, virus molecular mimicry, data integration, web visualization

Type 1 Diabetes is a growing disease impacting young children. The disease is caused by the pancreases producing insufficient amounts of insulin required for the body to convert sugar and starches into required energy. There is no known cause to Type 1 Diabetes though there are cited genetic and environment associations. An environmental implication is viruses disrupting host immune and regulatory systems using molecular mimicry inducing islet autoimmunity. One observed diabetes correlation is the presences of Picornavirus family, including enterovirus and poliovirus. Taking advantage of advances in the sensitivity and accuracy of Next Generation Sequencing, Diabetes Prediction and Prevention researchers have started a study to investigate the viral genome diversity and whether viral genomes can accelerate islet autoimmunity, when the host fails to produce insulin, triggering diabetes.

This application of Next Generation sequencing provides an unbiased study and essentially applying metagenomics techniques to disease research. Sequence analysis requires processing and storing of vast amount of information. The other challenges are data integration, interpretation and lack of reference viral genomes. This thesis describes an open sourced Web 2.0 project to integrate and manage analysis results upon application of Velvet, for *de novo* assembly and BLAST for viral strain identification.

Because of the lack of reference, majority of next generation reads are unmapped. Accounting for these unmapped reads and also investigating molecular mimicry, additional computational analysis includes translating raw sequence reads into protein fragments. Frames of these fragments were matched with islet autoimmunity protein markers. The preliminary matched results appear valuable with case affected samples in clear majority. The assembled results can be dynamically plotted on the web as aligned protein tracks for visual pattern and density inspection.

PREFACE

This thesis project and paper was done while working for the Professor Heikki Hyöty (Virology Group – School of Medicine@UTA). In addition, I am grateful to thank Dr. Ondrej Cinek for his help and patience with answering many questions about his existing sequence analysis scripts. I also would like to express my gratitude to my thesis advisors Dr. Reija Autio (School of Public Health@UTA) and Professor Matti Nykter (Computational Biology - BioMediTech@UTA) for their valuable advices and encouragements.

Tampere, 28.7.2015

Jake Lin

CONTENTS

1. INTRODUCTION.....	1
2. MOLECULAR DATA AND SEQUENCING.....	5
2.1 Biological background.....	5
2.1.1 Genomes.....	5
2.1.2 Gene expression and inheritance.....	6
2.1.3 Diabetes.....	8
2.1.4 Diabetes and environment.....	9
2.1.5 Virus and human health.....	9
2.2 Measuring Data.....	11
2.2.1 Microarrays.....	12
2.2.2 Sequencing Background.....	12
2.2.3 Next Generation Sequencing applications.....	13
2.2.4 Assembly.....	14
2.2.5 Alignment, Aberrations and Annotations.....	14
2.2.6 Genomic Visualization.....	15
2.3 Information Technology for Life Science.....	16
2.3.1 Internet Transformation.....	16
2.3.2 Relational databases.....	17
2.3.3 Cloud Computing.....	18
2.3.4 Computational Biology.....	18
2.3.5 Statistical Learning.....	19
2.3.6 Bioinformatics Platform.....	20
3. VIROME NGS ANALYSIS AND WEB INTEGRATION METHODS.....	21
3.1 Metagenomics.....	21
3.2 Experiment and NGS Preparation.....	21
3.3 Analysis Process.....	22
3.3.1 De Novo Assembly.....	22
3.3.2 BLAST commands.....	23
3.3.3 Viral identification with BLAST.....	24
3.3.4 Strain identification with BWA.....	25
3.3.5 Islet Autoimmunity Protein Assessment - Translation.....	26
3.3.6 Islet Autoimmunity Protein Assessment - Matching.....	28
3.3.7 Islet Autoimmunity Protein Align and View.....	31
3.4 Web Programming for Data Exploration and Management.....	31
3.4.1 LAMP architecture.....	31
3.4.2 Interactivity with JavaScript.....	32
3.5 Advanced visualization.....	35
3.6 Cloud Computing with CSC.....	36
4. VIROME NGS WEB APPLICATION RESULTS.....	37

4.1	Homepage and Layout.....	37
4.2	Interactive Interface.....	39
4.2.1	Selection and Sort, Filter and Refine	39
4.3	Visualization.....	39
4.3.1	Viral Reads Heatmap	39
4.4	Statistical Clustering.....	41
4.5	Islet Autoimmunity Markers Assessment.....	42
4.5.1	Insulin Case and Control Assessment	43
4.5.2	Case and Control Summary Counts	44
5.	DISCUSSION	46
5.1	Overview.....	46
5.2	Interface to Integrated Results and Interactive Visualizations	46
5.3	Molecular mimicry and islet autoimmunity	47
5.4	Limitation	47
5.5	Open source	48
6.	CONCLUSION	49
	REFERENCES.....	51

APPENDIX A: DNA CODON TO PROTEIN AMINO ACIDS

APPENDIX B: VIRAL HIT STRAINS

APPENDIX C: BACTERIAL HIT STRAINS

APPENDIX D: ISLET AUTOIMMUNITY PROTEIN MARKERS

TERMS, SYMBOLS AND ABBREVIATIONS

A	Adenine, DNA nucleotide, bonds to T
Aberrations	Structural abnormalities in DNA, can be small scaled like SNPs or indels or copy number events
ACID	Atomicity, consistency, isolation and durability
AIDS	Acquired Immune Deficiency Syndrome associated with HIV
AJAX	Asynchronous JavaScript and XML
Alignment	Computational process of arranging multiple stretches of DNA sequences to identify similarities in functional or evolutionary
Allele	Alternative forms of the same gene responsible for traits, sometimes call phenotype expressions
Amino acids	Organic chemical compounds that are the foundation elements of proteins
Annotation	Computational process of assigning functional roles in stretches of sequence
Antibody	A large Y-shaped protein used by immune system to identify and killed pathogens
Antigen	A biochemical structure that binds to a certain antibody, can be self or produced by foreign pathogen
Assembly	Computation task of putting together a genome sequence
BAM	binary sequence alignment mapping file format
BLAST	Basic Local Alignment Search Tool used for protein and DNA comparisons
Bowtie	Popular genomic aligner software known for speed and efficiency
BWA	Burros Wheeler Alignment
C	Cytosine, DNA nucleotide, bonds to G
CD4	A type of white blood cells protecting the body from infection
CGI	Common gateway interface
Cloud computing	Enabling the usage of network, nonlocal computation resources
Codon	DNA triplets that can be translated into an amino acid
Contig	Contiguous (non-interrupted) segment of DNA or protein sequence
Copy number	Not having two copies of a particular gene because of extra or loss of chromosome fragments
CSC	Finland's IT Centre for Science
CSV file	Comma separated value file
De novo assembly	Arranging the DNA or RNA of an organism linearly in correct order without the aid of a reference
Dendrograms	Tree diagram to visualized hierarchical relationships
Diabetes	Growing human disease involving high sugar levels in blood. Symptoms include increased hunger and frequent urination. Classified as type 1 and type 2
DIPP	Diabetes Prediction and Prevention project studying Type 1 diabetes, based in Finland and includes international collaborators
DNA	Deoxyribonucleic acid that carries genetic information for almost all organisms
DOM	Document Object Model refers to computational structure of HTML page
EMBL	European Molecular Biology Laboratory

ENCODE	Encyclopedia of DNA Elements is a consortium project follow up to Human Genome project to comprehensively annotate the human reference genome
Epigenetic	Changes in gene expression or protein function that are beyond changes in DNA or peptide sequence
Exon	DNA in chromosomes that code for genes
FASTA	File format produce by sequencing platforms, consisting of run identifier and quality info, then DNA fragment
G	Guanine, DNA nucleotide, bonds to C
GAD	Glutamate decarboxylase gene, a islet autoimmunity marker
Genomics data	Comprehensive term to include gene and DNA measurements across modern platforms such as microarrays and sequencing
Genotype	Genetic makeup of a cell
GWAS	Genome wide association study
Hamming distance	Measure distance that equals the number of positions that are different between two strings of equal lengths
Heatmap	Graphical representation of assigning values of a matrix to a color, typically range of gradual shades
HIV	Human immune virus
HLA	Human leukocyte antigen, locus where antibodies bind
HTML	Hypertext markup language
HTTP	Hypertext transfer protocol
H1N1	A variation of influenza virus from 2009 that originated from swine and considered an endemic resultant in 17,000 deaths
IA2	Islet antigen antibody
IGV	Integrated Genome Viewer software
Indel	Insertion or deletion aberrations in genome when compared to reference
INS	Insulin gene
Intron	Noncoding sequencing regions in chromosome DNA
In silico	Computer only biological simulation task
Islet autoimmunity	Term to describe beta cells, producing insulin, being targeted and killed by immune system prior to onset of diabetes
JOIN	SQL term to describe constructing a relationship between two tables
jQuery	Popular JavaScript library used in web programming
JSON	JavaScript Object Notation
LAMP	Popular open sourced web program platform that stands for Linux, Apache, MySQL, and Python/Perl
Levenshtein Distance	Metric to describe the difference between two strings
Mendelian	Form of inheritance proposed and confirmed by Gregory Mendel to include Law of Segregation (alleles), Independent Assortment and Dominance. His experiments invalidated the false gene blending principles believed by his contemporary biologists
Metagenomics	Science discipline to collect and measure samples from its native environment
Microarray	Technology to capture gene expressions
Mitosis	Cells divide and replicate
MRI	Magnetic Resonance Imaging
NCBI	National Center for Biotech Information

NGS	Next generation sequencing
NT	Nucleotides
Pathogen	Bacteria or virus that can produce a disease
Pathogenesis	Biological mechanisms leading to disease state
PCA	Principal component analysis
PCR	Polymerase Chain Reaction, a biotech method to amplify, create additional copies of a particular compound or structure
Phenotype	Observable trait characteristics, can include disease risks
PHRED score	A measure of quality for DNA bases from sequencing
Protein	Translated from RNA and large chemical responsible for all biological functions and structures
QC	Quality control
R	Open sourced and R statistical computer language
Reference	In bioinformatics and biology, reference is the accepted and representative DNA sequence of a particular species, such as human genome reference.
RNA	Ribonuclei Acid, transcribed from DNA and carries genetic information for some viruses
SAM	sequence alignment mapping file format
Sequencing	Capturing the DNA linear bases, refer to as chromosome or genetic sequence. Sequencing can be deep/full or exon (gene coding regions only)
SNP	Single nucleotide polymorphisms, can be non-synonymous (functional) or synonymous, not impacting resultant protein
SQL	Structured Query Language
T	Thymine, DNA nucleotide, bonds to A
TP53	Tumor protein 53, well study cancer suppressant gene
Transcription	DNA to RNA
Translation	RNA to Protein
TSV file	Tab separated value file
Type 1 diabetes	Diabetes impacting young children, no known cause
Type 2 diabetes	Diabetes impacting adults, insulin resistance resulting from diet, lifestyle and associated with obesity
T1D	Type 1 Diabetes, also known as juvenile Diabetes
U	Uracil, RNA nucleotide, bonds to A
UI	User interface
Velvet	Popular short read assembler software
Virome	Virus genome
Wnt Signaling	Genetic components responsible for passing messages from outside of cell to inside. Found to be integral to cancer and development
XML	Extensible markup language
W3C	World Wide Web Consortium
ZNT	Zinc transporter gene, an islet autoimmunity marker

1. INTRODUCTION

Diabetes is diagnosed when the pancreas does not produce enough insulin or when the cells are not responding to insulin resulting in failure to properly take up glucose and sugar. This failure can lead to many detriments such as heart diseases, stroke and kidney failures. Adult diabetes, also known as Type 2 is associated with obesity, diet and life style reasons; Type 1 Diabetes is characterized and impacting young children ages 0 to 13. A growing disease, it is cited in the International Diabetes Atlas [1] that around 8.3% of all adults have diabetes with an estimated global economic cost of diabetes in 2014 to be 614 billion US dollars and an estimated 20% of total US healthcare dollars go involves care for a diagnosed diabetic patient [2]. About 10% of diabetes is classified as Type 1 and there is no definite known cause though there has been published genetic and environment associations. Impacted children require regular injection of insulin. Functionally, the pancreases in the inflicted children produce none or insufficient amounts of insulin required for the body to convert sugar and starches into required energy [3,4]. One observed correlation is the presence of *enteroviruses* [5] and a well-known family member is *Poliovirus*.

Disturbingly, young children in Finland have the highest rate of having Type 1 Diabetes [3,6]. The virology group in School of Medicine at University of Tampere is part of the Diabetes Prediction and Prevention (DIPP) [7] consortium in screening and studying young children to further understand and find new treatment for the disease. The group is researching environmental causes, including viral and bacterial roles as a possible trigger and cause of the chronic disease. The group has partnered with hospitals in Tampere and Turku and innovatively collected control and afflicted stool samples from multiple time points, including before and after onset of disease where external insulin must be administered. It is well known that genes act in concerted pathways in determining disease state but what is unknown is the triggering of the innate immunity defense to messaging the beta cells to destroy the insulin producing pancreatic cells [6,8]. Developments and advances in Next Generation sequencing (NGS) technologies have enabled the increasing accuracy in the collection of DNA data [9]. A central theme in modern life science is that DNA is the blueprint for cells. Normal healthy cells within an organism have the exact DNA sequences and DNA mutation events, where a nucleotide base changes, while mostly mute can also be detrimental [10].

The central dogma of modern molecular biology describes the process of DNA being transcribed to RNA. Subsequently amino acids, translated from RNA, become the building blocks of different proteins. It is without exception that proteins are the enablers of vital organism traits and activities including reproduction and metabolism func-

tions. The different gene expressions, equivalent to different amounts of proteins within cell types are influenced by transcription factors, proteins themselves [11]. Many large-scaled projects are focusing on quantifying these gene expressions and particularly the regulating effect, promoting or repressing. Traditionally the expressions are measured using microarrays [12] but more labs are using NGS techniques due to its more sensitivity [9]. Along with such advances, though also brings new challenges and requirements: vast amount of information needs to be stored, analyzed and visualized. The bioinformatics methods and algorithms become the central components to information science pipelines. It is also acknowledge that software programs such as laboratory information management systems are vital in managing samples, annotations and also lab protocol steps and experimental designs. Moreover, the inputs, parameters, scripts and respected outputs must be tractable and repeatable. In all, bioinformatics, or the field of information aims to expertly manage, analyzed via application of statistical learning methods and integrate different modern instrument readings to benefit biomedical research [13, 14].

This thesis is part of Finland's Diabetes Prediction and Prevention study. There is intense focus in diabetes research as the incidence of Type 1 diabetes in Finnish children is the highest in the world. Started in 1994, DIPP is a comprehensive and time coursed project and some 150,000 children have been screened for Type 1 diabetes allelic markers [7]. Working together with researchers in studying the role of environment, particularly viral and bacterial, in triggering Type 1 Diabetes, the primary requirements of this thesis project are to enable investigators, including international collaborators, secured and interactive exploration and interpretation of analysis results from NGS sequencing of stool samples from selected diagnosed children across three time points together with unaffected matching controls. NGS assembled reads are *de novo* assembled and then analyzed and remapped for viral and bacterial matched content [8].

A web application using modern open source framework is developed to manage the analysis results from the NGS data analysis. As results from NGS data analysis tend to be complicated and its results hard to interpret, this web application needs to support flexible search functions to assist investigators with intuitive exploration of the results stored in a centralized database. An integrated and centralized solution also allows for efficient management of pipeline analysis jobs and tracking input parameters to outputs for data reproducibility. Web applications are the basis for secured client and server modern platforms. These browser interface solutions are installation free and capable of integrating annotations and results within interactive visualizations. Dynamic visualizations produced by learning methods can help investigators with insights into hidden patterns.

There is broad diversity among virus families and strains. It is suggested that there could be millions of viruses and likely these viruses have evolved to target different organisms [15]. Because of this vast flora and shorter viral generations, a major con-

cern in viral research is the lack of standard genomic reference and it was not a surprise that the majority, 74% of NGS outputs did not map to any known viral references [8]. A key contribution of this thesis is development of a novel analysis in assessing islet autoimmunity with molecular mimicry. Islet autoimmunity describes the stage where insulin-producing beta cells are being mistaken and targeted by the immune system and subsequently depleted. Occurring right before the onset of T1D symptoms, the cause of islet autoimmunity is unknown and the process is irreversible. Molecular mimicry is the collected term to describe protein sequence similarity from pathogen peptides triggering immunity response [16]. Antibodies identify antigens based on protein structures and the structure is based on linear chains of amino acids. Briefly the analysis comprehensively translates the NGS *contigs*, reads filtered for quality and assembled to exclude gaps, into multiple amino acid frames. The next step takes the translated amino acid files and matches sequence frames of 7 against human reference insulin (*INS*) and islet autoimmunity signature antibody gene markers (*GAD65*, *IH2*, *ZNT8A* and *ZNT8B*) [17]. Frame size of 6 to 8 is considered significant in amino acid identity assessment [17, 18]. The matched perfect and nearly perfect (1 missed) matches are stored. These results are further explored for hidden patterns on the web as reference aligned sequence tracks. Figure 1 below lists the major components of the software project. Source code and issue tracker are available at <https://sourceforge.net/projects/viromet1d>.

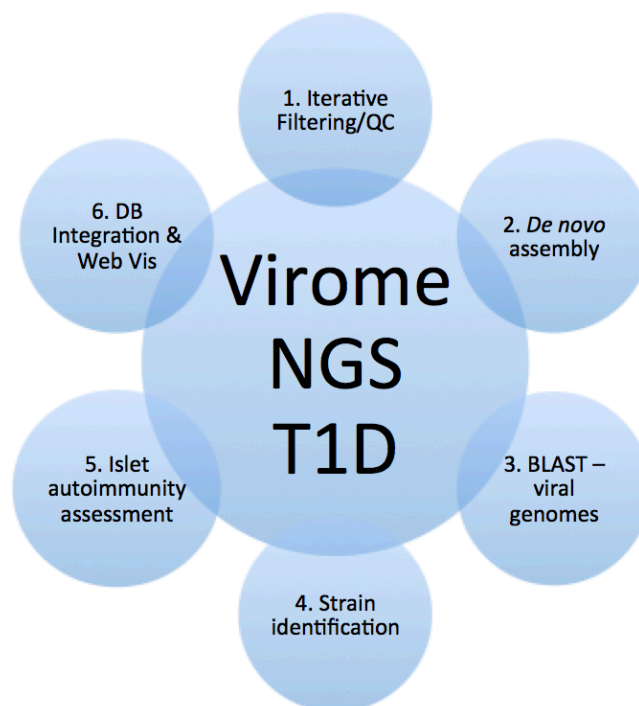


Figure 1. The main tasks of the Virome NGS analysis and integration pipeline consist of quality filtering and strain identification followed by assessment of islet autoimmunity antibody (molecular mimicry) signature markers. Results are accessible on web.

The thesis is structured into the following five succeeding chapters. The second chapter, *Molecular Data and Sequencing* will briefly discuss molecular central dogma of DNA to protein, diabetes biological underpinnings along with an overview of sequencing technology background and applications. Third chapter, *NGS Analysis Tools and Web Integration Methods* explains the required bioinformatics tools and databases relied on by researchers. This includes details of the implemented software steps required for quality control, taking raw reads to sequences and assembly, annotation and particularly database integration. The popular open sourced tools are also discussed. The next chapter *Virome NGS Application Results* provides details and features of the interactive visualizations presenting the analysis findings. The *Discussions* chapter discusses the summary of findings and limitations. This is followed by *conclusions* made from results and discussions.

2. MOLECULAR DATA AND SEQUENCING

Molecular biology is an information science dependent on accurate measurements. The advance and maturation of sequencing technology and instrumentation have allowed researchers to pose and probe unanswered questions from basic biology to human disease unknown underpinnings.

2.1 Biological background

2.1.1 Genomes

Modern biology defines genome as the genetic material of an organism. For almost all organisms, the genetic material consists of DNA, large and stable molecules located inside the nucleus and densely packaged and structured chromosomes. DNA is made of nucleotides Adenine (A), Thymine (T), Guanine (G), and Cytosine (C). It is important to note that DNA is double stranded and nucleotide properties of A exclusively bonding with T and G with C forms a complementary helix and allows for self replication, a key part of *mitosis*, when somatic cells replicated. This property forms the basis for sequencing instrumentation. RNA is slightly different in that thymine is replaced with Uracil (U) and less stable because it is single stranded. Different organisms have different copies and numbers of chromosomes. Normally, the human genome has two copies of 23 chromosomes that includes the sex chromosome of X and Y. The dual copies are inherited one each from mother and father. The process of crossover allows for some small exchanges of genetic material between mother and father chromosomes. Aneuploidy is the disorder that defines organisms with incorrect numbers and copies of chromosomes and this results in various diseases [11, 15]. Down's syndrome, also called trisomy 21 is when a human is born with an extra chromosome 21 [20]. Advanced human cancer cells have been observed with chromosomal structural aberrations also known as copy number variations. The reason that incorrect copies and numbers of chromosomes result in disease is that different genes are physically encoded in different chromosomes. For example, insulin, the diabetes gene responsible for regulating blood sugar in the human body, is located in discrete region of chromosome 11. Interestingly and actively research upon, most DNA segments in higher organisms do not code for genes. Only 3% of human chromosomes are known to code for the estimated 20,000 genes. The reason that the number of genes is an estimate is because of active discussions and disagreements on what constitutes a gene. Mechanically, gene start sites begin with nucleotides ATG, called an open reading frame and also stop sites of TAG, TAA or TGA; the full list of codons to amino acids are presented in Appendix A. The in be-

tween DNA segments are transcribed into RNA, transported into the ribosome where upon alternate splicing, the exon RNA are translated into functional proteins. Splicing removes segments of intragenic RNAs, commonly called introns and then combines the trimmed resultant exon. Splicing occurs in all domains of life though most common in eukaryote kingdoms, covering all higher order organisms. The process of transcription of DNA into RNA and subsequent translation of RNA into protein is the central dogma of molecular biology [11, 15].

2.1.2 Gene expression and inheritance

Another central idea of modern biology is that every cell of an individual organism has the same genetic material and yet different sets of genes, also called pathways, are expressed in different amounts. At the molecular scale and transcription level, the major controlling elements of gene expression are proteins called transcription factors. Transcription factors controlled the number of copies of RNA by binding to DNA promoter sequences usually upstream of the gene start site and can act as enhancers and repressors. The different amounts and rations of proteins produced are the chemical functional molecules responsible for all life functions. Coordination of proteins underscores all chemical functions and expressions, from digestion rates to physical movement to diseases [11, 15]. Proteins are large complex molecules responsible for all of the cells structure, function and various scales of biological regulations. Adapted from U.S. National Institute of Health (NIH) [21], Table 1 below lists the different function categories of proteins.

Table 1. Proteins are divided into different functions. Insulin, GADA and ZNT8 are important diabetes genes. Screenings for antibodies associated with these genes are used by clinics with higher risk children. Proteins often act in concert and especially messenger proteins are part of central biological pathways and can influence other pathways.

Function	Description	Example protein
Antibody	Antibodies help immune systems in identifying and destroying foreign particles and agents such as viruses and bacteria.	Immunoglobulins (<i>IG</i>) Y-shaped protein produced in plasma and beta cells against antigens
Enzyme	Enzymes carry out the numerous chemical reactions that take place in cells. They also assist with the formation of new molecules by reading the genetic	Glutamic acid decarboxylase (<i>GADA</i>) synthesizes <i>GABA</i> , an inhibitory neuron transmitter

	information stored in DNA.	
Messenger	Messenger proteins transmit signals between cells to coordinate biological processes. Hormones and growth factors are different classes of messengers.	Insulin (<i>INS</i>) allow cells to absorb glucose in the blood
Structural	These proteins provide structure and support for cells. Muscle and skin cells have more structural gene expressions.	Fibrillin (<i>FBN</i>) connective tissues
Transport/storage	These proteins bind and carry small molecules within cells and throughout the body. The transfer can be active and passive through cell membranes.	Zinc transporter (<i>ZNT8</i>) Cation transporter protein

Giving rise to variation, this theme is the premise and fundamental to the concept of genotypes to phenotype. Phenotypes are the observable traits, everything from growth rate of yeast in sugar to a person's eye color to complex disease states. Many organisms including human have two copies chromosomes, parental recessive allele genes can be expressed with their children having surprising traits, such as eye color and pigmentation. Dominance and recessive inheritance theories are credited justly to Gregor Mendel where in 1866 he published his findings on results on pod pea attributes such as height and seed color collected over years of experimentations and data tracking [22]. On a population level, the expressions of certain genes are defined in percentages; dominant genes have higher penetration. Related to penetration, alleles are different forms of a particular gene and can be dominant or recessive. A famous example of single allele dominant is Huntington's disease, the deadly deterministic neuron disorder involving cytosine-adenine-guanine repeats in the *HTT* gene impacting the protein and leading to gradual brain damage [23]. It is quite rare that genes or proteins are single dominant especially at complex traits such as disease, where genetics and epigenetic variables such as environment and life styles interact and contribute dynamically across individuals.

The action of transcription factors allows for unique variations but as they can promote or depressed, possibly leading to uncontrolled cell growth. Since cancer is defined by uncontrolled growth leading to malignant tumors, transcription factors are implicated and particularly their aberrations, that is mutations in the functional DNA where certain expressions of pathways are expressed uncontrollably. An often-cited example is mutations of *TP53* protein, on chromosome 17, as they are found frequently across different cancers [24]. Since *TP53* plays a large part in directing cell death, commonly referred to as apoptosis, and also part of the *Wnt Signaling* pathway, the mutation findings correlation is not surprising [24,25]. The inherent challenges of finding causality and also linearity, as most malignant cancer development occurs over long stretches of time are very complex. Cancer cells within the same locus are often heterogeneous with different gene expressions hence one reason drugs targeting cancer cells often have adverse side effects. The new discipline of personalized medicine aims to provide personal genomic analysis and in turn adjust cancer medicine based on individual differences, essentially catered to the individual cancer gene responses [26].

2.1.3 Diabetes

As stated in the opening chapter, Type 1 Diabetes (T1D), also known as juvenile diabetes, is diagnosed when the pancreas does not produce enough insulin or when the cells are not responding to insulin resulting in failure to properly take up glucose and sugar. Adult diabetes, also known as Type 2 is associated with obesity, diet and life style reasons. Adult diabetes impacts almost one in ten adults globally and the distribution is equal across men and women. Some common symptoms are extreme fatigue, thirst, hunger, frequent urination, and pain in hands/feet. Treatment includes improving life styles to include more exercises and better eating habits to maintain normal blood sugar level. Type 1 Diabetes is characterized as impacting young children ages 0 to 13. An estimated global economic cost of diabetes in 2014 is said to be 614 billion US dollars per year [2]. About 10% of diabetes is classified as Type 1 and these young children required daily injections of insulin. Mouth ingested insulin tablets do not work since the insulin must directly be in the blood stream [1,6].

This thesis work is a part of Finland's Diabetes Prediction and Prevention project and explicitly supports the exploration of NGS Viral study [8] results and implementation and integration of additional bioinformatics tools. There is intense focus in diabetes research as the incidence of Type 1 diabetes in Finnish children is the highest in the world [3,4]. Started in 1994, DIPP [7] is a comprehensive and time coursed project and some 150,000 children have been screened for T1D. Across 16 clinics, children and young adults with Type 1 diabetic relatives are eligible for DIPP studies. The screening aims to detect autoantibodies associated with diabetes from blood, it is noted that young children with a closed relative with T1D has a ten fold higher rate of getting the disease. A signature of T1D, islet autoimmunity is when beta cells in the pancreas islet are being

targeted and subsequently destroyed by the immune system. Before T1D symptoms occur, islet autoimmunity is marked by antibodies directed against *INS*, *GAD*, *IA-2*, *IA-2B* and *ZNT8* [17]. Only 3 to 4 individuals of out very 100 will test positive for autoantibodies [4,7] and detection of autoantibodies elevates the risk. The study provides close monitoring to include follow up appointments of at risk individuals. As stated previously, there is no definite known cause though there are published genetic and environment correlations. As an exception, Maturity onset diabetes of the young (MODY), or monogenic diabetes is a rare hereditary form of Type 1 diabetes caused by mutations in autosomal dominant diabetic related class of 12 genes, impacting around 1-2% of diagnosed individuals [28]. One main requirement of this thesis is to assist DIPP investigators, including international collaborators, with results interpretation and this is accomplished via browser-based exploration and advanced visualizations such as heatmaps. NGS results are hard to interpret because they are often heterogeneous, high in dimensions and density. Heatmaps can reveal hidden patterns within the results or relationships between different columns. In addition the viral results from the sequence data have been integrated with sample background annotation and supported with flexible search options.

2.1.4 Diabetes and environment

Type 1 Diabetes deviates asymmetrically from standard Mendelian inheritance laws [22]. This law implies that traits are passed from parents to off springs obeying laws of probability. However in T1D studies, a child with type 1 diabetic mother has a 1 in 25 chance if the mother is 25 or younger. The chance lowers to 1 in 100 if the mother is older than 25 [29]. An inflicted father gives the child 1 in 17 odds. The conclusive proof is that if one identical twin gets Type 1 diabetes, the other twin has less 50% incident rate and as identical twins have the same genes therefore the environment must be a contributor [5,8,29].

2.1.5 Virus and human health

Viruses occupy an unique place in the tree of life and being obligatory parasites, they rely on other creatures to reproduce. The virus attaches to a host cell, including bacterial, enters it, and releases its genetic material, DNA or RNA inside the cell. Viral genetic material can be DNA or RNA. All virus metabolisms stay inert until they infect their hosts. Viruses differ from each other in the growth properties in various human cell lines. It is reported that the growth rate differs for the same virus across different cell lines and some showing inert growth. Similar, viruses usually infect one particular type of cell. Common cold viruses infect only cells of the upper respiratory tract. It is this reason that researchers suspect that there are millions and millions of viral flora [11, 15]. One bioinformatics challenge is that there does not seem to be common signature.

At the host level, the severity of the symptoms of different people being inflicted with the common cold virus; this also undermines the different rates of effectiveness of vaccination. This supports the theory that every organism also appears to have its own set of viruses, including bacteria and drives the challenge of study of viruses. In addition to the cold virus, polio, herpes and human immuno virus (HIV) are famous examples of viruses negatively impacting human health [11]. Viruses usually infect one particular type of cell [15]. For example, common cold viruses infect only cells of the upper respiratory tract. Recent virus outbreaks include H1N1 and the bird viruses of South East Asia. It is largely unknown whether there can be beneficial coexisting properties since a virus must take control of a host to make more copies and the host cell dies. Because of the environment and epigenetic components of T1D, viruses have long been suspected of being an accomplice, possibly triggering autoimmunity where beta cells, producing insulin are mistaking, targeted and subsequently killed by the immune system. Figure 2 shows the major steps of the NGS T1D virome hypothesis undertaken by Cinek, Hyöty and their collaborators [5,8].

The Virus Pathogen Resource (<http://www.viprbrc.org>) is an excellent application to learn about genes and proteins of existing viral families and their targeted hosts. For example, many thousands of enterovirus genomes with their backgrounds are provided along with blastp results. The website provides online Basic Local Alignment Search Tool (BLAST) [30] search and also taxonomy browsing, along with other online applications. However for systematic purposes, it is not feasible to copy and paste some millions of sequence frames within this pathogen resource and moreover impossible to have an integrated and secured solution for custom data.

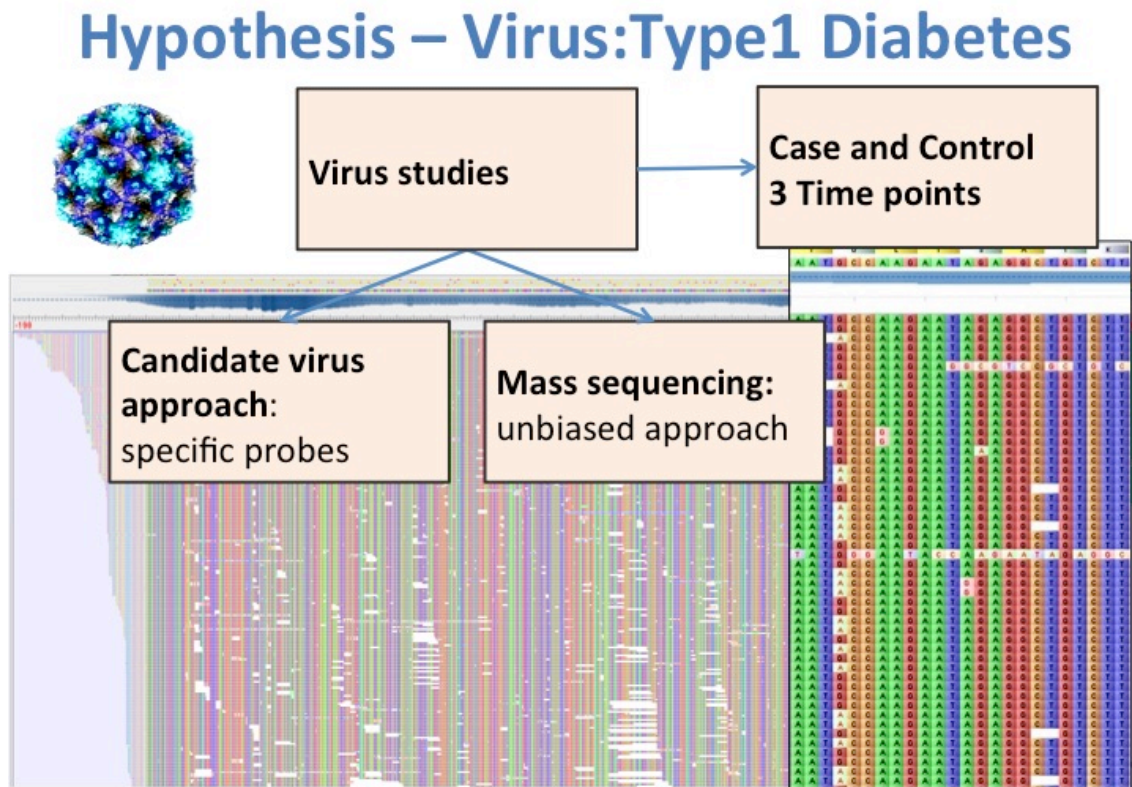


Figure 2. Researchers in The Diabetes Prediction and Prevention project are studying the diversity of virus in human gut using NGS and bioinformatics tools. This represents an unbiased approach for viral detection. Case and matching controls along with multiple time points allow comparative analysis.

2.2 Measuring Data

Data measurement instrumentation, particularly for life science needs to be sensitive and accurate. It is a high demand within molecular research as some of the measure targets are unknown. To compound the task, the research questions themselves are evolving. Advances in nanotechnology are allowing researchers measure expressions at nucleotide levels. Watson and Crick established the double helix structure of DNA in 1953 [31]. The intrinsic double helix strand and complementary bonding structure allows for self-replication. The canonical nucleotide bonding properties allows for specific microarray gene probes. The helix structure and canonical bonding properties are also the foundation basis for sequencing technologies. Discuss later, modern sequencing platforms offer researchers opportunities to pose novel and unaddressed hypotheses. Computational biologists are developing methods based on statistical learning to address accuracy and sensitivity. Interactive plotting libraries and web data exploration are often required to gain insights due to high data dimensions, density and heterogeneity.

2.2.1 Microarrays

Given the vast amounts of intrinsic genetic, regulatory and epigenetic measurements required for disease and basic biological research, many technologies have been developed and these technologies and more are constantly being refined and improved. The three most important attributes for scientific instruments are accuracy, cost and scalability. Microarray chips are sophisticated technology available for DNA, proteins and other biochemical measurements. These chips are capable of measuring many thousands of genes readings and can be custom and parallelized for specific studies. In brief, microarray chips are 2 dimensional array on a glass substrate studded with many DNA fragments. Each DNA fragment serves as a probe for specific targets. Recalled it was explained specific nucleotides have specific binding affinities. The captured complementary DNA is captured and labeled with fluorescent via the probe, upon washing; a laser is shined on the probes and the signals captured in indicating gene expression values [12]. Among one of the first technologies to be widely used by molecular and systems biology labs, large scaled genome wide association study (GWAS) and genotype to phenotype databases (<http://www.ncbi.nlm.nih.gov/gap>) have been developed for managing the results.

2.2.2 Sequencing Background

Though it can be scaled, it is the concept of sequencing that allows for reading the precise code of DNA, the blueprint of life. The first blueprint read was the bacteriophage *phi X 174* in 1977 by Fred Sanger and his team [32]. Also referred as chain termination sequencing, Sanger sequencing, is modeled after the natural DNA replication process and uses the enzyme DNA polymerase to access the separated helix strand, upon reading the nucleotide, the complement is determined. To put it simply, these experiments required many copies of the template DNA and their endings are added with modified fluorescent labeled nucleotides, the chain terminators. The fragmented copies vary in length are inserted into a plate and then onto the sequencing machine. Inside the machine, the copies are sorted by size and then transferred into thin glass capillaries where an electrical charge moves the DNA thru a laser capturing the fluorescent ends. The DNA fragment moves according to size, allowing the original DNA to be constructed [32]. Sanger was rewarded his second Nobel Prize in 1980 for this revolutionary invention. Though thru biochemistry and engineering breakthroughs the process is now capable of reading hundreds of thousands of bases per day; it would take years to sequence a human genome with some billions of nucleotides. At the cost of \$3 billion and 13 years, in 2003 the first human genome sequence [33] was published using advanced Sanger sequencer machines.

2.2.3 Next Generation Sequencing applications

The advanced Sanger sequencer machines used in the human genome project had 96 capillaries were accurate but limited in terms of parallelizing. Starting in 2005, several market leading companies released Next Generation Sequencing machines capable of massively parallelizing sequence capabilities. While it is not in the scope of this paper to discuss the technical differences, these machines were capable of producing gigabytes of DNA sequences and can sequence a human genome in weeks at the cost of fractions of the original human genome project. The parallelization is enabled because the DNA strand is randomly sheared into a short length (modern settings tend to be 36, 50, 120) and then usually read from both ends for quality reasons. This is called shotgun sequencing and common characteristics of these new sequencer platforms are that the output read lengths are shorter with lower accuracy. NGS machines also introduced pair end reads, meaning that the DNA bases in the fragments are read from both ends with some fixed window. The concern of shorter lengths and lower accuracy is resolved by high output, typically many gigabytes and that translates to higher coverage, defined as the number of times the particular nucleotides are read successfully. Higher coverage together with paired-end data allows for bigger piles of sequences, meaning longer consecutive DNA stretches, commonly called contigs and subsequent adequate quality consensus, final and correct ordering of the reads [9, 13].

Listed below are some of the common sequencing biological applications:

- Scan of genome wide variation
- Identification of protein binding sites (ChIP-seq)
- Quantitative analysis of transcriptome (RNA-seq)
- Genome wide methylation patterns
- Assembly of new genomes or transcriptomes
- Diversity and antigen analysis (metagenomics)

The advent of sequencing as a research tool has evolved where there are multiple sequencing service companies. Labs, free of high capital investment and infrastructure costs are able to ship their samples at the cost of around one thousand dollars per genome. RNASeq technology has all but replaced microarrays in measuring gene expressions as RNASeq is more sensitive and gives a more unbiased quantification of the transcriptome (genome wide RNA (gene) expressions) [34].

Although making sense of this blueprint of billions of letters requires different combination of analysis tools and vary dependent on the research hypothesis, the initial step in analysis translation of the machine laser readouts to DNA letters or protein amino acids along with a quality measure. This is sequencing platform specific and fortunately, the software required is bundled to the platform host computer and the DNA reads, usually available as a FASTA file together with summary statistical reports [9, 13].

2.2.4 Assembly

As mentioned above, modern NGS platforms are highly parallelized and the sample DNA is randomly broken into small pieces and then sequenced separately. A commonly used analogy is the cutting up of all the pages of a book into small pieces and then trying to put back the book together in the correct order. It can be a very complicated and computationally challenging task especially if there are many long repetitive regions. This process in bioinformatics is called assembly and the shorter the pieces, the higher coverage, the average number of times a base is read, it requires the number of times it requires. Assembly is fundamental as it takes the short reads and constructs the complete sequence in correct order. This implies reduction of repeats and is greatly influenced by the quality, or existence of a quality reference genome. *De novo* assembly is the challenging task term when the process is tried without a reference and often this is required in metagenomics. Metagenomics, discussed with more detail in chapter three, is the study of genetic material recovered directly from nature and hence the source organisms are unknown. Velvet [35] is the most cited tool for de novo assembly and pre-processing steps of quality inspection, filtering and format transformation are greatly aided by open sourced tools such as SAMtools [36] and FASTQC (www.bioinformatics.babraham.ac.uk/projects/fastqc).

2.2.5 Alignment, Aberrations and Annotations

Alignment, related to assembly is the process of arranging and comparing nucleotide or amino acids to identify similar regions across a set of sequences. Alignment can be global, or evolutionary where the compared sequences need to be similar or local in the cases where the input sequences are highly dissimilar. An example of global is placing a stretch of chromosome 1 of human and mouse reference DNA graphically on top of each other and the identical bases are the conserved mammalian regions. It is possible to think of assembly as aligning a unique genome against the reference and in many ways, that is how genetic aberrations are called. Single nucleotide polymorphisms (*SNPs*) are classified as single based differences. SNP variants largely are harmless if they occurred in noncoding regions or if they are synonymous, meaning that it does not impact the downstream protein. Protein variants from non-synonymous SNPs are responsible for genetic variations and they are useful as drug biomarkers; essentially these SNP impacted genes impact the way the drugs effectiveness and just as importantly the detriment side effects. SNPs can also be nonsense where the nucleotide change leads to an early stop codon resulting in a premature nonfunctioning protein. Other aberrations are *indels* (insertions and deletions) and chromosome structural aberrations that are responsible for copy number variations (*CNV*) meaning greater or less than the standard two copies of a particular gene [10, 11, 21]. Large scaled chromosome aberrations lead to abnormal development. For example Down syndrome Structural is identify by having an extra copy of chromosome 21 [20]. SNPs and indels can lead to gene fusion events

and a fusion event of *FGFR3-TACC3* [37] have been linked to glioblastoma, the most aggressive form of brain cancer.

Formally, annotation is the computational process of attaching biologically relevant info to genome sequence data. In practice and especially during the early stages of bioinformatics and human genome project, researchers were actively hunting for unknown genes and thereby looking for open reading frames described earlier [33]. Gene and SNP detections are structural annotation while functional annotation refers to attaching expression, interactions and regulations to genomic entities, essentially the dogma of genotype to phenotype. Genome annotation forms an active and challenging component of bioinformatics and recently, the worldwide consortium Encyclopedia of DNA Elements (ENCODE) [38] project released a comprehensive characterization of human genomic annotation database.

2.2.6 Genomic Visualization

Genome browsers are essential for visual inspections of sequencing data. There have been many advances in this class of software programs and many of these highly performing programs allowed for custom loading of genomic tracks and also support of different formats, including annotations. Different human referential annotations, including the ENCODE [38] project and also model organisms such as fly, mouse and yeast can be uploaded or configured. GBrowse [39] is particularly popular for viewing model organism sequence and annotations. Figure 3 below shows some of the other most popular genomic viewers, from the top: the UCSC Genomic browser (A), Circos (B), Broad IGV (C), Tablet (D), and the last is a custom visualization of human chromosome cytobands (E). Genomic visualizations, browsers as well as custom interactive plots, are important building blocks for the process of genomic insight discovery and exploration.

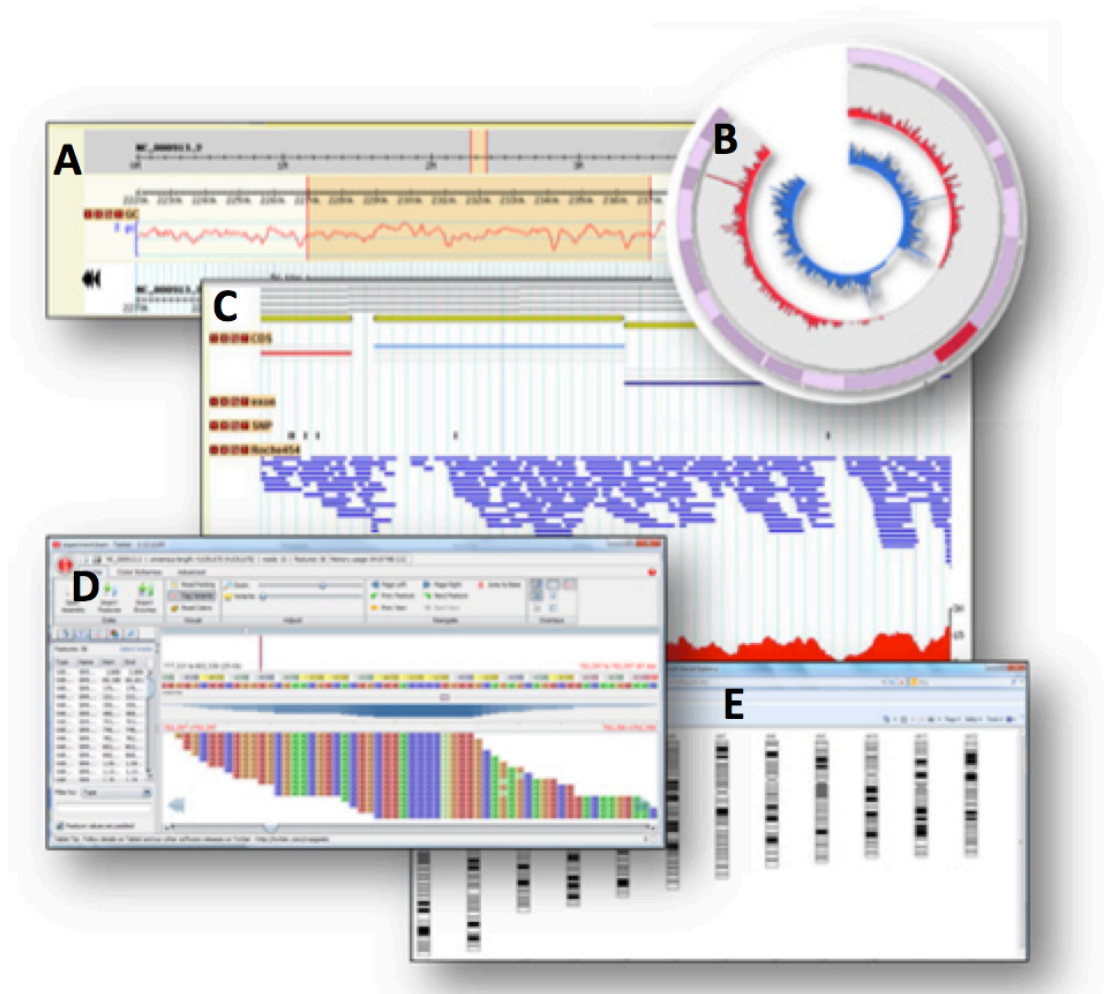


Figure 3. Genomic visualizations, browsers as well as custom interactive plots, are important building blocks for the process of genomic insight discovery and exploration.

2.3 Information Technology for Life Science

2.3.1 Internet Transformation

Computers and the Internet are playing important roles in all facets of human life and their impact on disease and basic biology research is very strong and central. At the core, biology is an information science and the computer enable efficient management and processing of bits. Beyond enabling efficient collaboration with emails and data sharing, the web is an unbiased medium for creating and distributing information. Benefiting from research advances from Internet consortiums like Web Consortium (W3C) and leading companies, browser based applications are installation free and offers dy-

dynamic and interactive interfaces to open sourced frameworks and advanced scalable databases. Particularly, W3C introduced new HTML5 and JavaScript that allows for seamless server calls and interactive dynamic plotting of the results.

2.3.2 Relational databases

Businesses, universities and nonprofit research organizations across industries and sizes store their data on database software. The dominant database technology implements a relational model where it consists of a schema that defines tables and columns. A relation is a set of tuples; commonly call rows or records within a particular table. To give uniqueness to the records unique keys are required. Foreign key columns can be defined to construct relationships across table entities, programmatically called a *join*. Reads and updates are called transactions. The primary benefits of relational database are the fulfillment of atomicity, consistency, isolation and durability (*ACID*). The heart of this promise is atomicity as it states that transactions must all complete or else fail and return all records back to unmodified state. Consistency is that the system produces the same results across platforms. Isolation implies that transactions are executed independent of other transactions, ensuring atomicity. Finally, durability guarantees that the data is safe and recoverable from hardware failures. ACID and relational schemas are powerful concepts and they imply that system changes and enhancements require a schema change. Therefore non-relational databases are becoming more popular. Often called No-SQL or Schema-Free, the data is store as key-value entities and in some cases as graph node and edge entities. In addition to schema free, another hope is to overcome the memory expensive tabular joins. No-SQL database is new and actively being research as it promises more scalability and robustness. It is beyond the scope of this thesis to debate their usability and ACID principles.

There are different technical relational database implementations and the commercial market leaders are Oracle, Microsoft and IBM. These companies offer software bundled database implementations and these products are sold and further customized to fit the needs of companies and organizations. Research groups often use free open sourced database technologies like MySQL (<https://www.mysql.com>) and PostGRES (<http://www.postgresql.org>). Data access and search are accomplished programmatically using a structure query language (SQL). Life science researchers search different database instances offered by large consortiums such as European Molecular Biology Laboratory (EMBL) and National Centre for Bioinformatics (NCBI) based in the United States. Gene Card from the Weizmann Institute in Israel is a comprehensive database of genes and proteins including information from EMBL and NCBI. These database resources are essential web applications and their implementations are usually independent of database technologies. More details are presented in the web architecture section of chapter 3.

2.3.3 Cloud Computing

The rise of the Internet disrupted the desktop and older standalone server mode of computing. Internet infrastructure employs the browser as the client and then a remote server. The main benefit is that applications do not need to be installed. Network programming was extended to load balancing meaning server and database farms duplicated across geographical sites to expedite transactions and load times. Taking network computer one-step further and at its heart, cloud computing allows for installation and infrastructure free software and development. They are all services, like utilities. Microsoft Office users can access Word directly and programmers can develop, test, and deploy applications onto Google App Engine or Amazon EC2 without buying and maintaining servers. There are some negatives to consider such as network availability, security and variable costs. Cloud computing in a sense is like renting, and just like renting the cost in the long run is higher than ownership [40]. There are excellent national and public computing infrastructures, for example Viral NGS analysis benefits from Finland's Computer Service Centre (CSC), available at <https://www.csc.fi/services-for-research-education-and-information-management>. CSC is highly communal and offers free cloud computing service and advice for nonprofit university research.

2.3.4 Computational Biology

Computational biology is a multidisciplinary field bringing together researchers and engineers from applied math, computer science and machine learning focusing and extracting meaningful relationships and insights from the abundance of biological measurements. This focus touches on development and application of statistical learning methods and models as *in silico* experiments are usually less costly and involved. Computational biology helps with gaining insights on data and this includes data and annotation integration. Dynamic querying and interactive visualizations allow experts to detect insights from large scaled heterogeneous data. Listed here are other possible roles of computer biology though often the responsibilities are shared:

- Building genomic and protein data bank databases and collaboration platforms.
- Analysis methods on assembling and searching gene sequences and alterations
- Protein structure and folding prediction and simulation.
- Imaging processing to include classification of organ and tissue from X-Rays and Magnetic Resonance Imaging (MRI) and morphology detection of cells and specific organelles such as mitochondria
- Cancer biomarker validation and detection along with drug repurposing.

2.3.5 Statistical Learning

Central to computational biology and NGS analysis tools are the application of statistical learning where its methods aim to address data prediction, inference and classification. Generally, these tools are divided into classes of *supervised* where the labeled inputs are defined and then the outputs are observed and assumedly caused by the inputs. *Unsupervised* methods assume that the output, or shape and distribution, is caused by hidden variables [39]. The field of statistical learning is broad and complex and though beyond the scope of this software thesis, it is important to discuss the main themes as biological instrumentation capturing, particularly heavily parallelized and Nano scaled measurements, has advanced greatly but it is not perfect and their output often includes missing data gaps and outliers. Linear regression, a classical supervised method is the most applied method aimed at prediction of missing data [41]. Data trimming or pruning is not trivial in biological research as sometimes the outliers carry significance. For example, Illumina Sequencing platforms produce outputs where the reads have a standardized quality score based on the *Phred* algorithm that reports the probability that the nucleotide base call is wrong [42]. The quality score, Q Score, is logged and ranges from Q10, Q20, ... and Q10 implies 90% accurate, Q20 being 99% and Q30 is 99.9%. Even though it is reported that modern sequencing platforms mostly produce Q20/99% accuracy calls, the 1% or even .1% error rate over many billions of bases can be significant and needs to be considered.

Statistical learning inference focuses on answering how an output, Y is affected as a set of inputs X_1, X_2, \dots, X_p changes. This includes finding the most important input variables since usually just a few of the input variables are responsible for the behaviors of Y and the application of Principal component analysis (PCA) is well established and practiced. Another focus is finding the general relationship of whether Y increases with increase to particular X variables relationship, and given this context, whether the relationship can be described as linear with one variable, one can think of fitting a straight line to the data or as usually in the case of biological data, more complex and consisting of multiple variables [41]. Gene expression studies will have more than twenty thousand measurements. Confounding cancer researchers are chromosome aberrations and nucleic base level methylation events; a phosphorylation measurement associated with cancer and this leads to millions and millions of variables against much smaller count of patients [13].

Given that biological measurements can be many millions of observations scored with statistical quality scores, there is high motivation for computational biologists to learn and apply cluster analysis, or clustering. These methods, supervised, unsupervised and also hybrids are essentially data mining applications and seek to place sets of observations into subgroups. As grouping of possible overlapping observations is relative and context sensitive, there is no absolute correct method and there are many types of clustering algorithms. The two best-known types are K-means and hierarchical clustering.

K-means clustering is iterative and seeks to partition the observation into a pre-defined K number of subgroups. Hierarchical clustering implies that the number of subgroups is unknown, unsupervised and results in a dendrogram, an attractive tree like visual representation. The branches and resultant nodes are split based a measure, usually Euclidean or correlation distances [41]. More details are discussed in the next chapter; it is a benefit to the bioinformatics community that there are active research into optimizing inference and clustering methods and that these implementations are available and supported in a variety of programming languages, including Matlab, R and Python.

2.3.6 Bioinformatics Platform

Given the importance of computation needs for data analysis, results management and integration for biological research, there have been valiant efforts at creating software platforms and ecosystems in fulfilling these requirements. Within bioinformatics, Taverna (<http://www.taverna.org.uk>) and Galaxy [43] are two popular solutions for managing workflows and managing scripts. A computation workflow can be thought of as an ordered set of steps or methods required for some complex scientific analyses. Taverna is widely adopted and excels at managing workflows. Although desktop based there is development to provide an online version, additionally, its status is influx, as it recently became an Apache incubator project. Impressively comprehensive, Galaxy is web based and offers free online computation with a wide list of tools. With a graphical workflow system, it can be used on the web or as a local instance. The problems with using a shared web system for NGS analysis is the need to upload gigabytes of raw files and of course exposing sensitive and unpublished data. Although Galaxy project has good documentation with setting up a local instance for genomics, it would have required drastic enhancements to integrate the annotations data and upgrade the interface. With regard to this thesis project focusing on specific analyses and visualization their results, it seems more appropriate to start from anew than extending Galaxy with its existing dependencies.

3. VIROME NGS ANALYSIS AND WEB INTEGRATION METHODS

DNA outputs from modern sequencing platforms result in hundreds of gigabytes. Computational biologists working closely with experimentalists must leverage different analysis tools and learning methods to glean insights into their studies. Often data annotation, interactive interfaces and data mining applications, such as clustering, are necessary for experimentalists to explore their data to gain insights from these large result sets.

3.1 Metagenomics

Along with sequencing as a service is becoming a viable business, sequencing instrument companies are also shipping more affordable mini/express machines in supporting different lab objectives, benefiting particularly metagenomics research, defined as the study of genetic material recovered directly from the environment. Metagenomics is a new and growing field driven by system biology principle concept of understanding the whole requires holistically studying all its components [44]. Refer as the human microbiota, the aggregate of microorganisms includes bacterial and yeast make up for a sizable part of the human body. It is estimated that their cells outnumbered the human 10 to 1 and that there are 10,000 different types of organisms typically found in a healthy individual [45]. As some bacterial and viral genetic materials cannot be cultured, that is grown outside of their natural environment; scientific experimentation and measurements must be directly taken from the acquired samples. This implies that sequencing instrumentation technologies must not only be accurate but very sensitive due to the extreme small genome sizes and in the context of composite fecal stool. This is an ongoing and variable challenge and discuss later is the lab clinical validation as well as computation processing. Recently submitted, this DIPP NGS T1D viral study of analyzing stool samples from stool content represents a novel approach of applying metagenomics methodology in directed to human disease research [8].

3.2 Experiment and NGS Preparation

Briefly this NGS viral study used stool samples from selected Finnish children enrolled in the DIPP project. The DIPP project screened and enrolled at risked Finnish families with diabetic history. Stool samples were taken at estimated 3, 6, and 9 months before initiation of autoimmunity, where the immune system targets and destroys insulin producing beta cells. As viral references are unknown and one of the goals of the study is

to gain insight into viral in the human gut genome, random DNA amplification was performed according specifications and library preparation was preprocessed using Illumina Preparation Kit in batches of 24 samples per run, performed on Illumina MiSeq platform. Further details of the DIPP study protocol are previously discussed [5, 7].

3.3 Analysis Process

To begin with, the Illumina processing pipeline outputs paired-end FASTA files containing raw DNA calls along with quality information. The components of the analysis process are presented in figure 1. Filtering is done using FASTQC courtesy of Babraham Bioinformatics (www.bioinformatics.babraham.ac.uk/projects/fastqc). Because of viral genome stipulations, the next steps require *de novo* assembly and then re-alignment for contig detection describe below. It should be said that in non-metagenomics sequencing analysis involving a suitable reference, such as human or mouse model, the search for functional variants, aberrations and other genomic annotation calls is an important task.

3.3.1 De Novo Assembly

Modern sequencing platforms are highly parallelized and random, genomic assembly is required and refers to process of taking many small random reads, stored in FASTA files and reconstructing them together in correct order for the original genome. It can be a complicated process because of nucleotide repeats and overlap nucleotide frames in the result sets. Nucleotide repeats, classified as identical long terminal, tandem and satellite repeats are a big problem as it has been estimated that more than two-thirds of the human genome contain repetitive elements [46]. While viral and bacterial genomes are much smaller and analysis programs do not need to account for long repeats, the massive diversity and unknown references drive the need for *de novo* assembly, the process of ordering without a reference.

The gold standard for *de novo* short read assembly is usage of the published method Velvet [35]. Based on De Bruijn directed graphs, Velvet efficiently constructs and uses compression to efficiently manipulate through these unique vertices, or nodes. Usage of Velvet requires defining a hash length, commonly known as k-mer in De Bruijn graph, and each unique k-mer sequence is considered a node with an unique identification. Nodes A and B are connected if A's first k-1 characters, in genomics nucleotides or amino acids, are the same as B's last k-1 characters. Figure 4 is a cartoon adapted from [47] and it describes k of 4 and the subsequent vertices and their nodes. It is important to point out the string GAC and ACT are repeated three times and that using compression and unique representation, magnitudes of memory are saved.

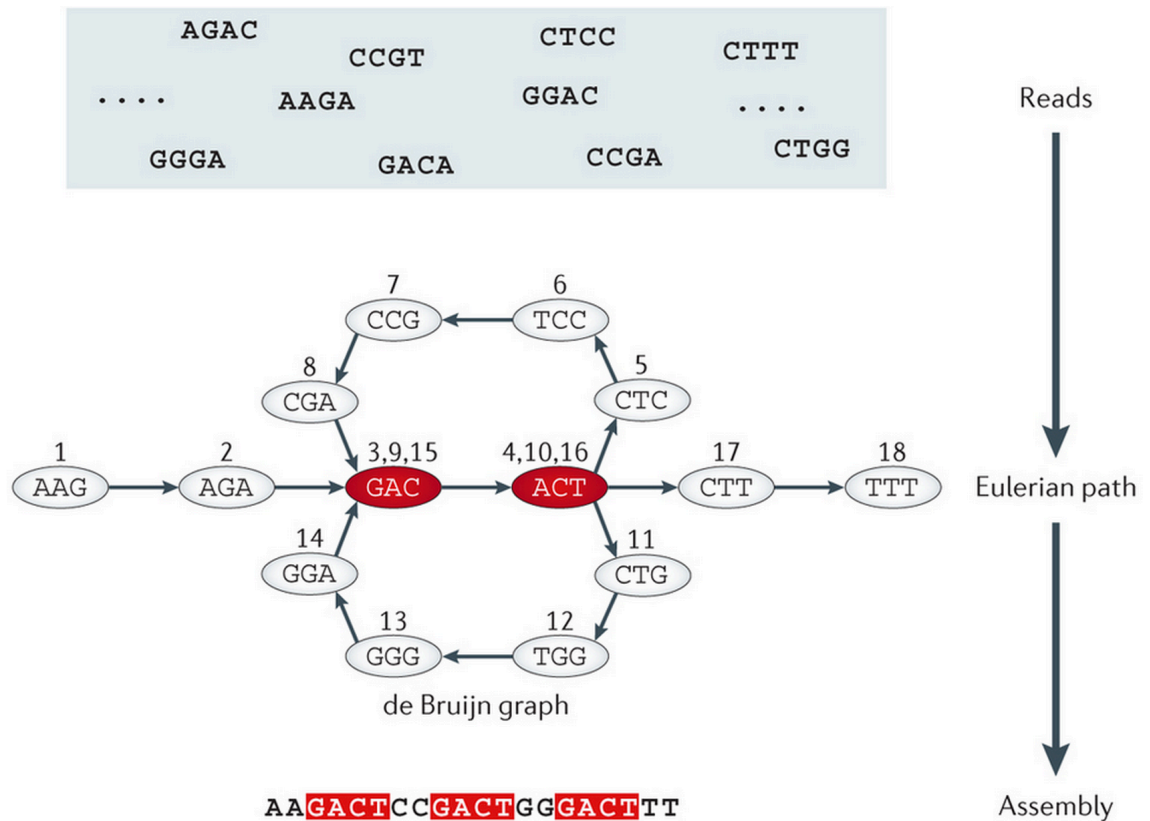


Figure 4. Velvet efficiently implements *de novo* assembly function based on construction of unique de Bruijn nodes and directed edges [47].

Coverage quality is stored as the number of times a vertex is repeated. At the practical level, there are two steps involved in velvet processing, first *velveth*, with *K* of 57 is called and this constructs the hash or *K*-mer nodes. Then the outputs of *velveth* are passed into *velvetg*. The key parameters for *velvetg* are *minimum contig length* of 150, with *insertSize* of 150 and *coverage* of 50. Both commands are repeated for paired and unpaired reads, majority reads are unpaired. Different parameters have been tried but the raw results stem from the given parameters. It is worth noting the unused velvet reads are saved that there is a plan to rerun and perform data comparisons. Appendix B outlines the properties of De Bruijn graphs and also has the detail *velveth* and *velvetg* invocation commands. The contig files produced by Velvet can be visualized in Tablet, as demonstrated in figure 3.

3.3.2 BLAST commands

The method Basic Local Alignment Search Tool (BLAST) [30] takes a string of nucleotides or protein sequences and compares the inputs to many compiled databases references. The full lists of specific BLAST commands, adapted from NCBI Blast page (<http://blast.ncbi.nlm.nih.gov/>) along with brief descriptions are listed in table 2.

BLAST commands, well established and supported by NCBI, have many parameters and they are comprehensively covered on the online manual. Statistical score of an *e value* or expectation, is included in the results and the lower the expectation score, the more significant the alignment. Investigators use BLAST, online on NCBI servers or as for this thesis a local installation with custom databases, in determining functional and evolutionary relationships [30].

Table 2. Adapted from BLAST manual, popular commands are listed. Some of these commands are computationally expensive and can take hours to complete. Local instances are required for performance and programmatic interfacing.

Program type	Command name	Description
Protein	blastp	Traditional BLASTP to compare a protein query to a protein database
Protein	blastp-short	BLASTP optimized for queries shorter than 30 residues
DNA	blastn	Traditional BLASTN requiring an exact match of 11
DNA	blastn-short	BLASTN program optimized for sequences shorter than 50 bases
DNA	megablast	Traditional megablast used to find very similar (e.g., intraspecies or closely related species) sequences
DNA	dc-megablast	Discontiguous megablast used to find more distant (e.g., interspecies) sequence
DNA	tblastn	Identifies db sequences encoding proteins similar to input
DNA	tblastx	Identifies db DNA sequences similar to the input coding potential

3.3.3 Viral identification with BLAST

The next step aims to determine what viruses reside in the stool samples. A local instance of BLAST is installed along with downloadable nt (DNA/RNA), nr (protein) and

16s (bacterial) databases. In addition, a custom *picornaviridae* viral family database is built from sequences downloaded from <http://www.ncbi.nlm.nih.gov/taxonomy> using taxonomy id 12058. Vertebrates are natural host of this family of RNA virus and listed are 50 species, including human diseases related *enterovirus* and *rhinovirus* species [11, 48]. The setup script is available in the util folder of source repository. It is noted that tblastx commands can take two hours and a powerful and/server or access to cloud computation are required. The other requirements for setting up a local BLAST instance are programmatic interface and also database integration.

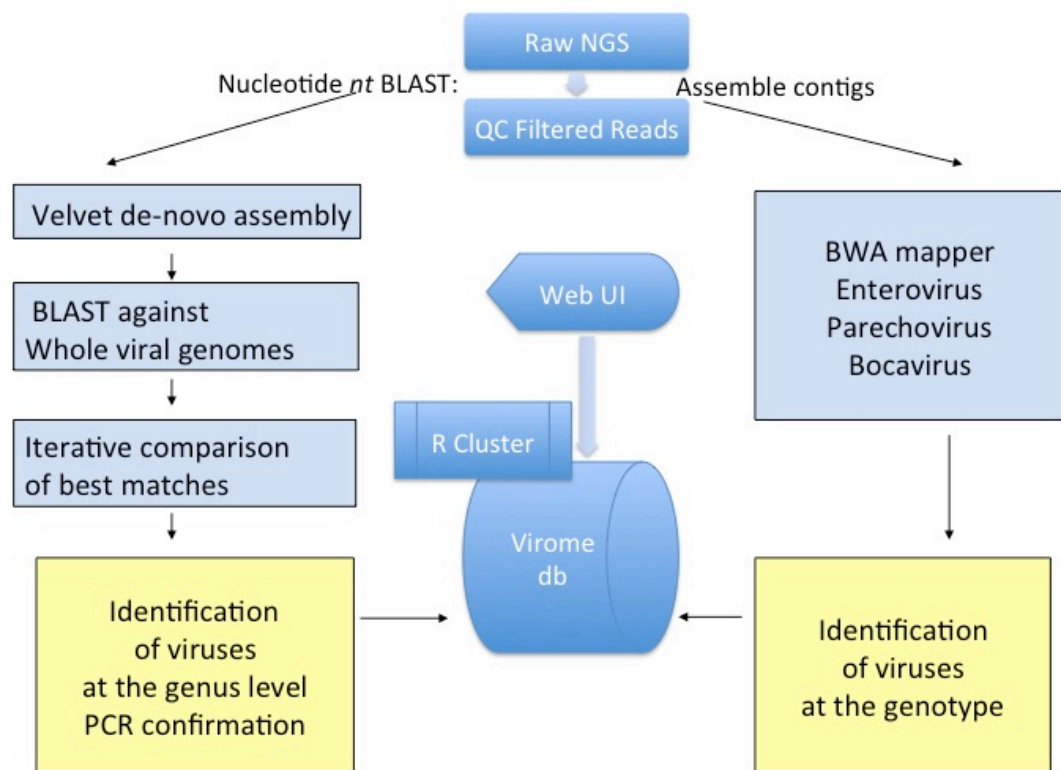


Figure 5. Pipeline workflow showing quality filtering and then assembly and alignment tasks, modified from [8].

3.3.4 Strain identification with BWA

The list of matches from the BLAST output along with the contig sequences are passed into Burrows-Wheeler Aligner (BWA) [49] tool. BWA is well cited and is an efficient program for aligning short reads to a reference or set of references. In this case, the BWA mapper function processes on the custom viral database population set. BWA remapping processing adds additional confidence as well as a list of matched viral ref-

erence sequences. This list of references allow for further investigation into evolving species and also determining polymorphisms. BWA aligner provides several variants and bwa-sw, describe in its manual, is selected as this option provides more error checking and though a bit slower than the standard operation, bwa-sw is more suitable for longer queries. The results include a standard sequence alignment mapped (SAM) file that is supported by most genomic browsers. Figure 5 above depicts the BLAST and BWA steps and then the storing of their results into a database accessible by the web interface.

3.3.5 Islet Autoimmunity Protein Assessment - Translation

As described earlier, a large majority (74%) of DNA sequence reads did not mapped to any known organisms. The likely reason is that viral populations are large and diverse. While some references are published and accessible, it is well known that virus adapt rapidly and have a relatively high mutation rate. It is estimated that RNA virus have more than one point mutation per genome per round of replication. As discussed by Domingo [50] in the Journal of Virology, RNA viral populations have been measured with much higher mutation rates and the population consists of extremely complex and dynamic mutant genomes. Hence, it is entirely possible that a consensus sequence cannot be constructed.

Because of the possible lack of consensus reference and the motivation to account for all the sequences to assess viral molecular mimicry of key diabetes genes (recall that one proposed method of virus taking control of a host cell is by mimicking the host protein structure to disrupt immune system, similar to Human Immune Virus mechanisms), an initiative was started to translate all the high coverage contigs, prior to BLAST and mapping, into amino acids. Stated earlier, proteins account for all the essential structure and function of an organism. Both Watson (forward) and Crick (backward) strands [31] were included in the translation. In addition, frame shifting was done, as shown in Program 1 seq2AA function (lines 13 to 27) implying that for each input DNA contig, 6 protein amino acid sequences will be produce and store. The coverage and run identifier annotations were captured. As it can be seen on line 3, the script imports SeqIO class from Biopython, the free standard bioinformatics BioPython library available at http://biopython.org/wiki/Main_Page.

```

1 import sys
2 import os
3 from Bio import SeqIO
4 bases = ['t', 'c', 'a', 'g']
5 amino_acids =
    'FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG'
6 codons = [a+b+c for a in bases for b in bases for c in bases]
7 codon_table = dict(zip(codons, amino_acids))
8 def readFasta(fain):
9     handle = open(fain, "r")
10    records = list(SeqIO.parse(handle, "fasta"))
11    handle.close()
12    return records
13 def seq2AA(seq):
14    ss0 = ""
15    ss1 = ""
16    ss2 = ""
17    for r in range(0, len(seq)):
18        if (r%3 == 0 and r > 0):
19            seq3a = seq[r-3:r]
20            ss0 = ss0 + codon_table[seq3a.lower()]
21        if (r%3 == 1 and r > 1):
22            seq3b = seq[r-3:r]
23            ss1 = ss1 + codon_table[seq3b.lower()]
24        if (r%3 == 2 and r > 2):
25            seq3c = seq[r-3:r]
26            ss2 = ss2 + codon_table[seq3c.lower()]
27    return (ss0, ss1, ss2)
28 if __name__ == "__main__":
29     label = sys.argv[1]
30     for fastafile in sys.argv[2:]:
31         records = readFasta(fastafile)
32         infastafile = fastafile.split("/")[-1]
33         if (not os.direxists('./aa_trans')):
34             os.makedirs('aa_trans')
35         outfa = open("./aa_trans/" + label + "_" + infasta-
file.split(".")[0] + "_0.aa", "w")
36         outfb = open("./aa_trans/" + label + "_" + infasta-
file.split(".")[0] + "_1.aa", "w")
37         outfc = open("./aa_trans/" + label + "_" + infasta-
file.split(".")[0] + "_2.aa", "w")
38         idoutfa = open("./aa_trans/" + label + "_" + infasta-
file.split(".")[0] + ".id", "w")
39         for ri in range(0, len(records)):
40             aarec = seq2AA(records[ri].seq)
41             outfa.write(aarec[0] + "\n")
42             outfb.write(aarec[1] + "\n")
43             outfc.write(aarec[2] + "\n")
44             idoutfa.write(records[ri].id + "\n")
45         outfa.close()
46         outfb.close()
47         outfc.close()
48         idoutfa.close()

```

Program 1: Raw nucleotide bases in a FASTA format are translated to amino acid for downstream protein structure similarity analysis. Forward and backward strands with shifting frames are processed.

3.3.6 Islet Autoimmunity Protein Assessment - Matching

The second part of the protein assessment involves matching the translated amino acids against key diabetes genes. The first obvious candidate is insulin (*INS*) and with help from DIPP investigators Dr. Cinek and Professor Hyöty, the other genes used in this assessment were glutamate decarboxylase (*GAD65*), tyrosine-protein phosphatase *IA-2* (3 isoforms) and *ZNT-8* (2 isoforms). The respected protein sequences were downloaded directly from NCBI protein database and listed in Appendix D. Upon checking literature and discussion with DIPP investigators, an amino acid frame window of seven was selected. Essentially 21 DNA bases, the motivation behind this selection was that the window size needed to be long enough to be meaningful cutoff but also inclusive to find as many possible matches. In striving for maximum flexibility and robustness, missed match functionality was implemented [17, 18]. Lines 8 to 17 of Program 2 below implements a function named Hamming two strings are compared with indexing, this allows for error reporting and also customization of number of missed matches allowed; essentially Hamming distance [51] between two strings. It is optimized to return once the error bits are greater than the missed matches allowed. The script also utilizes a Levenshtein distance python package available at <https://pypi.python.org/pypi/python-Levenshtein/>. Hamming distance used in telecommunication signal error corrections is implemented in program 2 lines 8-17. The distance plainly describes the number of characters that do not match between two equal length strings.

Levenshtein distance, also known as Edit Distance, defined as the minimum number of edits to change one word, or sequence, into another [52]. Levenshtein distance is more general than Hamming distance as it describes Hamming distance but it extends to comparisons of strings of different lengths. For instance, strings TCA and TCGA have a Levenshtein distance of 1 since TCA needs to insert a G to form TCGA. Highly optimized, it has been shown with benchmarking that this Levenshtein python library performs distance computing with the same accuracy as Hamming function while executing twice the speed. The matched results are stored as tab separated files while preserving id and quality measures.

```

1  from Bio import SeqIO
2  import sys
3  import argparse
4  import time
5  import util
6  import Levenshtein
7
8  def Hamming(genome_frame, sequence_frame, missedmatches = 1):
9      errors = 0
10     refaa = ""
11     for i in range(0,len(sequence_frame)):
12         if genome_frame[i] != sequence_frame[i]:
13             errors += 1
14         if errors >= missedmatches:
15             return errors

```

```

16         refaa = refaa + genome_frame[i]
17     return (errors, refaa, i)
18 if __name__ == "__main__":
19     parser = argparse.ArgumentParser(description='Process AA fasta files
20     to RefGene AA reference.')
21     parser.add_argument('--binx', type=str, default='bin83', help='bin
22     numbers such as bin83, multiple bins use comma, such as bin42,bin83')
23     parser.add_argument('--fs', type=int, default=7, help='AAFrameSize
24     default 7')
25     parser.add_argument('--mma', type=int, default=2,
26     help='MissedMatchesAllow default 2')
27     parser.add_argument('--reffa', type=str, de-
28     fault='./insulin_p1308.fasta', help='Fasta ref file containing protein se-
29     quence')
30     parser.add_argument('--joblabel', type=str, default="", help='Examples
31     are insulin, HLA, GAD65, ...')
32     parser.add_argument('--entropy', type=float, default=1.2,
33     help='Entropy of frame seq, on length of 8 repeats, 0, ... two distinct
34     AAs ~ default of 1.2')
35     parser.add_argument('--outdir', type=str, default="./out", help='./out
36     relative default out')
37     parser.add_argument('--test', type=int, default=0, help='0 implies
38     true, requires ./aa_test ./out_test inputs/outputs')

39     args = parser.parse_args()
40     print("\nStart " + time.strftime("%Y%m%d %H:%M:%S") + " " + args.reffa
41     + " " + args.joblabel)
42     readids = []
43     bin = args.binx
44     ridf = open("./aa_trans/" + bin + '_all_reads.id', 'r')
45     for l in ridf.readlines():
46         readids.append(l.strip())
47     ridf.close()
48     faref = readFasta(args.reffa)
49     allowmisses = args.mma + 1
50     framewin = args.fs
51     outdir = args.outdir
52     aa_dir = "aa_trans"
53
54     for fi in range(0, len(faref)):
55         aaref = faref[fi].seq
56         bin83_r0 = open("./" + aa_dir + "/" + bin + "_all_reads_0.aa", "r")
57         bin83_r1 = open("./" + aa_dir + "/" + bin + "_all_reads_1.aa", "r")
58         bin83_r2 = open("./" + aa_dir + "/" + bin + "_all_reads_2.aa", "r")
59         bin83_r0r = open("./" + aa_dir + "/" + bin + "_all_reads_r0.aa",
60         "r")
61         bin83_r1r = open("./" + aa_dir + "/" + bin + "_all_reads_r1.aa",
62         "r")
63         bin83_r2r = open("./" + aa_dir + "/" + bin + "_all_reads_r2.aa",
64         "r")
65         afiles =
66         (bin83_r0,bin83_r1,bin83_r2,bin83_r0r,bin83_r1r,bin83_r2r)
67         fout = open("./" + outdir + "/" + bin + "_" + args.joblabel
68         + ".matched", "w")
69         header_added = 0
70         processed_frame = {}
71         for f in afiles:
72             if (header_added == 0):

```

```

56             header_added = 1
57             fout.write("# " + faref[fi].id + " " + f.name + "\n" +
"id\taa_ngsframe\tmlmismatches\taa_ref\tmlmis_pos\tmlref_pos\n")
58             idc = 0
59             his_frame = {}
60             for l in f.readlines():
61                 aa_frame = l.strip()
62                 for myaa in range(0, len(aa_frame)):
63                     if (myaa >= framewin):
64                         _frame = aa_frame[myaa-framewin:myaa]
65                         if ((allowmisses) == 0 and processed_frame.get(_frame) != 1):
66                             processed_frame[_frame] = 1
67                             pp = Hamming_perfect(aaref, _frame)
68                             if (pp != -1):
69                                 fout.write(str(idc) + ":" +
f.name.split("_")[-1] + "\t" + _frame + "\t0" + "\t" + str(pp) +
"\tna\n")
70                                 elif (processed_frame.get(_frame) != 1):
71                                     processed_frame[_frame] = 1
72                                     for aai in range(0, len(aaref)):
73                                         if (aai >= framewin):# and aai%framewin ==
0):
74                                             aaref_frame = aaref[aai-framewin:aai]
75                                             if (method == "string_index"):
76                                                 if (hits_frame.get(_frame) == 1):
77                                                     continue
78                                                     rc = Hamming_1(aaref_frame,
_frame, hammisses)
79                                                 if (isinstance(rc, tuple) == True
and rc[0] <= allowmisses):
80                                                     fout.write(str(idc) + ":" +
f.name.split("_")[-1] + "\t" + _frame + "\t" + str(rc[0]) + "\t" +
str(aaref_frame) + "\t" + str(aai) + "\t" + str(rc[1]) + "\n")
81                                                     hits_frame[_frame] = 1
82                                                 elif (method == "levenshtein"):
83                                                     if (hits_frame.get(_frame) == 1):
84                                                         continue
85                                                         ld = Le-
venshtein.distance(str(aaref_frame), _frame)
86                                                         if (ld <= allowmisses):
87                                                             hits_frame[_frame] = 1
88                                                             fout.write(str(idc) + ":" +
f.name.split("_")[-1] + "\t" + _frame + "\t" + str(ld) + "\t" +
str(aaref_frame) + "\t" + str(aai) + "\n")
89                                                         idc = idc + 1
90             print(f.name + " finished " + args.joblabel + " file " +
time.strftime("%Y%m%d %H:%M:%S"))
91             f.close()
92             fout.close()
93             print("Completed " + bin + time.strftime("%Y%m%d %H:%M:%S"))

```

Program 2: Using string distance metrics, frames of translated NGS are compared with referential sequences. The program is suitable for parallel computing.

3.3.7 Islet Autoimmunity Protein Align and View

The outputs of the program 2, protein matching are further integrated with case/control and time point annotations and assembled with sample identifiers for appropriate visualization. The assemble python script is available within the util folder of project source repository. In addition to computing matched density distribution across case and control samples, this utility orders the matched protein outputs and produces data structure tracks suitable for web based sequence alignment viewer described in Table 3.

3.4 Web Programming for Data Exploration and Management

3.4.1 LAMP architecture

The virome NGS web app architecture consists of user interface built using HTML5 and modern JavaScript libraries with a Linux server running components of Apache and SQLite instances. Essentially a modified Linux Apache MySQL Python (LAMP) architecture favored by open source and research development, the SQLite instance can be replace with MySQL without too much complications as the their schemas and SQL syntax are similar. SQLite is ACID compliant and has the benefit of being lighter and easier for configuration. Figure 5 shows a cartoon of the browser-based client invoking HTTP request calls over the network. Also shown in the figure are Python common gateway interface (CGI) scripts handling request calls and in turn these scripts return Java Script Object Notation (JSON) objects as responses. Modern versions of python are packaged with SQLite bindings and this are used in scripts are also used for schema building, data transformation and loading of the analysis results. These scripts are available via source forge repository.

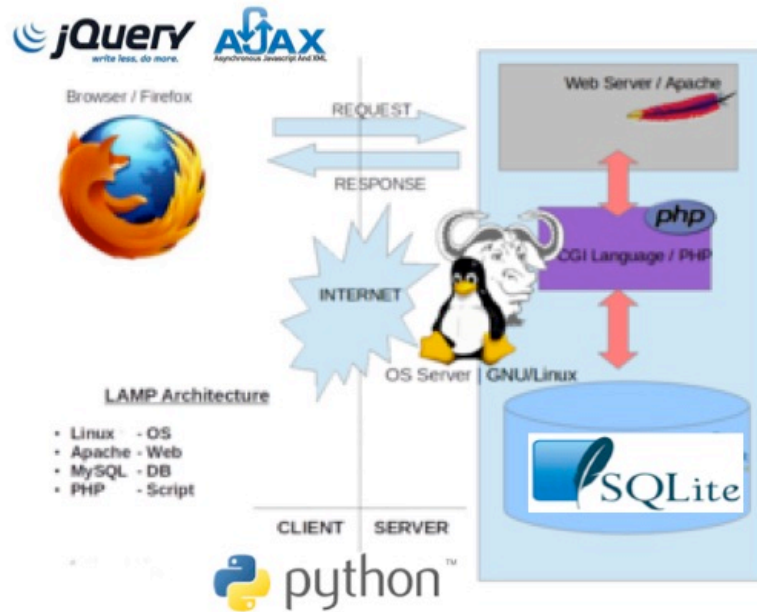


Figure 6. Viral NGS web application adheres to standard client server programming architecture. Web based user interface invokes an AJAX HTTP call to CGI handler on the server, and the CGI script hits the database and return the results via HTTP. jQUERY and HTML5 process the data and dynamically reloads the page components. Program 3 `initBase()` below has the implementation details.

3.4.2 Interactivity with JavaScript

In addition to LAMP components, Figure 6 also shows AJAX and jQuery trademarks. jQuery, freely available at <https://jquery.com>, is a fast and comprehensive JavaScript framework library for data manipulation and event handling. It is widely used and many other libraries are built on top of jQuery. Table 3 below lists the other JavaScript libraries required by this web application.

Table 3. Web 2.0 programs rely on JavaScript for DOM manipulation and event management. Below is the full list of libraries required by the NGS Virome web application for interactivity and visualizations.

Library name	Source	Description
jQuery	https://jquery.com	Framework required by other libraries. Used for AJAX calls and DOM manipulation and events
jQuery-ui	https://jqueryui.com	Tab interface and selection
bootstrap	http://getbootstrap.com	Type Ahead and multi select dropdowns
DataTables	https://www.datatables.net	Dynamic tables with sort and refine functions
Highcharts	http://www.highcharts.com	Exportable heat maps
MessiJS	https://github.com/MessiJS/MessiJS	Progress messages and user dialogs
Sequence Alignment Viewer	https://github.com/AndrewCRMartin/JSAV	Adapted for visual data mining of matched amino acids
simple-statistics	http://simplestatistics.org	Percentile operations for advanced search

Asynchronous Javascript and XML (AJAX) defines a set of technologies to send and retrieve data via the network without disturbing the parent page. The signature technologies behind Web 2.0 application wave in the early 2000s are AJAX and HTML and Cascade Style Sheets (CSS) advances. The key concept underlying AJAX is upon a remote network call, the request itself is asynchronous and independent of other calls; upon success or error or timeout, the program can invoke callbacks or within the success or error blocks. A callback function or the success/error blocks are used to manipulate the page dynamically as it acts on the HTML5 document object model (DOM).

While XML was used in the initial generation of AJAX technologies, JSON objects are the specification format for data exchange. JSON is native JavaScript and extremely portable. Essentially defining attribute-value pairs and the values can be nested attribute value pairs supporting list and objects, it can be made to represent complex data structures while remaining quite light. Program 3 below outlines an ajax call and the manipulation DOM with the returned json data. Line 13 demonstrates direct style updates while line 14 adds a dynamic button and also associates an onclick event. It should be noted that the color type within that same block of code for the mincolor and maxcolor fields are only supported in HTML5 compatible browsers, older Internet Explorer programs will display the color picker widget/color bar as hexadecimal text values.

```

1  function initBase(){
2      if (backgroundTable != null) { return; }
3      hideContigs();
4      var ds = $("#dsset").val();
5      $("#core_dialog").html("Processing.. <img src='images/progress.gif'
/>");
6      $.ajax({
7          type: "POST",
8          url:  "/cgi-bin/virome/select_fullbase_results.cgi",
9          data: {'dsset': ds, 'column': annocol},
10         success: function(json){
11             $("#dsinfo").html("Active set " + ds);
12             var _data = $.parseJSON(json)["baseResults"];
13             $("#core_dialog").html("<font color='blue'>" +
_data.length + " retrieved results from " + ds + "</font>");
14             $("#baser_mapctl").html("&nbsp;<button id='baser_hbt'
title='Heatmaps with all samples and result are memory intensive; plotting
can take some seconds' onclick='javascript:prepPlotBaser()'>Plot
heatmap</button> min:<input type='color' id='mincolor' on-
change='updateColor()' style='width:40px' value='#d3d3d3'> shade:<input
type='color' id='shadecolor' onchange='updateColor()' style='width:40px' val-
ue='#0000FF'> max:<input type='color' id='maxcolor' onchange='updateColor()'
style='width:40px' value='#1F1F99'>");
15             backgroundTable = $('#baseresults_table').dataTable( {
16                 "bProcessing": true,
17                 "bDestroy": true,
18                 "iDisplayLength": 100,
19                 "oLanguage": { "sSearch": "<font color='blue'>Refine
results</font>" },
20                 "aaData": _data,
21                 "aoColumns": aoc});
22             //register advance filtering, supporting < and > percent
operations
23             backgroundTable.dataTableExt.afnFiltering.push( function(
oSettings, aData, iDataIndex ) {
24                 if ($('#coladvin').val() == "") { return; }
25                 var advfil = parseInt($('#coladvfil').val());
26                 if (_metric == null){ _metric = par-
seInt($('#coladvin').val());}
27                 var coladvin = parseInt($('#coladvin').val());
28                 var _op = document.getElementById('coladvop').value;
29                 var passedb = false;
30                 if (_op == "<="){
31                     // _metric threshold dynamically set on search input

```

```

32             if (aData[advfil] < _metric){
33                 passedrows = passedrows + 1;
34                 passedb = true;
35             }
36         }else{
37             if (aData[advfil] >= _metric){
38                 passedrows = passedrows + 1;
39                 passedb = true;
40             }
41         }
42         if (!quantmode) { _metric = null; }
43         return passedb; //end of inline filtering function
44     });
45     $("#op_container").css("float", "right");
46     $("#columns").val(annocol);
47 },
48 error: function(){
49     new Messi('Data retrieval error possibly because of network or
50 server. If problem continues, please contact help.', {title: 'Server er-
51 ror'}));}
52     });
53 }

```

Program 3. The initBase function implements an AJAX call retrieving virome data from the server. Upon success, the DOM is manipulated with new fields and associated events.

3.5 Advanced visualization

The analysis results include more than six hundred viral strain counts and almost eight hundred unmapped contigs. Though flexible search and refine functions help users with data exploration, intuitive visualization can assist with visual data mining. Heatmaps are ideal in picking out clusters and this thesis implements an in browser heatmap plotting function using JavaScript HighCharts library, described in Table 3. Existing logic allows for custom selection of columns (annotation counts) to set of sample bins.

Protein mimicry assessment results are aligned and visually accessible also directly on the web. Utilizing the Sequence Alignment Viewer, investigators can interact with all the match groups of the study, contrast by case and control, with multiple time points reflecting sample date collection prior to islet autoimmunity.

Discussed earlier, appropriate application of statistical based methods allows investigators to detect relationships, such as the most important variables. Cluster methods can reveal subgroups and hierarchical clustering produces attractive dendrograms and combined with annotations can reveal hidden patterns and also distributions. This thesis has integrated custom R scripts that supports Euclidean and correlation distance measures. The R scripts has dependencies of the following libraries available at Bioconductor:

gplots, vegan, RColorBrewer and Heatplus. No installation is required as the scripts can be called directly from the web application.

The results chapter and appendix will include visualizations of the viral hit count heatmap, Sequence to insulin gene assessment results and dendrogram with case and control annotations.

3.6 Cloud Computing with CSC

Because of the vast amounts of digital data produced in today's connected world, cloud computing is important and deserves investigation. While big US based companies such as Google, Microsoft and Amazon are offering cloud solutions; their costs and complexity cannot be underestimated. Finland's IT Centre for Science (CSC) provides free cloud computation and advice to nonprofit and university research. Helpful documentation is available at <https://www.csc.fi> and CSC cloud servers include many popular bio-informatics tools, including BLAST. Stated earlier, some BLAST commands usually execute for hours and CSC cloud servers allow parallelizing and executing these jobs simultaneously. CSC also provides free training and dependent on the level of collaboration, its team of experts can assist with custom tool dependencies.



4. VIROME NGS WEB APPLICATION RESULTS

In supporting exploration viruses associated with triggering Type 1 Diabetes, this thesis has implemented a modern web resource to access results from the described analysis scripts integrated with annotations from database. Existing analysis has revealed 10.2% of the samples across all match groups contain some viral content [8]. Accounting for control and case, the findings are not conclusive concerning viral genomes triggering islet immunity. It was revealed with laboratory experiments that current sequencing technology sensitivity is still an issue because of the minute viral content within stool. The NGS results are still valuable and particularly the large unmapped reads. The initial analysis of the reads thru *in silico* molecular mimicry assessment together with sequence alignment might prove to be a valuable visual mining resource. Virome NGS web application is available at <http://compbio.uta.fi/virology> and its access securely administered. Open sourced and compatible with all modern browsers, bugs and ongoing enhancement issues are tracked on <https://sourceforge.net/p/viromet1d/>.

4.1 Homepage and Layout

Upon authentication, the homepage loads the virome NGS analysis tabular results along with instrumentation run and sample background annotation. Built using jQuery UI, the homepage consists of six tabs as shown in Figure 7A. Beginning from the left, the ‘Results’ tab list known viral results, including enterovirus, parechovirus, sapovirus and bocavirus families; the full list is included in Appendix B. Next to it is the ‘Contigs’ tab, this includes results to assembled contigs. These contigs are unknown. After contigs tab is ‘16s’ where it presents bacterial content. Following 16s is the ‘Pipeline’ and this is the container for analysis components and their parameters. Next, ‘Clustering’ tab allows viewing of clustered results as well as new submissions with custom inputs and parameters. The rightmost tab is ‘Project’ and includes contact, project and cross reference ids between bacterial and viral projects.

A

  **Virome NGS**
SCHOOL OF MEDICINE

Results Contigs 16s Pipeline Clustering Projects

Results: All Remapped Columns: tot_reads >= 96 retrieved results

from remapped_full
[Manage visibility:](#)

Samples(*case): ALL Read results: None selected Plot heatmap min: shade: max:

Show 100 entries Refine results

Bin	Sample	Individual	SampleDate	MatchGroup	Case	Gender	TimePoint	s_year	s_month
1	091717A_2001-07-02	091717A	2001-07-02	36	1	f	1	2001	07
2	091868A_2001-07-02	091868A	2001-07-02	36	0	f	1	2001	07
3	039959A_1999-04-12	039959A	1999-04-12	38	0	m	3	1999	04

B

Results Contigs 16s Pipeline Clustering Projects

Results: All Remapped Columns: tot_reads >= 60% 39 passed

Manage visibility:

Samples(*case): ALL Read results: None selected Plot heatmap min: shade: max:

Show 100 entries Refine results

Bin	MGTimepoint	Case	Gender	tot_reads	enterovirus_pos	enterovirus_reads
93	87_2	0	m	859134	1	1063
89	35_1	1	f	730316	0	111
83	15_1	1	m	1323010	0	35
1	36_1	1	f	409095	0	0
2	36_1	0	f	422648	0	0
3	38_3	0	m	408046	0	0
4	38_2	0	m	541452	0	0

Figure 7. Home page includes interactive tabs with sample background annotations and integrated with viral analysis results from NGS sequencing. Visibility controls, filtering and sorting are built in for easy data exploration.

4.2 Interactive Interface

Advanced filtering and annotation selection features are built to support easy exploration. Viral compositions of different samples can be plotted as heat maps and statistically clustered, particularly suitable for sample time points and controls. All the matched protein matching results from the molecular mimicry analysis of key diabetes genes are available as assembled interactive stacked bar plots.

4.2.1 Selection and Sort, Filter and Refine

As there are many hundreds of annotations, the page includes custom settings for visibility management for all columns via simple toggling clicks. Any combination can be achieved by toggling on the provided links. Sorting can be done on any column, ascending is the default and clicking it again reverses the order to descending. The initial page loads the full sample list. Investigators can submit custom filtering using the dropdown and then input condition. Advanced searches include percentile operations. Further refine can be done on the active result set by typing in match condition in the Refine results input field on the right side of the page. Figure 7B shows search input of the top 60% total NGS reads resulting in 39 (cropped for visibility and note that sequencing run index and collection date related fields were toggled using Manage visibility functions), the result set is then sorted by enterovirus reads. It can be seen that enterovirus content is not specific to case samples and not correlated to total read counts. Figure 7B shows the top enterovirus reads from a control sample but the next two are cases with a mix ordering of total reads.

4.3 Visualization

4.3.1 Viral Reads Heatmap

Implemented using HTML5 and Highcharts, the web application allows for heatmap plotting with support of sample and annotation subset selections. Heatmap allows fast and intuitive visual inspection of the respected NGS analysis results. Particularly, the plots can be downloaded to for comparison, such as sets of individuals by year of birth and also match groups of control versus case. Investigators can adjust the representative min and max color selection, along with defining a custom outlier. The virome heatmap function can be used across viral results, contigs and 16s tabs. Figure 8A shows heatmap results of 12 affected samples from year 2000 and the hits of the enterovirus, pareechovirus, sapovirus and bocavirus strains. Figure 8B shows 11 affected samples collected in 2001 (8) and 2002 (3) plotted to the same columns. It can be seen that possibly bocavirus, the last two columns, is not found in 2001 and 2002 but cap-

[illegible]

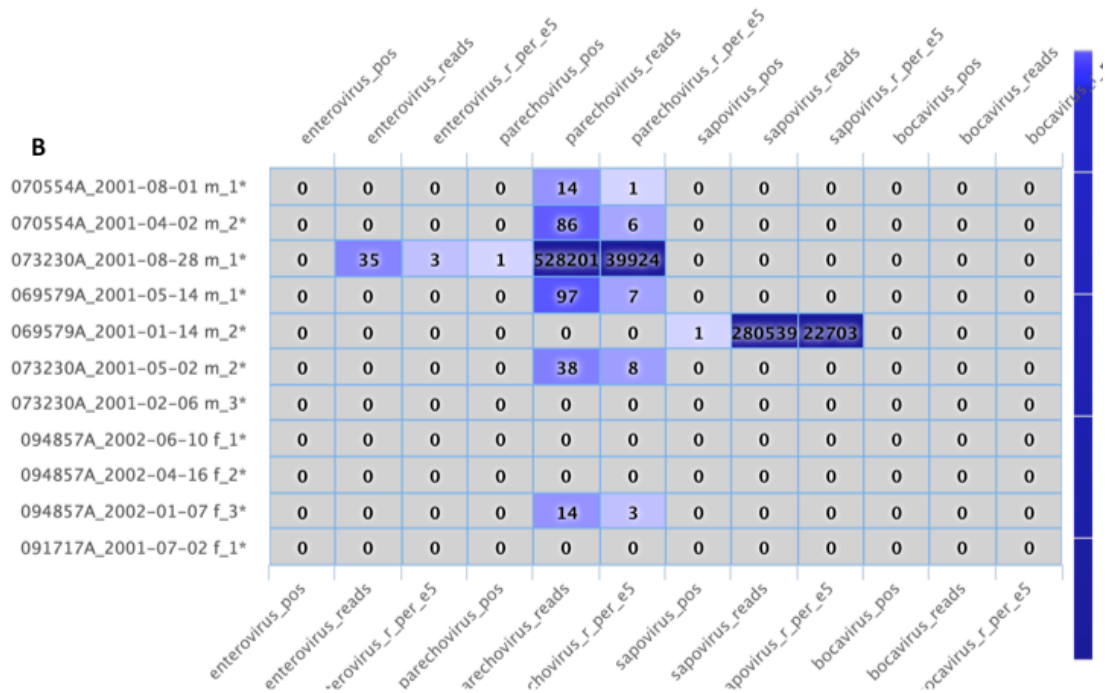


Figure 8. Heatmaps of case samples in 2000 (top) and 2001 and 2002 with difference in bocavirus, rightmost two columns, signals.

The heatmap function is available for the contig results. DIPP investigators can freely select any combination of samples to sets of analysis reads and export the image using a modern browser.

4.4 Statistical Clustering

Statistical learning methods are central to NGS analysis as NGS data are often high dimension and heterogeneous. To the benefit of the community, many learning method implementations are freely available in popular bioinformatic languages such as Matlab, python and R. Figure 9 shows a hierarchical clustered heatmap of all contig reads with annotation bars plotted in R. Contigs are overlapping sequences assembled from NGS raw reads with sufficient coverage and quality. Correlation distance, measuring statistical independence, is used for the clustering of rows and columns. The distance between clusters uses a linkage method of complete, implying the max or full distance [53, 54]. The analysis resulted in approximately 650 contigs and as most of these did not mapped to reference, this clustered plot was done to investigate possible relationship among the variables and subgroups. As some viruses are seasonal, the annotation bar chart on the right draws the samples to the month and year but there does not appear to be any obvious patterns. The output image format is pdf with zoom support to maintain resolution. Contigs with counts that are in less than 5% of the samples are pruned. Currently the

plotting can be called directly from the web interface without any desktop installations though the parameterization of this feature is still being refined.

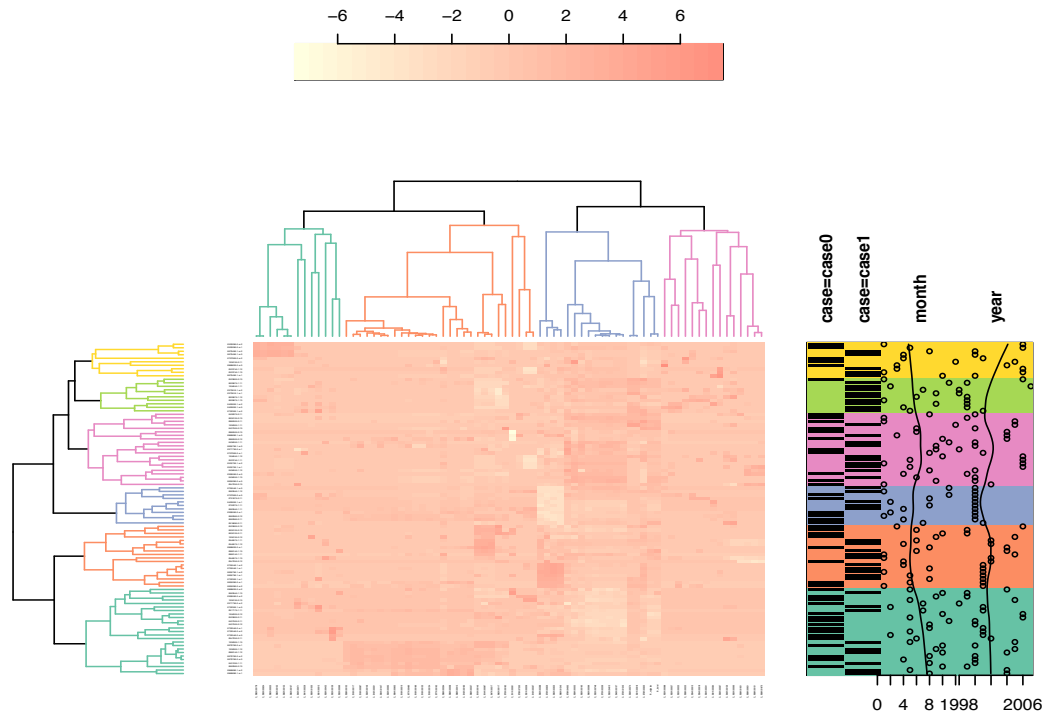


Figure 9. Heatmaps of case samples in 2000 (top) and 2001 and 2002 with difference in bocavirus, rightmost two columns, signals.

4.5 Islet Autoimmunity Markers Assessment

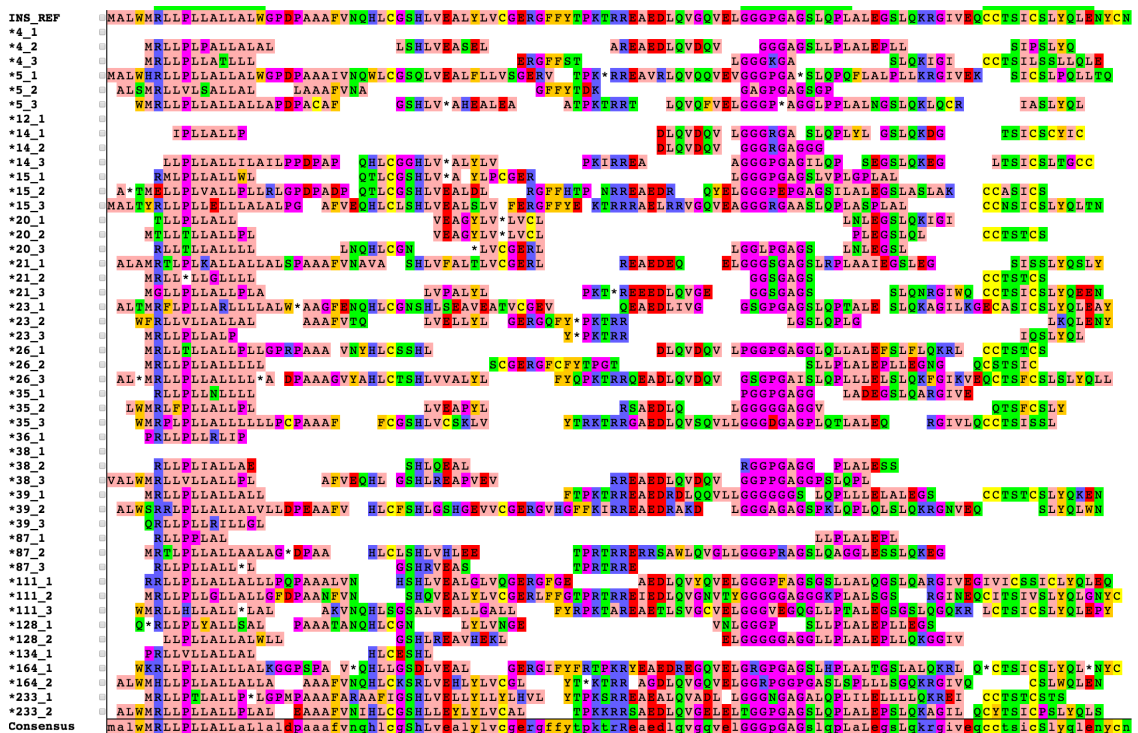
Islet autoimmunity, defined insulin producing beta cells being targeted by the immune system, is the key step before T1D symptoms [5, 8, 17]. Essentially the immune system treats the beta cells as a foreign antigen and antibodies bind and kill the cells. Islet autoimmunity can be detected when antibodies directed against genes *INS*, *GAD65*, *IA-2* and *ZNT8* [17] are found to be elevated in the bloodstream. These gene reference sequences downloaded from NCBI are provided in Appendix D. The triggering of islet autoimmunity is unknown and it has been raised that molecular mimicry, a theory that a foreign peptide (protein fragment) similarity can activate antibodies. It has been observed that the AIDS virus is capable of molecular mimicry similar to *CD4* binding

[55]. This underscores the importance of investigating protein homology especially given that 74% of NGS reads from the study mapped to unknown organisms.

4.5.1 Insulin Case and Control Assessment

As discussed in methods chapter, these reads are being translated forward and backward with sliding codon frames, resulting in six protein files per sample unmapped file. Figure 10 below shows the collected comparison results between NGS raw contig sequences and insulin (*INS*) reference. It can be seen that the case, affected samples, on top are much denser as there are more matches. The bottom figure B, shows the sparser control results. As the reference track is shown on top, the misaligned bases can be detected though a systematic analysis to also include sub-region density would be valuable.

A



B

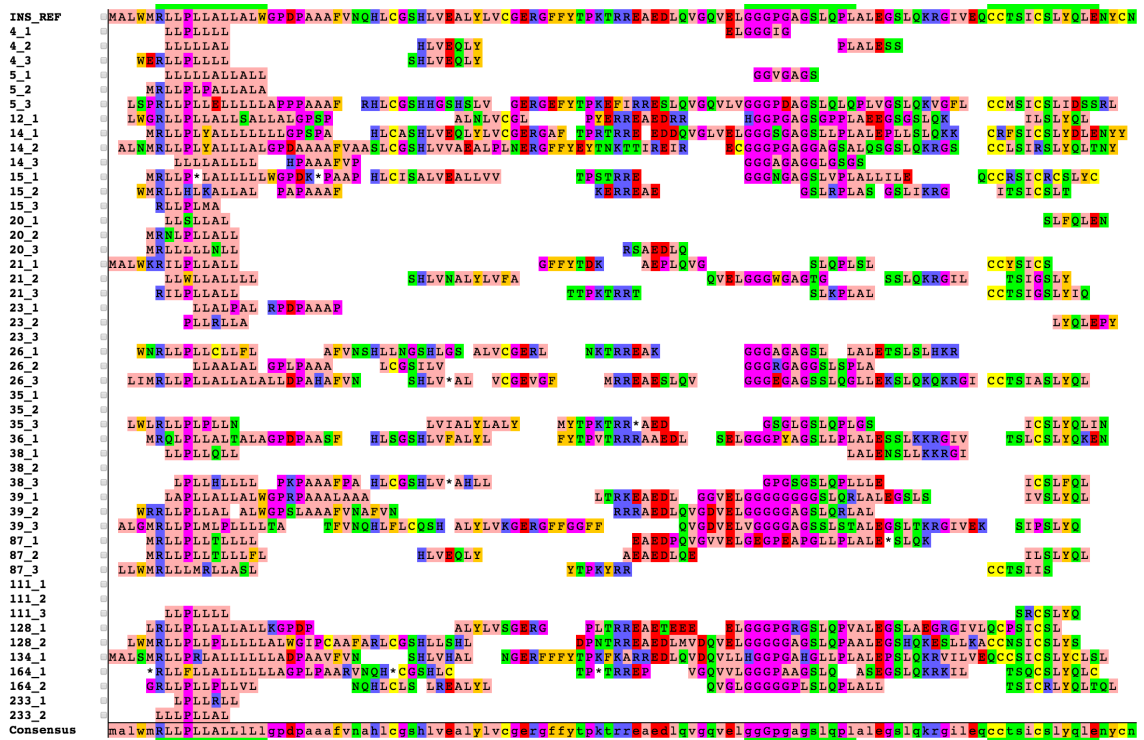


Figure 10. Assessment of Virome NGS translated contigs matched and aligned with Insulin. It can be seen that the case samples, top figure A, have more matches and denser compared to bottom control samples. Colors can be selected from popular bioinformatic schemes.

Not surprisingly, very small numbers of matches are perfect. Perfect matches are defined to be 7 perfect amino acids translated from 21 nucleotides while almost perfect allows for one mismatch. The number of perfect matches and their sequences are not listed because of the ongoing nature of the study. Along with the other gene images, the matched sequences are available upon request.

4.5.2 Case and Control Summary Counts

The summary counts across the islet autoimmunity markers are collected in table 4. The short gene length of insulin makes it reasonable that its match counts are lower than the other islet autoimmunity gene markers. For example, IA2 isomer 1 has nearly 9 times more amino acids and the total number of matches is consistent with that ratio. These findings are quite preliminary though the statistics for the other genes (accession number and full sequences listed in appendix D) are inline with *INS*, with case matches always in the clear majority. It can also be observed visually within Figure 10 top and bottom that the green and purple base segments to the right of the graph reveals particular density difference. The next stage of the project analysis will need to consider the regional distribution and take into account perfect matches, along with coverage and quality and run bias along with annotation data such as collection time points.

Table 4. Islet autoimmunity marker genes are listed with protein sequence match results. The total matches include perfect (0 mismatches) and almost perfect (1 mismatch). A match indicates 0 or 1 mismatch of 7 amino acids between translated NGS read and indicated gene protein sequence. Gene protein lengths as shown by INS with 110 and IA2_1 with 979 heavily impact the counts. It is observed that case (affected) samples are the majority for all marker genes.

Gene	Protein Length	Total Matches	Case	Case Percent
INS	110	3180	1987	62%
GAD65	585	7178	4387	61%
ZNT8A	369	9183	5333	58%
ZNT8B	320	8691	5002	58%
IA2_1	979	28881	17029	59%
IA2_2	950	27863	16471	59%
IA2_3	889	22232	13150	59%

5. DISCUSSION

5.1 Overview

The motivation and methods of an integrated web system to manage and visualize NGS virome analysis results have been discussed in prior chapters. This thesis was driven by the increasing bioinformatic needs of the ongoing Diabetes Prediction and Prevention Virome NGS project. Supported by well-known tools such as Velvet and BLAST, existing scripts were in place for processing viral identifications though the output was accessed only on spreadsheets. Through construction of HTML5 based web interface and centralized database, DIPP investigators and international collaborators can securely search and visualize the NGS analysis results. Furthermore, a new analysis was conducted in assessing the NGS contig reads towards genetic markers of islet autoimmunity, the irreversible condition before Type 1 Diabetes symptom onset. It appears that the preliminary matched results can be potentially revealing and a valuable data resource.

5.2 Interface to Integrated Results and Interactive Visualizations

The analysis pipeline to detect viral signatures have yielded that 10.4% (10/96) of the samples contained human viral strain content. This is not statistically conclusive to have an impact on T1D islet autoimmunity after including control and case data points. However, web interface and database schemas have been built to store the results integrated with the background annotations. Flexible filtering searches and view options have been implemented to explore the hundreds of columns. Custom selected heatmaps are available for visual pattern detection. In addition, steps were taken to integrate R based scripts that produced dendrograms with hierarchical clustering abilities. These tree diagrams are good method to visualize the diverse and broad family of virus. The goal is that these interactive visualization and exploration tools are helping investigators gain new insights and viewpoints for discussion, similar to matured sequencing technologies allowing measurements being taken at nucleotide base levels and also at comprehensive large sample collections [56].

The computational scripts responsible for calling the viral hits are available within the source repository. These scripts take advantage of well-cited and popular libraries and tools. Still, genomic analysis tools, particularly involving statistical learning are complex and the output unsymmetrically impacted by their parameter inputs. This implies that there are good reasons to rerun the analysis with various combinations, implying

additional computational resources and also challenge to systematically analyze and compare results across the parametric runs.

5.3 Molecular mimicry and islet autoimmunity

It has been theorized that virus can perform molecular mimicry in disrupting and eventually taking over its host immune system. Molecular mimicry alluding to sequence similarity pertains that the destruction of beta cells is caused by viral genomes containing similar sequences that can trigger antibodies. One driving challenge in studying viromes is the lack of references as evident by 74% unmapped reads regarding total NGS outputs. New analysis was performed to take these raw reads, translating them to proteins and mapped them against islet autoimmunity genetic markers. Perfect and almost perfect, with 1 mismatch, were collected and visualized. The preliminary results are promising as it is showed that for every genetic marker, the clear majority of the matches are with the affected T1D case samples.

The sequence findings and statistics are available to DIPP investigators and an interactive novel protein alignment visualization pages have been developed. Particularly interesting are subsequence regions that seem to have a density visual pattern. Coverage and platform run information are being validated. The comparably small number of perfect matches will be realigned to viral references and further analyze. It appears that within literature, the prior viral molecular mimicry has been focused on HIV immune related receptors: cytokines and chemokine [55, 60]. There are plans to extend this novel application of NGS assisted to assess the raw reads to human leukocyte antigen (HLA). These are the locus of genes responsible for regulation of human immune system. It is not straightforward, as it is known that these HLA genes located on chromosome 6 have large variation across populations. Clearly, large amount of additional analysis and verification are needed, including subdividing this assessment into collection time points and also directly building interactive visualizations for interested peptide sequences. Protein folding structure is an active and challenging medicinal and disease branch of modern biological research as it is known that their shapes impact protein functions and identifies. Recall that the antibodies are Y shaped and binds to specific shaped antigens. Advances in existing protein three-dimensional visualization tools are necessary as presently they are mostly desktop or static thus making them difficult to integrate within network based solutions.

5.4 Limitation

Sequencing technologies have matured and advanced greatly. Together with the drastic reduction in costs and establishment of sequencing service companies, the accessibility and convenience to use NGS as a research tool is within the means of many labs. Recently in the Science journal, Funari and Canosa discussed the importance and pressing

need of more bioinformatics tools for NGS analysis [13]. NGS data analysis, similar to other big data domains, is difficult because of the high computational resources needed. Biomedical research, particularly for viral and metagenomics, compounds the difficulty and opportunities, since little is known, the data might be noisy and the processing requires *de novo* assembly. Data interpretation can be even more challenging as the analysis results are high dimensional and heterogeneous. Development of advanced visualization will help researchers with picking out hidden patterns but confirmation and further analysis of these patterns are challenging due to instrument accuracy, reproducibility and missing data [9, 13].

Applicable to this study, a combination of NGS sensitivity, virome genome size (they are 20-50 times physically smaller than bacteria cell [11, 15]) and lack of viral references lead to 74% of the resultant reads being unmapped. The lack of viral reference is well known as viruses have shorter generations and fast evolving. A lab PCR experiment revealed a slight accuracy limitation. It is not surprising since extracting viral content using metagenomics and NGS technology is a novel approach. Sequencing companies, similar to computer and semiconductor firms are evolving and becoming more specialized. PacBio Sciences (<http://www.pacificbiosciences.com>) have introduced single cell and single molecule sequencing. These new advances, while demanding additional computational needs and bioinformatics development, are also providing researchers with exciting possibilities, old and new questions within basic biochemical and disease disciplines.

5.5 Open source

This thesis has been greatly assisted by the availability of open sourced bioinformatics tools, databases and nonprofit computation resources such as Finland's IT Center for Science. Academic biomedical research and open sourced development shared similar ideals. The analysis server side scripts are implemented in Python using Biopython and other freely available libraries. Together with web client written in HTML5 and JavaScript, the code base and issue trackers are available on: <https://sourceforge.net/projects/viromet1d/>.

6. CONCLUSION

The rapid and maturation of next generation sequencing (NGS) platforms and technology services have enabled the collection of omics data at nucleotide scales. The decrease in cost together with increase sensitivity has allowed diabetes prediction and prevention (DIPP) researchers undertake unbiased study into the gut viral community using stool samples from affected and unaffected children across multiple time points. While several projects have cited the presence of enterovirus with Type 1 diabetes [5, 58, 59], this DIPP study is novel at applying NGS technology in determining the role of virus in triggering or advancing islet autoimmunity, when the host body immune system begins targeting and killing insulin producing beta cells residing in the pancreas islet.

Initial analysis results have found approximately 10.4% (10/96) of samples with human viral content though the results are not proven conclusive when considering cases and controls. Still, the comprehensive human gut virome is a valuable resource and in order for DIPP researchers to further make sense of this NGS study, it requires bioinformatics integration of the additional background annotations to the diverse viral flora families. These results and annotations are now integrated and securely accessible from a browser based interface. The developed intuitive search forms allow investigators flexibility and advanced filtering options. The application of statistical learning methods is required to detect hidden patterns particularly within diverse large scaled result sets. Dendrogram plots, depicting hierarchical relationships, produced by clustering and learning methods implemented in R are accessible to users without any installations. Along with the importance to integrate advanced learning methods, there is also the need to help researchers with improved understanding of the statistical and visual outputs. Interactive heatmap plotting function has been developed allowing for the plotting of any viral strain columns to custom selection of samples, such as contrasting match groups across months before onset of Type 1 diabetes.

It appears that preliminary results derived from the new computational analysis assessing NGS raw reads towards islet autoimmunity genetic markers are valuable and warrants further investigation. The exact cause of T1D is unknown and it is likely that there are environmental reasons. Irreversible, T1D symptoms onset prior to islet autoimmunity, where insulin producing beta cells are being targeted and subsequently destroyed by the immune system. It has been observed with HIV virus and white blood cell membrane *CD4* gene [55, 60], that this virus can utilize molecular mimicry during pathogenesis, to compromise, disrupt and eventual takeover the host immune system mechanisms. The implementation of this islet autoimmunity antibody marker assessment analyses structural similarity comparison of insulin, insulin antibody and zinc

transport isomers and glutamine acid zinc transport, cited to be significant markers [17], with the translated DIPP Virome NGS fragments. The integrated resultant matches can be visualized directly on the browser as selectable aligned stacked protein tracks, here order by control and case samples and respective time points. Through direct visual data mining, the case matched results show a clear majority, as confirmed by summary statistics across all the genes and their isomers in Table 4. Also interesting and possibly revealing is that within each gene marker, there are sub-regions where case samples are especially denser than controls.

Clearly additional analysis work is required and especially challenging will likely be experimental confirmation. Though the frame size of 7 is cited to be significant [17, 18], a method validation effort has been started to rerun the analysis for smaller and larger frame sizes. The new results will be compared though it can be seen that the thesis initial findings demonstrate the value of bioinformatics tools together with NGS measurements, as a research tool. The findings also reiterate the value of interactively visualizing genomic results within an integrated framework.

REFERENCES

- [1] International Diabetes Federation, IDF Diabetes Atlas, 6th Edition 2013 ISBN: 2-930229-85-3, Available: <http://www.idf.org/diabetesatlas>
- [2] American Diabetes Association, The Cost of Diabetes, 2013, accessed 21 July 2015 [Online] Available: <http://www.diabetes.org/advocacy/news-events/cost-of-diabetes.html>
- [3] Harjutsalo V, Sjöberg L, Tuomilehto J, Time trends in the incidence of type 1 diabetes in Finnish children: a cohort study, *Lancet*. 2008 May 24;371(9626):1777-82. doi: 10.1016/S0140-6736(08)60765-5.
- [4] Diabetes Epidemiology Research International Group: Geographic patterns of childhood insulin-dependent diabetes mellitus. *Diabetes* 37:1113–1119, 1988
- [5] Oikarinen M, Tauriainen S, Oikarinen S, Honkanen T, Collin P, Rantala I, Mäki M, Kaukinen K, Hyöty H, Type 1 diabetes is associated with enterovirus infection in gut mucosa, *Diabetes*, 2012 Mar;61(3):687-91. doi: 10.2337/db11-1157, Epub 2012 Feb 7.
- [6] Rewers M, LaPorte RE, King H, Tuomilehto J for The Diabetes Epidemiology Research International Study Group (DERI): Trends in the prevalence and incidence of diabetes: insulin-dependent diabetes mellitus in childhood. *World Health Stat Q* 41:179–189, 1998
- [7] Diabetes Prediction and Prevention, The DIPP Project, accessed 01 July 2015, [Online] Available: <http://dipp.utu.fi/index.php>
- [8] Kramná L, Kolarova K, Oikarinen S, Persiheimo JP, Ilonen J, Simell O, Knip M, Veijola R, Hyöty H, Cinek, O Gut virome and the development of islet autoimmunity leading to early-onset type 1 diabetes, Submitted, 2015
- [9] Metzker ML, Sequencing technologies - the next generation, 2010, *Nat Rev Genet*. 2010 Jan;11(1):31-46. doi: 10.1038/nrg2626
- [10] Genetics Home Reference, What kinds of mutations are possible? accessed 20 July 2015, [Online] Available: <http://ghr.nlm.nih.gov/handbook/mutationsanddisorders/possiblemutations>
- [11] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P *Molecular Biology of the Cell: Reference Edition*, 5th ed. New York: Taylor & Francis Group, 2008.

- [12] Maskos U, Southern, EM Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ, *Nucleic Acids Res*, 1992, 20 (7): 1679–84. doi:10.1093/nar/20.7.1679
- [13] Funari V and Canosa SJ, The Importance of Bioinformatics in NGS: Breaking the Bottleneck in Data Interpretation, *Science* 9 May 2014, Vol. 344 no. 6184 p. 653 DOI: 10.1126/science.344.6184.653-c
- [14] Wikipedia, Bioinformatics, accessed 01 July 2015, [Online] Available: <https://en.wikipedia.org/wiki/Bioinformatics>
- [15] Hunter L, *The Processes of Life An Introduction to Molecular Biology*, MIT Press ISBN: 9780262013055 January 2009 1-12.
- [16] Wikipedia, Molecular mimicry, accessed 01 July 2015, [Online] Available: https://en.wikipedia.org/wiki/Molecular_mimicry
- [17] Eringsmark Regnéll S, Lernmark A, The environment and the origins of islet autoimmunity and Type 1 diabetes, *Diabet Med*. 2013 Feb;30(2):155-60. doi: 10.1111/dme.12099.
- [18] Herman RA, Song P, ThirumalaiswamySekhar A, Value of eight-amino-acid matches in predicting the allergenicity status of proteins: an empirical bioinformatic investigation *Clin Mol Allergy*. 2009; 7: 9. Published online 2009 Oct 29. doi: 10.1186/1476-7961-7-9
- [19] Crick FH, On Protein Synthesis, 1956, *Symp. Soc. Exp. Biol.* XII, 139-163
- [20] Wikipedia, Down syndrome, accessed 01 July 2015, [Online] Available: https://en.wikipedia.org/wiki/Down_syndrome
- [21] National Institute for Health, What are proteins, accessed 01 July 2015, [Online] Available: <http://ghr.nlm.nih.gov/handbook/howgeneswork/protein>
- [22] Mendel, G., 1866, Versuche über Pflanzen-Hybriden. *Verh. Naturforsch. Ver. Brünn* 4: 3–47 (in English in 1901, *J. R. Hort. Soc.* 26: 1–32).
- [23] Wikipedia, Huntington's disease , accessed 01 July 2015, [Online] Available: https://en.wikipedia.org/wiki/Huntington%27s_disease
- [24] Isobe M, Emanuel BS, Givol D, Oren M, Croce CM, Localization of gene for human p53 tumour antigen to band 17p13, *Nature* 1986, 320 (6057): 84–5. doi:10.1038/320084a0. PMID 3456488

- [25] Wrighton KH, Mechanisms of disease: p53 puts a damper on WNT signalling, *Nature Reviews Molecular Cell Biology* 12, 770 (December 2011), doi:10.1038/nrm3238
- [26] Salwitz JA, The Future is Now: Personalized Medicine, American Cancer Society 2012, accessed 20 July 2015, [Online] Available: <http://www.cancer.org/cancer/news/expertvoices/post/2012/04/18/the-future-is-now-personalized-medicine.aspx>
- [27] Koivisto VA1, Pelkonen R, Cantell K Effect of interferon on glucose tolerance and insulin sensitivity *Diabetes*. 1989 May;38(5):641-7
- [28] Wikipedia, Maturity onset diabetes of the young accessed 02 July 2015. [Online] Available https://en.wikipedia.org/wiki/Maturity_onset_diabetes_of_the_young
- [29] American Diabetes Association, Genetics of Diabetes, accessed 10 July 2015. [Online] Available <http://www.diabetes.org/diabetes-basics/genetics-of-diabetes.html>
- [30] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, Basic local alignment search tool. 1990 *J. Mol. Biol.* 215:403-410
- [31] Watson JD, Crick FH, Molecular structure of nucleic acids; a structure for deoxy-ribose nucleic acid, *Nature*, 1953, 171 (4356): 737–738. Bib-code:1953Natur.171..737W. doi:10.1038/171737a0. PMID 13054692
- [32] Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison CA, Slocombe PM, Smith M Nucleotide sequence of bacteriophage Φ X174 DNA. *Nature* 265 (5596): 687–95. 1977
- [33] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, *Nature*, 2001, 409 (6822): 860–921. doi:10.1038/35057062. PMID 11237011.
- [34] Wang Z, Gerstein M, Snyder M, RNA-Seq: a revolutionary tool for transcriptomics, *Nat Rev Genet.* 2009 Jan; 10(1): 57–63. doi: 10.1038/nrg2484
- [35] Zerbino DR, Birney E Velvet: Algorithms for de novo short read assembly using de Bruijn graphs, *Genome Research* 2008 18 (5): 821–829. doi:10.1101/gr.074492.107. PMC 2336801. PMID 18349386
- [36] Li H A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011 Nov 1;27(21):2987-93. Epub 2011 Sep 8. [PMID: 21903627]

- [37] Parker BC, Annala MJ, Cogdell DE, Granberg KJ, Sun Y, Ji P, Li X, Gumin J, Zheng H, Hu L, Yli-Harja O, Haapasalo H, Visakorpi T, Liu X, Liu CG, Sawaya R, Fuller GN, Chen K, Lang FF, Nykter M, Zhang W The tumorigenic FGFR3-TACC3 gene fusion escapes miR-99a regulation in glioblastoma *J Clin Invest*. 2013 Feb;123(2):855-65. doi: 10.1172/JCI67144
- [38] The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004
- [39] Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S The generic genome browser: a building block for a model organism system database. *Genome Res*. 2002, Oct;12(10):1599-610.
- [40] Linthicum D, Do the math: The true cost of cloud computing, *InfoWorld*, 2014 accessed: 20 July 2015 [Online] Available: <http://www.infoworld.com/article/2841806/cloud-computing/true-cost-cloud-computing.html>
- [41] James G, Witten D, Hastie T, Tibshirani, *An Introduction to Statistical Learning*, Springer, 2012, ISBN 978-1-4614-7138-7, pp. 1, 15-29, 385-400.
- [42] Technical Informatics Illumina, Understanding Illumina Quality Scores, accessed 21 July 2015. [Online] Available: http://www.illumina.com/documents/products/technotes/technote_understanding_quality_scores.pdf
- [43] Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. 2010 Aug 25;11(8):R86.
- [44] Handelsman J, Metagenomics: Application of Genomics to Uncultured Microorganisms, *Microbiol Mol Biol Rev*. 2004 Dec; 68(4): 669–685. doi: 10.1128/MMBR.68.4.669-685.2004
- [45] Mundas S, Human Microbiome Project Reveals largest microbial map, *BBC News* 2012, accessed 21 July 2015, [Online] Available: <http://www.bbc.com/news/health-18422288>
- [46] de Koning AP, Gu W, Castoe TA, Batzer MA, and Pollock DD Repetitive Elements May Comprise Over Two-Thirds of the Human Genome, *PLoS Genet*. 2011 Dec; 7(12): e1002384.
- [47] Berger B, Peng J, Singh M Computational solutions for omics data *Nature Reviews Genetics* 14, 333–346 fig. 1 (2013) doi:10.1038/nrg3433

- [48] Wiki, Picornavirus, Wikipedia 2015, access 21 July 2015. [Online] Available: <https://en.wikipedia.org/wiki/Picornavirus>
- [49] Li H. and Durbin R. To cite BWA: Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25:1754-60. [PMID: 19451168]
- [50] Domingo E, Quasispecies Theory in Virology, *J. Virol.* January 2002 vol. 76 no. 1 463-465, doi: 10.1128/JVI.76.1.463-465.2002
- [51] Wiki, Hamming distance, access 21 July 2015. [Online] Available: https://en.wikipedia.org/wiki/Hamming_distance
- [52] Levenshtein VI, Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Physics Doklady*, 1966, 10 (8): 707–710
- [53] MiniTab, Distance measures for cluster observations, accessed 20 July 2015, [Online] Available: <http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/multivariate/item-and-cluster-analyses/distance-measures-for-cluster-observations/>
- [54] MiniTab, Linkage methods, accessed 20 July 2015, [Online] Available: <http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/multivariate/item-and-cluster-analyses/linkage-methods/>
- [55] Bisset LR, Molecular mimicry in the pathogenesis of AIDS: the HIV/MHC/mycoplasma triangle, *Med Hypotheses*. 1994 Dec;43(6):388-96.
- [56] Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, Berlin AM, Malboeuf CM, Ryan EM, Gnerre S, Zody MC, Erlich RL, Green LM, Berical A, Wang Y, Casali M, Steeck H, Bloom AK, Dudek T, Tully D, Newman R, Axten KL, Gladden AD, Battis L, Kemper M, Zeng Q, Shea TP, Gujja S, Zedlack C, Gasser O, Brander C, Hess C, Gunthard HF, Brumme ZL, Brumme CJ, Bazner S, Rychert J, Tinsley JP, Mayer KH, Rosenberg E, Pereya F, Levin JZ, Young SK, Jessen H, Altfeld M, Birren BW, Walker BD, Allen TM(2012) Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection. *PLoS Pathogens* 8(3): e1002529.
- [57] Huson DH and Mitra S (2012) Introduction to the Analysis of Environmental Sequences: Metagenomics with MEGAN In: *Evolutionary Genomics: Statistical and Computational Methods*, ed. by Maria Anisimova. Springer, chap. 17, pp. 415-429.

- [58] Hyöty H, Enterovirus infections and type 1 diabetes, *Ann Med.* 2002;34(3):138-47.
- [59] Lin HC, Wang CH, Tsai FJ, Hwang KP, Chen W, Lin CC, Li TC, Enterovirus infection is associated with an increased risk of childhood type 1 diabetes in Taiwan: a nationwide population-based cohort study, *Diabetologia*, January 2015, Volume 58, Issue 1, pp 79-86
- [60] Alcami A, Viral mimicry of cytokines, chemokines and their receptors, *Nat Rev Immunol.* 2003 Jan;3(1):36-50.

APPENDIX A: CODON TO AMINO ACIDS

Amino Acid	SLC	DNA codons
Isoleucine	I	ATT, ATC, ATA
Leucine	L	CTT, CTC, CTA, CTG, TTA, TTG
Valine	V	GTT, GTC, GTA, GTG
Phenylalanine	F	TTT, TTC
Methionine	M	ATG
Cysteine	C	TGT, TGC
Alanine	A	GCT, GCC, GCA, GCG
Glycine	G	GGT, GGC, GGA, GGG
Proline	P	CCT, CCC, CCA, CCG
Threonine	T	ACT, ACC, ACA, ACG
Serine	S	TCT, TCC, TCA, TCG, AGT, AGC
Tyrosine	Y	TAT, TAC
Tryptophan	W	TGG
Glutamine	Q	CAA, CAG
Asparagine	N	AAT, AAC
Histidine	H	CAT, CAC
Glutamic acid	E	GAA, GAG
Aspartic acid	D	GAT, GAC
Lysine	K	AAA, AAG
Arginine	R	CGT, CGC, CGA, CGG, AGA, AGG
Stop codons	Stop	TAA, TAG, TGA

Figure A1: Amino acid compounds are translated from the listed DNA codons.

ATG is also the start codon in addition for coding for M. RNA codons would be replacing nucleotide Thymine (T) with Uracil (U), the translated Amino Acids are the same.

APPENDIX B: DE BRUIJN GRAPH AND VELVET

De Bruijn graphs are directed graphs representing overlapped characters of particular size and dimension. Directed graphs contain direction, therefore $A \rightarrow B \neq B \rightarrow A$.

Properties from en.wikipedia.org/wiki/De_Bruijn_graph:

If $n=1$ then the condition for any two vertices forming an edge holds vacuously, and hence all the vertices are connected forming a total of m^2 edges.

Each vertex has exactly m incoming and m outgoing edges.

Each n -dimensional De Bruijn graph is the line digraph of the $(n - 1)$ -dimensional De Bruijn graph with the same set of symbols.

Each De Bruijn graph is Eulerian and Hamiltonian. The Euler cycles and Hamiltonian cycles of these graphs (equivalent to each other via the line graph construction) are De Bruijn sequences.

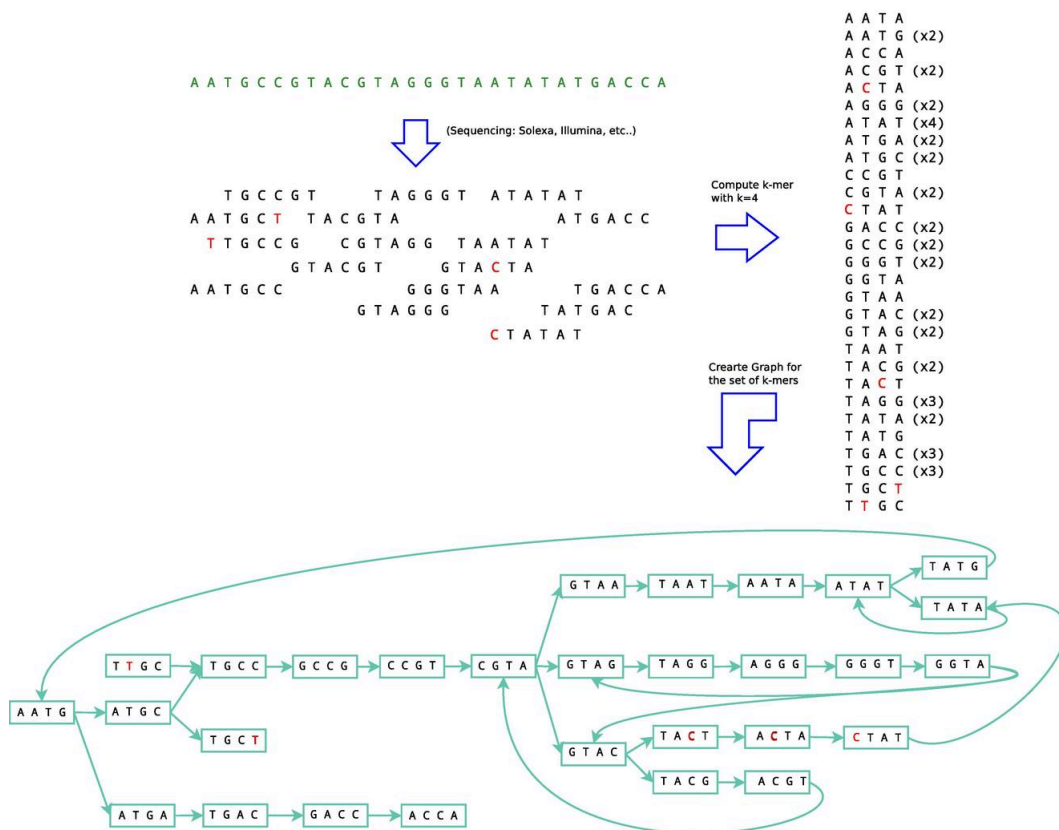


Figure A2: Velvet nodes are linked if first $k-1$ characters matches last $k-1$. Taken from https://en.wikipedia.org/wiki/Velvet_assembler#cite_note-zerbino-4.

APPENDIX C: VIRAL, BACTERIAL AND MEGAN [57] STRAINS

VIRUS: enterovirus_pos enterovirus_reads enterovirus_r_per_e5 parechovirus_pos
parechovirus_reads parechovirus_r_per_e5 sapovirus_pos sapovirus_reads
sapovirus_r_per_e5 bocavirus_pos bocavirus_reads bocavirus_r_per_e5 rhinovirus_pos
rhinovirus_reads rhinovirus_r_per_e5 anellovirus_pos anellovirus_reads anellovi-
rus_r_per_e5

BACTERIAL: 16s_ct ent_lib_ct boca_lib_ct parecho_lib_ct ent_ct boca_ct sapo_ct
adv_ct parecho_ct sapo_lib_ct Burkholderia Lactococcus Lactobacillus Escherichia
Clostridium Enterococcus Streptococcus Salmonella Shigella Bacteroides Pseudomonas
Acinetobacter Propionibacterium Bifidobacterium Leuconostocaceae Vibrio Steno-
trophomonas Ruminococcus Roseburia faecis Faecalibacterium prausnitzii Bacterio-
phage Lactococcus phage Lactobacillus phage Escherichia phage Klebsiella phage
Clostridium phage Enterococcus phage Streptococcus phage Salmonella phage Shigella
phage Bacteroides phage Pseudomonas phage Acinetobacter phage Propionibacterium
phage Bifidobacterium phage Leuconostoc phage

MEGAN: Megan_root; Megan_root;Not assigned; Megan_root;No hits; Me-
gan_root;cellular organisms;Bacteria; Megan_Bacteroidetes; Megan_Firmicutes; Me-
gan_Actinobacteria; Megan_root;cellular organisms;Eukaryota; Megan_Viridiplantae
Megan_Homo sapiens Megan_Viruses Megan_dsDNA viruses, no RNA stage Me-
gan_dsRNA viruses Megan_Retro-transcribing viruses Megan_ssDNA viruses Me-
gan_ssRNA viruses Megan_unclassified viruses Megan_Enterovirus Me-
gan_Parechovirus Megan_Bocavirus Megan_Sapovirus Megan_Rhinovirus Me-
gan_Adenoviridae Megan_Anelloviridae

APPENDIX D: ISLET AUTOIMMUNITY MARKER GENES

INS: >gi|631226408|ref|NP_001278826.1| MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN

GAD: >gi|4503875|ref|NP_000809.1| MASPGSGFWSFGSEDGSGDSENPGTARAWCQVAQKFTGGIGNKLCALLYGDAEKPAESGGSQPPRAAARKAACACDQKPCSCSKVDVNYAFLHATDLLPACDGERPTLAFLQDVMNILLQYVVKSFDRSTKVIDFHYPNELLQEYNWELADQPQNLEEILMHCQTTLKYAIK-TGHPRYFNQLSTGLDMVGLAADWLSTANTNMFTYEIAPVFVLLLEYVTLLKKMRI-IGWPGGSGDGIFSPGGAISNMYAMMIARFKMFPEVKEKGMAALPR-LIAFTSEHSHFSLKKGAAALGIGTDSVILIKCDERGMIPSDLER-RILEAKQKGFVPFLVSATAGTTVYGAFDPLLAVADICKKYKIWMHVDAAWGG-GLLSMRKHKWKLSGVERANSVTWNPHKMMGVPLQCSALLVREE-GLMQNCNQMHASYLFQQDKHYDLSYDTGDKALQCGRHVDVFKLWLMWRAGTTGFETH-VDKCLELAELYLYNIKNREGYEMVFDGKQPHTNVCFWYIPPSLRTLEDNEERMSRLSKVAPVIKARMMEYGTMTVSQPLGDKVNFFRMVISNPAATHQDIDFLIEEIERLGQDL

ZNT8A: >gi|64762489|ref|NP_776250.2| MEFLERTYLVNDKAAKMYAFTLESVELQQKPVNKDQCPREPEREESGGMYHCHSGSKPTEKGANEYAYAKWKLCSA-SAICFIFMIAEVVGGHIAGSLAVVTDAHLLID-LTSFLLSLFSLWLSSKPPSKRLTFGWHRAEILGALLSILCIWVVTGVLVYLACERLLY-PDYQIQATVMIIIVSSCAVAANIVLTVVLHQR-CLGHNHKEVQANASVRAAFVHALGDLFQSSISVLISALIIYFKPEYKIADPICTFIFSIL-VLASTITILKDFSILLMEGVPKSLNYSVGKELILAVDGVLSVHSLHIWVSLTMNQVILSAHVATAASRDSQVVRREIAKALSKSFTMHSLTIQMESPVDQDPDCLFCEDPCD

ZNT8B: >gi|289803003|ref|NP_001166282.1| MYHCHSGSKPTEKGANEYAYAKWKLCSA-SAICFIFMIAEVVGGHIAGSLAVVTDAHLLID-LTSFLLSLFSLWLSSKPPSKRLTFGWHRAEILGALLSILCIWVVTGVLVYLACERLLY-PDYQIQATVMIIIVSSCAVAANIVLTVVLHQR-CLGHNHKEVQANASVRAAFVHALGDLFQSSISVLISALIIYFKPEYKIADPICTFIFSIL-VLASTITILKDFSILLMEGVPKSLNYSVGKELILAVDGVLSVHSLHIWVSLTMNQVILSAHVATAASRDSQVVRREIAKALSKSFTMHSLTIQMESPVDQDPDCLFCEDPCD

IA2: >gi|4506321|ref|NP_002837.1| receptor-type tyrosine-protein phosphatase-like N isoform 1 precursor [Homo sapiens]
MRRPRRPGLGGSGGLRLLCLLLSSRPGGCSAVSAHGCLFDRRLCSHLEVCIQDGLFGQCQVGVGQARPLLQVTSPVLQRLQGVLRQLMSQGLSWHDDLTYQVISQEMERIPRLRPPEPRPRDRSGLAPKRPGPAGELLQDIPTGSAPAAQHRLPQPPVGKGGAGASSLSPLQAELLPLLEHLLLPQPPHPSLSYEPALLQPYLFHQFGSRDGSRVSEGSPGMVSVGPLPKAEAPALFSRTASKGIFGDHPGHSYGDLPGPSAQLFQDSGLLYLAQELPAPSRARVPRLPEQGSSSRAEDSPEGYEKEGLGDRGEKPASPAVQPDAAQLRLAAVLAGYGVLELRQLTPEQLSTLLTLQLLPKGAGRNPGGVVNVGADIKTMEGPVEGRDTAELPARTSPMPGHPTASPTSSEVQQVPSPVSSEPPKAARPPVTPVLLEKKSPLGQSQPTVAGQPSARPAEEYGYIVTDQKPLSLAAGVKLL EILAEHVHMSSGSFINISVVGPAALTFRIRHNEQNLSLADVTQQAGLVKSELEAQTGLQILQTGVGQREEAAAVLPQTAHSTSPMRSVLLTLVALAGVAGLLVALAVALCVRQHARQQDKERLAALGPEGAGHDTTFEYQD

LCRQHMA TKSLFNRAEGPPEPSRVSSVSSQFSDAAQASPSSHSTPSWCEEPAQANMDISTGHMILAYME
 DHLNRDR LAK EWQALCAYQAE PNTCATAQGE GNIKKNRHPDFLPYDHARIK LKVESSPSRSDYINASPI
 IEHDPRMPAYIATQGPLSHTIADFWQMVWESGCTVIVMLTPLVEDGVKQCDRYWPDEGASLYHVYEVNVLV
 SEHIWCEDFLVRSFY LKNVQTQETR TLTFHFLSWPAEGTPASTRPLLD FRRKVNKCYGRGRSCPIIVHCS
 DGAGRTGT YILIDMVLNRMAKGVKEIDIAATLEHVRDQRPGLVRSKDQFEFALTAVAE EVNAILKALPQ

IA2_2: >gi|315113878|ref|NP_001186692.1| receptor-type tyrosine-
 protein phosphatase-like N isoform 2 precursor [Homo sapiens]
 MRRPRRPGGLGGSGGLRLLLCLLLLSSRPGGCSAVSAHGCLFDRRLCSHLEVCIQDGLFGQCQVGVGQAR
 PLLQVTSPVLQRLQGVLRQLMSQGLSWHDDL TQYVISQEMERIPRLRPPEPRPRDRSGLAPKRPGPAGEL
 LLQDIPTGSAPAAQHRLPQPPVKGKGAGASSLSPLQAELLPLLEHLLLPPQPPHPSLSYEPALLQPYL
 FHQFGSRDGS RVSEGSPGMVSVGPLPKAEAPALFSRTASKGIFGDHPGHSYGDLPGPSAQLFQDSGLLY
 LAQELPAPSRARVPRLPEQGSSSRAEDSPEGYEKEGLGDRGEKPASPAVQPDAA LQRLA AVLAGYGV ELR
 QLTPEQLSTLLTLLQLLPKGAGRNPGGVVNVGADIKKMEGPVEGRDTAELPARTSPMPGHPTASPTSSE
 VQQVPSPVSS EPPKAARPPVTPV LLEKKSP LGQSQPTVAGQPSARPA AEYGYIVTDQNVVGPALTFRIR
 HNEQNLSLADVTQQAGLVKSELEAQTGLQILQTGVGQREEAAVLPQTAHSTSPMRSVLLTLVALAGVAG
 LLVALAVALCVRQHARQQDKERLAALGPEGAHGDTTFEYQDLCRQHMA TKSLFNRAEGPPEPSRVSSVSS
 QFSDAAQASPSSHSTPSWCEEPAQANMDISTGHMILAYMEDHLNRDR LAK EWQALCAYQAE PNTCATA
 QGE GNIKKNRHPDFLPYDHARIK LKVESSPSRSDYINASPIIEHDPRMPAYIATQGPLSHTIADFWQMVW
 ESGCTVIVMLTPLVEDGVKQCDRYWPDEGASLYHVYEVNVLVSEHIWCEDFLVRSFY LKNVQTQETR TLTFHFLSWPAEGTPASTRPLLD FRRKVNKCYGRGRSCPIIVHCS
 DGAGRTGT YILIDMVLNRMAKGVKEIDIAATLEHVRDQRPGLVRSKDQFEFALTAVAE EVNAILKALPQ

IA2_3: >gi|315113881|ref|NP_001186693.1| receptor-type tyrosine-
 protein phosphatase-like N isoform 3 [Homo sapiens]
 MSQGLSWHDDL TQYVISQEMERIPRLRPPEPRPRDRSGLAPKRPGPAGELLLQDIPTGSAPAAQHRLPQP
 PVGKGAGAGASSLSPLQAELLPLLEHLLLPPQPPHPSLSYEPALLQPYL FHQFGSRDGS RVSEGSPGMV
 SVGPLPKAEAPALFSRTASKGIFGDHPGHSYGDLPGPSAQLFQDSGLLYLAQELPAPSRARVPRLPEQG
 SSSRAEDSPEGYEKEGLGDRGEKPASPAVQPDAA LQRLA AVLAGYGV ELRQLTPEQLSTLLTLLQLLPKG
 AGRNPGGVVNVGADIKKMEGPVEGRDTAELPARTSPMPGHPTASPTSSEVQQVPSPVSS EPPKAARPPV
 TPV LLEKKSP LGQSQPTVAGQPSARPA AEYGYIVTDQKPLSLAAGVKLLEILA EHVHMSSGSFINISVV
 GPALTFRIRHNEQNLSLADVTQQAGLVKSELEAQTGLQILQTGVGQREEAAVLPQTAHSTSPMRSVLLT
 LVALAGVAGLLVALAVALCVRQHARQQDKERLAALGPEGAHGDTTFEYQDLCRQHMA TKSLFNRAEGPPE
 PSRVSSVSSQFSDAAQASPSSHSTPSWCEEPAQANMDISTGHMILAYMEDHLNRDR LAK EWQALCAYQ
 AEPNTCATAQGE GNIKKNRHPDFLPYDHARIK LKVESSPSRSDYINASPIIEHDPRMPAYIATQGPLSHT
 IADFWQMVWESGCTVIVMLTPLVEDGVKQCDRYWPDEGASLYHVYEVNVLVSEHIWCEDFLVRSFY LKNVQ
 TQETR TLTFHFLSWPAEGTPASTRPLLD FRRKVNKCYGRGRSCPIIVHCS
 DGAGRTGT YILIDMVLNRMAKGVKEIDIAATLEHVRDQRPGLVRSKDQFEFALTAVAE EVNAILKALPQ