BIOINFORMATICS ANALYSIS OF MITOCHONDRIAL DISEASE

By

Kieren Lythgow
Bsc (Hons), MRes

Thesis submitted to the University of Newcastle upon Tyne in candidature for the degree of Doctor of Philosophy

2010

Abstract

Several bioinformatic methods have been developed to aid the identification of novel nuclear-mitochondrial genes involved in disease. Previous research has aimed to increase the sensitivity and specificity of these predictions through a combination of available techniques. This investigation shows the optimum sensitivity and specificity can be achieved by carefully selecting seven specific classifiers in combination. The results also show that increasing the number of classifiers even further can paradoxically decrease the sensitivity and specificity of a prediction. Additionally, text mining applications are playing a huge role in disease candidate gene identification providing resources for interpreting the vast quantities of biomedical literature currently available. A workflow resource was developed identifying a number of genes potentially associated with Lebers Hereditary Optic Neuropathy (LHON). This included specific orthologues in mouse displaying a potential association to LHON not annotated as such in humans.

Mitochondrial DNA (mtDNA) fragments have been transferred to the human nuclear genome over evolutionary time. These insertions were compared to an existing database of 263 mtDNA deletions to highlight any associated mechanisms governing DNA loss from mitochondria. Flanking regions were also screened within the nuclear genome that surrounded these insertions for transposable elements, GC content and mitochondrial genes. No obvious association was found relating NUMTs to mtDNA deletions. NUMTs do not appear to be distributed throughout the genome via transposition and integrate predominantly in areas of low %GC with low gene content. These areas also lacked evidence of an elevated number of surrounding nuclear-mitochondrial genes but a further genome-wide study is required.

Contents

List of Figures

List of Tables

Acknowledgements

This work was funded by the Medical Research Council.

Firstly I would like to thank Professor Patrick Chinnery and Dr Peter Andras for their expert supervision throughout the whole of my studies. I would like to thank Patrick for the fantastic opportunities he has provided me with and the opportunity to study a PhD. Patrick has been a great person to work for, extremely helpful, thoughtful and wise whilst maintaining a wicked sense of humour, someone I will always admire and respect. In addition, Peter provided me with continued help throughout my studies and provided intelligent suggestions and advice. Peter co-supervised myself throughout my PhD and his help was crucial. I enjoyed the weekly conversations, both extremely helpful and interesting allowing me to progress with confidence whilst maintaining a sharp determination on the project. I thank Patrick and Peter for making this PhD a deeply interesting and pleasurable experience.

I would like to extend a special thanks to Dr Gavin Hudson who has proof read this thesis several times and been a great help throughout my time in the MRG and a very good friend. He is sadly one of the only blokes left from the old school, when Thursday night footy was followed by cheap pints in The Hancock. These went slightly downhill when the Scottish contingent left, one who sadly gave up scientific research to become a pantomime cat. Dr Lynsey Cree and Dr Angela Pyle for all their help throughout my PhD and for being really good friends, their support is really appreciated. Dr Brendan Payne for his help with statistics in the midst of his indecipherable mumbling. The rest of the PFC lab both past and present who made my time there really enjoyable. All the lads who played Thursday night football, Graham, Paul, Mateuz, Pierre.... I would also like to thank Dr Daniel Swan for his useful advice and tips throughout my time studying bioinformatics. He has always been very responsive and helpful. Dr Michael Barton for all his advice and hilarious conversations to keep the spirits up when we were flagging. Mike ran the best bioinformatics blog at bioinformaticszen.com which gained loads of popularity and was very informative and aided my research in a number of aspects. Everyone in the MRG for the 4 years I was there for making it such a friendly and enjoyable place to work. MRG nights out were legendary especially the all day sessions during the Christmas parties.

All my friends in Newcastle who have been really supportive, who have continually helped me out when times were hard and gave me a place to stay towards the end of my time in Newcastle. Thanks Dave and Karen for everything you did for me. Also Dave

and Ruth, Glenn and Gillian, Fletch and Andrea, Sagey and Lisa, Dan and Joleen.

I would like to extend a very special thank you to my parents David and Judith, whose continued love and support has helped me through all the difficult times during my studies. I am extremely grateful for all their help and certainly could not have done it without their support. Tiffany, Rowan, James and Anne-Marie for all their love and support throughout.

I would like to extend another special thank you to my girlfriend Jo who has been amazing throughout the final year of my studies and whose help and support has been incredible. A big thank you to Jo's family as well Jane, Geraint, Nikki and Sian.

Author's declaration

This thesis is submitted to the degree of Doctor of Philosophy in the University of Newcastle upon Tyne. The research detailed within was performed in the Institute for Ageing & Health under the supervision of Professor Patrick Chinnery and Dr Peter Andras between October 2007 and October 2010 and is my own work unless otherwise stated. I certify that none of the material offered in this thesis has been previously submitted by me for a degree or any other qualification at this or any other university. This copy has been supplied in the understanding that it is copyright material and that no quotation from the thesis may be made without proper acknowledgement.

List of Publications

Kieren T. Lythgow, Gavin Hudson, Peter Andras, Patrick F.Chinnery. In Press. A critical analysis of the combined usage of protein localization prediction methods: Increasing the number of independent data sets can reduce the accuracy of predicted mitochondrial localization. *Mitochondrion*. In Press.

Craig Kate, Takiyama Yoshihisa, Soong Bing-Wen, Jardim Laura B, Saraiva-Pereira Maria Luiza, **Lythgow Kieren**, Morino Hiroyuki, Maruyama Hirofumi, Kawakami Hideshi, and Chinnery Patrick F: Pathogenic expansions of the SCA6 locus are associated with a common CACNA1A haplotype across the globe: founder effect or predisposing chromosome?, *Eur J Hum Genet* 16(7), 841-7, July 2008

Abbreviations

LHON - Leber's Hereditary Optic Neuropathy
NUMTs - Nuclear Mitochondrial Sequences
mtDNA - Mitochondrial DNA
nDNA - Nuclear DNA
KEGG - Kyoto Encyclopedia of Genes & Genomes
BLAST - Basic Local Alignment Search Tool
SOAP - Simple Object Access Protocol
REST - Representational State Transfer
NCBI - National Center for Biotechnology Information
EBI - European Bioinformatics Institute
XML - Extensible Markup Language
WSDL - Web Service Description Language
HTTP - Hypertext Transfer Protocol
EMBL - European Molecular Biology Laboratory
EMBOSS - The European Molecular Biology Open Software Suite
DDBJ - DNA Databank of Japan
SGML - Standard Generalised Markup Language
HTML - Hypertext Markup Language
JDBC - Java Database Connectivity
SQL - Structured Query Language
API - Application Programming Interface
SVM - Support Vector Machine
OMIM - Online Mendelian Inheritance in Man
HMPDb - Human Mitochondrial Database
PDB - Protein Data Bank
mtDB - Human Mitochondrial Genome Database
LC/MS/MS - Liquid Chromotography & Mass Spectrometry
TPR - True Positive Rate
FPR - False Positive Rate
FDR - False Discovery Rate
cFDR - Corrected False Discovery Rate
SGD - Saccharomyces Genome Database
GO - Gene Ontology
GFP - Green Fluorescent Protein
SRS - Sequence Retrieval System

MSI - Mass Spectrometry Instruments
NMR - Nuclear Magnetic Resonance
SD - Standard Deviation
DSBs - Double Strand Breaks
NHEJ - Non-Homologous End Joining
GC - Guanine/Cytosine
AT - Adenine/Thymine
rCRS - Revised Cambridge Reference Sequence (mitochondrial)
SNP - Single Nucleotide Polymorphism
TE - Transposable Element
CSV - Comma Seperated Values
LINES - Long Interspersed Elements
SINES - Short Interspersed Elements
ERVs - Endogenous Retroviruses
LTR (retrotransposons) - Long Terminal Repeat

Chapter 1

Introduction

1.1 Biological background

1.1.1 Mitochondrial evolution

Mitochondria are present in most eukaryotic cells and are commonly thought to have originated as free-living prokaryotes. This is believed to have been the result of a single bacterial endosymbiosis around 2 billion years ago (Gabaldón and Huynen, 2004). Traditionally mitochondria are considered as small separate organelles within the eukaryotic cell but, more accurately described as complex branching networks (Schapira, 2006). Specialised cells involved in high energy dependence such as neurones, cardiac and skeletal muscle cells contain higher levels of mitochondria.

The main role of mitochondria involves ATP production and energy metabolism. Due to the acquisition of ATP/ADP translocase mitochondria have the ability to exchange ATP with the cells cytoplasm (Gabaldón and Huynen, 2004). In addition to energy metabolism mitochondria are involved in various cellular processes and defects have been associated with apoptosis, ageing and a broad range of human diseases including Alzheimers, Parkinsons and diabetes mellitus (Cotter *et al.*, 2004). Human mitochondrial DNA is also responsible for protein synthesis but is totally dependent on the nuclear genome to provide enzymes governing replication, repair, transcription and translation (Schapira, 2006).

1.1.2 The mitochondrial genome

The mitochondrial genome is self replicating and contains double-stranded, circular DNA which in humans is 16,569bps long (Cotter *et al.*, 2004). The majority of proteins involved in mitochondrial function are nuclear encoded, synthesised in the cytosol and then targeted to mitochondria (Elstner *et al.*, 2009). It is estimated that more than 1000 mitochondrial proteins are derived from nuclear genes. Human mitochondrial DNA (mtDNA) is only responsible for encoding 37 genes overall. These include a 12S and 16S rRNA, 22 tRNAs required for protein synthesis and 13 essential genes for oxidative phosphorylation (OXPHOS) polypeptides (Chinnery, 2003; Brandon *et al.*, 2005). Figure 1.1 displays the mitochondrial genome and Figure 1.2 illustrates the intercommunication between the nucleus and mitochondria.

Oxidative phosphorylation is responsible for most of the cellular energy and generates most of the endogenous reactive oxygen species (ROS). This process also regulates apoptosis through the mitochondrial permeability transition pore (mtPTP) (Brandon *et al.*, 2005). Complexes of the OXPHOS system are composed of subunits encoded by both nuclear DNA (nDNA) and mitochondrial DNA. The nuclear-encoded system is responsible for mitochondrial transport and maintenance. The most common respiratory

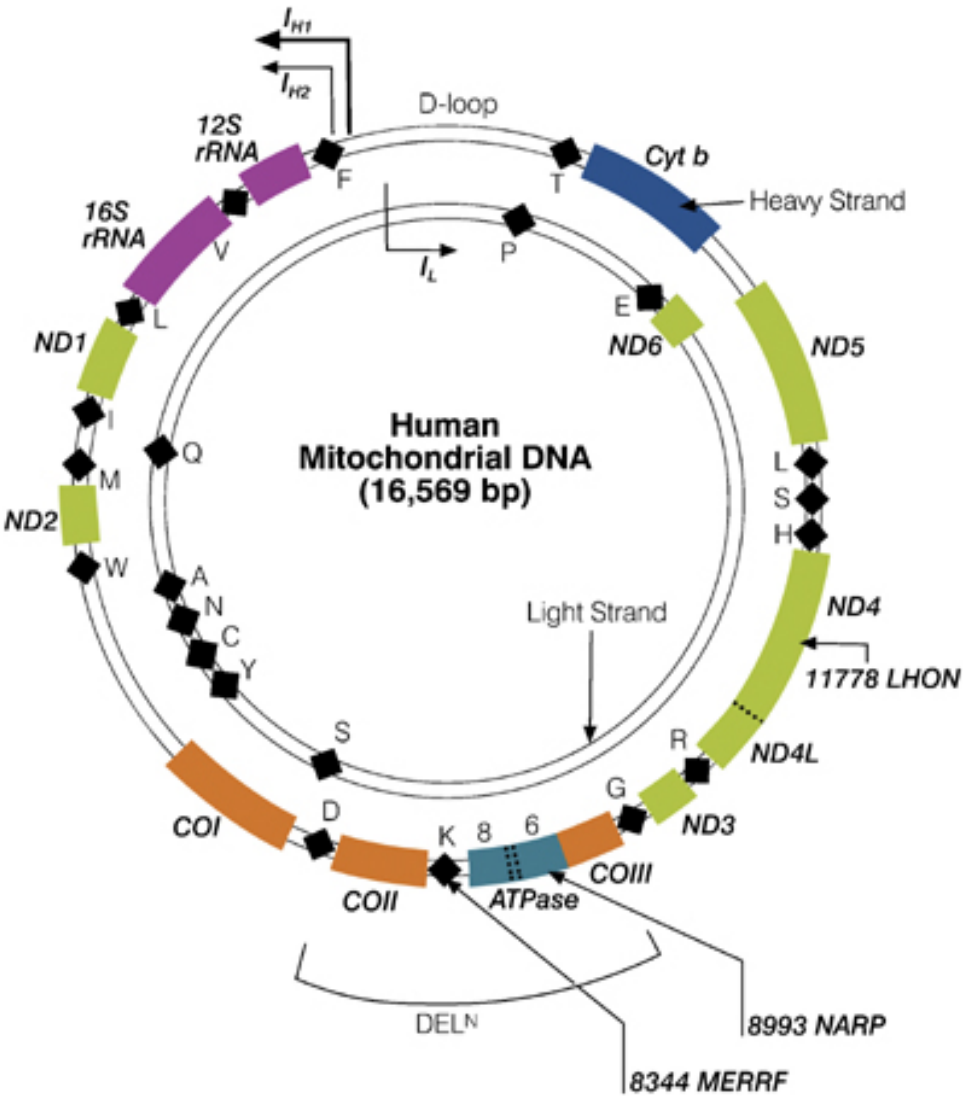

Figure 1.1: The mitochondrial genome.

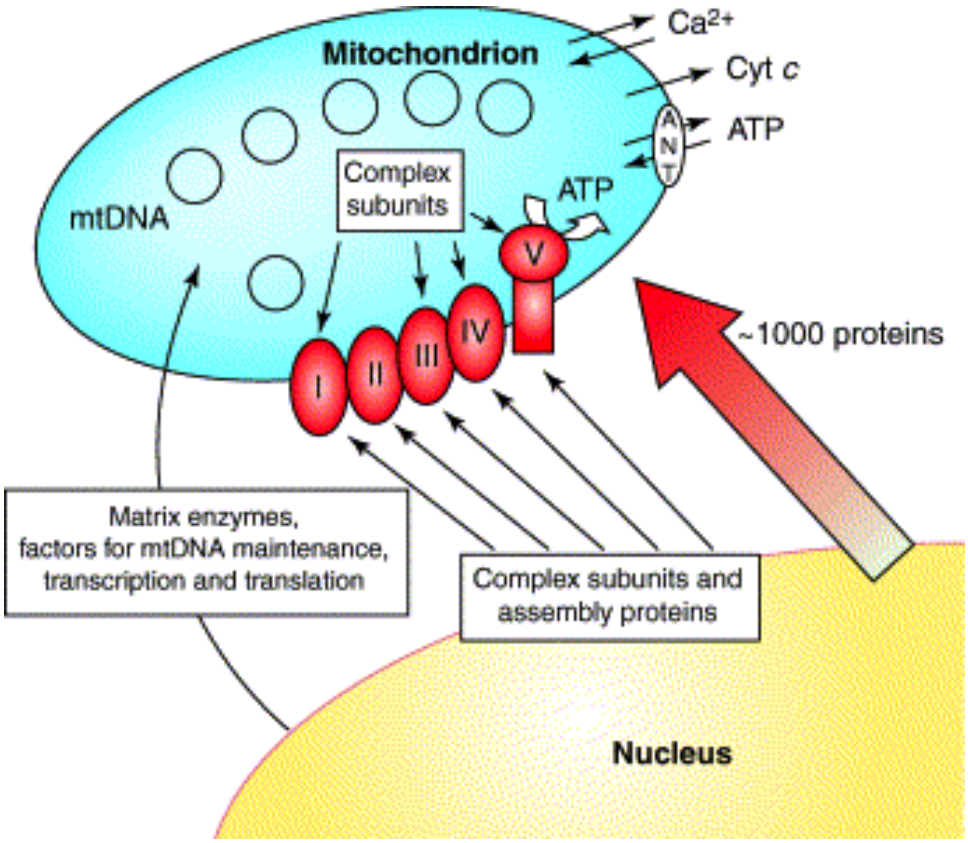

Figure 1.2: Thirteen essential polypeptide respiratory chain complex subunits are synthesised in the mitochondrial matrix from mtDNA. More than 1000 different mitochondrial proteins are synthesised in the cytosol from nuclear DNA taken from Chinnery (2003).

chain defects arise from problems in complex 1 (Petruzzella and Papa, 2002).

1.1.3 Mitochondrial function & disease

Recent epidemiological studies have provided evidence that disorders of the mitochondrial respiratory chain affect at least 1 in 5000 of the population (Schaefer *et al.*, 2004). These figures make these disorders some of the most common genetic diseases (Chinnery, 2003). Currently, there is no effective treatment for these diseases making prevention a priority. Nuclear-mitochondrial disorders are still misunderstood due to the complexities of nuclear-mitochondrial genetic mechanisms. Nuclear-mitochondrial disease genes have been difficult to identify due to a number of problems. In childhood-onset autosomal recessive disease the severe clinical phenotype is rapidly progressive leading to infantile death greatly reducing the size of affected families (Chinnery, 2003).

The potential for conventional gene mapping is limited due to clinical heterogeneity and phenocopies. These factors make it difficult in determining sporadic cases to having the same genetic disorder. Other approaches, including cell complementation techniques and micro-chromosomal transfer are very effective but technically demanding and laborious. Following the mapping of the human genome the list of potential candidates is huge with no apparent indication of which genes are most suitable for further research (DiMauro and Schon, 2001).

Mitochondrial disorders are likely to arise in mutations within genes of unknown function. Disorders could also arise due to mutations in genes with well established function but not thought to affect mitochondria (DiMauro and Schon, 2001). Current lab-based techniques have been used to identify mitochondrial proteins in a number of different organisms. These procedures have contributed a great deal to mitochondrial research. However, each technique is labour intensive and time consuming. This is a major set back due to the rapid developments and exponentially increasing genomic data in the available databases.

Complementary bioinformatics techniques are required for sequence annotation to augment our understanding of mitochondrial disease, protein localisation and protein structure and function. A variety of resources are available that search for mitochondrial targeting signals in sequence data. Comparing the results obtained from these software packages will provide an accurate method for ranking these proteins regarding mitochondrial-relatedness.

1.1.4 Nuclear-mitochondrial diseases

The mitochondrial respiratory chain is the only metabolic pathway in the cell that is under the dual control of the mitochondrial genome and the nuclear genome (DiMauro and Schon, 2001). Mitochondrial DNA mutations are associated with impairment of protein synthesis. However nDNA mitochondrial disorders are much more abundant due to nDNA being responsible for several processes (Dimauro and Davidzon, 2005). These processes are (i) synthesis of assembly proteins; (ii) intergenomic signaling; (iii) mitochondrial import of nDNA-encoded proteins; (iv) synthesis of inner mitochondrial membrane phospholipids; (v) mitochondrial motility, fission and fusion.

1.1.5 Nuclear DNA mutations causing disease

Mitochondria have lost most of their autonomy through evolution and are now largely nuclear encoded. Nuclear DNA is responsible for numerous factors that are critical for mitochondrial transcription, translation and replication (van den Heuvel and Smeitink, 2001). Table 1.1 displays a variety of nDNA defects caused by mutations arising in genes associated with the complexes of the OXPHOS system. Nuclear DNA is essential for the correct assembly of respiratory chain complexes and is associated with a number of disorders including Leigh syndrome (Tiranti *et al.*, 1995). Various factors encoded by nDNA are essential for mtDNA integrity and replication. Mitochondrial diseases can arise due to defects in protein transport of nDNA-encoded proteins from the cytoplasm into mitochondria (Zeviani *et al.*, 2003). However the defects in protein transport is an area that is not well understood. Figure 1.3 displays a diagram of the OXPHOS pathway taken from the KEGG database, which illustrates the respiratory chain is embedded in the lipid bilayer of the inner mitochondrial membrane. Mitochondria are motile within the cell and can divide by fission or fuse with other mitochondria. Diseases such as autosomal dominant optic neuropathy can arise due to defects that affect these essential processes (Yu-Wai-Man *et al.*, 2009). Mutations in mtDNA associated with the OXPHOS system are only encountered in approximately 5% of patients with mitochondrial disease (van den Heuvel and Smeitink, 2001). The majority are caused by nDNA defects within the respiratory chain. Complex II subunits are exclusively encoded by nDNA. Subunits I, III and V are encoded by either mtDNA or nDNA. The 70 nuclear gene products associated with the OXPHOS system together with mtDNA products combine to form the five complexes.

Structural components of RC complexes

Protein	Function	Phenotype
NDUFS4	Complex I	Atypical Leigh s.
NDUFS8	Complex I	Leigh s.
NDUFVI	Complex I	Leukodystrophy, myoclonus, Leigh s.
NDUFS1	Complex I	Leigh s.
NDUFS7	Complex I	Leigh s.
Flavoprotein	Complex II	Leigh s.
SDHD, SDHC	Complex II	Hereditary paraganglioma
Synthesis of CoQ10	Complex I, II, III	Ataxia, myopathy, seizures

Factors controlling OXPHOS or mtDNA metabolism

Protein	Function	Phenotype
SURF1	COX assembler	Leigh s.
SCO1	COX assembler, copper metabolism	Infantile encephalopathy
SCO2	COX assembler, copper metabolism	Infantile cardiomyopathy
COX10	COX assembler heme A synthesis	Infantile encephalopathy
BCS1	Complex III assembly	Infantile encephalopathy, tubulopathy, hepatopathy
ANT1	Nucleotide pool	adPEO, chr 4q
Twinkle	Helicase/primase nucleotide pool?	adPEO, chr 10q
Thymidine phosphorylase	Nucleoside pool	MNGIE

Mitochondrial proteins indirectly related to OXPHOS

Protein	Function	Phenotype
Tim 8/9	Transporter of carrier proteins	X-linked deafness-dystonia s. (Mohr-Tranebjaerg s.)
ABC7	Iron exporter	X-linked ataxia/sideroblastic anaemia s.
Frataxin	Iron storage protein	Friedreichs ataxia
Paraplegin	Metalloprotease, involved in protein turnover	Hereditary spastic paraplegia
OPA1	Dynamin-related protein, possibly involved in mitochondrial division and biogenesis	Autosomal dominant optic atrophy

Table 1.1: Nuclear mitochondrial proteins causing disease taken from Zeviani (2001)

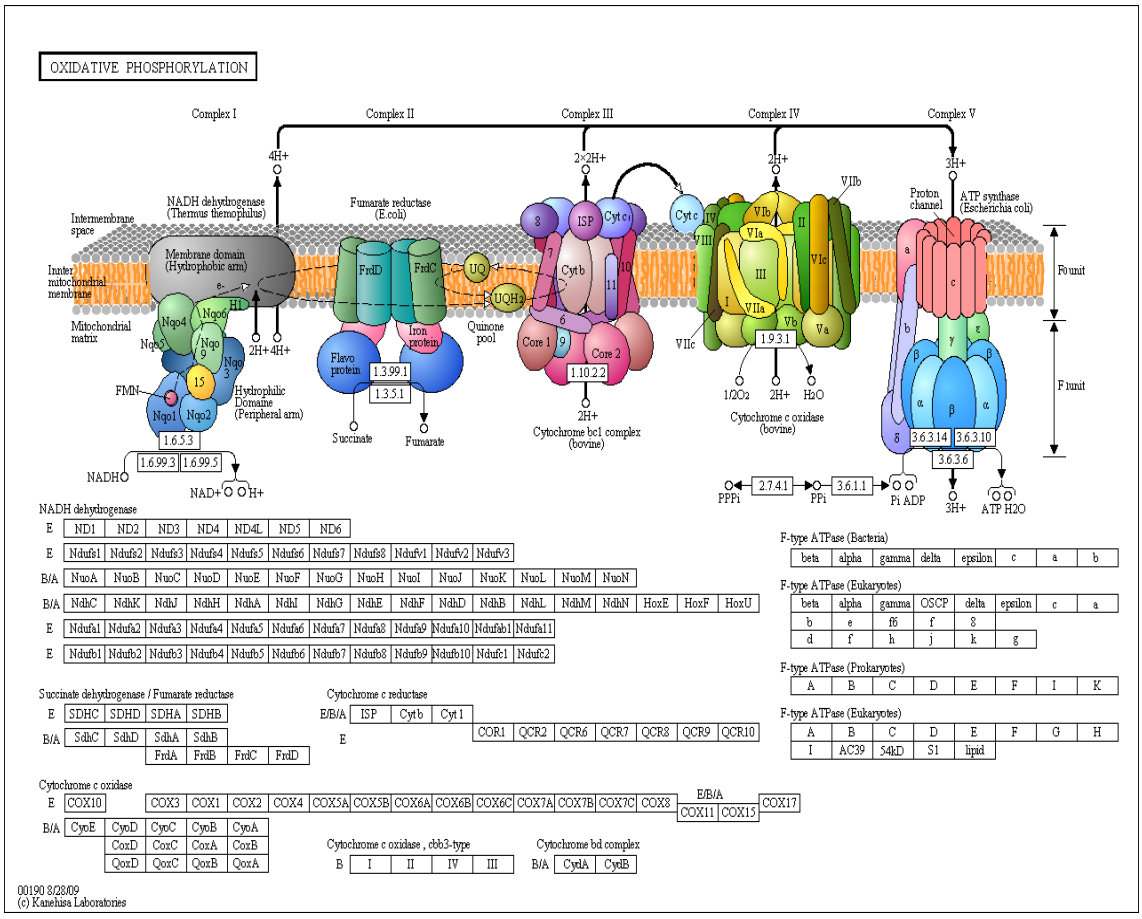

Figure 1.3: The OXPHOS system consists of electron acceptors, coenzyme Q, cytochrome C and five multisubunit protein complexes (I-V). Around 70 nuclear gene products are associated with the OXPHOS system (Petruzzella and Papa, 2002).

Most of these nuclear-encoded genes have been characterised in humans and appear to display random distribution over the chromosomes with no evidence of clustering (van den Heuvel and Smeitink, 2001).

1.2 Technical background

1.2.1 E-science and bioinformatics

Bioinformatics has revolutionised modern biology increasing the rate of biological analysis considerably resulting in large quantities of genomic and proteomic data being produced. A multitude of services are available on the web for bioinformatics analysis such as tools to predict sequence similarity (BLAST), protein structure/function (InterProScan) and multiple sequence alignments (ClustalW). A lab-based biologist can now fully appreciate the wealth of information bioinformatics can generate providing more time for actual scientific investigation and planning (Wolstencroft *et al.*, 2007). e-Science developments have greatly facilitated bioinformatics analysis considerably generating vast amounts of data. This provides bioinformaticians with more time to effectively analyse the data being produced from large *in silico* experiments (Wroe *et al.*, 2004). Even the simplest workflows can produce vast quantities of information that must be analysed effectively. It is at this point a considerable amount of time must be spent on effectively analysing the relevant data and allowing scientists to direct their efforts into their main areas of interest. Using workflows to generate data is very effective. The main problems lie with the reliability of distributed open software and the varying data formats that exist (Goble and Stevens, 2008).

1.2.2 Web services

Web services allow programmatic access to tools available on the web. These have become crucial applications in bioinformatics research enabling scientists to build analysis protocols consisting of distributed resources (Pettifer *et al.*, 2009). These resources consist of many diverse biological analysis programs and biological databases available through web interfaces. A modern definition of the term *web service* refers to any method of programmatic access over the underlying technologies of the web (Stockinger *et al.*, 2008). Various web service technologies have recently been introduced including Web 2.0, Service Oriented Architectures (SOA), SOAP (Simple Object Access Protocol) and REST (Representational State Transfer) (McWilliam *et al.*, 2009; Pettifer *et al.*, 2009). A number or projects are dedicated to the documentation and management of available bioinformatics web services that are currently available such as the

EMBRACE Registry and BioCatalogue that are now merged (Belhajjame *et al.*, 2008). The BioCatalogue web service registry can be seen in Figure 1.4. Common bioinformatics web services consist of BLAST based at the NCBI and InterProScan at the EBI.

Apache Axis and Tomcat

Apache Axis is an implementation of SOAP (Simple Object Access Protocol). SOAP is an XML-based information exchange protocol consisting of three parts: An envelope defining a framework for describing the contents of a message and its processing procedure; a set of encoding rules for expressing instances of application-defined data types and a convention for representing remote procedure calls and responses (APACHE AXIS). Axis also includes a simple standalone server; a server that plugs into servlet engines such as Tomcat; support for the Web Service Description Language (WSDL) tool that generates Java classes from WSDL documents; sample programs and a tool that monitors TCP/IP packets (APACHE AXIS). Apache Tomcat is a standalone web server for web service development. Tomcat provides HTTP functionality needed for web services and redirects requests for the running of web services to the axis installation.

1.2.3 The myGrid Project

The large increase in scientific data requires distributed global collaborations enabled by the internet. Large scale computing resources are required for the integration of this data held across multiple sites around the world (Stevens *et al.*, 2003). The myGrid consortium was founded in 2001 based on a collaboration comprising five Universities (Manchester, Southampton, Newcastle, Nottingham and Sheffield), the European Molecular Biology Laboratory and European Bioinformatics Institute (EMBL-EBI) and various industrial partners consisting of GlaxoSmithKline, Merck KGaA, AstraZeneca, Sun Microsystems, IBM, GeneticXchange, Epistemics and Cerebra. This consortium is currently funded by the EPSRC and comprises the Universities of Manchester and Southampton. The myGrid team consists of bioinformaticians, computer scientists and software engineers. A large proportion of the bioinformaticians have a life science background as myGrid focuses on biological research as a use case. However, much of the middleware developed is generically applicable across the scientific domains including astronomy (AstroGrid, HELIO), chemistry (CDK, CICC) and medical imaging (MIASGrid). myGrid aims to develop generic middleware to allow biologists to perform in silico experiments including the investigation of several diseases.

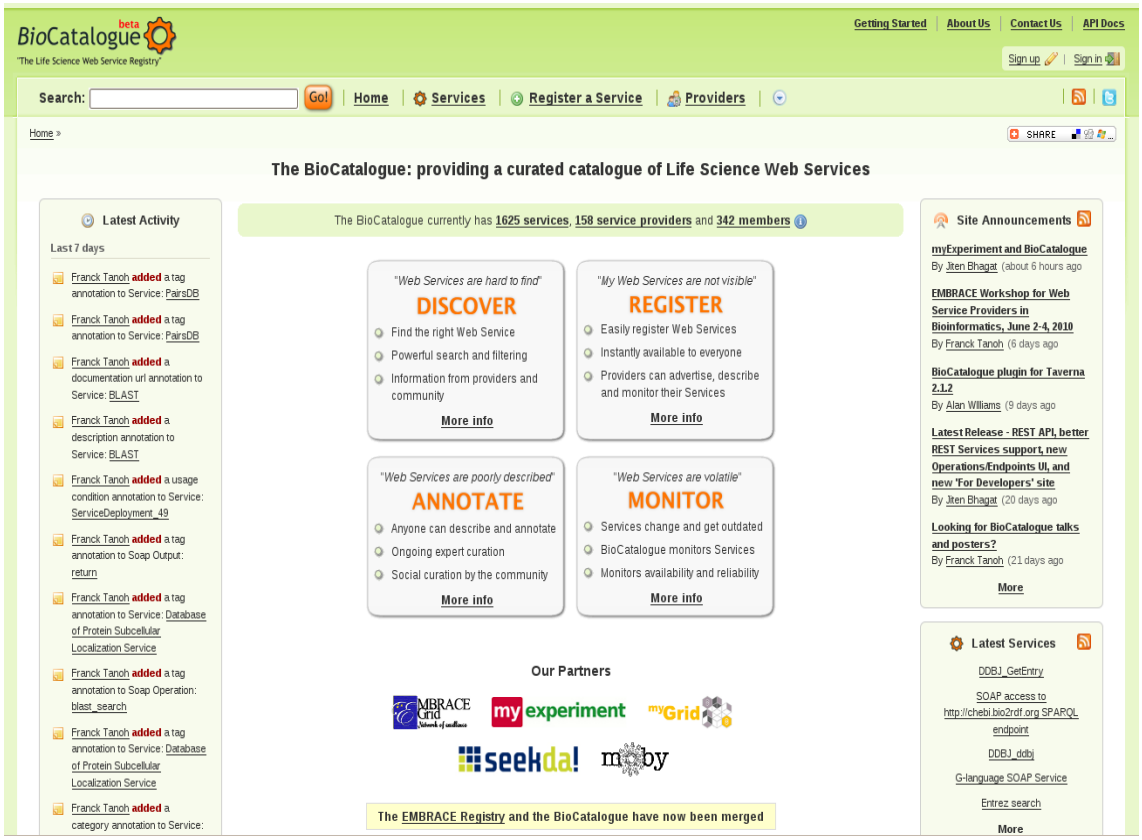

Figure 1.4: BioCatalogue registry for the discovery, registry, annotation and monitoring of life science web services.

At Newcastle University researchers have used myGrid technology to study the genetic mechanisms of Grave's Disease (Li *et al.*, 2003). The University of Manchester have focused on Williams-Beuren Syndrome (WBS) (Stevens *et al.*, 2004) and Liverpool University performed candidate gene analysis for trypanosomiasis in cattle (Fisher *et al.*, 2009). Numerous projects are currently implementing myGrid methods including ONDEX, a project that aims to solve technical and semantic heterogeneities between diverse biological datasets for integration and visualisation purposes (Köhler *et al.*, 2006). Others examples include Obesity e-Lab which aims to develop an e-science toolkit to investigate obesity, REFINE which aims to produce text mining procedures applied to systems biology, and e-LICO that aims to provide a virtual laboratory for data mining. Until recently, in silico experiments were time-consuming processes that commonly suffered from incompatibility problems. Results and methods were often poorly documented causing a lack of reproducibility. The myGrid project consists of a toolkit comprised of core components allowing the formation, execution, management and sharing of discovery experiments. These experiments are developed by bioinformaticians using the Taverna Workbench which enables the user to create and run workflows using the Freefluo enactor engine (Oinn *et al.*, 2004b). Workflows are the main mechanism for conducting the experiments. These can incorporate many services, Java applications and database software. myGrid has developed open source high-level service based middleware consisting of technologies from the semantic web (Wolstencroft *et al.*, 2007). A collection of tools have been developed by myGrid most notably Taverna which allows users to develop workflows to conduct large scale *in silico* experiments and myExperiment - a social networking utility for sharing, reusing and augmenting workflows.

1.2.4 The Taverna workbench

Taverna 1.7 includes a workbench application that can be used to create workflows written in a language called Scufl. Scufl is an abbreviation of Simple conceptual unified flow language and each step in the workflow executes a specific function (Oinn *et al.*, 2004b). Workflows allow bioinformaticians to conduct large in silico experiments drawing upon a variety of local and distributed web resources. These allow the storing and tracking of provenance data and can be repeatedly executed or manipulated. Taverna allows a range of abstraction levels allowing users to interact with individual services in detail Hull *et al.* (2006). Using the Scufl workflow language allows workflows to be assembled and refined for a particular experiment, enacted using the relevant services and recorded for provenance alongside the experimental data. Many of the available resources are from a number of databanks based at the EBI, EMBOSS, NCBI and

DDBJ. The EBI has developed SoapLab, which is a set of web services providing access to many applications on remote computers (Senger *et al.*, 2003). SoapLab has several advantages its uniform method for describing analyses and their input/output data by an XML-based metadescription.

XML (eXtensible Markup Language) is derived from SGML (Standard Generalised Markup Language). XML overcomes most of the limitations experienced with HTML (HyperText Markup Language). Data in bioinformatics is complex to model due to the multitude of data formats and their numerous relationships (Szomszor *et al.*, 2005). New data types emerge at a constant rate including genome sequences, microarrays, protein-protein interaction maps and proteomics data (Achard *et al.*, 2001). Inferring data creates new data that also needs to be integrated and raw data must be stored. Bioinformatics databases such as the NCBI and EMBL are updated frequently with a constant flow of data exchange. Therefore a language must express power and scalability at run time and be flexible (Achard *et al.*, 2001). XML is highly flexible and internet-oriented with the ability to successfully combine data. This is highly useful for database interconnection. However, not all biological databases provide an XML view of their data. Concerns have arisen about the technological scalability of XML and if this is sufficient for molecular biology (Achard *et al.*, 2001). Overall it appears XML has many advantages compared to disadvantages. It is essential to store data outputs from a workflow into a database and the same outputs are required as inputs for further analysis programs in the workflow.

The Taverna workbench provides an interface for creating and enacting scientific analysis workflows. For biological purposes, a workflow can be exemplified by an *in silico* experiment designed to perform sequential tasks involving data retrieval, analysis, integration and storage. Taverna allows researchers to access numerous distributed services and database repositories enabling the construction of large analysis pipelines. Currently, there are in excess of 3500 services available through the Taverna workbench with a rapid increase in computational applications dedicated to DNA, RNA and protein analysis (Hull *et al.*, 2006; McWilliam *et al.*, 2009). In bioinformatics, resources can consist of information repositories such as EMBL and SwissProt or computational analysis tools including BLAST and ClustalW. An *in silico* experiment frequently involves a combination of these resources linked in a specific order, thus forming a workflow process (Oinn *et al.*, 2004b).

A typical bioinformatics experiment could consist of the following protocol:

1. **Retrieve a DNA sequence in FASTA format from an online repository such as EMBL.**
2. **Perform a gene identification analysis on the DNA sequence using the gene prediction software GENSCAN.**
3. **Any resulting amino acid sequence can be compared to a database of existing proteins using BLASTP to investigate potential homology revealing possible roles of protein function.**
4. **Perform a multiple sequence alignment on specific sequences of interest against the query sequence using ClustalW.**
5. **Produce a graphical display of a phylogenetic tree visualising the evolutionary relationships between the sequences using Phylip.**
6. **Store relevant results into a relational database using JDBC (Java Database Connectivity) and SQL (Standard Query Language).**

Performing the aforementioned procedure manually would take a bioinformatician a long time, especially for multiple sequence queries. This approach would also be difficult regarding repeatability and prone to human error. This would typically involve manually transferring results between services by noting values produced and re-keying them into a new interface by cutting and pasting. Although problematic and error prone, this method facilitated scientific investigation through experimentation and underpins web service and grid technologies (Wroe *et al.*, 2004). New types of services arise rapidly in the bioinformatics community. Oinn *et al.* (2004b) developed the Taverna project as an open source software tool enabling scientists to orchestrate bioinformatics web services and existing bioinformatics applications into workflows for the life sciences community.

The Taverna interface (Figure 1.5) consists of several frames representing specific functionality. The Available Processors window lists all the services that can be incorporated into a workflow. These are distributed resources residing at various research institutes including the EBI, NCBI, KEGG and DDBJ and numerous local processors involving string manipulation and conditional statements. Various scripting shells are available including facilities for writing beanshells and R scripts to include into an analysis pipeline.

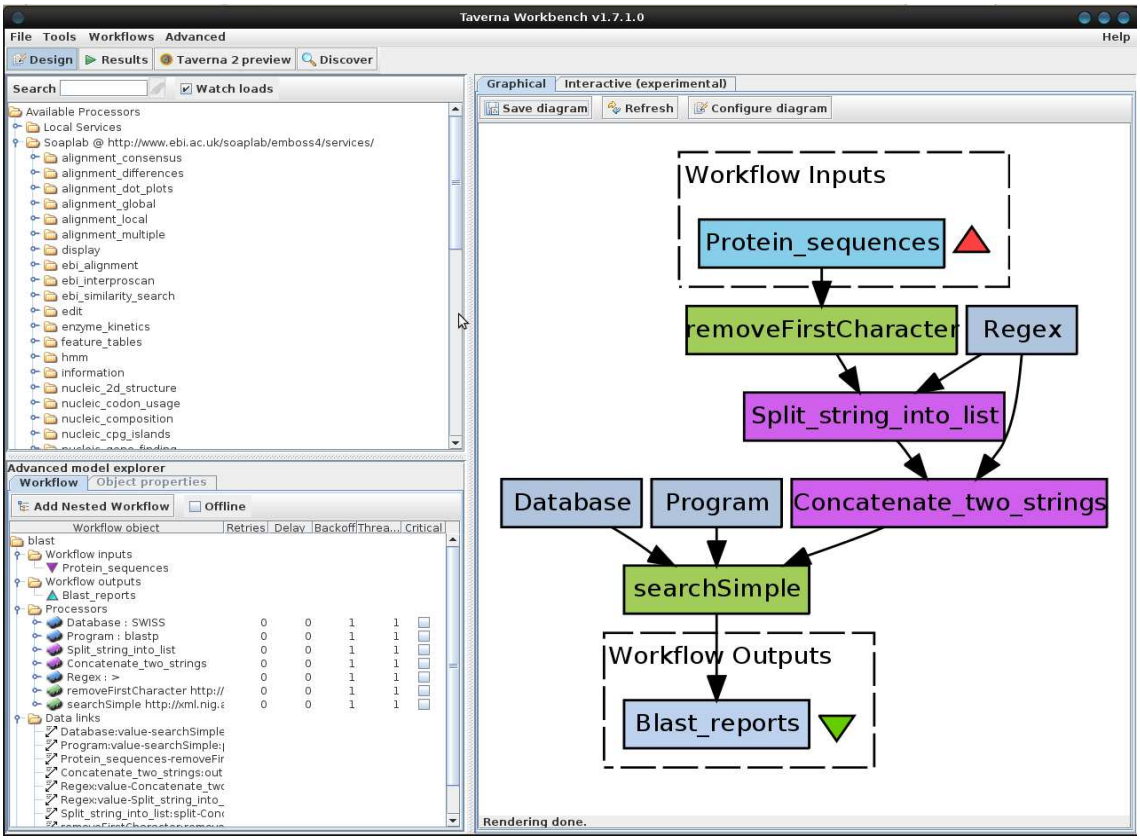

Figure 1.5: The Taverna Workbench interface enabling the development of workflow experiments.

In addition, external services can be introduced into Taverna by scavenging APIs, WSDL documents or workflow XML files enabling the import of predefined workflows. Oinn *et al.* (2004a) discuss the architecture of Taverna in detail and document the functionality of the workflow software. The Advanced Model Explorer window provides the user with the ability to physically construct the workflow by connecting the relevant input and output ports for each processor. For example, one processor make retrieve a nucleotide sequence from a database and pass this result to a processor designed to convert this into an amino acid sequence (Figure 1.6). Tracking provenance data is essential when conducting in-silico experiments as this identifies the data origin by recording the metadata and intermediate results associated with the workflow (Oinn *et al.*, 2004b). This allows scientists to track down any anomalies in their results and find information on how the data was processed. Due to the vast quantities of web-based tools and databases, workflows are highly useful resources.

1.2.5 MyExperiment - Sharing workflows

Another aspect of the myGrid consortium is myExperiment which is a virtual research environment allowing users to publish and share scientific workflows developed within the Taverna Workbench. The myExperiment repository is also available as a in-built service within the most recent versions of Taverna 2. The myExperiment environment now has over 3000 users and over 1000 workflows aiming to reduce reinvention, propagate best practice and allow scientists to concentrate on important research (De Roure *et al.*, 2009; De Roure and Goble, 2009). De Roure *et al.* (2009) claims the myExperiment draws parallels with other familiar social networking websites publicly available allowing users an immediately understandable interface for ease of workflow sharing and reuse. Figure 1.7 displays the user interface enabling the deposition of personally developed workflows and ability to provide workflow descriptions and user instructions.

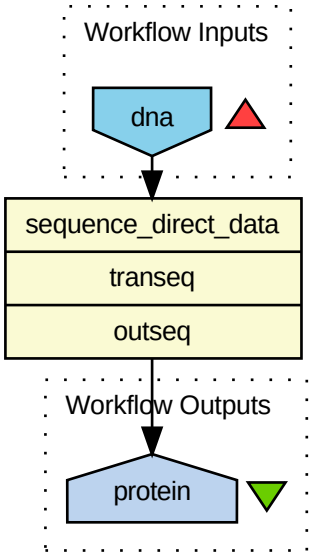

Figure 1.6: A basic Taverna workflow that converts a DNA sequence into protein.

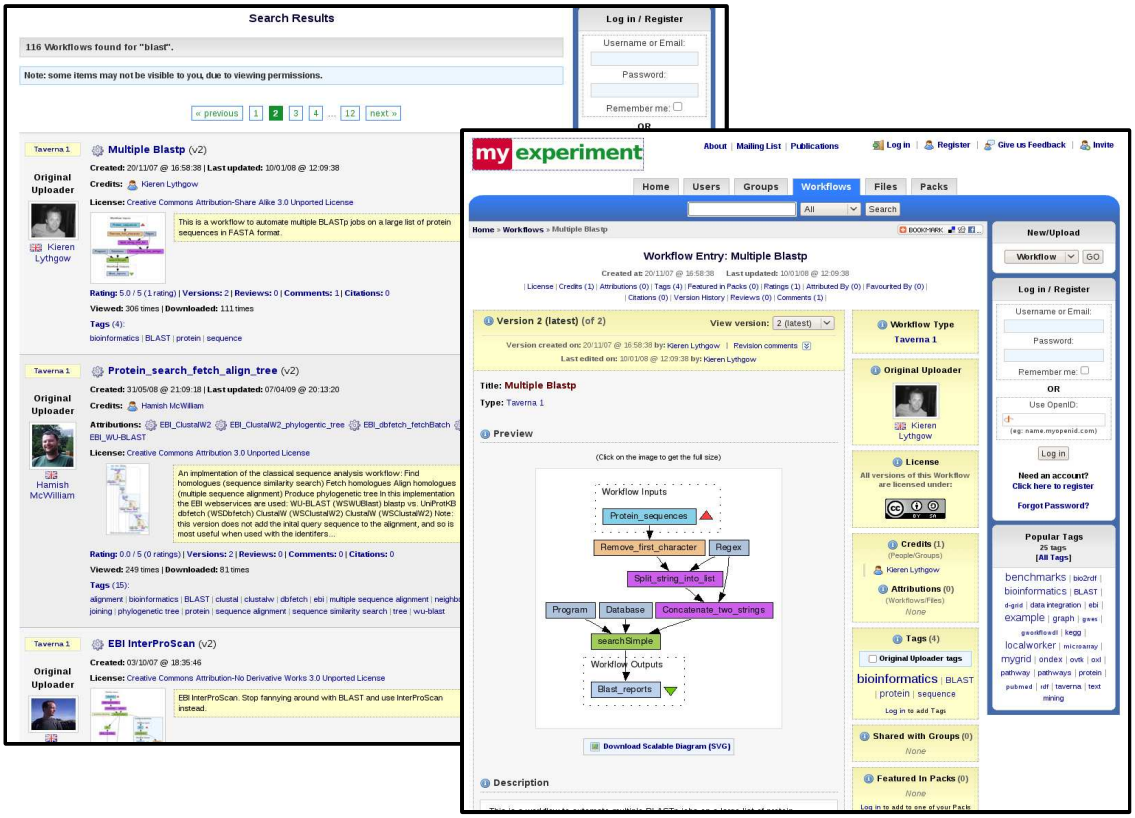

Figure 1.7: The myExperiment social networking platform. Users can search pre-existing workflows and upload workflows to share with the myExperiment community.

1.3 Aims and Objectives

The aims and objectives of this thesis are:

- To systematically analyse all combinations of suitable mitochondrial prediction methods to highlight the most complementary techniques. Machine learning will applied in the form of a trained support vector machine (SVM) on a dataset of identified mitochondrial and non-mitochondrial proteins. An automated workflow using the Taverna workbench 1.7 will be used to implement repetitive testing to assess the sensitivities and specificities of all the combinations within the investigation. Previous research has failed to provide evidence of rigorous multiple testing and the standard deviations produced when this is applied. The aim of this investigation is to perform large scale multiple testing of all combinations of appropriate datasets and the standard deviations produced between each test.

- Identify candidate genes involved in LHON by interrogating mitochondrial prediction databases MitoCarta and MitoSVM. In addition, a text mining algorithm will be developed to screen gene ontology records and OMIM documents for specific keywords related to LHON. Orthologues of closely related species will also be text mined for novel candidates using innovative techniques not previously utilised for this disease.

- Compare a large existing database of known mitochondrial DNA (mtDNA) deletions to mtDNA fragments that reside in the nuclear genome. This will potentially reveal any correlation regarding the mechanisms governing mtDNA deletions and NUMTs within the nuclear genome. Flanking regions will be analysed for transposable elements, nuclear mitochondrial genes and GC content. This aims to reveal novel information about the mechanisms governing mtDNA integration and the evolutionary processes involved.

Chapter 2

Systematic evaluation of mitochondrial protein prediction methods

Abstract

Background

Disorders of the mitochondrial respiratory chain affect approximately 1 in 5000 of the population making these disorders some of the most common genetic diseases. The mechanism of nuclear-mitochondrial interactions is still misunderstood. The implementation of bioinformatics techniques is essential to further our understanding of mitochondrial disease. Recent advances in e-science technology have provided mechanisms for collating and interrogating large volumes of distributed data across the World Wide Web. More specifically, workflow technology has experienced a huge increase in support and allowed developers to construct sophisticated in silico analysis pipelines for scientific research. Machine learning applications have provided a powerful mechanism for analysing heterogeneous biological data. The development of a workflow comprising these methods for candidate gene analysis has proved to perform well in comparison to classical techniques. Previous research has suggested using an integrated genomics approach for the identification of mitochondrial proteins. However, these applications of machine learning methods did not put an emphasis on evaluating the validity of their predictions and in particular, did not assess the variance (or standard deviation) of prediction performance as data sets and model parameters change. From this analysis of all combinations of mitochondrial classifiers from the integrated approach a reduction appears to augment sensitivity and specificity. This suggests certain classifiers are more influential than others.

Results

An optimum combination of prediction methods was found consisting of 7 specific classifiers achieving the highest mean sensitivity without compromising specificity. Incorporating additional classifiers reduced the mean sensitivity highlighting the inadequacies of previous research. This was also subject to rigorous multiple testing using machine learning to reveal the standard deviations for each combination. Mitodomain was found to be the strongest classifier in contrast to ancestry which was found to be the weakest. The study emphasises the importance of carefully selecting prediction methods to achieve the most accurate prediction results for mitochondrial genes.

Conclusion

The combination achieving the highest mean sensitivity was used to construct a model that was implemented in a genome-wide prediction for human mitochondrial genes. This allows clinicians to focus on high scoring candidates within their region of interest. This may lead to the discovery of novel mitochondrial genes involved in disease.

2.1 Introduction

Mitochondrial dysfunction is a major cause of human genetic disease, affecting at least 1 in 5000 of the population (Schaefer *et al.*, 2004). There is also emerging evidence that a genetic predisposition of a more subtle nature contributes to the risk of developing a number of complex human traits, particularly neurodegenerative diseases such as late-onset Parkinsons disease (Chinnery, 2003). Experimental approaches to these biological problems are currently restricted by rudimentary methods for analysing and interpreting existing and newly acquired data sets. The challenge of this project is to harness currently available bioinformatic and computational tools to develop new analytical approaches in a number of related areas. Although the last decade has seen major advances in molecular diagnosis of families with mitochondrial disorders, this is not possible in a large proportion of cases where the inheritance pattern implicates a nuclear genetic cause (Parfait *et al.*, 1997). Mitochondria are thought to contain in excess of 1000 proteins, but an accurate mitochondrial proteome has not been determined experimentally (Chinnery, 2003). A number of bioinformatic tools have been developed to help predict whether a protein localises to mitochondria, but the sensitivity and specificity of different combinations of these prediction tools has not been investigated (Calvo *et al.*, 2006; Shen and Burger, 2007; Pagliarini *et al.*, 2008). Determining the best combination of existing tools, and constructing a continually developed database of likely mitochondrial proteins with the genetic location of their corresponding genes, would provide an invaluable tool for identifying novel nuclear-mitochondrial disease genes.

2.1.1 Machine learning

2.1.2 Machine learning applications in bioinformatics

Machine learning is an extremely powerful method for the analysis of large complex datasets. This is especially useful when applied to molecular biological data. The Human Genome Project has resulted in an unprecedented amount of biological information that is impossible to analyse manually (Venter *et al.*, 2001). This information is hugely varied ranging from protein and DNA sequences to biomedical literature. The method of machine learning provides a mechanism for training an algorithm to recognise specific biologically meaningful patterns within the data. This trained system can then be tested for predictive accuracy when analysing unseen data.

Numerous examples exist regarding machine learning applications in bioinformatics. Figure 2.1 displays the six main biological domains where these techniques are being applied consisting of genomics, proteomics, microarray data, systems biology, evolution and text mining (Larrañaga *et al.*, 2006). The learning term refers to the invocation of an algorithm or computer program resulting in a predictive model by incorporating training data or past experience. Specific problems in bioinformatics require different techniques in machine learning for analysis. A variety of machine learning methods are available, popular in the field of bioinformatics research including Hidden Markov Models, decision trees and support vector machines (Ling *et al.*, 2005).

2.1.3 Support vector machines

Support vector machines (SVMs) are a group of supervised machine learning methods used for classification and regression developed by Vladimir Vapnik (1999). An SVM aims to construct a separating hyperplane within high dimensional data. SVMs are also known as maximum margin classifiers as they attempt to maximise the distance between two labelled groups (Figure 2.2). The margin is defined by any positive distance from the decision hyperplane (Larrañaga *et al.*, 2006). Support vectors are the training samples that define the optimal separating hyperplane. The use of SVMs are particularly widespread and recent applications include isolated handwritten digit recognition, object recognition, speaker identification, charmed quark detection, face detection in images and text categorisation (Burges, 1998).

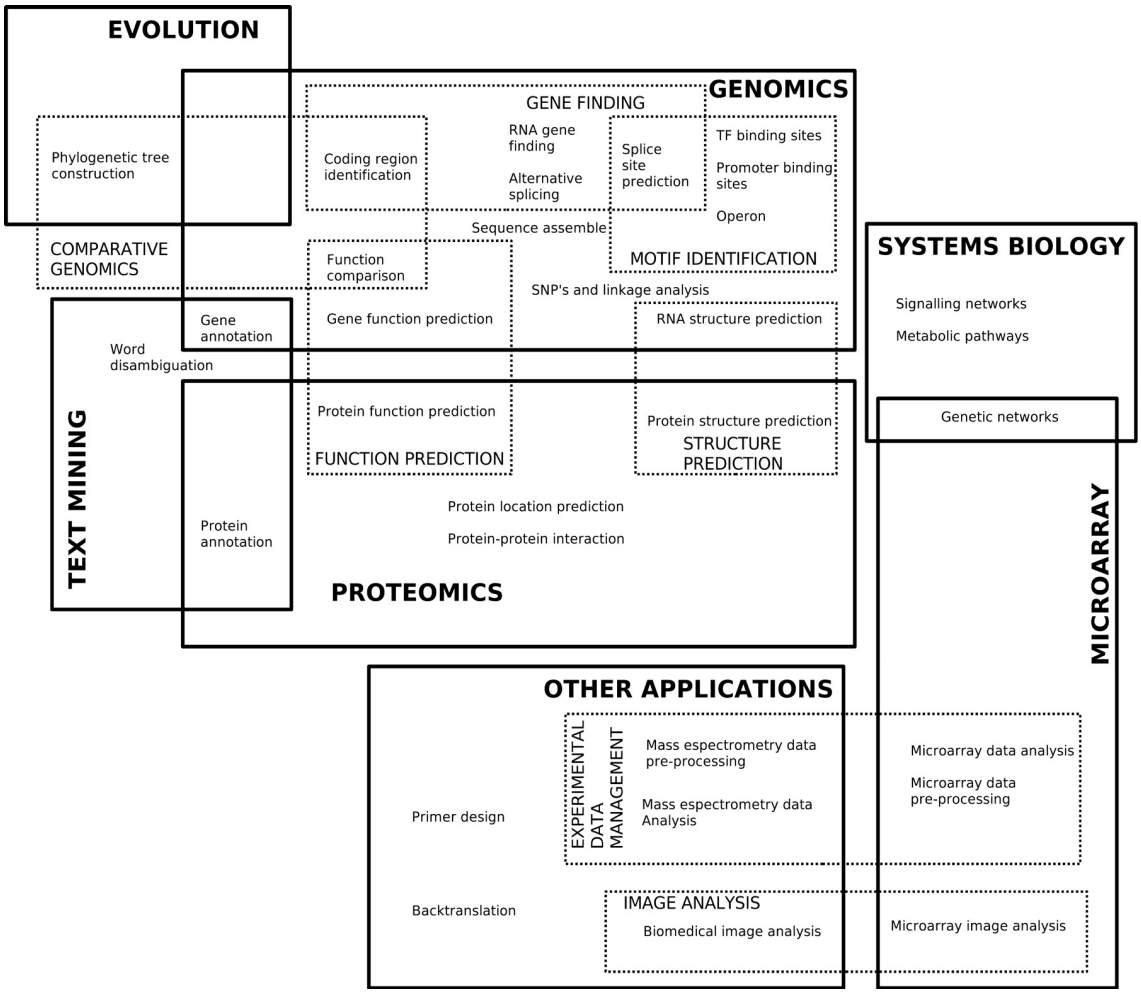

Figure 2.1: Classification of areas machine learning has been applied in bioinformatics from Larrañaga *et al.* (2006)

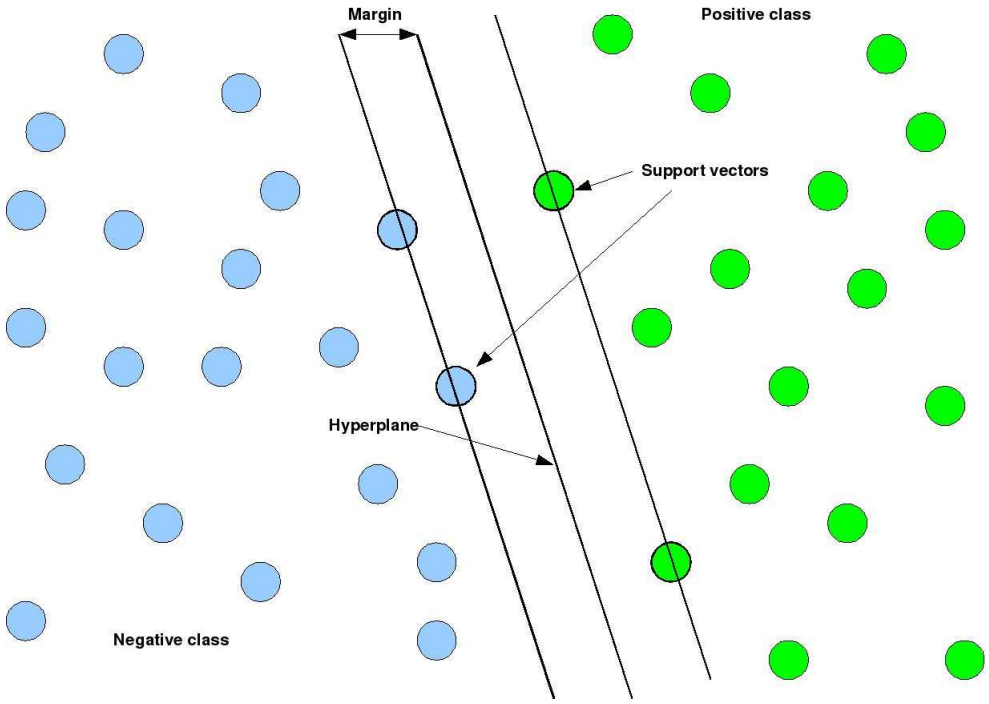

Figure 2.2: Support vector machines construct a separating hyperplane between two distinct groups of data in a high dimensional feature space. The margins are set by support vectors comprising the largest distance from the separating hyperplane in order to maximise their separation.

Many biological problems involve high dimensional, noisy data and SVMs have proven to perform well in comparison to other statistical or machine learning methods (Noble, 2006). In addition, kernel-based methods including SVMs have the ability to handle non-vector inputs including variable length sequences or graphs. SVMs offer several advantages as opposed to previous methods such as hierarchical clustering and self-organising maps (Brown *et al.*, 2000). Brown *et al.* (2000) found that SVMs employ distance functions in extremely high dimensional feature space allowing these to implicitly measure correlations within gene expression data. These characteristics of SVMs make them highly suited for classification of heterogeneous biological data. More specifically examples of very recent applications of SVMs in bioinformatics include biological sequence classification (Sonnenburg *et al.*, 2008), active compound detection for drug discovery (Cao *et al.*, 2008), identification of gene-disease associations (Ozgür *et al.*, 2008), protein fold and superfamily recognition (Melvin *et al.*, 2007), disease-related nsSNP detection (Capriotti *et al.*, 2006) and protein-protein binding site prediction (Bradford and Westhead, 2005).

2.1.4 Sublocalisation prediction software

Various software programs have been developed to predict sublocalisation of proteins within the cell based on novel algorithms analysing amino acid sequences (Claros and Vincens, 1996; Emanuelsson *et al.*, 2000). Most proteins in the eukaryotic cell are encoded in the nucleus and synthesised in the cytosol requiring additional sorting regarding their organelle destination (Kurland and Andersson, 2000). A protein import mechanism is required to enable sublocalisation of proteins within the range of subcellular compartments. If the destination is the mitochondrion, chloroplast or secretory pathway, sorting usually relies on the presence of an N-terminal targeting presequence recognised by the translocation machinery (Emanuelsson *et al.*, 2000). N-terminal sequence prediction programs have been developed using a variety of different methods. MitoprotII developed by Claros and Vincens (1996) is an algorithm that facilitates the detection of the transit peptide localisation. The software derives empirical rules from sequences that contain descriptions determining that their mitochondrial precursor proteins are encoded in the nucleus and product resides in the mitochondria. MitoprotII consists of four main steps including an initial screening procedure that aims to eliminate proteins that present physical constraints on protein import regardless of the existence of a targeting presequence (Claros and Vincens, 1996). The developers incorporated a mechanism to detect these physical constraints based on hydrophobicity scales with limiting functions that if exceeded present physical barriers for protein import. The second step of the algorithm aims to detect the presence of an N-terminal

targeting sequence with specific characteristics including being enriched in positively charged residues and having a minimum amino acid sequence length of 12. Thirdly, parameters are used to assess the amphiphilicity of the targeting sequences followed by the final classification step involving the determination of cleavage site presence, total net charge for the complete sequence and amino acid composition (Claros and Vincens, 1996).

TargetP is another N-terminal sequence prediction tool based on neural networks which aims to assign proteins to one of four localisation compartments (mitochondrion, chloroplast, ER/golgi/secreted and other) developed by Emanuelsson *et al.* (2000). This software is based on two layers of neural networks with the first layer containing a specific network regarding the type of presequence (chloroplast targeting peptides, mitochondrial targeting peptides and signal peptides abbreviated to cTP, mTP and SP, respectively). The second is an integrated network responsible for producing the actual prediction. The non-plant version of TargetP aims to distinguish only between mTPs, SPs and other. The first layer of the network is implemented using a logarithmic error function and sparsely encoded sliding windows for input data encoding (Emanuelsson *et al.*, 2000). This input sequence window consists of 20 nodes representing the relevant amino acid residues and if a particular amino acid is present the relevant node is set to 1 as opposed to the default 0. The network is then trained to recognise if the residue is part of a targeting sequence (Emanuelsson *et al.*, 2000). Implementation of the second integrating network based on a quadratic error function considers the outputs from the first network corresponding to the 100 N-terminal positions of the query sequence. The top layer generates scores per targeting peptide with the highest score determining the localisation prediction.

However, methods involved in determining sorting signals in amino acid sequences rely heavily on the quality of the gene 5'-region or N-terminal target sequence. (Nakai and Horton, 1999) claim that subcellular localisation prediction methods which depend on sorting signals will be inaccurate in the absence of these signals or if only partially present and these signals are not general enough to include each organelle's resident proteins. A program known as SubLoc provides a unique method based on machine learning to assess amino acid composition that can be used as a complementary technique alongside sorting signal detection algorithms. SubLoc developed by Hua and Sun (2001) implements a support vector machine to perform supervised pattern recognition of high dimensional data. The algorithm consists of an input vector of 20 with each input representing an amino acid. These input vectors can then be mapped into a high dimensional space separated by an optimal separating hyperplane (see section 2.2.2).

For eukaryotic sequences the classification consists of four categories (mitochondrial, cytoplasmic, extracellular and nuclear). The classification system results in a prediction based on the highest scoring compartment from the SVM output (Hua and Sun, 2001).

Mitopred has been employed for the detection of nuclear-encoded mitochondrial proteins, incorporating Pfam domain occurrence and amino acid compositional differences between mitochondrial and non-mitochondrial proteins. Mitopred was developed by Guda *et al.* (2004) to rectify the limitations presented by previous prediction methods based on a single criterion. Mitochondrial protein import consists of myriad complex mechanisms based on several features (Schatz, 1996). The software involves a unique method that analyses the amino acid sequence for the presence of Pfam domains. Each domain is classified into one of three groups (Mito only - exclusive mitochondrial domains; Non-mito only - exclusive non-mitochondrial domains and Shared - domains found in both of the groups).

2.1.5 Mitochondrial databases

A number of mitochondrial databases are available dedicated to the storage of mitochondrial protein data. The Human Mitochondrial Database (HMPDb) is a repository of mitochondrial and nuclear-mitochondrial proteins that combines distributed information from external resources including SwissProt, Protein Data Bank (PDB), Online Mendelian Inheritance in Man (OMIM) and mitochondrial specific databases such as the Human Mitochondrial Genome Database (mtDB) and MITOMAP. The HMPDb currently has 1465 proteins recorded that consist of experimentally determined proteins and computationally predicted mitochondrial proteins. MitoProteome is an object-relational mitochondrial protein sequence database and annotation system, again consisting of careful curation of existing databases (Cotter *et al.*, 2004). Initially it contained 847 human mitochondrial protein sequences but was extensively revised in 2009. MitoProteome currently contains 780 mitochondrial proteins (both mitochondrial encoded and nuclear encoded) 175 of these proteins were determined experimentally by mass spectrometry (LC/MS/MS). The MitoP2 database focuses on the nuclear-encoded proteome of mitochondria and aims to provide a comprehensive list of mitochondrial proteins in humans (Elstner *et al.*, 2009). MitoP2 also interrogates other species for orthology including *Saccharomyces cerevisiae*, *Mus musculus*, *Arabidopsis thaliana* and *Neurospora crassa*. Elstner *et al.* (2009) have employed an SVM to assess the likelihood a protein is mitochondrial or not.

2.1.6 Integrative methods for sublocalisation prediction

Due to the abundance of sublocalisation prediction tools available employing a range of unique methods for analysing amino acid sequences, an emphasis has been made to integrate these methods in order to augment the overall accuracy. Certain techniques are perceived to be complementary in determining the location of nuclear-encoded proteins. Shen and Burger (2007) aimed to integrate several tools in two ways including a decision tree algorithm and majority-win voting attempting to harness the strengths of each prediction method. Results from majority-win voting proved to be less successful than the decision tree algorithm. The mitochondrial prediction tools investigated were TargetP, SubLoc, SherLoc (Shatkay *et al.*, 2007), pTARGET (Guda, 2006), Predotar (Small *et al.*, 2004), Protein Prowler (Bodén and Hawkins, 2005), PASUB (Lu *et al.*, 2004), Mitoprot and CELLO (Yu *et al.*, 2006).

A decision tree was constructed that integrated four mitochondrial target peptide (MTP) based tools (TargetP, Mitoprot, Predotar and Protein Prowler) referred to as MTP-DT. This was further combined with the five remaining tools in a stacked decision tree referred to as STACK-DT. As the localisation tools were less efficient at predicting membrane proteins as opposed to matrix proteins a further integration was developed incorporating four additional transmembrane prediction methods named Phobius (Käll *et al.*, 2004), TMHMM (Krogh *et al.*, 2001), HMMTOP (Tusnády and Simon, 2001) and SOSUI (Hirokawa *et al.*, 1998). The most successful decision tree was an incorporation of all these tools known as STACK-mem-DT expressing the highest true positive rate (TPR) against false positive rate (FPR). Figure 2.3 compares the outcome of all these integrated methods. The protein sequences used for training were downloaded from SwissProt and selected based on specific criteria. All proteins were nuclear-encoded, experimentally verified regarding subcellular location and did not contain ambiguous annotation such as probable or possible (Shen and Burger, 2007).

A systematic identification of mitochondrial genes was performed by Calvo *et al.* (2006) in an attempt to expand the catalogue estimated to be in excess of 1500 genes. This investigation aimed to combine established methods for protein sequence analysis with the addition of recent progress applied to the ancestry and transcriptional regulation of the organelle.

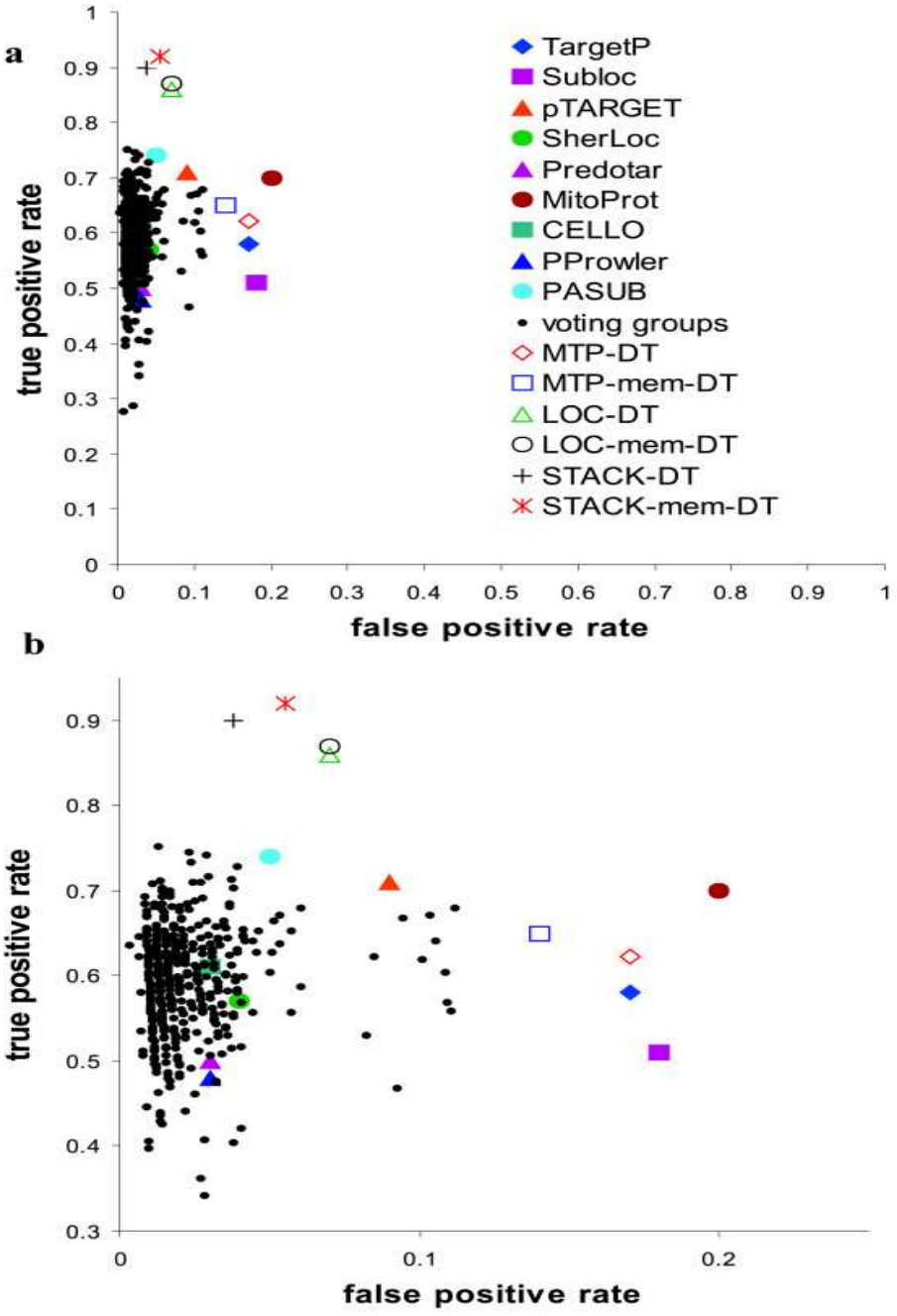

Figure 2.3: Prediction performance of individual and integrated tools on human mitochondrial proteins taken from Shen and Burger (2007). Filled symbols: individual localisation tools; Dots: voting groups (tools integrated by majority-win voting); Open symbols: decision trees. The top left hand corner of the plot displays the most successful results, representing high true positive rate and lowest false positive rate (FPR). a. Full scale result. b. Magnified region with FPR 0 0.25 and TPR 0.3 0.95

Eight genome-scale datasets were incorporated into the systematic analysis using a naive Bayes classifier consisting of the following:

1. Targeting signal based on TargetP prediction of N-terminal target peptide.
2. Protein domain score reflecting presence of mitochondrial protein domains based on SwissProt annotation.
3. *cis*-motif score measuring evidence for conserved transcriptional regulatory elements previously discovered to be upstream of mitochondrial genes.
4. Yeast homology score derived from the detection of *Saccharomyces cerevisiae* orthologues based on experimental evidence annotated in SGD (Saccharomyces Genome Database).
5. Ancestry score based on sequence similarity to *Rickettsia prowazekii* proteins.
6. Coexpression score measuring transcriptional coexpression with known mitochondrial genes using genome-scale atlases of RNA tissues across diverse tissues based on a neighbourhood metric.
7. MS/MS score reflecting the number of tissues a protein was detected from previous proteomics research.
8. Induction score measuring the upregulation of mRNA transcripts in a cellular model of mitochondrial biogenesis.

These methods were assessed using datasets consisting of 654 mitochondrial proteins derived from MitoP2 and 2847 non-mitochondrial proteins with GO annotations specifying organelles apart from the mitochondrion derived from Ensembl. The performance of Maestro compared to the individual methods is displayed in Figure 2.4. Maestro was trained on gold standard positive and negative datasets and was implemented in a genome-wide scan of the entire 33,860 Ensembl human proteins using a scoring threshold of 5.65 corresponding to a false discovery rate of 10%. Maestro achieved a sensitivity of 71% and a specificity of 99.4% revealing 368 previously unassociated genes believed to be involved in mitochondrial function. These novel predictions expressed considerable overlap with Mitopred, the best existing computational prediction algorithm, but with greater sensitivity and specificity on the training data (Calvo *et al.*, 2006).

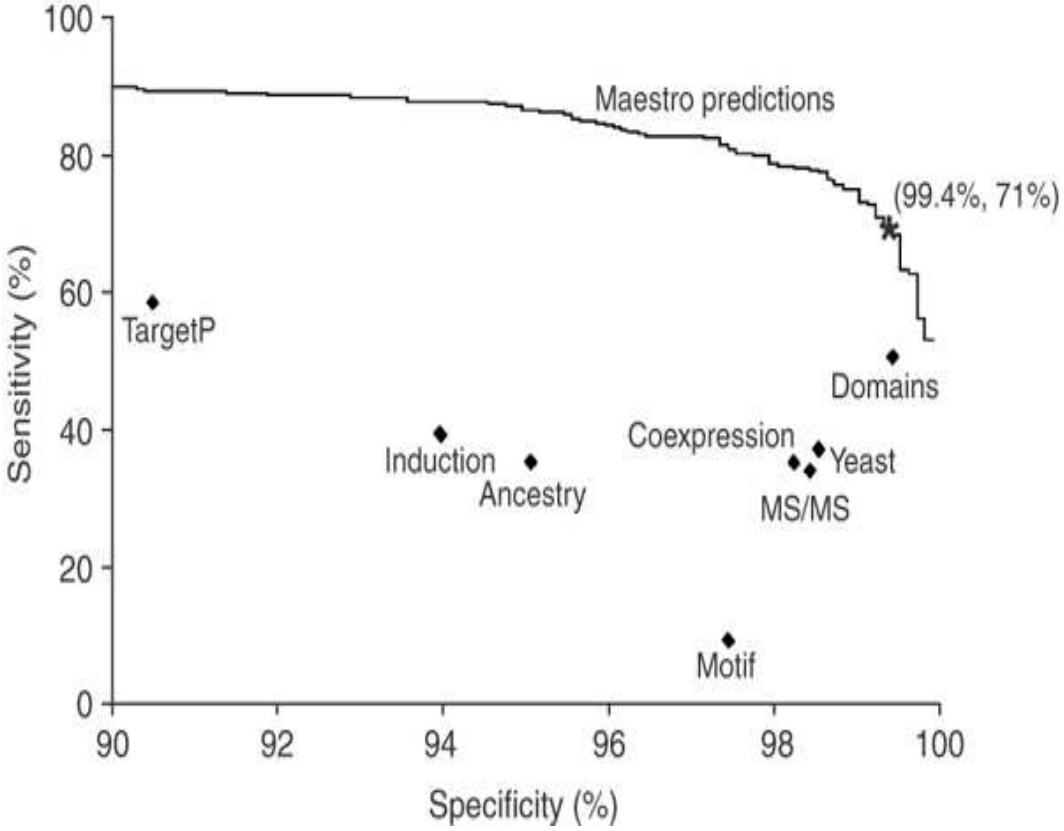

Figure 2.4: Sensitivity and specificity of genome-scale prediction methods taken from Calvo *et al.* (2006). Maestro displays a range of thresholds with the selected threshold being marked by an asterisk resulting in a sensitivity of 71% and specificity of 99.4%.

A follow up analysis was implemented by Pagliarini *et al.* (2008) to generate a mitochondrial protein compendium involving an in depth analysis combining results from protein mass spectrometry (MS), microscopy and machine learning. MS proteomics was performed on highly purified and crude mitochondrial preparations in order to discover genuine mitochondrial proteins and distinguish them from contaminants based on enrichment. These experiments consisted of a discovery phase involving the isolation of highly purified mitochondria from 14 different tissues and a subtractive phase based on the observation that mitochondrial proteins should become enriched during the purification process with contaminants being depleted (Pagliarini *et al.*, 2008). Training sets were compiled consisting of 591 known mitochondrial genes and 2519 non-mitochondrial genes excluding proteins characterised previously by proteomic studies. This enabled the genes to be assessed for mitochondrial association derived from log likelihood ratios. The improved MS data was integrated again using a Bayesian framework with six other genome-scale datasets applied by Calvo *et al.* (2006) with the exclusion of *cis*-motifs. Using the Maestro program a new threshold was set to 4.56 corresponding to a 10% false discovery rate achieving a sensitivity of 84%. The resulting protein compendium named MitoCarta consists of 1013 genes for human and 1098 for mouse displaying protein expression across 14 different tissue types. These techniques are illustrated in Figure 2.5.

2.1.7 Sensitivity and Specificity

Sensitivity and specificity test the accuracy of a prediction by comparing this to known data. Sensitivity is a measure of the number of correctly identified positive results (true positives) from a test. Specificity is the measure of correctly identified negative results (true negatives) from a prediction. The following equations are used to calculate sensitivity and specificity:

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{2.1}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{2.2}$$

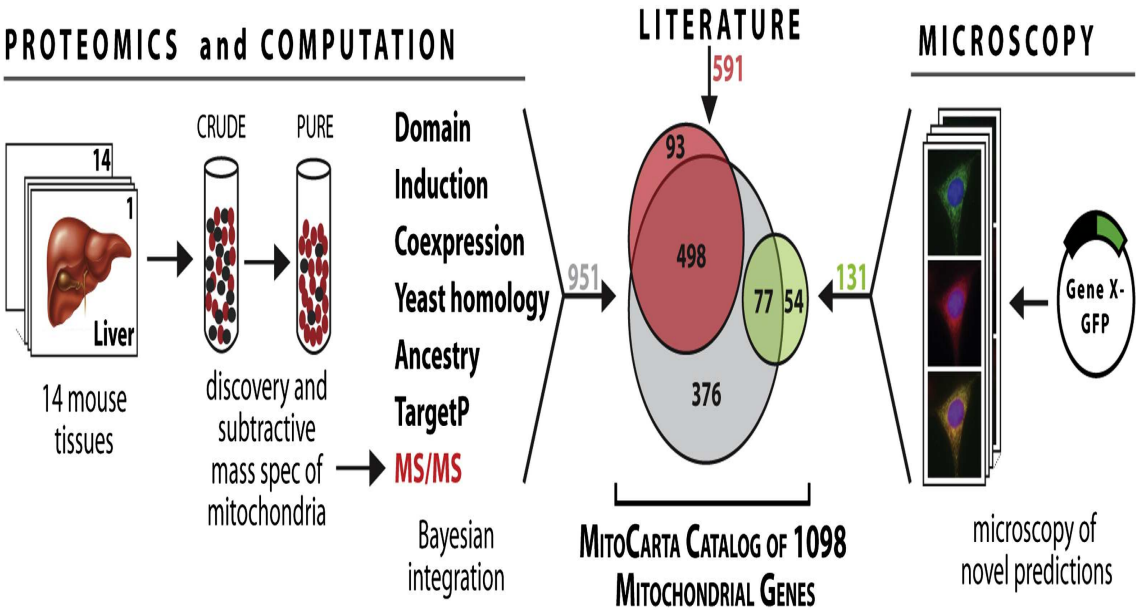

Figure 2.5: MitoCarta protein compendium taken from Pagliarini *et al.* (2008) illustrating the combination of several approaches including: 1) An integrated analysis of seven genome-scale datasets including MS data (grey circle), 2) large scale GFP tagging and microscopy (green circle), and 3) prior biological knowledge from literature (red circle). MitoCarta consists of a union of all these genes.

2.1.8 Proposed Approach

The aim of this study is systematically assess all combinations of suitable mitochondrial prediction methods in an attempt to qualify the most complementary techniques. This would also reveal prediction methods that were negatively affecting overall prediction accuracy. Machine learning was applied implementing a trained SVM on a dataset of known mitochondrial and non-mitochondrial proteins. Sensitivity and specificity were recorded for every combination, and each combination was interrogated 100 times each consisting of a randomised training and testing dataset. This would allow an assessment to be made regarding the standard deviations for each combination when analysed with random unseen testing data and varying candidates in the training data. Different combinations subjected to repeated testing should reveal important information relating to confidence intervals. It is expected that merely increasing the number of independent datasets will negatively affect sensitivity and specificity whereby an optimum set of classifiers will be revealed. Datasets will be interrogated and organised for SVM training and testing using a workflow constructed in the Taverna Workbench. All data will be automatically stored into a relational database for further assessment.

2.2 Methods

2.2.1 Protein sequence retrieval

A reference collection of proteins were downloaded from SwissProt using SRS (Sequence Retrieval System) using specific filters (Figure 2.6). Mitochondrial proteins were retrieved specifying 'Homo sapiens NOT hypothetical NOT uncharacterised AND mitochondrial AND Protein Existence: 1: Evidence at protein level'. For non-mitochondrial proteins the following filter was applied 'Homo sapiens NOT mitochondrial NOT hypothetical NOT uncharacterised AND Protein Existence: 1: Evidence at protein level'. These queries enable the retrieval of experimentally determined proteins by specifying the correct level of protein existence. SwissProt records each contain a PE (Protein Existence) line with the following format:

PE *Level: Evidence;*

The PE line has the following levels:

1: Evidence at protein level
2: Evidence at transcript level
3: Inferred from homology
4: Predicted
5: Uncertain

The queries specified that each protein required "1: Evidence at protein level" as this indicates clear experimental evidence for the existence of a protein. This may have been quantified in various ways including clear identification by mass spectrometry (MSI), X-ray or NMR structure or detection by antibodies. In addition, the queries were also specified to remove any hypothetical or uncharacterised proteins. Only proteins with clear annotation and sublocalisation were required in order to perform accurate testing of predictive software. The sublocalisation was determined by the SUBCELLULAR LOCATION paragraph from the CC line in the SwissProt record. This allowed determination of the specific subcellular compartment the particular protein was associated with. This also provided details about the experimental procedures used in determining the presence of the protein. For each protein the amino acid sequence, SwissProt accession number, Ensembl gene id, gene name, description and chromosomal coordinates were extracted.

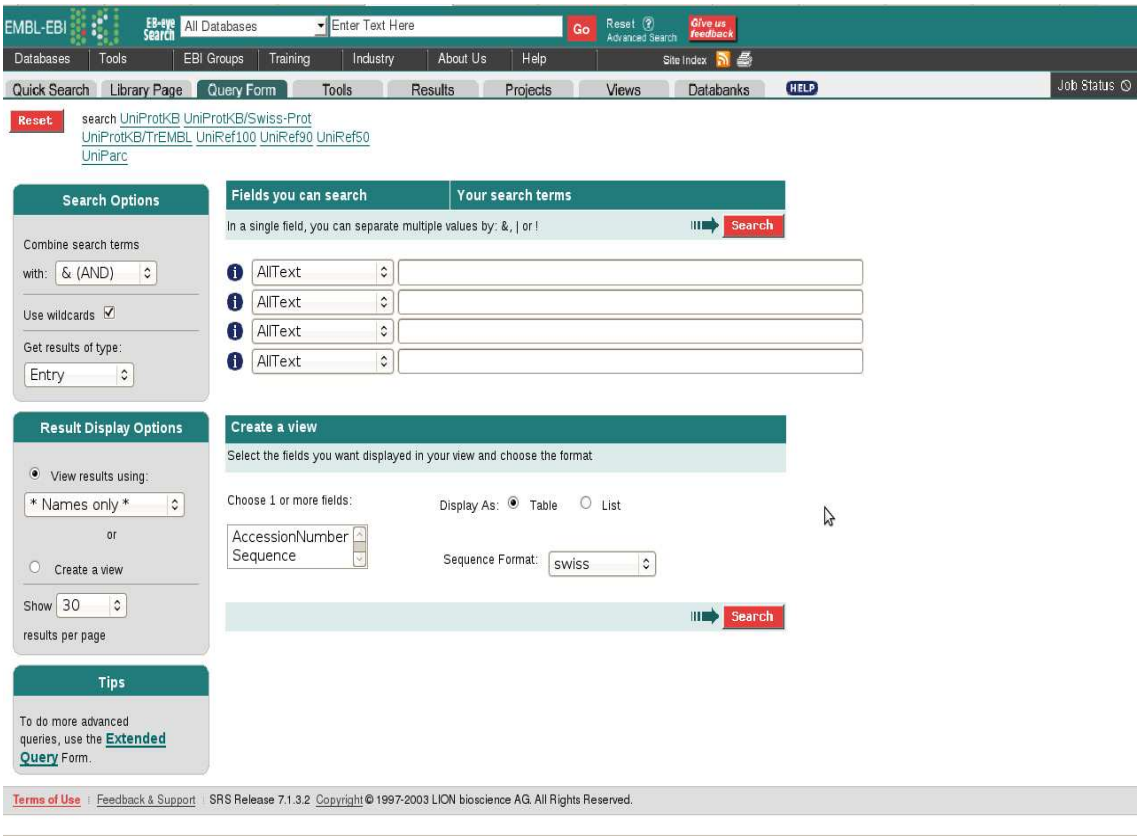

Figure 2.6: Sequence Retrieval System (SRS) based at the European Bioinformatics Institute. This allows the user to administer specific queries to retrieve sequence information.

The entire protein dataset was analysed with a workflow designed to run each protein sequence through sublocalisation software including MitoprotII, TargetP and SubLoc and stored into a relational database (Figure 2.7). These specific prediction tools were chosen due to the range of methods employed including target sequence prediction, neural networks, amino acid composition and Pfam domain assessment. All the results were stored into a relational database along with all the relevant information from the data retrieval procedure. The entire protein dataset consisted of 467 mitochondrial and 6352 non-mitochondrial proteins.

2.2.2 Data integration

Supplementary information was downloaded from the results section of the genome-wide analysis performed by Calvo *et al.* (2006) (Table 2.1). A local relational database was constructed using SQL designed to store all the results from this genome-wide analysis. This data representing 33,860 human proteins was extracted and stored by exporting the data from an excel spreadsheet into the database and named calvo_genome. This database contained all the relevant information including Ensembl gene ids, protein ids, gene names, descriptions and results from all the independent analyses. The Ensembl protein id was used as the foreign database key to allow the database to be joined with other databases containing a matching Ensembl protein id column for further analysis. The calvo_genome database consisted of the architecture displayed in Table 2.2. The results from the MitoFlow workflow displayed in Figure 2.7 were automatically stored into a database using Java Database Connectivity (JDBC). Finally, a database that consisted of all the calvo_genome data and MitoFlow data was combined into a single database named reference_proteins. This database consisted of columns and associated datatypes displayed in Table 2.3. This reference database could then be implemented in the training and testing of the SVM.

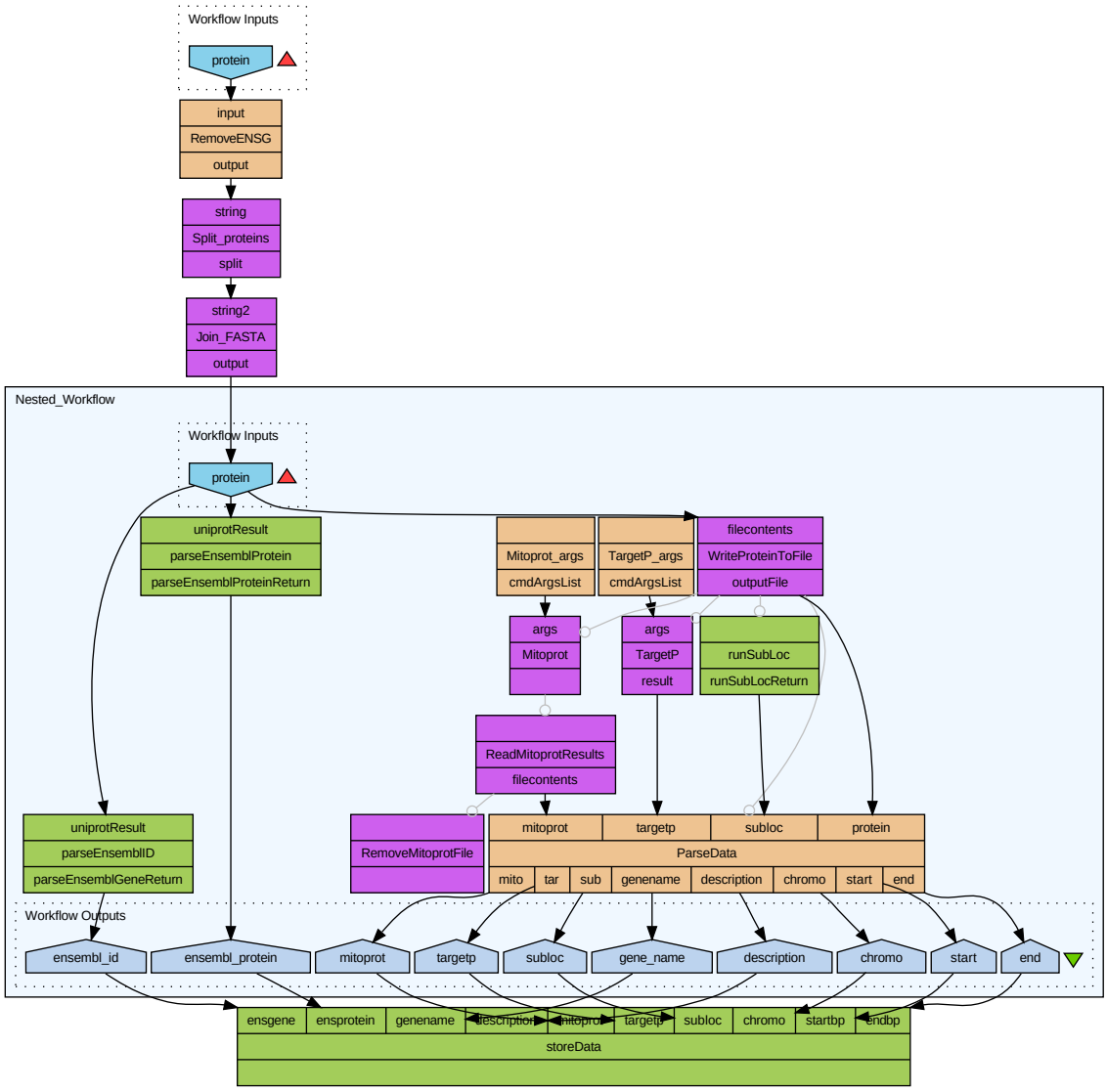

Figure 2.7: MitoFlow. A workflow to analyse the human proteome with sublocalisation software Mitoprot, SubLoc and TargetP. A list of protein sequences are submitted in FASTA format and then split into separate jobs for iterative analysis. Each sequence is sent to the relevant prediction programs and the results are extracted, alongside Ensembl gene information and stored into a relational database.

Eleven genome-scale data sets used to predict mitochondrial localization	
Method	Genome-scale dataset
1. Protein domain	Pfam domain found only in eukaryotic mitochondrial proteins (SwissProt)
2. Cis motif	Erra motif in human/mouse promoters
3. Yeast homology	S. cerevisiae mitochondrial orthologue
4. MS/MS	Mouse mitochondria (brain, heart, liver, kidney)
5. Induction	Difference in gene expression during mitochondrial biogenesis induced by PGC-1a
6. Mitopred	Pfam domains
7. Targeting signal	TargetP on human/mouse orthologues
8. Ancestry	*R. prowazekii* orthologue
9. Coexpression	Coexpression with known mitochondrial genes in human/mouse tissue atlases
10. Targeting signal	Mitoprotii
11. Amino acid composition	SubLoc

Table 2.1: Genome-scale datasets implemented in the analysis performed by Calvo *et al.* (2006). Additional datasets Mitoprot and SubLoc were added to the investigation.

2.2.3 Support vector machine training and optimisation

The parameters for the SVM learning algorithm were optimised to yield the highest accuracy of prediction with the lowest number of support vectors (<10% of the overall dataset). An accuracy of 95.76% was achieved using specific parameters available as part of the svm^{light} software. From an overall dataset of 6819 candidates (467 mitochondrial and 6352 non-mitochondrial), 730 were extracted from the database at random to form the test set (100 mitochondrial and 630 non-mitochondrial), leaving the 6089 candidates for training. The SVM was trained and tested 100 times for each one of the 2047 possible combinations of the eleven prediction parameters. Each run used a different randomly selected training and test set, allowing calculation of the mean and standard deviation of sensitivity and specificity values.

column	datatype
id	integer
ensgene_id	character varying(100)
ensprotein_id	character varying(100)
gene_name	character varying(200)
description	text
mitodomain	integer
erra_motif	integer
yeast_homologue	integer
targetp	double precision
induction	double precision
rp_homologue	integer
coexpression	double precision
msms_tissues	integer
mitopred	double precision

Table 2.2: Architecture of the calvo_genome database storing all the information from a genome-wide analysis of 33,860 human proteins performed by Calvo *et al.* (2006). The Ensembl protein id was used as the foreign key to enable joining other tables containing the same protein id. The id column formed the primary key to ensure each row was unique.

column	datatype
id	integer
acc_no	character varying(10)
job_id	character varying(100)
ensgene_id	character varying(100)
evidence	text
sublocalisation	text
mitodomain	character varying(10)
erra_motif	character varying(10)
yeast_homologue	character varying(10)
msms_tissues	character varying(10)
induction	character varying(10)
mitopred	character varying(10)
targetp	character varying(10)
rp_homologue	character varying(30)
coexpression	character varying(10)
mitoprotii	character varying(10)
subloc	character varying(50)

Table 2.3: Architecture of the reference_genome database combining all the information from the calvo_genome database of 33,860 human proteins and the MitoFlow database of MitoProt and SubLoc results. The Ensembl protein id was used as the foreign key to enable joining other tables containing the same protein id. The id column formed the primary key to ensure each row was unique.

2.2.4 Combination analysis workflow

A program was developed to extract a random 730 protein candidates from the integrated reference dataset (100 mitochondrial and 630 non-mitochondrial). The program was designed to create 100 examples and stored as a file to be incorporated in the combination analysis pipeline. Each example was structured so the first 100 proteins were mitochondrial followed by the remaining 630 non-mitochondrial proteins.

Workflow user query

The workflow was designed to receive a list of all the 2047 combinations (Figure 2.8). Each combination consisted of datasets represented by a string of comma separated keywords such as mitodomain, yeast_homologue, targetp, subloc. The second input required was an identical list but referring to the required database table names reflecting each combination such as do_ye_tp_sb. This ensured that the results for each combination could be stored into a unique relational table required for further analysis later in the automated pipeline. In order for the workflow to analyse each combination correctly the workflow must consist of a nested architecture in Taverna. This means the nested section of the workflow needs to have completed for each combination before the next one is analysed. Therefore an example input list for the workflow is as follows:

Input 1) Example list of five unique combinations:

- mitodomain, erramotif, yeasthomologue
- mitodomain, erramotif, msmstissues
- mitodomain, erramotif, induction
- mitodomain, erramotif, mitopred
- mitodomain, erramotif, moothtargetp

Input 2) Example list of five unique relational table names corresponding to the above list of combinations:

- xdo_er_ye
- xdo_er_ms
- xdo_er_in
- xdo_er_mp
- xdo_er_tp

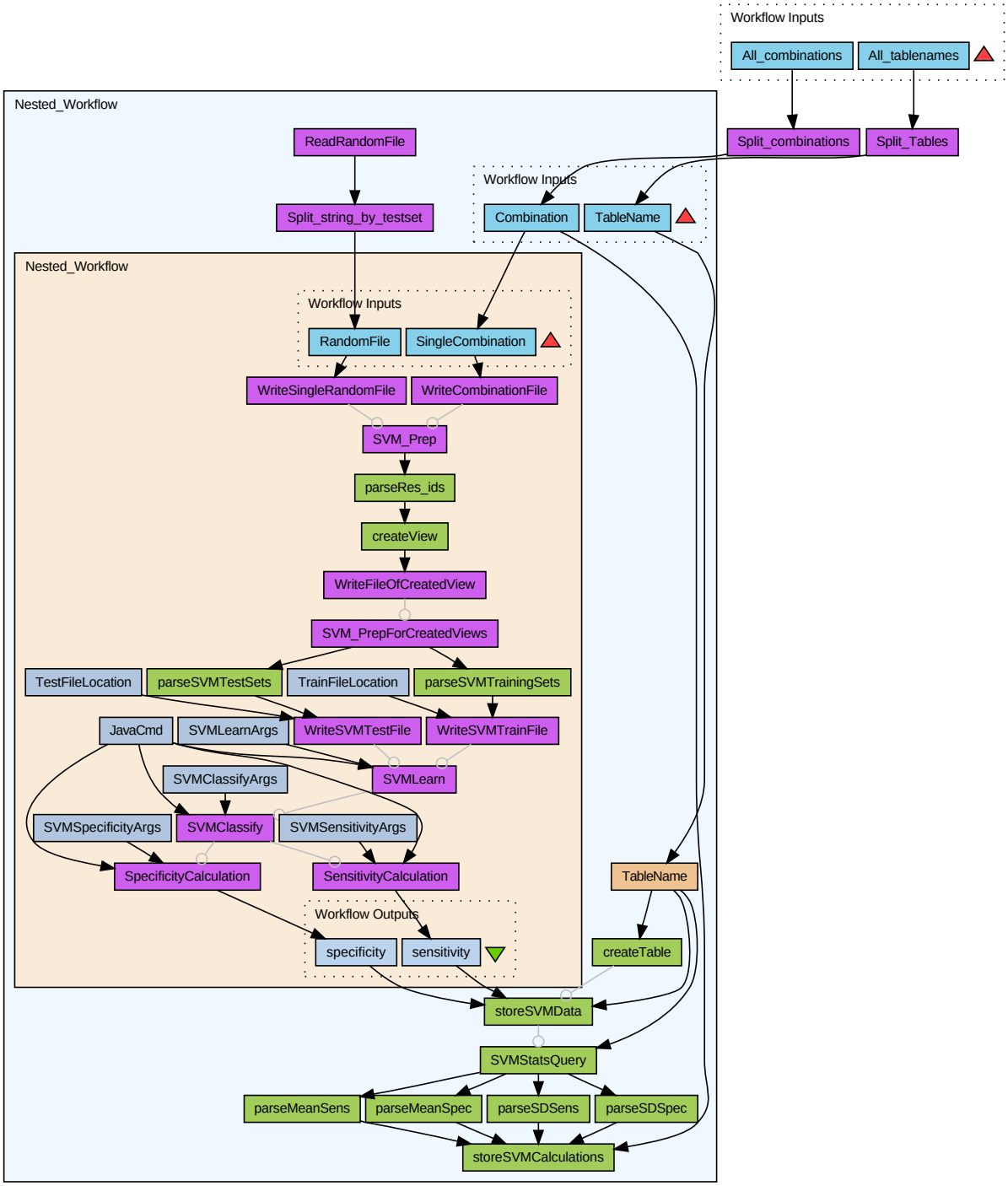

Figure 2.8: Combination analysis workflow. This is a nested within a nested workflow allowing each combination of classifiers to be individually analysed 100 times.

Combination generation

Every combination of the datasets was investigated in order to measure sensitivity and specificity for each. This could potentially highlight which datasets could contribute to an increase in the sensitivity of predictions as opposed to datasets that were disadvantageous when involved in an integrated genomic approach. The complete list of the 2047 unique combinations was generated using the following equation:

$$\frac{n!}{r!(n-r)!} \tag{2.3}$$

This formula was applied to calculate the total number of combinations 11 different datasets would produce where n is the total number of datasets and r is the number selected for a specific combination. The total number of combinations amounted to 2047 as this is different to a permutation. Within a combination the order the classifiers appear is not important as long as they are present. Order is important for a permutation resulting in for example Mitodomain, TargetP and induction being classed as different when compared to TargetP, induction and Mitodomain. With this important difference taken into account the calculation for the 11 classifiers was based on combinations and not permutations. The same equation was applied for the corresponding relational table names in order to store the results generated by each combination in the structure mentioned above. These lists were stored as text files in preparation for input into the first section of the combination analysis workflow. The lists are sent to the first nested workflow which is illustrated by the pale blue box in Figure 2.8. The input lists are split into separate queries by carriage returns using local processors (purple) available in Taverna. The nested workflow is then able to process each individual query separately. In order for each separate combination input to be correctly correlated with its unique relational table name the implicit iteration functionality of Taverna is utilised. This is to ensure that the first input of the combinations list is sent to the workflow coupled with the first input of the relational tables list and the second with the second and so on. Taverna enables this allowing developers to manipulate the iteration strategy of a workflow. The options consist of cross product and dot product whereby cross product reflects an all against all strategy which is unsuitable for this workflow. The dot product option ensures first against first, second against second which is illustrated in Figure 2.9. The nested architecture means that each individual combination is fully analysed and results stored before the next combination proceeds into the workflow.

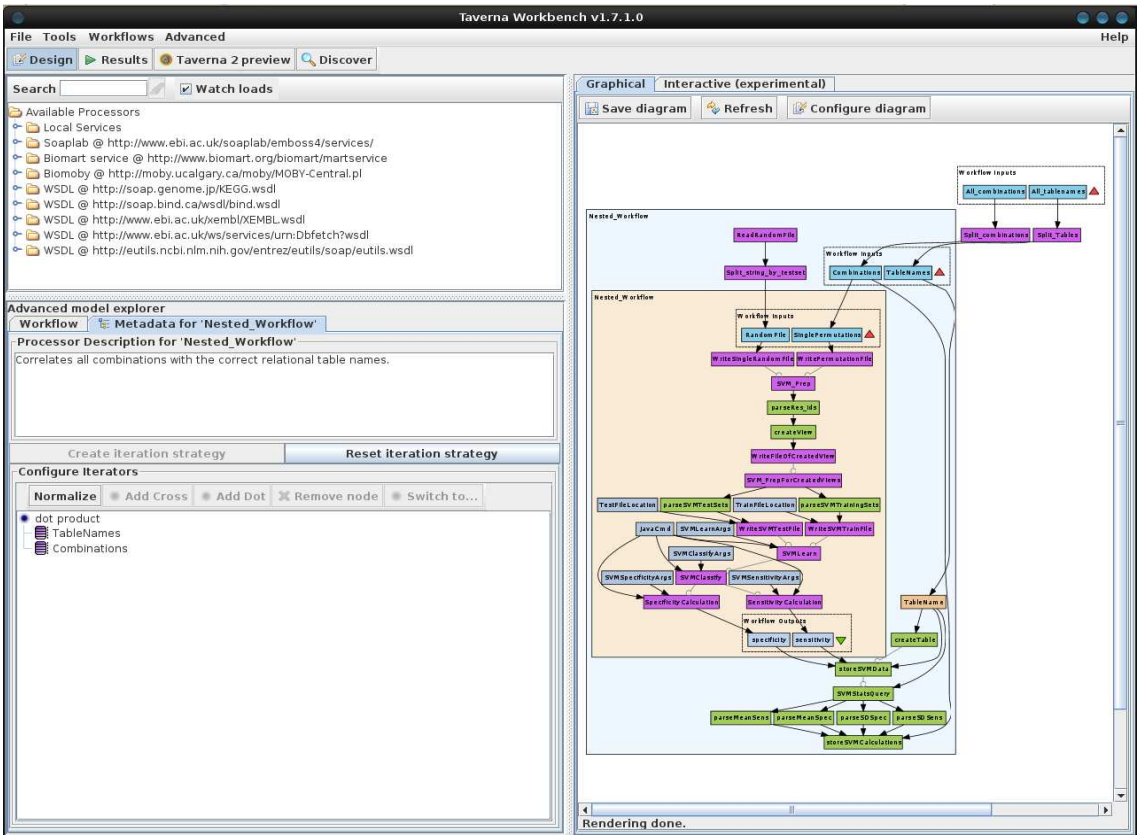

Figure 2.9: Configuration of the iteration strategy in the Taverna workbench. The selection of dot product ensures the correlation of the two inputs from their relative positions in the input lists.

Process 1: nested workflow 1

The first nested workflow (pale blue) takes a query and its related table name as input such as:

- Combination: mitodomain, targetp, induction
- TableName: do, tp, in

Using the table name this is instantly used to create a unique relational table in preparation for storage of the resulting data towards the final part of the workflow. This consists of a Java program implementing JDBC with the following embedded SQL:

```
s.executeUpdate("CREATE TABLE " + table + "
(id serial, sensitivity double precision,
specificity double precision)");
```

This is implemented by the createTable program (green processor) creating a relational database table named, in this example, do_tp_in. This table name is also directed as input into other programs within this section namely storeSVMData and SVMStatsQuery. The combination is also directed as input, in this example, mitodomain, targetp, induction to a data storage program storeSVMCalculations. These processors will be explained in more detail towards the end of this section as these require the completion of the second nested workflow (beige).

Process 2: Random testset generator

A Java program was developed to extract 730 random proteins (100 mitochondrial and 630 non-mitochondrial) from the reference database created previously. This required the name of the reference database as input into the program to allow successful execution. The program relied on Java Database Connectivity (JDBC) which contained embedded SQL statements to allow communication with the database to allow data retrieval and storage. The following SQL code was implemented in the program:

```
ResultSet rs1 = s.executeQuery("(SELECT * FROM " + table + " " +
"WHERE job_id = 'mito' ORDER BY RANDOM() LIMIT 100) UNION ALL
(SELECT * FROM " + table + " " + "WHERE job_id = 'nonmito'
ORDER BY RANDOM() LIMIT 630)");
```

The first part of the statement extracts all the necessary data from the reference database (using the database name received from the program input) where the id is mitochondrial. This is retrieved at random with a limit set to 100. This returns 100 random mitochondrial protein candidates from the reference database. This is attached by the UNION ALL statement to the results from the final part of the SQL statement. This part retrieves all the necessary data from the reference database where the id is non-mitochondrial and extracts this at random with a limit of 630. This returns 630 random non-mitochondrial candidates appended to the 100 mitochondrial candidates resulting in a random testset of 730 proteins. This program contains a looping construct to allow the procedure to be automatically executed 100 times. The result is written to a text file containing 100 different examples of the random testsets. This text file forms an essential part of the combination analysis workflow. A local processor (purple) in Taverna allows developers to read from and write to text files. The processor ReadRandomFile retrieves the randomly generated test sets produced by the aforementioned Java program and using another local processor splits these into separate entries each consisting of 730 protein examples. Again this requires an iteration strategy to be implemented to allow each individual combination to be analysed 100 times. This involves using a strategy known as a cross product. This is an all against all strategy resulting in the one single combination being fed into the second nested workflow (beige processor) alongside each of the 100 random testsets. Therefore this allows each combination to be analysed with all the 100 randomly generated testsets. The second nested workflow needs to have completed a full analysis of one testset before the next one proceeds.

Process 3: Nested workflow 2

The second nested workflow is the large beige processor that is contained within the larger blue processor in Figure 2.8. This receives two inputs consisting of a single combination such as mitodomain, targetp, induction and one random testset file containing all data for 730 random proteins. Each input is then written to a file and saved to allow the following program SVM_Prep to execute successfully as this program requires this data to function. The program SVM_Prep is conditionally linked to the file writing processors (purple) meaning these processors need to have completed before SVM_Prep is invoked. This can be seen in Figure 2.10 displaying the physical linking of the processors in the workflow diagram and below how this is configured in the Advanced Model Explorer window in the Taverna workbench.

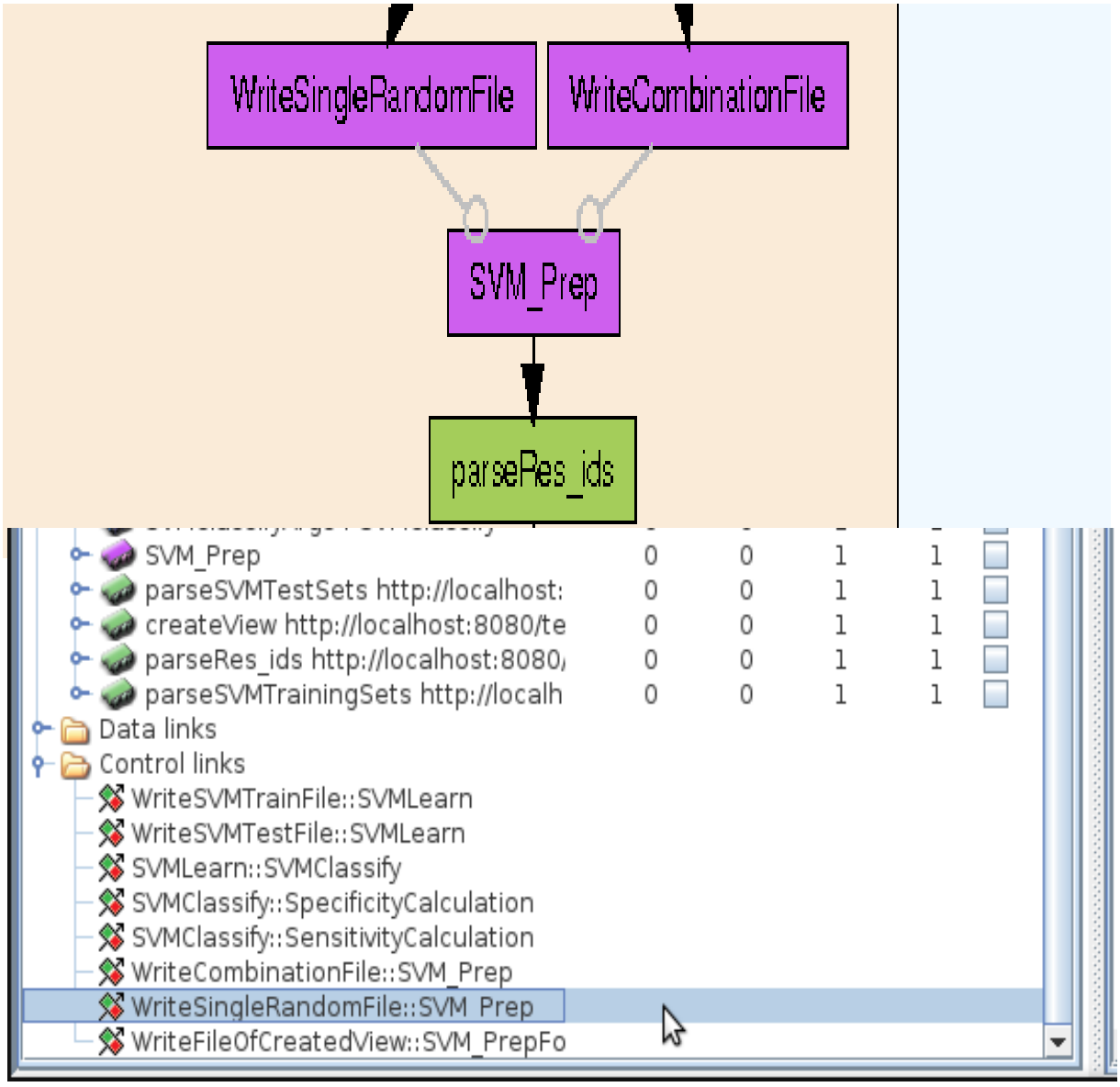

Figure 2.10: Control link nodes in Taverna allow the user to ensure certain processors have fully completed before the next one proceeds.

Process 4: SVM Training and test file preparation

The next stage of the workflow is a Java program named SVM_Prep. This is accessed via a Taverna local processor that allows the user to execute a command line application from inside the workflow. In this instance the processor requires the following command line argument as a permanent input:

```
java SVM_Prep
```

The processor sends this to the command line and returns the program output. SVM_Prep requires two input files in order to function which are specified by filepaths within the program. These are the files constructed from the previous step in which one file contains the individual combination and the second file contains data from the random 730 proteins. The random dataset contains all the information regarding the 11 classifiers for each of the 730 proteins forming the testset. The architecture of this data is illustrated below:

SVM testset example

```
6807 1 1:1 2:0 3:1 4:0 5:2.9 6:100 7:0 8:-130 9:5.5 10:0 11:0
6624 1 1:1 2:0 3:1 4:4 5:4.21 6:100 7:2 8:-130 9:33.5 10:0.9531 11:0
6749 1 1:-2 2:0 3:0 4:4 5:2.74 6:0 7:0 8:-130 9:27.5 10:0.6221 11:0
6687 1 1:0 2:0 3:0 4:1 5:1.4 6:100 7:2 8:-130 9:9.5 10:0.9448 11:1
6804 1 1:1 2:0 3:0 4:0 5:1.6 6:73.6 7:2 8:-020 9:7.0 10:0.9342 11:1
4601 -1 1:-1 2:0 3:0 4:0 5:0 6:0 7:1 8:0 9:0.5 10:0.1483 11:0
684 -1 1:-1 2:0 3:0 4:0 5:0 6:0 7:0 8:0 9:0.5 10:0.1873 11:0
1400 -1 1:-2 2:0 3:0 4:0 5:0 6:0 7:0 8:0 9:0.5 10:0.0679 11:0
783 -1 1:-1 2:0 3:0 4:0 5:0 6:0 7:0 8:-03 9:4.0 10:0.0007 11:0
2876 -1 1:-2 2:0 3:0 4:0 5:0 6:0 7:0 8:0 9:0.5 10:0.0502 11:0
```

This represents five mitochondrial candidates (1) and five non-mitochondrial candidates(-1) merely for visualisation purposes. The first number is the unique primary key id from the database record, the second is the job_id denoting whether the candidate is experimentally determined to be mitochondrial (1) or non-mitochondrial (-1). The proceeding 11 numbers followed by semi-colons are the values generated from each classifier (mitodomain, cis motifs, yeast homologues, etc). SVM_Prep contains all the classifier names in its memory so when it reads the combination file it maps this to the random testset and retrieves only those columns of data.

The combination mitodomain, targetp, induction would return the following:

Combination specific SVM testset example

```
6807 1 1:1 5:2.9 7:0
6624 1 1:1 5:4.21 7:2
6749 1 1:-2 5:2.74 7:0
6687 1 1:0 5:1.4 7:2
6804 1 1:1 5:1.6 7:2
4601 -1 1:-1 5:0 7:1
684 -1 1:-1 5:0 7:0
1400 -1 1:-2 5:0 7:0
783 -1 1:-1 5:0 7:0
2876 -1 1:-2 5:0 7:0
```

This is due to the fact that mitodomain is the first column in the reference database, induction is the fifth and targetp is the seventh. This now represents 730 protein candidates with information only from these specific classifiers. This forms the SVM testing data, but an SVM training set was required which did not contain the proteins contained within the testset. Following the completion of SVM_Prep the output was sent to parseRes_ids (green processor) which extracted all the unique primary key ids from the above testset in preparation for the program createView. This returned a comma separated string of unique ids (e.g '6807', '6624', '6749', '6687'...) that could be incorporated directly into an SQL statement contained within the createView Java program. As each testset contains 730 randomly extracted proteins this needs to be trained against the remaining 6089 proteins from the reference database with these specific testset candidates removed. The program createView achieves this by again, using the JDBC and SQL methods mentioned previously to extract the relevant training candidates from the reference database. This was implemented using the following SQL code:

Step 1:

```
s.executeUpdate("DROP VIEW correlated_set");
```

Step 2:

```
s.executeUpdate("DROP VIEW random_selection");
```

Step 3:

```
s.executeUpdate("CREATE VIEW random_selection
AS SELECT * FROM protein_reference_set
WHERE uid in (" + res_ids + ") ORDER BY job_id desc");
```

Step 4:

```
s.executeUpdate("CREATE VIEW correlated_set
AS SELECT * FROM protein_reference_set
EXCEPT (SELECT * FROM random_selection)
UNION ALL SELECT * from random_selection");
```

Step 5:

```
ResultSet rs = s.executeQuery("SELECT * FROM correlated_set");
```

This program created SQL views which represent reduced portions of a complete database as they were only required temporarily and creating new relational tables was deemed unnecessary. Step 1 and 2 of the SQL statements remove the previous views produced from previous jobs acting as a cleaning procedure. Step 3 creates a view named random_selection containing the random 730 protein candidates using the unique primary ids extracted prior to this process by the parseRes_ids program. Step 4 creates a another view named correlated_set that extracts all proteins from the reference database with the exception of the random candidates contained in the random_selection view created in step 3. The final part of step 4 appends the random candidates to the end of the correlated view. The resulting view contains the full complement of 6819 proteins whereby the first 6089 are training candidates and the final 730 are for testing. Step 5 retrieves all the entries within the correlated view that are organised by training set followed by the testset.

The output of createView now contains all 6819 reference proteins containing only information for the specific combination in question and ordered by training set (6089) followed by testset (730). This output is then written to a file using a local processor (purple) in preparation for the next program.

The file is now in the appropriate format for svm^{light} analysis shown below:

Combination specific created view example

```
1 1:1 5:2.9 7:0
1 1:1 5:4.21 7:2
1 1:-2 5:2.74 7:0
1 1:0 5:1.4 7:2
1 1:1 5:1.6 7:2
-1 1:-1 5:0 7:1
-1 1:-1 5:0 7:0
-1 1:-2 5:0 7:0
-1 1:-1 5:0 7:0
-1 1:-2 5:0 7:0
```

Following this procedure two Java programs were developed to extract the testset and training set from the created view file. These programs parsed the output of the created view. The training candidates were positioned at the top of the created view file and could be filtered out using the specific line numbers (1-6089). The random testset was appended to the bottom of the file and therefore could be filtered using line numbers (6090-6819). Theses outputs were then written to two separate files named testset.txt and trainingset.txt that could be used by svm^{light}. Once these datasets have been filtered they are stored as text files in preparation for the svm^{light} software analysis. The first step of the svm^{light} procedure is the invocation of the program svm^{learn} which performs the SVM training. This is conditionally linked to the file writing processors as this step can only proceed once the training and testing files are available.

Process 5: SVM learning and classification

The software package svm^{light} v6.02 was implemented to perform the training and classification machine learning procedure using the relevant prepared files. The svm^{light} package consists of two scripts, svm^{learn} and svm^{classify} responsible for training the algorithm against the training dataset provided producing a model file and classifying the proteins based on predicting which candidates belong to the positive(1) or negative (-1) class, respectively. Both scripts are accessed through local Taverna processors that communicate with the command line. Each script is invoked via a Java wrapper that executes the specific algorithm. svm^{learn} was invoked using a sophisticated Java program enabling piping information to and from the command line that was not achievable through the local Taverna processor. The local processor invoked the

Java wrapper on the command line as opposed to directly invoking the command line svm^{learn} executable. The specific commands within the wrapper enabled the configuration of the svm^{learn} parameters selected through the optimisation procedure performed at the outset. These parameters were implemented within the Java code as follows:

svm^{learn} parameter configuration
Trade-off between training error and margin

```
command.add("-c");
command.add("25");
```

Using a biased hyperplane

```
command.add("-b");
command.add("0");
```

Using RBF (Radial basis function)

```
command.add("-t");
command.add("2");
```

Setting gamma

```
command.add("-g");
command.add("0.1");
```

Epsilon allowing error for termination criterion

```
command.add("-e");
command.add("0.001");
```

Next command line argument: svm directory

```
command.add(dbFile.getPath());
```

The model file

```
command.add(modelFile.getPath());
```

Finally the sequence file

```
command.add(seqFile.getPath());
```

Give the command line argument to the builder

```
builder.command(command);
```

This list of commands is sent from the Java wrapper to the command line specifying all the arguments required followed by the location of the training file and where the model file should be written (see Figure 2.11 for an illustration of this procedure). The command line receives the following arguments:

```
./svm_learn -c 25 -b 0 -t 2 -g 0.1 -e 0.001
svmlight/trainingset.txt svmlight/model.txt
```

Once this process is complete a model file is produced that can be tested against with the relevant random testset in order to make predictions that can be assessed for sensitivity and specificity regarding proteins being mitochondrial or not. The model file is required by the algorithm $\text{svm}^{classify}$ to perform the predictions and is therefore conditionally linked to the svm^{learn} processor. The $\text{svm}^{classify}$ script is then invoked using a modified version of the Java wrapper that was used to execute svm^{learn}. Unlike the svm^{learn} program, $\text{svm}^{classify}$ did not require any arguments apart from the file locations for the testing data, the model file produced previously and the location of where the predictions should be written. Following production of the prediction file this could be used to calculate sensitivity and specificity of the specific combination under investigation. This program pipes the following command to the command line:

```
./svm_classify svmlight/testset.txt
svmlight/model.txt svmlight/predictions.txt
```

Process 6: Sensitivity and specificity calculations

Two Java programs were developed that calculated sensitivity and specificity. Again, these were invoked on the command line via Taverna's local processors (purple) and results returned to the workflow output. These programs were conditionally linked to the $\text{svm}^{classify}$ processor as they required the completion of the prediction file in order to perform the calculations. Due to the organisation of the testset file (first 100 mitochondrial and last 630 non-mitochondrial) calculating the sensitivities and specificities could be based on line numbers within the prediction file. The SensitivityCalculation program determined sensitivity by calculating the percentage of positive scoring candidates in the first 100 examples. Therefore the SpecificityCalculation program determined specificity by calculating the percentage of negative scoring candidates in the last 630 examples.

Due to the nested architecture of the workflow, 100 sensitivity and specificity results are produced per combination (refer to Process 3). This generates 100 sensitivity and

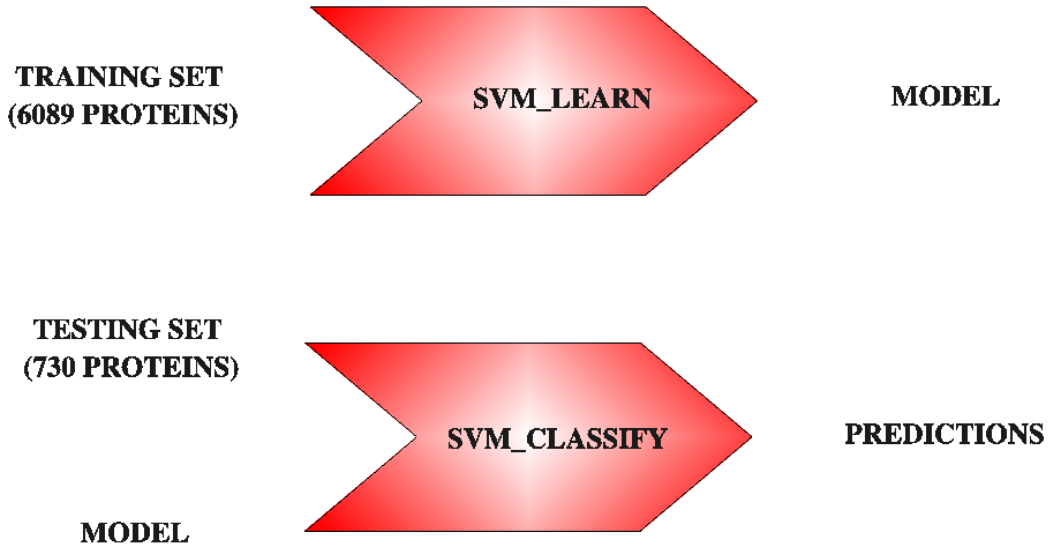

Figure 2.11: The architecture of the svm^{light} pipeline. A training set is used by the program svm^{learn} to construct a model file. A testing set is then queried against the model file classifying the candidates into a positive or negative class using the program svm^{classify}. The final output is a prediction file of values used to determine whether the candidates are classified as mitochondrial or non-mitochondrial.

specificity values for each specific combination as an output from the beige processor. These values can then be stored into the unique tables specified at the start of the workflow created simultaneously with each combination run (refer to Process 1). Each unique combination has a unique relational table containing 100 sensitivity and specificity values. The relational tables are created by the createTable processor which uses the specific combination as the table name. Prior to this the beanshell processor TableName merely adjusts the input for the datasets into syntax accepted by PostgreSQL. For example the input do, tp, in is converted to do_tp_in using underscores instead of comma separated values as this is not acceptable format for PostgreSQL.

Once these values have been stored the following processor SVMStatsQuery performs mean and standard deviation calculations for the sensitivity and specificity of each particular combination. These results are then extracted by the four parsing processors parseMeanSens, parseSDSens, parseMeanSpec and parseSDSpec and passed to the final processor storeSVMCalculations.

On completion of the entire workflow a database is produced containing the means and standard deviations for sensitivity and specificity for all 2047 combinations.

2.2.5 False Discovery Rate Calculations

The false discovery rate (FDR) is the proportion of all the false predictions generated during the analysis. This is represented by the following equation whereby FP = false positives and TP = true positives:

$$FDR = \frac{FP}{(FP + TP)} \tag{2.4}$$

A difference in the proportion of mitochondrial and non-mitochondrial proteins in the training set when compared to the actual proportion in the human genome can bias the FDR. This can be accounted for by scaling the FDR based on the estimated number of mitochondrial proteins encoded by the nuclear genome (1,500 of 21,000 genes). Using this approach, the corrected FDR can be calculated as below whereby TN = true negatives and FN = false negatives.

$$cFDR = \frac{(1 - specificity)}{(1 - specificity + sensitivity)} x \frac{1500}{21000} \tag{2.5}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{2.6}$$

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{2.7}$$

2.2.6 Genome wide analysis using MitoSVM

Following the determination of the best combination achieving the highest sensitivity without compromising specificity, the resulting SVM model file was used to classify all human proteins contained within the Ensembl Human database. An SVM testfile was created containing values for the best combination of classifiers for all 63,271 human transcripts contained within the Ensembl database. These could then be ordered by chromosomal location and descending SVM score to highlight the strongest candidates for further investigation based on genomic region of interest.

2.3 Results

2.3.1 Performance of the combination workflow

Each individual workflow run completed in 5 seconds. As each combination was analysed 100 times this resulted in each job requiring approximately 8 minutes for completion. Analysing the entire set of 2047 combinations required 273 hours (approximately 11.5 days) of automated processing time. The workflow automates a large analysis protocol that would require several months to complete using a manual approach. This effectively eradicates human error and is simple regarding method repeatability. All the resulting data is automatically stored and organised into a relational database providing an appropriate architecture for analysis.

2.3.2 Sensitivity for all combinations

A barcode plot was produced displaying all the sensitivity results for the 11 different combinations from the genome-scale datasets generated by (Calvo *et al.*, 2006) in descending order of score (Figure 2.12). This is a visual display illustrating the strongest classifiers in contrast to the weakest. The top of the plot contains the datasets that were most consistent in achieving a high sensitivity and lack the presence of the weakest predictors. Towards the bottom of the plot the weakest predictors are most abundant with the absence of the strongest classifiers. Coloured arrows indicate the location of the prediction methods in isolation highlighted by a colour coded key allowing for the evaluation of the importance of using combinations of datasets as opposed to one method alone. Figure 2.12 also highlights the top 100 combinations to the left of the image providing a magnified view displaying which classifiers are involved in the highest sensitivities.

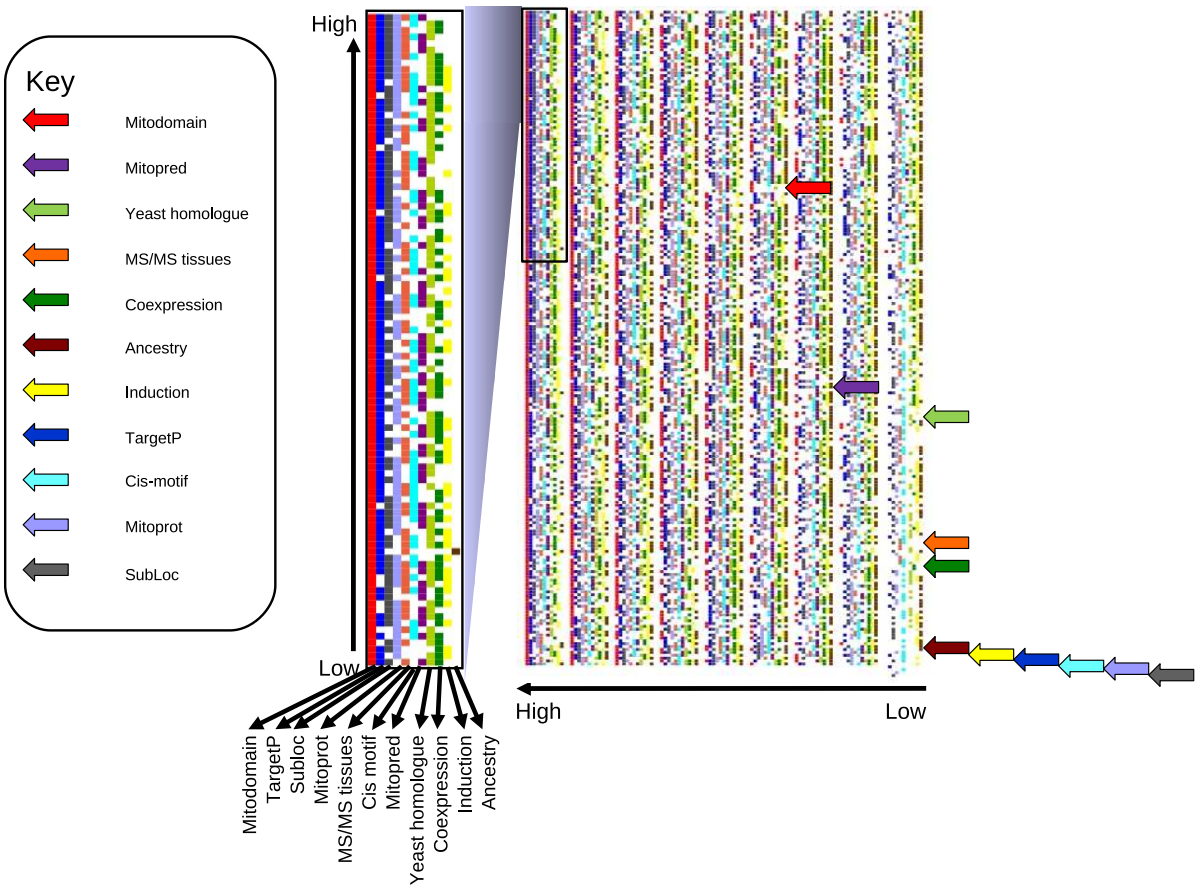

Figure 2.12: Graphical display of all combinations in order of descending sensitivity (top to bottom moving from the left column to the right) with a magnified view of the top 100 combinations to the far left. Coloured arrows point to the location of the individual methods in isolation. Each specific combination is represented by a sequence of coloured bars reflecting the specific classifiers the combination contains.

Boxplots were produced displaying the mean sensitivities and specificities along with the range for differing numbers of classifiers (Figures 2.13 and 2.14). The range for sensitivity decrease as more classifiers were involved reaching an optimum number of classifiers. This pattern was also expected to be reflected regarding the mean values. For specificity the values for the range and mean values were not expected to fluctuate and remain fairly consistent regardless of the number of classifiers involved in the prediction. The probability of an independent dataset contributing to a sensitivity prediction was calculated using the results from the entire combination analysis. Figure 2.15 displays a heat coloured contour plot of results ordered along the x axis starting with the dataset expressing the highest probability of generating high sensitivity values. Predictors contributing to high percentages of high sensitivity values are displayed by red colour. Low percentage contribution is displayed in yellow allowing for a clear contrast between the performance of the 11 prediction methods. The same data is displayed in Figure 2.16 in the form of a line graph with each individual line representing an independent dataset. Strong predictors migrate to the top right of the graph reflecting high percentage contribution. Low percentage contribution to high sensitivity values is reflected by lines migrating downwards to the bottom right of the graph. The same analysis was performed for specificity and can be seen in Figures 2.17 and 2.18. Values for specificity displayed negligible differences between the 11 prediction methods. This resulted in a clustering of values around 99%.

2.3.3 Sensitivity and specificity of all combinations

A scatterplot was produced displaying the sensitivities and specificities of all 2047 combination results. This was expected to display a clustering of results in the top right hand corner of the graph pertaining to 7 an 8 classifiers, corresponding to the best prediction for mitochondrial localisation. This would clarify that this number of datasets were required to achieve optimal sensitivity without reducing the specificity (Figure 2.19).

2.3.4 Standard deviations of sensitivity and specificity

Figure 2.20 displays the standard deviations for all 2047 combinations. Each combination was tested with random testsets 100 times generating standard deviations for each individual combination. An expected trend would be a stable value for standard deviation for both sensitivity and specificity. It was hypothesised that as the mean sensitivity increased the standard deviation decreased and stabilised around a reasonable threshold. Figure 2.21 displays the standard deviation values against the sensitivity of

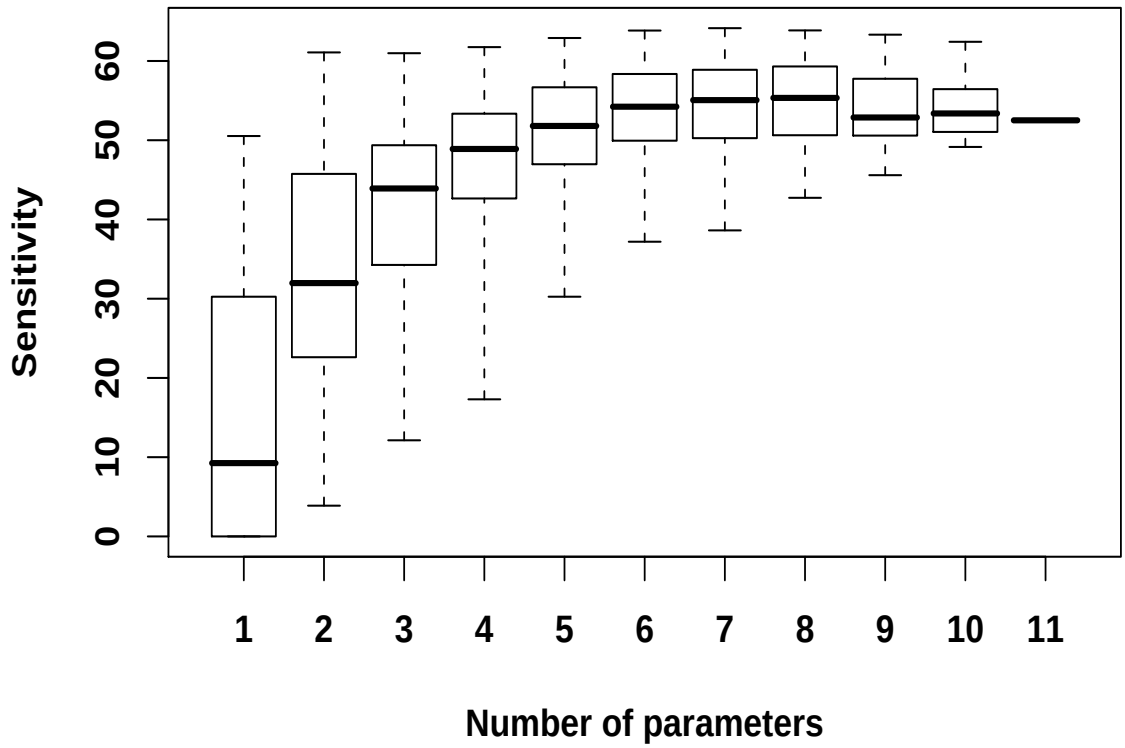

Figure 2.13: Boxplots illustrating the statistics of the sensitivity values for all 2047 combinations of the 11 different prediction tools. The first boxplot on the far left represents all the sensitivity values when one classifier is involved in the prediction. Each boxplot moving to the right displays the statistics when an additional classifier is added with the far right boxplot displaying the result when all 11 classifiers are involved in the prediction. Horizontal bar = mean, box = standard deviation and whiskers = range.

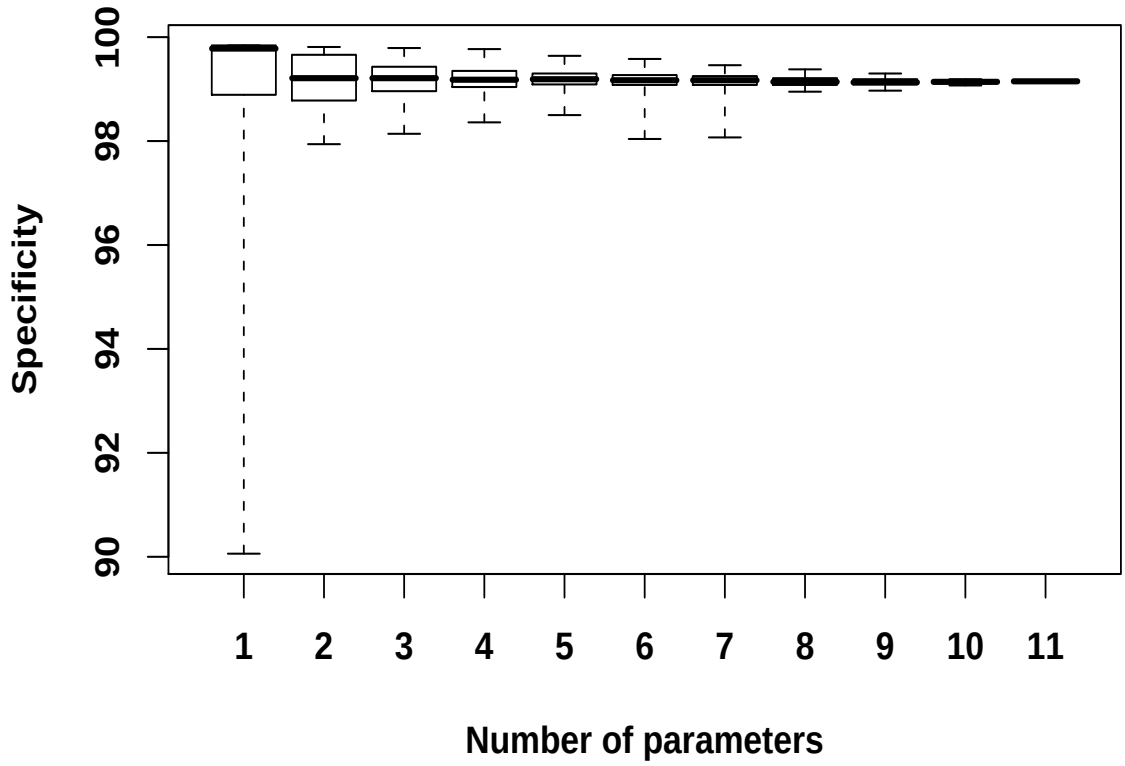

Figure 2.14: Boxplots illustrating the statistics of the specificity values for all 2047 combinations of the 11 different prediction tools. The first boxplot on the far left represents all the specificity values when one classifier is involved in the prediction. Each boxplot moving to the right displays the statistics when an additional classifier is added with the far right boxplot displaying the result when all 11 classifiers are involved in the prediction. Horizontal bar = mean, box = standard deviation and whiskers = range.

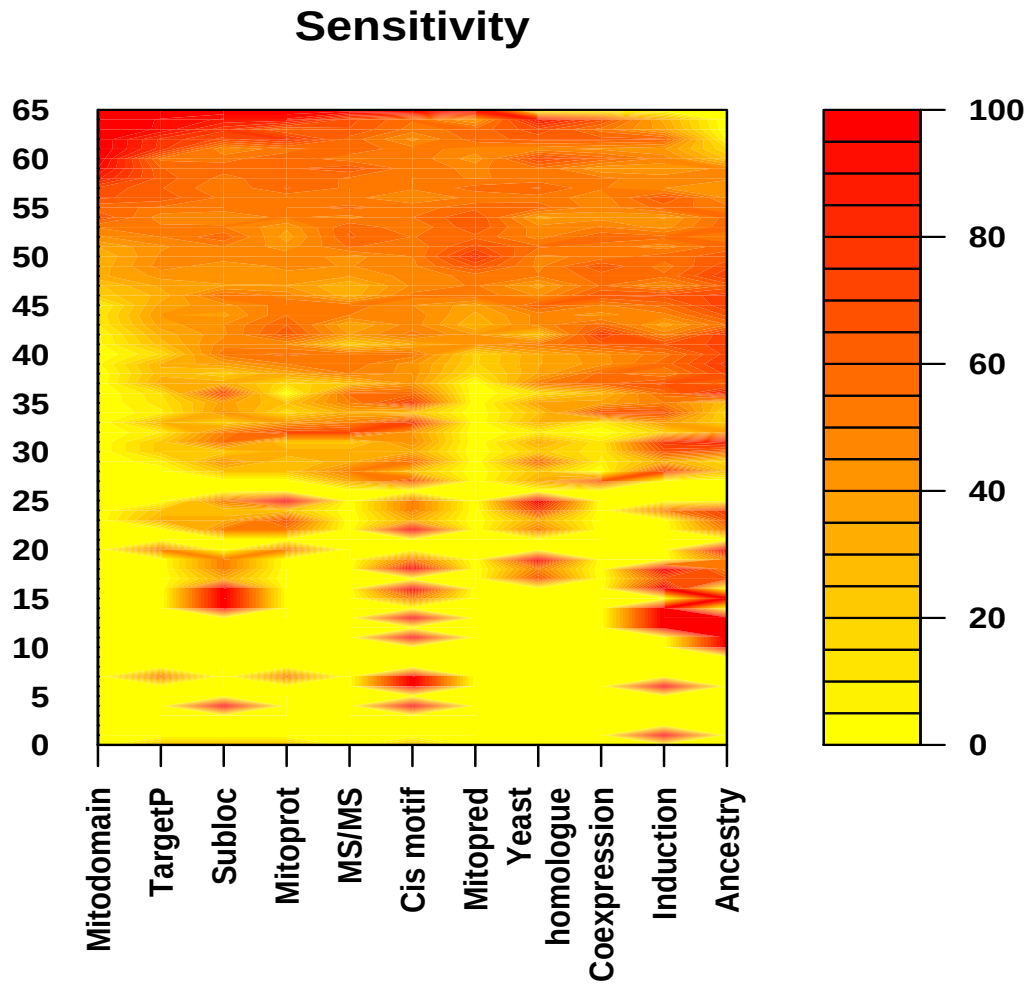

Figure 2.15: A heat coloured contour plot representing the percentage contribution of each prediction tool to a given level of mean sensitivity. Red reflects 100% contribution to that specific level of sensitivity and yellow reflects 0%. The colour is a gradient represented by the key on the far right of the diagram. The classifiers range in predictive strength from left to right.

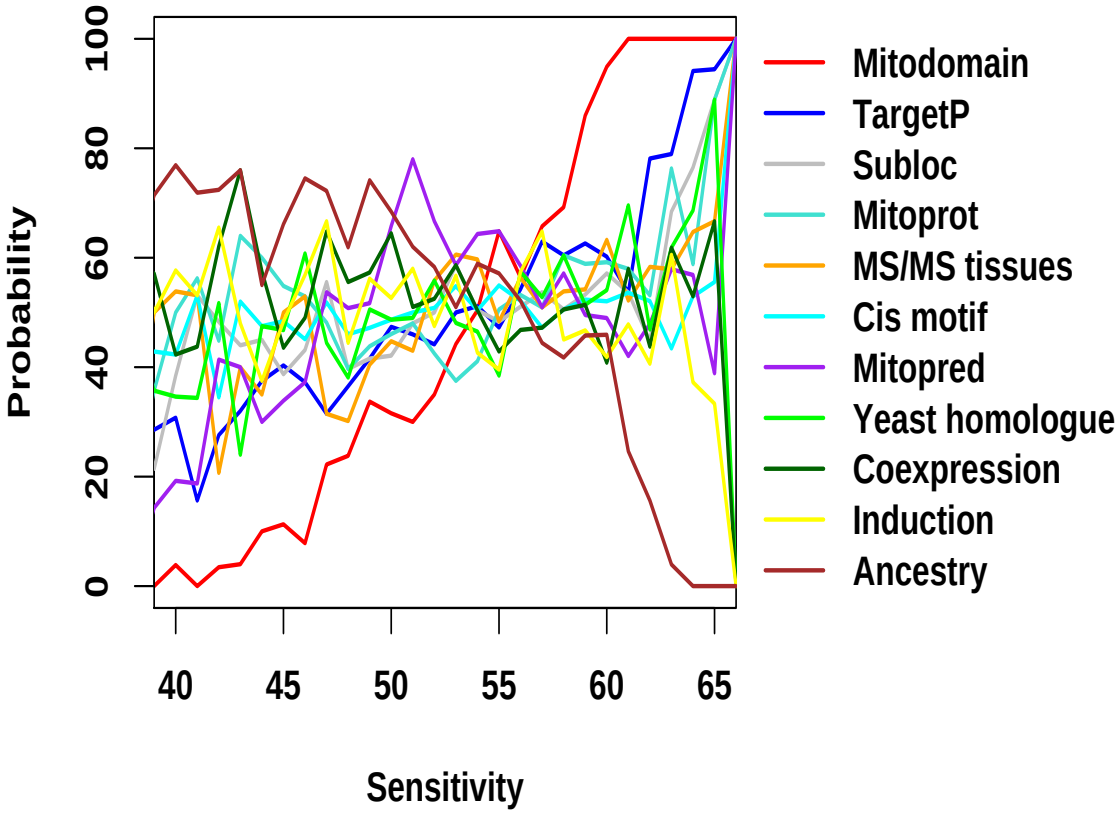

Figure 2.16: A line plot representing the percentage contribution of each prediction tool to a given level of mean sensitivity. Strong classifiers are reflected by lines that move from bottom left to top right as this displays low contribution to low sensitivities and high contribution to high sensitivities. An opposite trend is displayed for weak classifiers.

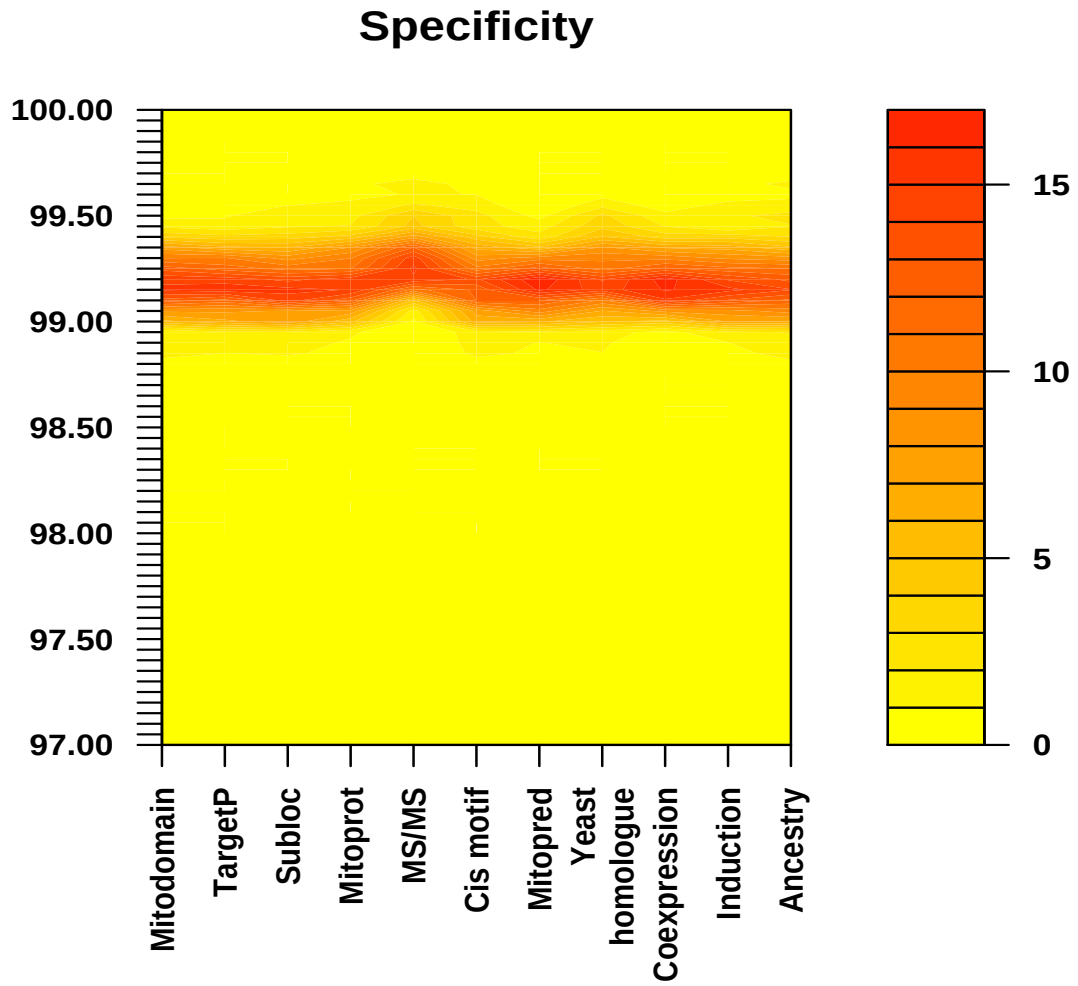

Figure 2.17: A heat coloured contour plot representing the percentage contribution of each prediction tool to a given level of mean specificity. Red reflects 100% contribution to that specific level of specificity and yellow reflects 0%. The colour is a gradient represented by the key on the far right of the diagram. The classifiers range in predictive strength from left to right.

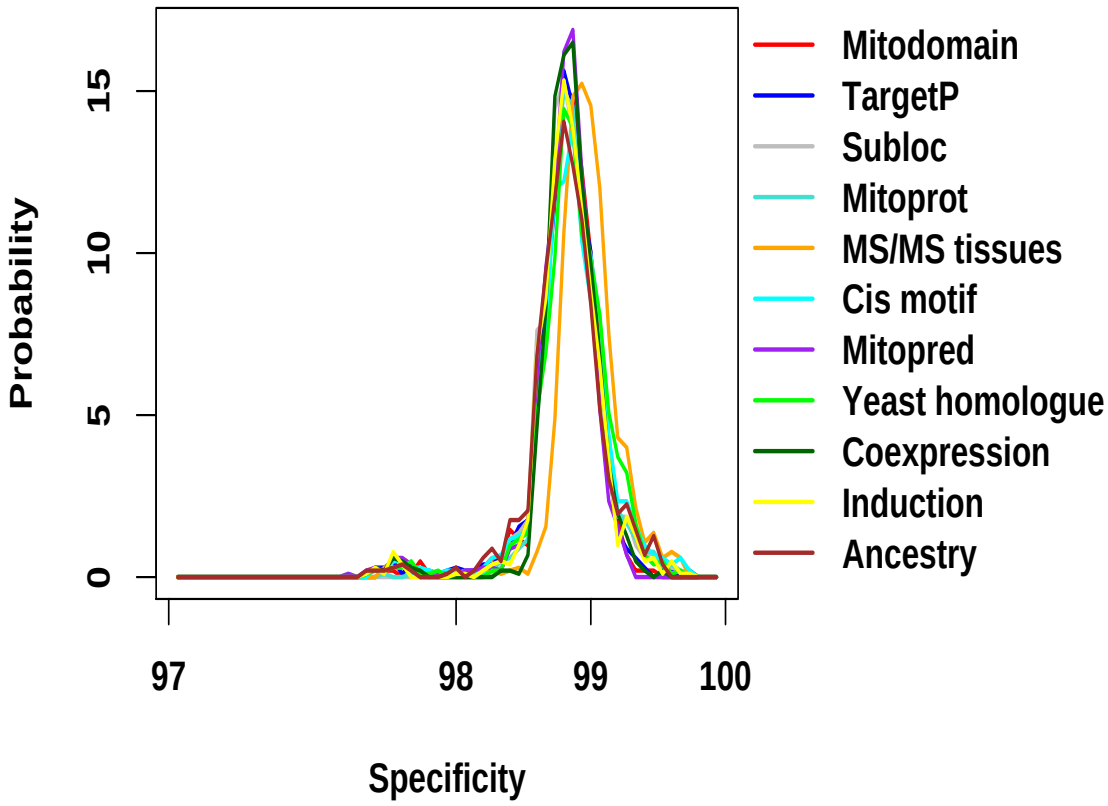

Figure 2.18: A line plot representing the percentage contribution of each prediction tool to a given level of mean specificity. Strong classifiers are reflected by lines that move from bottom left to top right as this displays low contribution to low specificities and high contribution to high specificities. An opposite trend is displayed for weak classifiers.

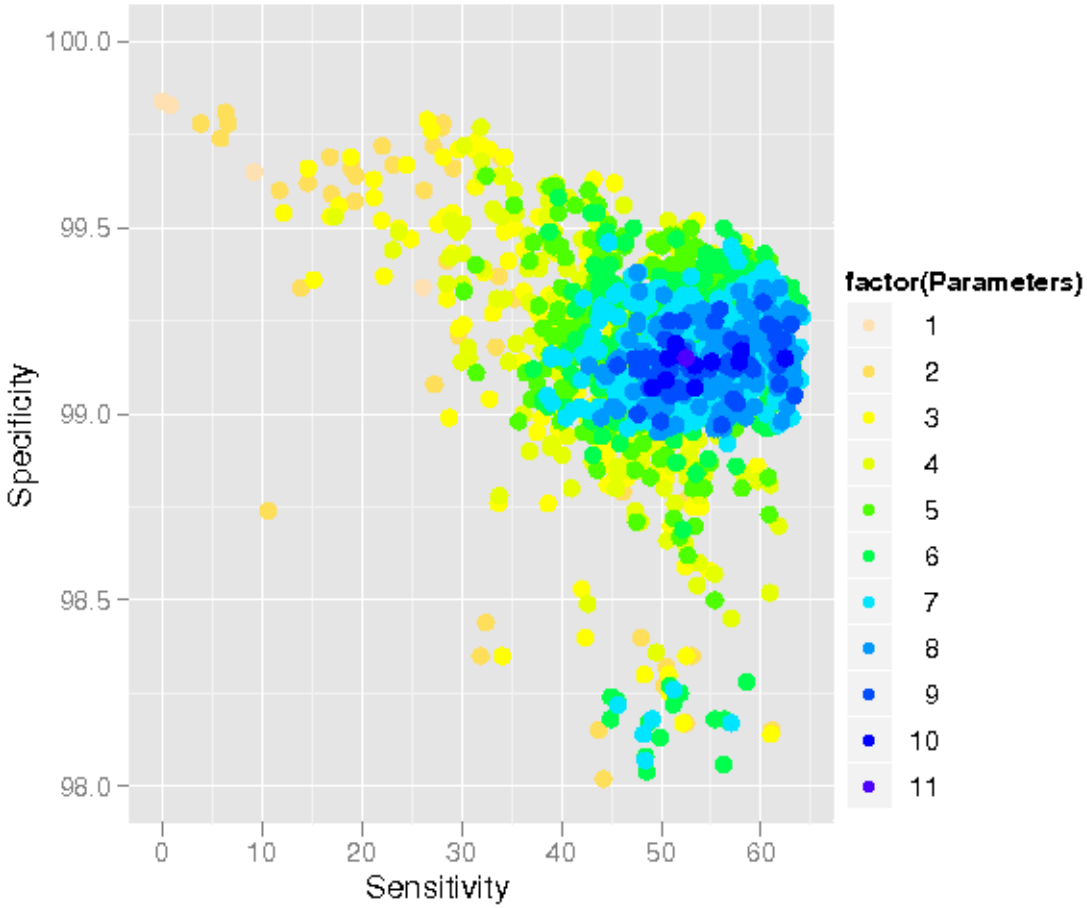

Figure 2.19: A scatterplot displaying sensitivity against specificity for all 2047 combinations comparing the differences when changing the number of classifiers involved in a prediction. The colour key refers to the number of prediction methods involved in that particular combination. Combinations clustering in the top right hand corner of the plot represent the highest sensitivities without compromising specificity. Each colour represents a different number of classifiers.

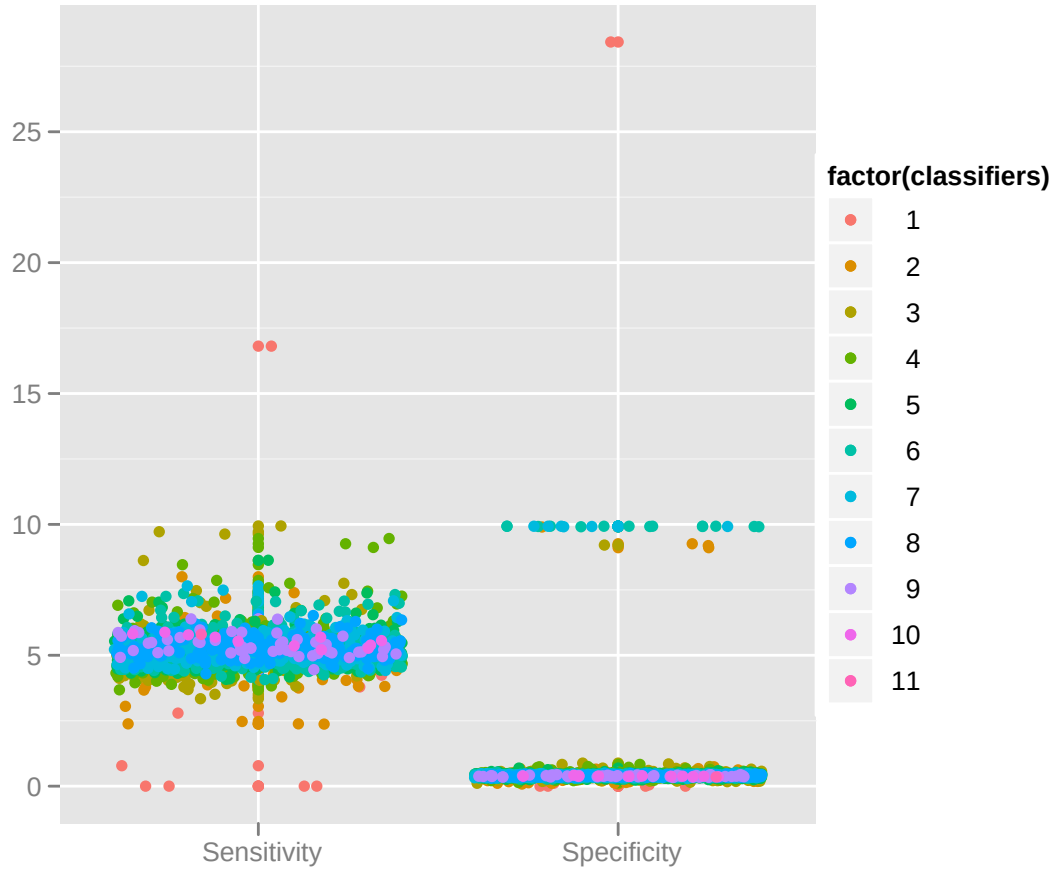

Figure 2.20: Standard deviations for sensitivity and specificity for all 2047 combinations comparing the differences when changing the number of classifiers involved in a prediction. The colour key reflects the number of prediction methods involved in that particular combination. The left plot displays all the standard deviations for the sensitivity results and the right plot displays all the standard deviations for the specificity results.

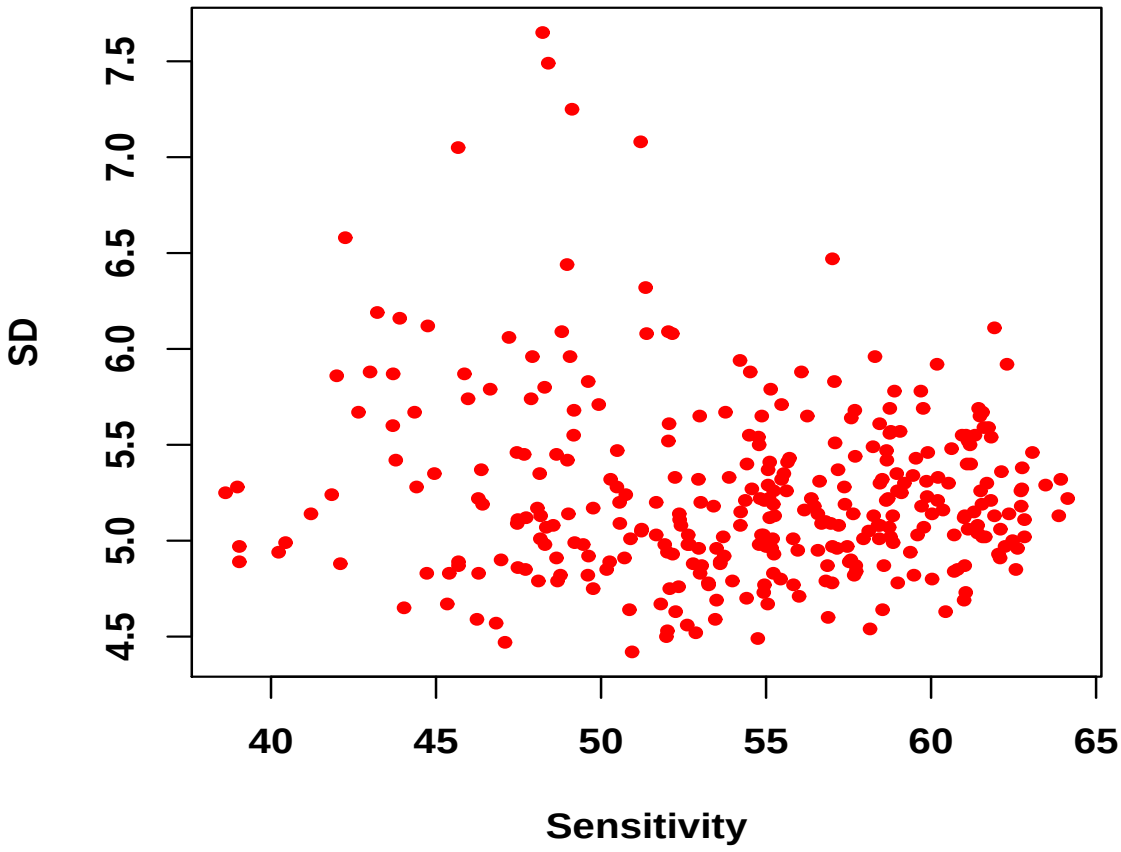

Figure 2.21: A scatterplot of standard deviations against sensitivity for all combinations involving 7 classifiers. This is to illustrate the trend of standard deviations as sensitivity increases.

all combinations that involved 7 classifiers. This would represent a higher mean having a lower standard deviation pertaining to a more reliable prediction accuracy.

2.3.5 False discovery rate

The false discovery rate (FDR) and corrected false discovery rate (cFDR) is an important assessment of all the false predictions generated during the investigation. These rates are expected to decrease as the number of parameters approaches the optimum number. Figure 2.22 displays the results of these rates in relation to one another. Boxplots displaying these values are displayed in Figures 2.23 and 2.24 with the addition of the means and ranges of values generated.

2.3.6 Genome wide analysis using MitoSVM

An SVM model file was produced based on the best combination of classifiers and implemented using the SVM to classify all proteins in the human genome contained in the Ensembl database. Table 2.4 displays the top 20 highest scoring mitochondrial candidates in the human genome based on the SVM model that achieved the highest sensitivity and specificity.

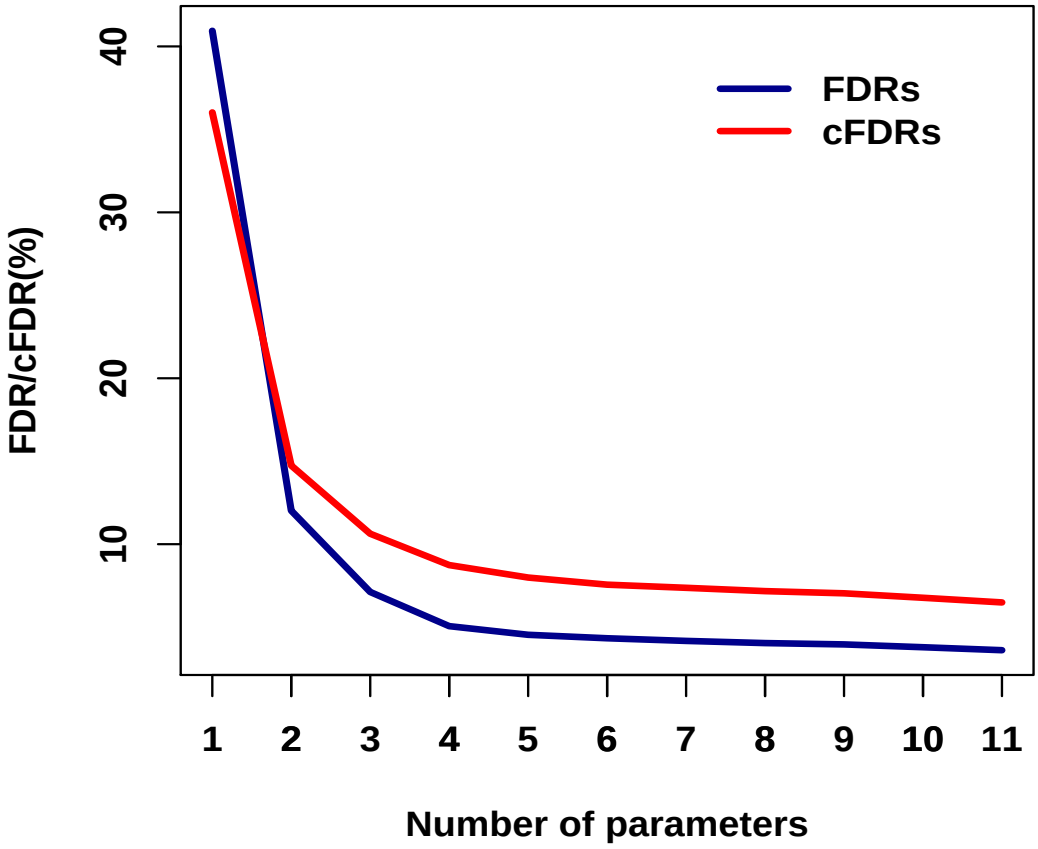

Figure 2.22: A plot displaying the false discovery rates (FDR) and corrected false discovery rates (cFDR) when increasing the number of parameters involved in a prediction.

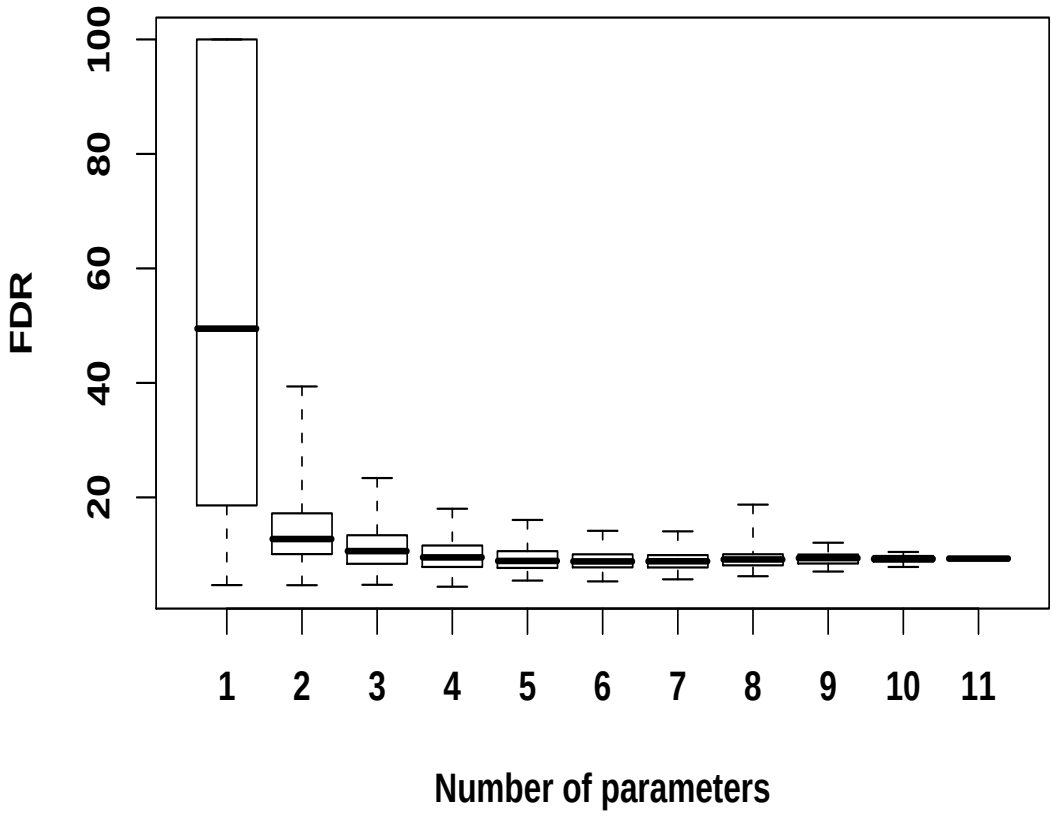

Figure 2.23: Boxplots illustrating the statistics of the FDR values for all 2047 combinations of the 11 different prediction tools. The first boxplot on the far left represents all the FDR values when one classifier is involved in the prediction. Each boxplot moving to the right displays the statistics when an additional classifier is added with the far right boxplot displaying the result when all 11 classifiers are involved in the prediction. Horizontal bar = mean, box = standard deviation and whiskers = range.

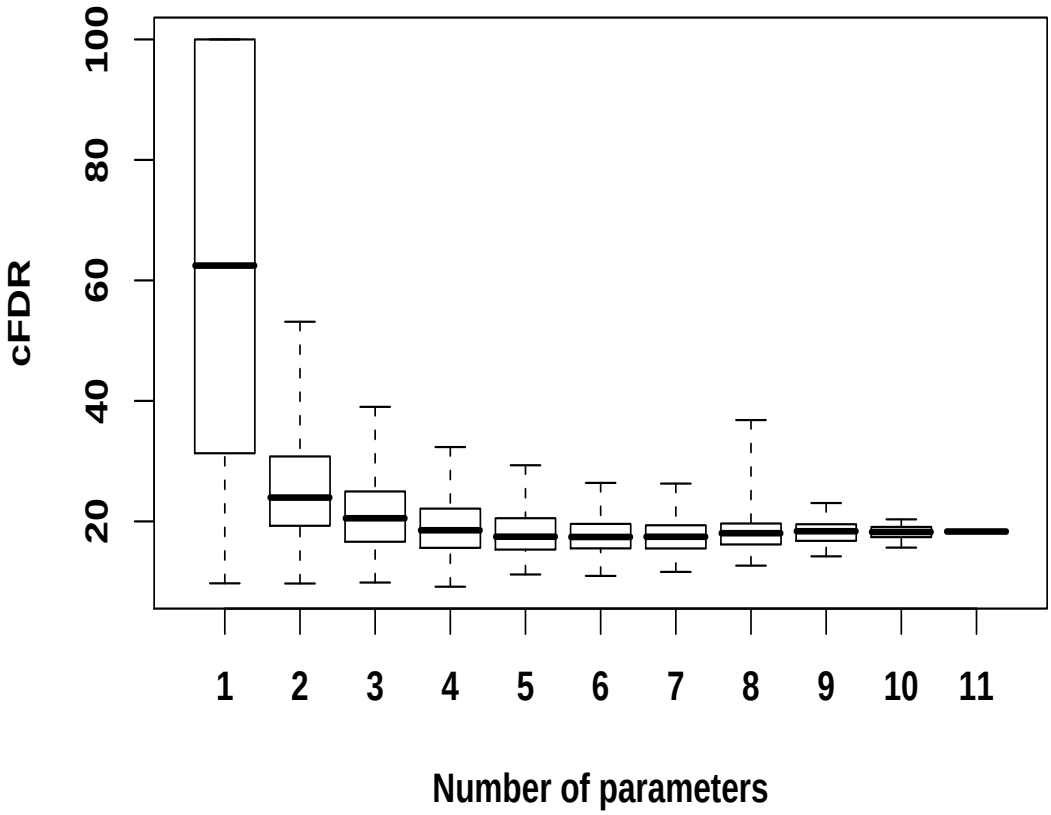

Figure 2.24: Boxplots illustrating the statistics of the cFDR values for all 2047 combinations of the 11 different prediction tools. The first boxplot on the far left represents all the cFDR values when one classifier is involved in the prediction. Each boxplot moving to the right displays the statistics when an additional classifier is added with the far right boxplot displaying the result when all 11 classifiers are involved in the prediction. Horizontal bar = mean, box = standard deviation and whiskers = range.

Ensembl Gene	Ensembl Protein	Gene	Chromo	Start bp	End bp	SVM Score
ENSG00000166998	ENSP00000300283	CKMT1A	15	41772376	41778712	2.701
ENSG00000168775	ENSP00000338496	CKMT1B	15	41672544	41678896	2.701
ENSG00000010256	ENSP00000203407	UQCRC1	3	48611436	48622102	2.320
ENSG00000182199	ENSP00000333667	SHMT2	12	55909786	55914981	2.306
ENSG00000110717	ENSP00000315774	NDUFS8	11	67554670	67560686	2.298
ENSG00000131730	ENSP00000254035	CKMT2	5	80564895	80597970	2.271
ENSG00000138095	ENSP00000260665	LRPPRC	2	43968391	44076648	2.270
ENSG00000136521	ENSP00000353026	NDUFB5	3	180805269	180824981	2.109
ENSG00000183044	ENSP00000268251	ABAT	16	8675928	8785933	2.027
ENSG00000167969	ENSP00000301729	DCI	16	2229901	2241604	1.992
ENSG00000126432	ENSP00000265462	PRDX5	11	63828023	63845858	1.964
ENSG00000156709	ENSP00000316320	AIFM1	X	129091018	129127489	1.942
ENSG00000136521	ENSP00000259037	NDUFB5	3	180805269	180824981	1.896
ENSG00000137513	ENSP00000281038	NARS2	11	77824895	77963474	1.860
ENSG00000157326	ENSP00000311993	DHRS4	14	23492805	23508326	1.858
ENSG00000146701	ENSP00000327070	MDH2	7	75515329	75533864	1.858
ENSG00000165672	ENSP00000349432	PRDX3	10	120917205	120928335	1.845
ENSG00000114686	ENSP00000264995	MRPL3	3	132663736	132704519	1.826
ENSG00000184117	ENSP00000216121	NIPSNAP1	22	28280800	28307244	1.795
ENSG00000181378	ENSP00000295729	CCDC108	2	219575820	219614489	1.772

Table 2.4: Top 20 highest scoring mitochondrial candidates in the human genome generated using the strongest combination of classifiers following the systematic analysis MitoSVM. The results are displayed in order of descending SVM score.

2.4 Discussion

2.4.1 Biological discussion

Increasing the number of independent datasets to accurately determine the mitochondrial proteome does not contribute to an increase in mean sensitivity. Specific classifiers have been shown to make a positive contribution to a high value for mean sensitivity whereas others are unable to achieve high scores. A support vector machine was trained with an established set of experimentally determined proteins extracted from the SwissProt database. This reference set consisted of 6819 proteins consisting of 467 mitochondrial and 6352 non-mitochondrial candidates. These were then analysed implementing their values from the 11 independent datasets testing every combination of these classifiers(n=2047) and repeated 100 times. Approximately 90% of the reference dataset was used as training data and the remaining 10% as testing data.

Mitodomain is present in all of the top 100 combinations reflecting the importance of this dataset in contributing to the highest sensitivities. The presence of mitochondrial domains appears to be a highly successful method for mitochondrial protein determination. However, ancestry is absent in all but one of the top 100 combinations contributing to lower sensitivity values. TargetP was also present in a high percentage of the top 100 combinations revealing the importance of implementing neural networks for N-terminal target sequence prediction.

Using a small number of prediction tools (<4) resulted in considerable variation for both mean sensitivity and specificity expressing a large range of mean values. Increasing the number of prediction tools reduced the range considerably and increased the mean sensitivity without compromising specificity. However, using more than 8 prediction tools resulted in a decrease in mean sensitivity thus reflecting the optimal number of prediction tools being exceeded for accurate mitochondrial protein determination. The mean specificity of the SVM tests was high (>98%) regardless of the number of prediction tools implemented. Similar results were obtained for the false discovery rate (FDR) and corrected false discovery rate (cFDR) which incorporates prior probabilities in the calculations.

Another method for comparing the classifiers regarding sensitivity and specificity was to determine the probability of any one classifier being involved in a prediction with a specific value for sensitivity and specificity. Each classifier had an equal likelihood of contributing to a high specificity. However, only specific prediction tools had a high probability of generating high sensitivity scores. Mitodomain, TargetP and SubLoc are more abundant when the sensitivity is >57% than the other classifiers and consistently associated with high sensitivity predictions. In contrast, ancestry involving the presence

of *R. prowazekii* orthologues expressed a low probability of <20% when contributing to a sensitivity of >60% and had an 80% probability of generating a sensitivity value of <40%. In addition, these results are confirmed when the whole result set was ranked in order of descending sensitivity. The poor performance of *R. prowazekii* homology is probably due to several factors limiting the ability for a successful mitochondrial prediction, including the limited number of orthologous proteins discovered between *H. sapiens* and *R. prowazekii* and the poor overall homology between the two distantly related species which shared a common ancestor over 2.5 billions years ago (Kurland and Andersson, 2000).

The highest sensitivities were achieved when specific bioinformatic methods were involved in unison which included Mitodomain, TargetP and SubLoc reflecting the complementary nature of these approaches based on the presence of protein domains only present in mitochondrial proteins; the implementation of neural networks for N-terminal target sequence prediction; and amino acid composition. In order to achieve optimal sensitivity the addition of four other classifiers was required. These were Mitoprot for target sequence prediction; mouse mitochondria detected in brain, heart, liver and kidney; *cis*-regulatory motifs present in human/mouse orthologues; and Mitopred (a method combining the occurrence patterns of Pfam domains, amino acid composition, and pI value differences between mitochondrial and non-mitochondrial locations).

The optimum number of classifiers was determined for both mean sensitivity and specificity for all 2047 combinations. This displays a clustering in the top right hand corner of the plot reflecting 7 and 8 classifiers achieving the highest sensitivity and specificity values. This corroborates with the earlier results displayed in the boxplots displaying the range of mean values for sensitivity and specificity. Incorporating extra classifiers above this threshold compromised both sensitivity and specificity confirming earlier hypotheses.

Omitting the following classifiers was necessary for optimal sensitivity: *S. cerevisiae* homology, gene expression difference during mitochondrial biogenesis induced by PGC-1a, coexpression with known mitochondrial genes in human/mouse tissue atlases, and *R. prowazekii* homology. The best 7 prediction methods achieved a mean sensitivity of 64.14% with an SD of 5.22 and was significantly greater than the values generated when implementing all 11 classifiers (52.51, SD=5.80).

The same pattern was observed following the addition of further experimentally derived predictions of mitochondrial localisation. Recent research conducted by (Pagliarini *et al.*, 2008) involved the addition of a log-likelihood prediction of mitochondrial local-

isation involving subtractive MS/MS following mitochondrial enrichment. Adding this to the 11 initial prediction methods produced a combined mean sensitivity result of 50.89 (SD=4.54) and specificity of 98.89 (SD=0.44). The addition of the subtractive MS/MS enrichment dataset did not produce significantly different results. However, removing the four prediction methods shown previously to compromise sensitivity led to an improved prediction (mean sensitivity = 69.87, SD = 4.40, specificity = 99.23, SD = 0.34). This result was not a significant improvement when compared to the result for the best 7 predictors without the subtractive MS/MS enrichment. However, this does demonstrate the importance of selecting the optimum combination of prediction methods based on systematic analysis.

2.4.2 Technical discussion

Each time the machine learning algorithm was executed for any combination, considerable variability was apparent regarding sensitivity. This feature appears to be absent from previous research. As the number of prediction tools increased a general trend emerged resulting in a decreased standard deviation. However, the standard deviation did not decrease with increasing mean sensitivity as the majority of mean sensitivity predictions resulted in an SD of 5%. Therefore, the 95% confidence intervals for mean sensitivity of any combination was +/-10% of the mean value highlighting the inadequacies of even the most optimal bioinformatic approaches for mitochondrial localisation prediction.

The approach used for this investigation differs from previous research involving mitochondrial protein determination in several ways. Firstly, a support vector machine was used for multidimensional classification and explicitly optimised to reduce the number of support vectors to 10%. If the number is greater than 10% of the number of proteins in the reference dataset, then the performance of the SVM will result in overfitting the data leading to an artificial elevation in performance. This would mean the chances for novel predictions are dramatically reduced. Secondly, the results are harnessed from a rigorous statistical comparison of each prediction tool. Considerable variation between each run regarding a specific combination reveals the possibility that previously reported integrated approaches claiming to achieve high sensitivity and specificity values may be erroneous. Specific combinations may perform poorly over repeated runs with independently sampled datasets. Thirdly, the entire analysis was designed utilising current workflow technology. This has several advantages as the design of the workflow is modular and specific elements can be easily added, substituted or removed allowing the pipeline to remain contemporary. As tools are enhanced and new

experimental datasets are produced these can be incorporated into the workflow. As the procedure is fully automated this is time-efficient and repeatable regarding the methodology. This e-science based approach means the list of mitochondrial proteins will be constantly revised and updated, including revisions to the human genome sequence. In addition, this will allow the incorporation of genomic variation in the predictions with the increasing depth of next generation sequencing. Using the most complementary bioinformatic methods will accelerate the reliable identification of human disease candidate genes responsible for novel mitochondrial disorders, providing a molecular diagnosis for families, and reveal novel disease mechanisms.

An important limitation is apparent when collating training data for machine learning classification. The lack of verified negative data regarding non-mitochondrial proteins is an issue. The non-mitochondrial examples were experimentally verified to localise to a specific subcellular compartment but there is an absence of information pertaining to their absolute disassociation from mitochondria. It is evident that various proteins can co-localise to more than one compartment within the cell. SwissProt lacks this information as the documentation for each protein reflects positive information garnered through directed investigation. It is crucial for any training algorithm to contain experimentally validated true negatives as well as true positives.

However, mitochondrial genes are believed to comprise only 4% of the human genome. As the training set was acquired randomly the non-mitochondrial dataset will be fairly enriched with the desired negative examples with only a small background of false negatives. To achieve the most robust 'gold standard' training dataset, negative data must be verified and the evidence has to reflect no biological association with mitochondria to achieve high confidence regarding the true negative examples.

Another issue is that of data circularity with regards to the mitodomain dataset. Mitodomain consists of protein domains annotated as mitochondrial within the Pfam database. Pfam entries are derived from an underlying sequence database known as Pfamseq which is built from the most recent release of UniProt and SwissProt. This may be the reason for mitodomain being the strongest classifier as the SVM training dataset was extracted from SwissProt. A common problem found in training machine learning algorithms in bioinformatics is the issue of training datasets consisting of bioinformatically-predicted entries. This was avoided by specific filters during extraction from SwissProt which specified direct experimental evidence of protein localisation.

Chapter 3

Identification of nuclear-mitochondrial genes involved in LHON

Abstract

Background

Text mining applications are playing a huge role in the bioinformatics field providing resources for interpreting the vast quantities of biomedical literature currently available. The construction of a workflow designed to mine keywords and phrases from gene ontology records has produced interesting disease candidates potentially involved in Lebers hereditary optic neuropathy (LHON).

Results

Strong candidates have been identified that may be involved in the pathology of LHON. MitoCarta and MitoSVM have been interrogated for genes that score high regarding the prediction scores and also reside in the specific linkage region on the X chromosome. Additionally, text mining has revealed candidates that may have an indirect association to mitochondria possessing phenotypic associations closely resembling LHON disease characteristics. Orthologues were also investigated in mouse, rat and chimpanzee for eye-related genes that were not documented as such in humans potentially revealing novel LHON disease genes.

Conclusion

Several interesting candidates were generated following text mining analysis of genes potentially associated with LHON. These lists contained unique genes and mouse orthologues phenotypically related to eye disorders. These genes can be further investigated to highlight any relationship with LHON.

3.1 Introduction

3.1.1 Leber hereditary optic neuropathy

Leber hereditary optic neuropathy (LHON, MIM: 535000) is a mitochondrial genetic disease first recognised as a familial optic neuropathy in 1871 by the German ophthalmologist, Theodor Leber. LHON is a common cause of inherited blindness most prevalent in young adult males affecting at least 1 in 30,000 individuals in the UK (Yu-Wai-Man *et al.*, 2009). LHON is primarily characterised by bilateral subacute loss of central vision resulting from a focal degeneration of the retinal ganglion cell layer. This occurs within the papillomacular bundle responsible for sending information to the optic nerve (Man *et al.*, 2002; Yu-Wai-Man *et al.*, 2009). It is a maternally inherited disease that is usually painless but visual loss is typically rapid progressing from one eye to both from a few weeks to months later (Levin, 2007). Over 95% of cases arise from one of three pathogenic mtDNA point mutations: m.3460G>A, m.11778G>A and m.14484T>C (Kirkman *et al.*, 2009b). LHON affected patients have been assessed for quality of life using a VF-14 questionnaire that measures an individuals ability to perform 14 vision-dependent activities. The results of this study indicated that the visual impairment in LHON had a severe impact on the quality of life of these patients in comparison to other inherited and acquired ophthalmic disorders (Kirkman *et al.*, 2009a). The mean VF-14 score in LHON sufferers is the worst yet determined at a level of 25.1 (SD=20.8; range=0-95) which is significantly lower than the mean VF-14 score for unaffected LHON carriers of 97.3 (SD=7.1; range=25-100) (Kirkman *et al.*, 2009a).

Previous research has suggested potential nuclear genetic involvement in the aetiology of LHON and determined an X-chromosomal locus harbouring a susceptibility allele (Hudson *et al.*, 2005; Shankar *et al.*, 2008). Evidence has indicated that nuclear modifying genes and environmental factors may be required in addition to mtDNA mutations to cause optic neuropathy (Shankar *et al.*, 2008). LHON-associated mutations are maternally transmitted to all offspring but most do not develop the disease, even in homoplasmic individuals. LHON exhibits a male gender bias and variably reduced penetrance (Howell, 1998; Puomila *et al.*, 2007).

3.1.2 GoPubMed

Generating new research ideas requires a detailed knowledge and awareness of the subject area. Thankfully, due to the availability of text mining services, literature can be intelligently mined for relevant articles. GoPubMed allows you to mine PubMed abstracts using specific search terms. Four options are available including an ontology-based literature search that annotates the PubMed abstracts with your keywords and then groups

the articles into a hierarchy based on the gene ontologies (Doms and Schroeder, 2005). An advanced semantic search allows you to search using actual GO terms such as "protein biosynthesis" or "apoptosis". The HotTopic feature is a really useful system for retrieving statistical data based on the biomedical literature. This feature performs a bibliometric analysis graphically displaying the growth of literature chronologically. Key authors and journals are highlighted along with geographic information based on the publications.

3.1.3 Gene Ontology

The Gene Ontology Consortium aims to provide a comprehensive vocabulary to unite all genes across different database repositories when involved in the same molecular functions and biological processes. This provides a framework that can allow software to automatically clarify distributed information regarding a gene. Gene ontologies consist of 3 species-independent categories: 1) Cellular component describing the organelle or subcellular structure the gene is associated with (e.g. mitochondria, endoplasmic reticulum), 2) Biological process indicating the specific phases a gene is involved in (e.g calcium ion transport) and 3) Molecular function that describes the activities the occurs at the molecular level (e.g protein binding) (Ashburner *et al.*, 2000). Gene and protein functions are recognised as being evolutionary conserved in most living cells, allowing biologists to assign functional roles to uncharacterised genes using knowledge from sequence homology to genes in related organisms. The Gene Ontology Consortium is currently comprised of several organism databases including the Mouse Genome Informatics (MGI) database, Saccharomyces Genome Database (SGD) and several repositories regularly annotated by the European Bioinformatics Institute (EBI) including chicken, cow and human (Ashburner *et al.*, 2000).

3.1.4 Homology

Homology describes a gene or protein that is derived from a common ancestor and conserved over evolutionary time. This term consists of distinct types including orthology and paralogy. An orthologue is evolutionary conserved and related via speciation retaining the same biological function (vertical descent) (Koonin, 2005). Orthologues provide crucial information that can be extrapolated among many different species and provide important information when performing sequence similarity studies using software such as BLAST (Altschul *et al.*, 1990). These homologous sequences can reveal important information about uncharacterised genes and proteins established in other organisms. A paralogue is a gene that shares a common ancestral form but has undergone various changes resulting in a change of function. Paralogues belong to the same

organism and usually arise following a gene duplication event (Koonin, 2005).

3.1.5 Online Mendelian Inheritance in Man

Online Mendelian Inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov/omim/) is a clinical database containing detailed information regarding Mendelian disorders in over 12,000 human genes. The database focuses on genotype and phenotype correlations and the disease manifestations garnered through genetic and clinical evidence.

3.1.6 Proposed Approach

In order to investigate candidate genes potentially involved in LHON the MitoSVM and MitoCarta databases will be queried. These can then be ordered by the relevant scores pertaining to mitochondrial relatedness. In addition. a text mining workflow will be developed in Taverna 1.7 to mine UniProt and Gene Ontology records for disease-related terms and phrases for all genes on the X chromosome. The workflow will also cross-reference every candidate and mine literature associated with mouse, rat and chimpanzee orthologues within these regions. This may potentially reveal unique orthologue candidates expressing eye-related disease phenotypes that are not documented in the human homologue.

3.2 Methods

3.2.1 Application of MitoSVM and MitoCarta to LHON candidate gene analysis

MitoSVM

Following the systematic evaluation of mitochondrial protein prediction methods to define the best combination of independent datasets for mitochondrial gene prediction, a model file was produced using machine learning reflecting the highest mean sensitivity and mean specificity. The best combination resulting in the highest mean sensitivity with no significant loss in mean specificity was achieved using seven specific classifiers (see Chapter 2 section 2.4). Following the genome wide analysis performed for all proteins contained in the Ensembl human genome database, these results were stored into a local relational database. By specifying the chromosomal coordinates defined through the linkage analysis on the X chromosome an ordered list was extracted in descending order of SVM score. These results were exported into a spreadsheet for further investigation revealing predicted LHON candidates generated using the most sensitive combination of prediction methods.

MitoCarta

The Human MitoCarta database was downloaded and stored locally from the research department's webpage (http://www.broadinstitute.org/pubs/MitoCarta/) selecting the HumanMitoCartaAll.sql for the relational database schema and HumanMitoCartaAll.txt for the data. The relevant data was then exported into a spreadsheet using the chromosomal coordinates as before to highlight potential candidates involved in LHON and ordered by descending Maestro score. These candidates could then be cross referenced with the MitoSVM predicted genes to further the investigation to define a holistic candidate gene list.

3.2.2 Text mining of Gene Ontologies and OMIM

Gene ontologies have specific information regarding the cellular compartment, molecular function and biological processes a gene is involved in. These ontologies harness valuable information collated through laboratory investigation. Associations can be found within these ontologies pertaining to specific disease processes. A recent extension to the PubMed medical bibliographic database is GoPubMed. Specific gene ontologies were required for a text mining workflow to generate candidates involved in eye-related diseases and biological processes. A list of these ontologies was confirmed

using an extension to PubMed that performed a bibliographic analysis of the literature and categorises the results by gene ontologies, known as GoPubMed. The terms "LHON", "eye" and "blindness" were applied in GoPubMed to reveal the most common ontologies associated with these searches. GoPubMed produced the following most common gene ontologies from searching these key terms:

optic nerve development
optic nerve formation
optic nerve maturation
optic nerve morphogenesis
visual perception
retinal ganglion cell axon guidance

These specific terms could then be applied within the text mining workflow to highlight other genes and orthologues that contain these ontologies. Other keywords and terms were required to extend the vocabulary of the text mining analysis and were collated using common eye-related terms alongside advice from clinical experts. The following list of keywords and terms was determined:

vision
visual loss
optic atrophy
optic neuropathy
optic nerve
eye
blindness
ganglion
retina
retinal
retinal ganglion
retinopathy

Text mining could then be applied using the vocabulary determined above for all OMIM and UniProt records associated with genes within the candidate region of interest. A text mining program was then developed that aimed to scan through gene ontology and UniProt records searching for these disease-related terms. A workflow was developed requiring UniProt accession numbers as input that extracted all the relevant gene ontology records from the UniProt database and passed each one to the text mining program. The results were then analysed for hits >= 1. This method can also be

applied to OMIM records to highlight any previous evidence of disease relatedness.

3.2.3 Text mining workflow

A large text mining application was developed in the form of a multiple nested workflow that incorporated several Biomart queries and text mining Java web services (Figure 3.1). The primary focus of this pipeline was to perform text mining of specific keywords, terms and gene ontologies from OMIM and UniProt records. Mining these records would reveal potential candidates that expressed an eye-disease related phenotype or biological process. In addition, text mining was applied to UniProt records for any mouse (*Mus musculus*), rat (*Rattus norvegicus*) and chimpanzee (*Pan troglodytes*) orthologues generated from the human candidates. The workflow was developed in Taverna 1.7 and designed to accept chromosomal coordinates, keywords and terms, and specific gene ontologies to search for against every candidate within the specific chromosomal region. Orthology mining would potentially reveal genes expressing an eye-related phenotype or biological process in the mouse, rat or chimpanzee not yet discovered in the human orthologue.

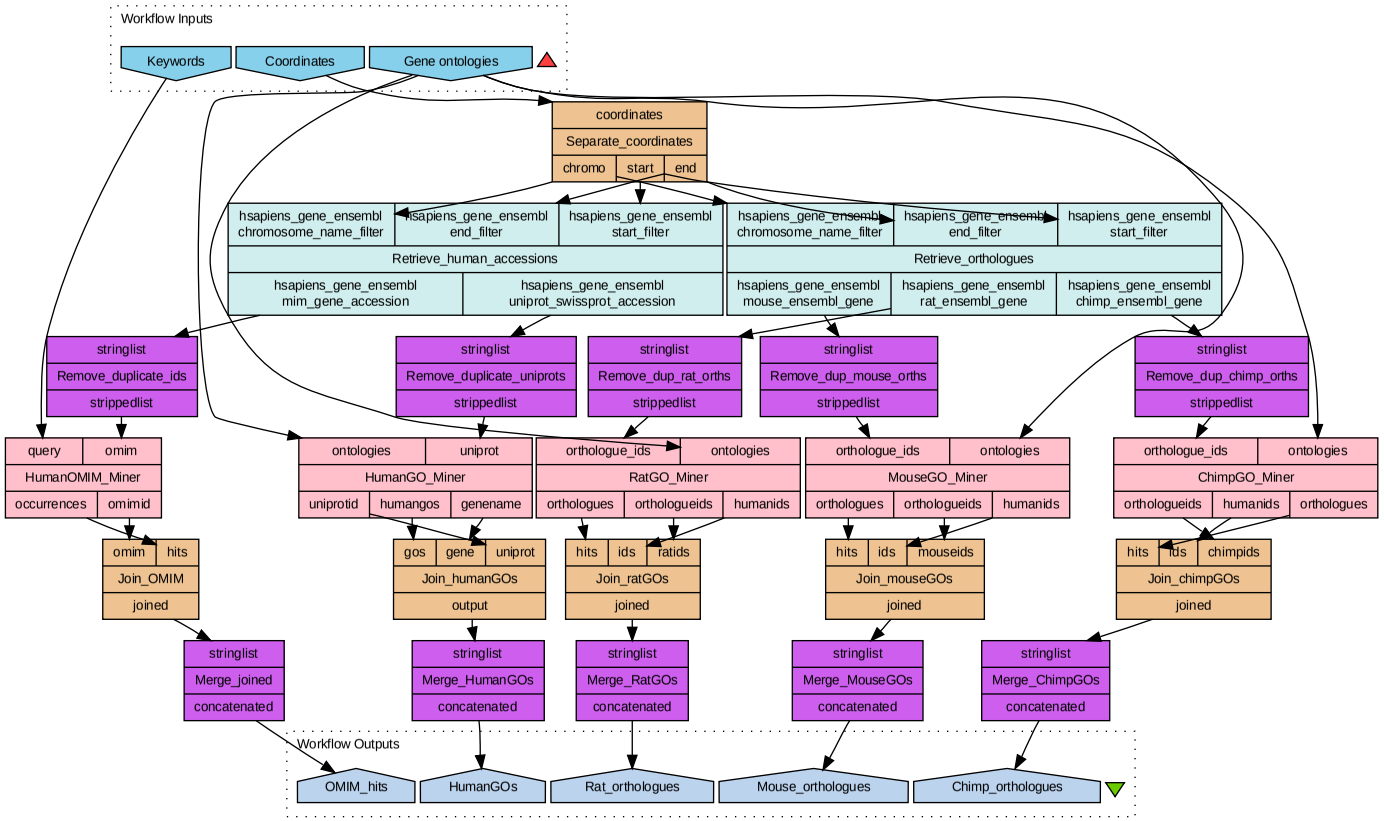

Figure 3.1: Text mining workflow that requires chromosomal coordinates, specific keywords and gene ontologies to allow text mining of each gene's OMIM and UniProt records. Orthologues for mouse, rat and chimpanzee are also queried in addition to highlight potential novel candidates.

Workflow user query

The workflow required 3 inputs that consisted of 1) chromosomal coordinates, 2) specific gene ontologies and keywords and 3) keywords alone for OMIM record mining. An example query can be seen below:

Chromosomal coordinates: X 48000000 52000000
Keywords:
blindness
optic atrophy
Gene ontologies:
optic nerve development
visual perception
blindness
optic atrophy

Initially all the search terms were sent to the four nested workflows OMIMMiner, HumanGO_Miner, MouseGO_Miner, RatGO_Miner and ChimpGO_Miner for use later in the workflow process. The chromosomal coordinates were then separated into separate inputs by the beanshell SeparateCoordinates in preparation for the following Biomart queries using the script below:

```
//Separation of chromosomal coordinates
StringBuffer result1 = new StringBuffer();
StringBuffer result2 = new StringBuffer();
StringBuffer result3 = new StringBuffer();

String[] elements = coordinates.split("\\s+");

String chromo = result1.append(elements[0] + "\n").toString().trim();
String start = result2.append(elements[1] + "\n").toString().trim();
String end = result3.append(elements[2] + "\n").toString().trim();
```

This processor consumes the chromosomal coordinates and splits them into separate entities by converting them into an array using whitespace as the delimiter. This produces 3 separate outputs that can be sent to the Biomart processors in the following procedure.

Process 1: Retrieve human accession numbers

A biomart was designed to consume the chromosomal coordinates and retrieve the corresponding OMIM ids and UniProt accession numbers from the Ensembl 56 Genes (Sanger UK) database specifying the *Homo sapiens* genes ensembl (GRCh37) subset. Filters were selected within the REGION section specifying chromosome and base positions. The following attributes were selected within the features section under the EXTERNAL heading:

Biomart EXTERNAL attributes configuration
MIM Gene Accession
UniProt/SwissProt Accession

This returned all the OMIM ids and UniProt accession numbers for all of the genes within the specified chromosomal region. These two lists were then sent to local processors (purple) to remove any duplicate entries generated from the Biomart query. Following the removal of any duplicate entities the list of unique OMIM ids were sent to OMIM_Miner and the unique UniProt accession numbers were sent to HumanGO_Miner for analysis further within the workflow explained in further detail in Processes 3 an 4 respectively.

Process 2: Retrieve orthologues

Text mining of UniProt records containing gene descriptions and gene ontologies was required for any existing orthologues relating to human genes found within the specified chromosomal region. Therefore a specific biomart was constructed to extract the relevant orthologue Ensembl gene ids. A biomart was configured using the same database implemented in Process 1 but configured to return orthologue information. The exact same filters were applied regarding chromosomal coordinates, but within the attributes the Homologues section was selected. Within this section under the CHIMP ORTHOLOGUES, MOUSE ORTHOLOGUES and RAT ORTHOLOGUES headings, the relevant Ensembl gene ids were selected:

Biomart CHIMP ORTHOLOGUES attributes configuration
Chimp Ensembl Gene ID

Biomart MOUSE ORTHOLOGUES attributes configuration
Mouse Ensembl Gene ID

Biomart RAT ORTHOLOGUES attributes configuration
Rat Ensembl Gene ID

All orthologous Ensembl gene ids for chimpanzee, rat and mouse were retrieved as two separate lists and sent to local processors (purple) as in Process 1 to remove any duplicate entities. The unique mouse Ensembl gene ids were then sent to MouseGO_Miner, rat Ensembl gene ids to RatGO_Miner and chimp gene ids sent to ChimpGO_Miner. This is further described in the Nested workflow 3 section.

Nested workflow 1: OMIM Miner

A nested workflow was created to retrieve OMIM records for text mining analysis. This consumed a list of OMIM ids and a list of the gene ontologies and keywords required for text mining (Figure 3.2). The keywords were split into separate queries by the local processor Split_queries and sent to the Java text mining web service mineRecord to be implemented further down the workflow. The OMIM ids were sent individually to another Java web service FetchOMIM that retrieved the relevant OMIM records from the OMIM database. Each OMIM record was written to a file by the local processor WriteOMIMFile and stored in preparation for the next procedure. The text mining program mineRecord was conditionally linked to the WriteOMIMFile processor as this required the existence of the OMIM record before execution.

Text mining web service: mineRecord

The mineRecord program then accessed the file and and analysed the OMIM records for all the gene ontologies and keywords returning the each term followed by the number of times the word was found. The Java program used the following regular expression to mine for keywords and phrases within a text document:

```
\\Specific keyword or phrase to mine against OMIM and GO records
final Pattern pattern = Pattern.compile
("\\b+" + query + "+\\b", Pattern.CASE_INSENSITIVE);
```

Contained within the line of code is the word query which was the program input pertaining to the specific keyword or phrase entering the program. This was flanked by boundary classifiers (`"\\b+"`) whereby the specific keyword or phrase would only produce a hit if it was present as an entire word and not just part of a longer string.

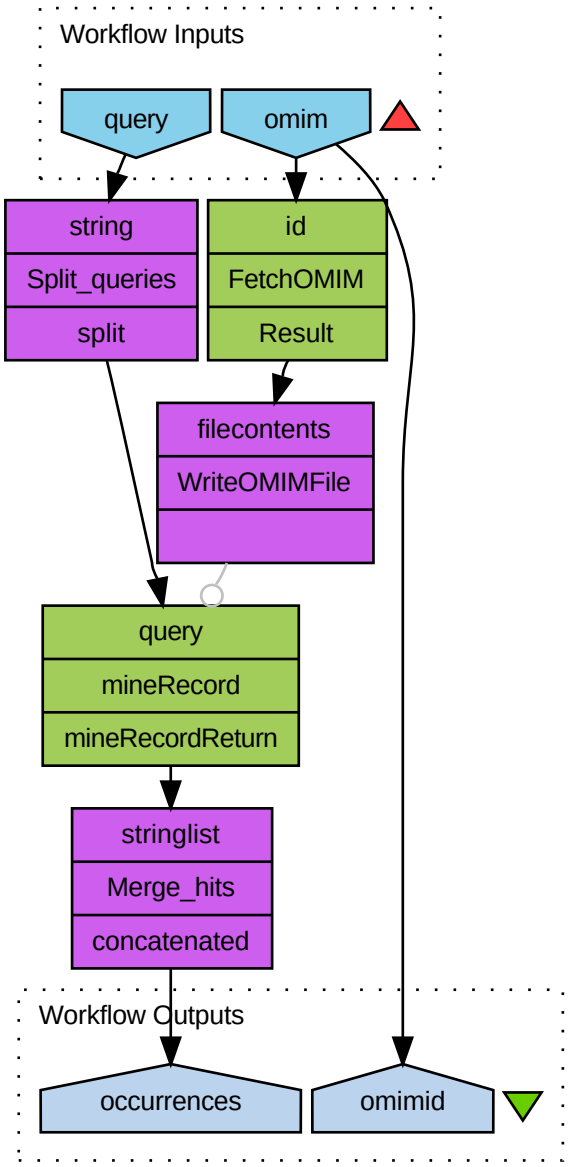

Figure 3.2: A nested workflow that consumes OMIM ids, gene ontologies and keywords. Each OMIM record is retrieved and analysed for hits relating to the lists of ontologies and keywords aiming to reveal potential candidates.

For example this meant the keyword vision would not produce a spurious result from the word visionary. Another important aspect was the final part of the expression that makes the match case insensitive. This allows any keyword or phrase to produce a hit if capitalised either in full or just the first letter due to the word being at the start of a sentence. An example of the output can be seen below:

Example output from the mineRecord program
eye: 8
optic atrophy: 3
optic neuropathy: 1

Following this the text mining results for the specific OMIM id being investigated are merged from a line separated list into a single line separated by a semi colon by the processor Merge_hits . This converts the above output into the following result:

Example output after merging
eye: 8 : optic atrophy: 3 : optic neuropathy: 1

Each OMIM id is also outputted alongside its relevant hit count result from the text mining program.

Nested workflow 2: Human Gene Ontology Miner

This nested workflow consumes the list of gene ontologies and keywords alongside a list of UniProt Accession numbers (Figure 3.3). The ontologies are split into separate entries by the local processor Split_ontologies and sent to the mineRecord program. In addition, the UniProt accession numbers are sent to a Java web service called getUNIPROTentry which returns the file from the UniProt database. The UniProt file is then written to the local machine by the processor WriteGOFile in preparation for the mineRecord program. The UniProt accession number is sent to a second service consisting of a human biomart query Retrieve_UniprotGene that consumes the accession number and returns the related Ensembl Gene ID and associated gene name. These outputs are then concatenated into single semi colon separated line by the beanshell Join_gene_name and merged by the Merge_genename processor. As before, the text mining service is implemented and the UniProt file is searched for all keyword and gene ontology occurrences generating a hit count for each term. This workflow generates 3 outputs comprised of 1) the UniProt accession number, 2) the hit counts from the text mining program and 3) the related gene information.

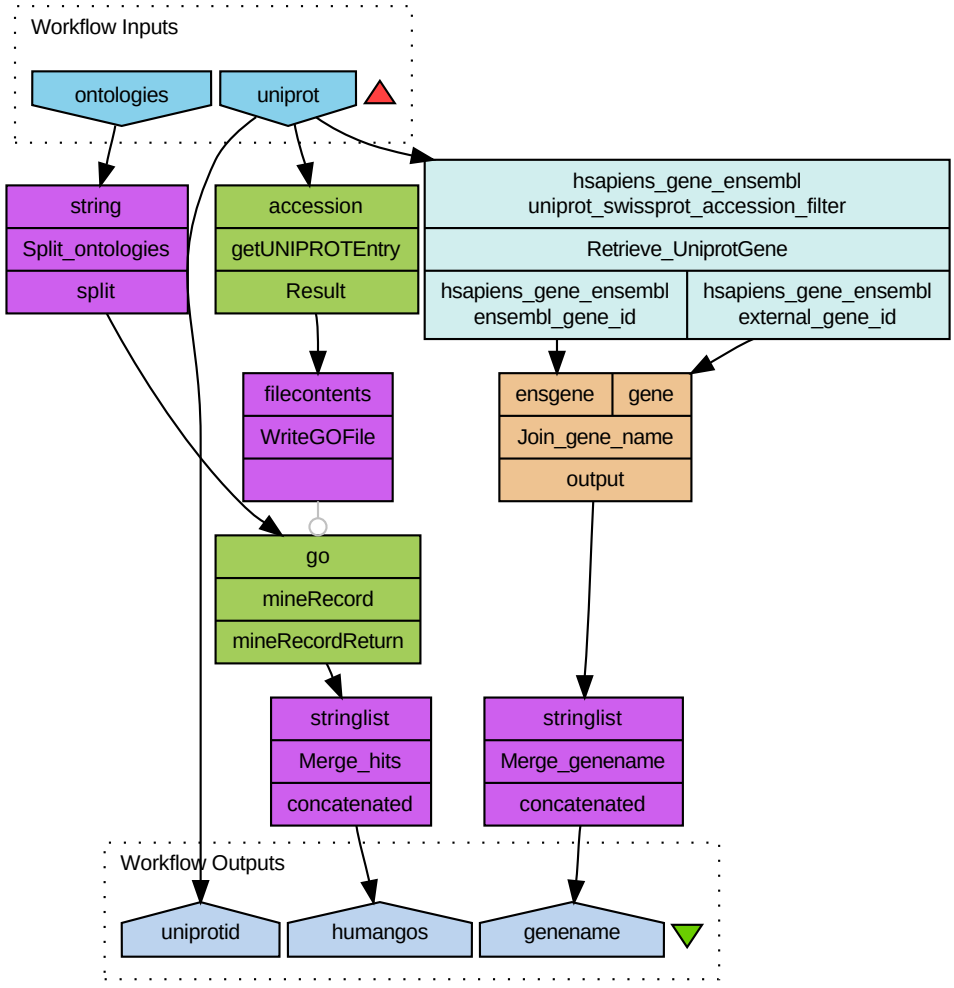

Figure 3.3: A nested workflow that consumes UniProt accession numbers, gene ontologies and keywords. Each UniProt record is retrieved and analysed for hits relating to the lists of ontologies and keywords aiming to reveal potential candidates.

Nested workflow 3: Mouse, Rat and Chimpanzee Gene Ontology Miner

Following removal of any duplicate Ensembl gene ids the nested workflows MouseGO_Miner, RatGO_Miner and ChimpGO_Miner were constructed to consume these orthologous gene ids for analysis alongside the keywords and ontologies (Figure 3.4). As before the gene ontology terms were separated into a list of queries by the processor Split_ontologies and sent to the mineRecord web service. The ensembl gene ids were queried with species-specific biomart queries using the Ensembl 56 Genes (Sanger UK) database specifying the species subset *Mus musculus* genes (NCBIM37), *Rattus norvegicus* genes (RGSC3.4) and *Pan troglodytes* genes (CHIMP2.1). This consumed the Ensembl gene id and returned the related UniProt accession number. Using the accession number the getUNIPROTentry web service was queried that returned the specific UniProt file relating to the species orthologue. This was written and stored by the processor WriteGOFile in preparation for the mineRecord service. Another species biomart query Retrieve_human_ids was configured for returning the gene name and human ensembl gene id to allow cross referencing when analysing the final candidate list. A beanshell was constructed (Join_human_ids) to concatenate the gene information. The final results were all merged from both the text mining application and biomart query by the merge processors. The workflow produced 3 outputs which were 1) the orthologous ensembl gene ids, 2) the text mined hit counts from the species-specific UniProt records, and 3) the related human gene information for cross-referencing.

Final process: Merging data for analysis

All the resulting data produced from each nested workflow is concatenated by specific beanshell processors such as Join_mouseGOs and Join_OMIM to combine all the relevant data into lines and merged into a single document to allow further analysis for disease candidates potentially involved in LHON. The results were exported into spreadsheets and categorised by scores relating to the number of keywords present in the relevant gene related UniProt and OMIM files.

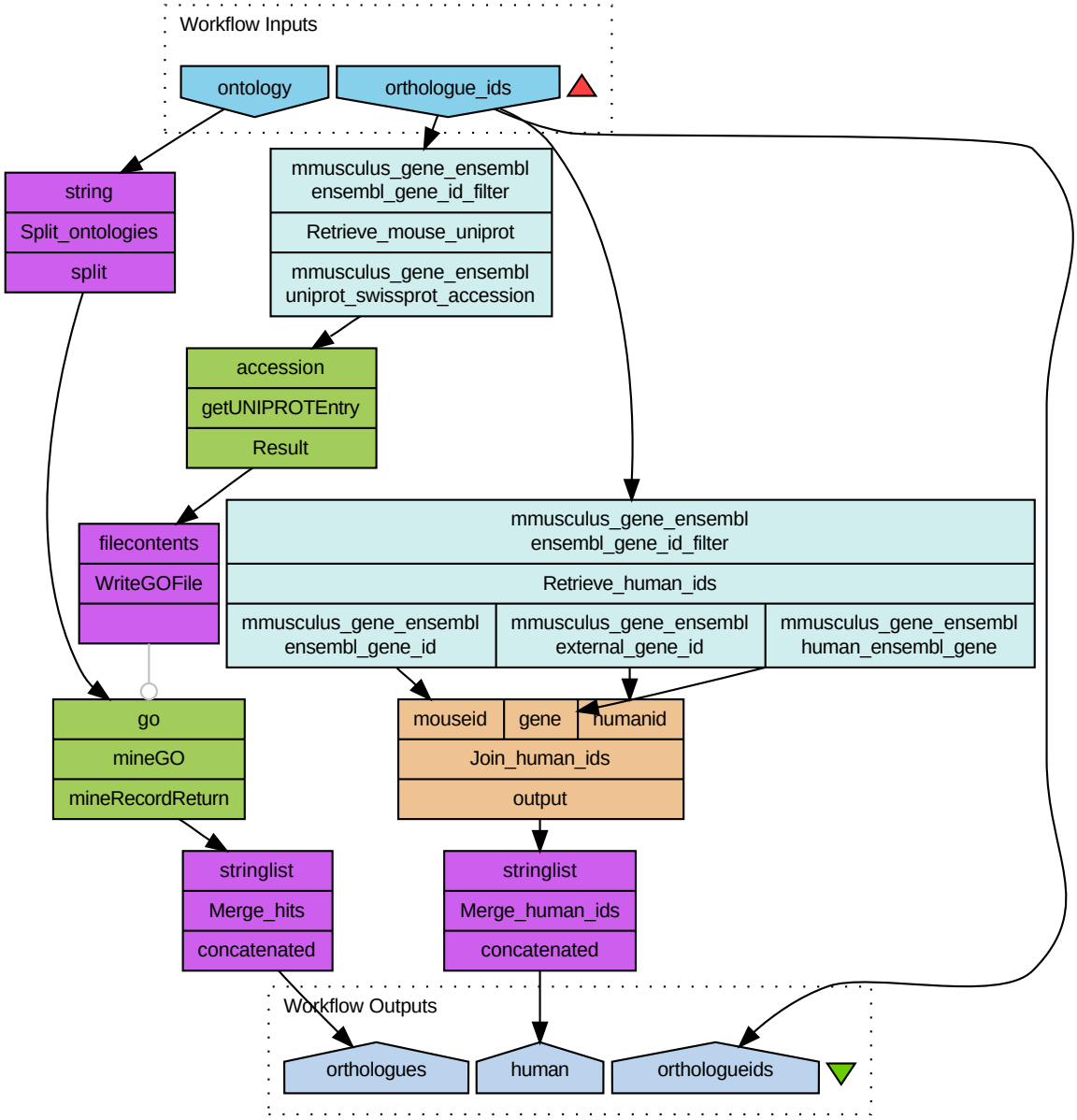

Figure 3.4: A nested workflow that consumes orthologous Ensembl gene ids, gene ontologies and keywords. Each UniProt record is retrieved and analysed for hits relating to the lists of ontologies and keywords aiming to reveal potential candidates.

3.3 Results

3.3.1 Application of MitoSVM and MitoCarta to LHON candidate gene analysis

In order to define a candidate list of mitochondrial related proteins for the chromosomal region implicated from the LHON linkage analysis, MitoSVM and MitoCarta were queried. The MitoSVM database containing results of a genome-wide analysis using a combination of classifiers achieving the highest sensitivity was queried. MitoSVM contained 13 candidates on the X chromosome with 7 being contained within the linkage region (Table 3.1). In addition, querying the locally stored MitoCarta database revealed 25 candidates on the X chromosome achieving a Maestro score >5. This revealed 11 candidates within the linkage peaks (Table 3.2). Each method revealed several candidates present above both thresholds for MitoSVM and MitoCarta with the exception of specific genes only being found using one method (Table 3.3).

3.3.2 Text mining of Gene Ontologies and OMIM

OMIM candidates

The entire human X chromosome was queried and whereby every gene's OMIM record was automatically text mined using a workflow. A hit count for each gene was generated based on the number of keyword or phrase matches. The resulting list of hits was analysed in a spreadsheet and given a category score. This reflected the number of separate keywords found as opposed to how many times each word occurred. If a word occurred just once this would reveal a positive association regardless of how many times it appeared in the document. The number of keywords and phrases mined for in OMIM_Miner was 12 and this represented the maximum category score whereby a hit is achieved for every keyword or phrase. A total of 78 genes had at least one association. Table 3.4 shows the top scoring candidates from the OMIM analysis achieving a category score of 3 or above.

Ensembl GeneId	Ensembl ProteinId	Gene Name	Chromo	Start Bp	End Bp	SVM Score
ENSG00000156709	ENSP00000316320	AIFM1	X	129091018	129127489	1.941537500000
ENSG00000036473	ENSP00000039007	OTC	X	38096680	38165650	1.417221700000
ENSG00000102078	ENSP00000354455	SLC25A14	X	129301699	129335014	1.307115200000
ENSG00000072506	ENSP00000168216	HSD17B10	X	53474931	53478048	1.235455300000
ENSG00000158578	ENSP00000337131	ALAS2	X	55052213	55074136	1.186321000000
ENSG00000077713	ENSP00000338628	SLC25A43	X	118397679	118472459	0.936613280000
ENSG00000182890	ENSP00000327589	GLUD2	X	120009143	120011475	0.906491310000
ENSG00000123130	ENSP00000336580	ACOT9	X	23631698	23694513	0.815902850000
ENSG00000004961	ENSP00000326579	HCCS	X	11039342	11051122	0.710626070000
ENSG00000131269	ENSP00000253577	ABCB7	X	74189834	74292857	0.638949810000
ENSG00000165349	ENSP00000298085	SLC7A3	X	70062163	70067700	0.187230920000
ENSG00000101986	ENSP00000218104	ABCD1	X	152643517	152663410	0.145616960000
ENSG00000178605	ENSP00000316598	GTPBP6	X	160025	170886	0.022343984000

Table 3.1: All mitochondrial candidates found on the X chromosome scoring 0 or above for the SVM score contained within the genome-wide MitoSVM database.

Entrez id	Gene Name	Chromo	Start Bp	End Bp	Maestro score
5160	PDHA1	X	19271971	19287885	41
3421	IDH3G	X	152704414	152713160	30
1349	COX7B	X	77041616	77047536	25
9131	AIFM1	X	129091017	129127488	22
54539	NDUFB11	X	46886562	46889380	21
1678	TIMM8A	X	100487305	100490342	19
79979	CXorf34	X	100151186	100193725	18
3028	HSD17B10	X	53474930	53478047	17
292	SLC25A5	X	118486436	118489306	15
5009	OTC	X	38096301	38165552	15
5165	PDK3	X	24393474	24462462	15
3052	HCCS	X	11039372	11051121	14
4694	NDUFA1	X	118889761	118894645	14
2710	GK	X	30581460	30658645	13
293	SLC25A6	X—Y	1465044	1470993	12
139322	FAM121A	X	84145560	84229429	12
10245	TIMM17B	X	48635675	48640369	10
11238	CA5B	X	15666345	15712578	9
212	ALAS2	X	55052212	55074135	9
9016	SLC25A14	X	129301727	129335015	8
644310	LOC644310	X	51683387	51683578	8
215	ABCD1	X	152643529	152663374	7
23597	ACOT9	X	23631697	23671327	7
56474	CTPS2	X	16516042	16640979	7
203427	SLC25A43	X	118417050	118472461	6

Table 3.2: All mitochondrial candidates found on the X chromosome scoring >5 for the Maestro score contained within the MitoCarta database.

MitoSVM	MitoCarta
GLUD2	SLC25A6
ABCB7	CA5B
SLC7A3	CTPS2
GTPBP6	PDHA1
	PDK3
	GK
	NDUFB11
	TIMM17B
	LOC644310
	COX7B
	FAM121A
	CXorf34
	TIMM8A
	SLC25A5
	NDUFA1
	IDH3G

Table 3.3: MitoSVM and MitoCarta found specific genes that were unique to these applications.

OMIM Id	Ensembl GeneId	Gene Name	vision	visual loss	optic atrophy	optic neuropathy	optic nerve	eye	blindness	ganglion	retina	retinal	retinal ganglion	retinopathy	Category Score
300658	ENSG00000124479	NDP	1	0	0	0	0	1	1	1	1	1	1	1	8
312700	ENSG00000102104	RS1	1	1	0	0	0	1	1	1	1	1	0	0	7
300078	ENSG00000125356	NDUFA1	0	0	0	1	1	0	0	1	1	1	1	0	6
312610	ENSG00000156313	RPGR	1	0	0	0	0	1	1	0	1	1	0	1	6
300110	ENSG00000102001	CACNA1F	0	0	0	0	0	1	1	1	1	1	0	0	5
303800	ENSG00000147380	OPN1MW	1	0	0	0	0	1	1	0	1	1	0	0	5
300500	ENSG00000101850	GPR143	1	0	0	0	0	1	1	0	0	1	0	0	4
300035	ENSG00000090776	EFNB1	0	0	0	0	0	1	0	0	1	1	0	0	3
300041	ENSG00000101890	GUCY2F	0	0	0	0	0	1	0	0	1	1	0	0	3
300144	ENSG00000182890	GLUD2	0	0	0	0	0	0	1	0	1	1	0	0	3
300278	ENSG00000188937	NYX	0	0	0	0	0	0	1	1	1	0	0	0	3
300356	ENSG00000126953	TIMM8A	1	1	0	0	0	0	1	0	0	0	0	0	3
300377	ENSG00000198947	DMD	0	0	0	0	0	0	0	1	1	1	0	0	3
303900	ENSG00000102076	OPN1LW	1	0	0	0	0	0	1	0	0	1	0	0	3
309550	ENSG00000102081	FMR1	0	0	0	0	0	1	0	0	1	1	0	0	3
311850	ENSG00000147224	PRPS1	1	0	1	1	0	0	0	0	0	0	0	0	3

Table 3.4: Top scoring genes on the human X chromosome achieving a category score of 3 or above following text mining of OMIM records.

Human gene ontology candidates

In addition to the OMIM records all gene ontologies were queried through the text mining of UniProt files automatically downloaded from the UniProt database. The same procedure was applied as before for the analysis with category scores being applied. The 12 keywords were sent to the Human_GOMiner in addition to 6 specific gene ontology terms reflecting a maximum category score of 18. A total of 73 candidates revealed at least one association. 14 candidates scored 3 or above and are shown in Table 3.5.

Mouse, Rat and Chimpanzee orthologues

Specific sections of the text mining workflow aimed to reveal potential candidates within closely related species that have not been determined in humans. The mouse, rat and chimpanzee genomes were selected for analysis applying species specific biomart queries to return the orthologues Ensembl gene ids and UniProt accession numbers. These results were exported into spreadsheets for further analysis and ordered by category scores. As in the human gene ontology analysis a maximum category score was 18. The chimpanzee analysis returned no candidates relating to any of the gene ontologies or keywords. However, the mouse orthologue analysis returned 111 candidates with at least one association and 15 candidates scored 2 or above (Table 3.6). The rat analysis returned 4 candidates with 2 having one keyword hit and 2 candidates scoring 4 (Table 3.7.

UniProt Accession	Gene Name	Ensembl GeneId	visual perception	vision	eye	blindness	ganglion	retina	retinal	retinopathy	Category Score
Q00604	NDP	ENSG00000124479	1	1	0	1	1	1	1	1	7
Q9GZU5	NYX	ENSG00000188937	1	1	0	1	1	1	1	0	6
O75695	RP2	ENSG00000102218	1	1	1	1	0	1	1	0	6
P24386	CHM	ENSG00000188419	1	1	0	1	0	1	1	0	5
Q92834	RPGR	ENSG00000156313	1	1	0	1	0	1	1	0	5
O60840	CACNA1F	ENSG00000102001	0	1	1	1	0	1	1	0	5
P36575	ARR3	ENSG00000120500	1	1	1	0	0	1	1	0	5
P51841	GUCY2F	ENSG00000101890	1	1	0	0	0	1	1	0	4
O15537	RS1	ENSG00000102104	1	1	0	0	0	1	1	0	4
P51810	GPR143	ENSG00000101850	1	0	1	0	0	1	1	0	4
P04000	OPN1LW	ENSG00000102076	0	1	0	0	0	1	1	0	3
Q6T4R5	NHS	ENSG00000188158	0	0	1	0	0	1	1	0	3
Q13796	SHROOM2	ENSG00000146950	0	0	1	0	0	1	1	0	3
P04001	OPN1MW	ENSG00000147380	0	1	0	0	0	1	1	0	3

Table 3.5: Top scoring human genes achieving a category score of 3 or above on the X chromosome following text mining of UniProt records for specific keywords and gene ontologies

MouseID	HumanID	Gene	visual perception	vision	eye	blindness	ganglion	retina	retinal	retinal ganglion	retinopathy	Category Score
ENSMUSG00000051228	ENSG00000188937	Nyx	1	1	1	0	1	1	1	0	0	6
ENSMUSG00000031142	ENSG00000102001	Cacna1f	1	1	0	1	0	1	0	0	0	4
ENSMUSG00000042282	ENSG00000101890	Gucy2f	1	1	0	0	0	1	1	0	0	4
ENSMUSG00000060890	ENSG00000120500	Arr3	1	1	0	0	0	1	1	0	0	4
ENSMUSG00000031402	ENSG00000130830	Mpp1	0	0	1	0	1	1	0	0	0	3
ENSMUSG00000025592	ENSG00000126733	Dach2	0	0	1	0	0	1	1	0	0	3
ENSMUSG00000031217	ENSG00000090776	Efnb1	0	0	0	0	1	0	1	1	0	3
ENSMUSG00000031302	ENSG00000196338	Nlgn3	0	0	0	0	1	1	1	0	0	3
ENSMUSG00000031293	ENSG00000102104	Rs1	0	0	1	0	0	1	1	0	0	3
ENSMUSG00000031394	ENSG00000166160	Opn1mw	0	1	0	0	0	1	1	0	0	3
ENSMUSG00000002012	ENSG00000130822	Pnck	0	0	0	0	1	1	0	0	0	2
ENSMUSG00000040138	ENSG00000124479	Ndp	0	0	0	0	1	1	0	0	0	2
ENSMUSG00000045103	ENSG00000198947	Dmd	0	0	0	0	0	1	1	0	0	2
ENSMUSG00000045180	ENSG00000146950	Shroom2	0	0	1	0	0	0	1	0	0	2
ENSMUSG00000038482	ENSG00000198176	Tfdp1	0	0	0	0	1	1	0	0	0	2

Table 3.6: Mouse orthologues achieving a category score of 2 or above on the X chromosome following text mining of UniProt records for specific keywords and gene ontologies for mouse.

RatID	HumanID	Gene	visual perception	vision	eye	blindness	ganglion	retina	retinal	retinal ganglion	retinopathy	Category Score
ENSRNOG00000019086	ENSG00000101890	Guc2f	1	1	1	0	0	0	1	0	0	4
ENSRNOG00000002904	ENSG00000120500	Arr3	1	1	0	0	0	1	1	0	0	4
ENSRNOG00000003812	ENSG00000196338	NLGN3_RAT	0	0	0	0	1	0	0	0	0	1
ENSRNOG00000024322	ENSG00000146950	SHRM2_RAT	0	0	0	0	0	0	1	0	0	1

Table 3.7: Rat orthologues achieving a category scores. Only 4 candidates achieved positive scores on the X chromosome following text mining of UniProt records for specific keywords and gene ontologies for rat.

3.4 Discussion

3.4.1 Biological discussion

Several studies have determined an X-chromosomal haplotype associated with LHON but as yet no nuclear modifier has been determined. Bioinformatics analysis has been applied to refine candidate gene lists and prioritise these based on their mitochondrial involvement. Integrated techniques consisting of various mitochondrial prediction methods including MitoSVM and MitoCarta have been utilised in this search and several genes have yet to be discarded. However, no causative nuclear gene has been identified thus far. Various reasons may exist that account for the lack of evidence for an X-chromosomal LHON candidate disease gene. In addition, text mining OMIM and UniProt records for specific keywords and phrases expressing LHON related phenotypes broadens the search for LHON disease genes. This same procedure was applied to UniProt records for any orthologues in mouse, rat and chimpanzee in order to reveal any novel candidates associated with LHON related phenotypes not documented in human.

Mitochondrial gene candidates

Querying the MitoSVM database of mitochondrial predicted genes, a total of 13 scored above the threshold. Nine of these candidates were also found in the MitoCarta database scoring >5 using the Maestro score. Therefore, 4 genes were unique to the MitoSVM database. MitoCarta generated a total of 25 genes above the Maestro threshold whereby 16 of these were unique to MitoCarta. The strongest candidates are the 9 genes occuring in both databases.

Text mined candidates

Text mining for medical records and gene ontology information comprises a unique method for determining genes potentially associated with LHON. This method is not primarily focused on mitochondrial-related candidates but potential indirect associations sharing phenotypes for eye-related diseases. Using highly relevant and specific keywords and ontologies as a vocabulary for mining through gene-centred and disease-centred literature provides a powerful mechanism for discovering phenotypically related genes manifesting retinal neuropathies. Automated text mining procedures can rapidly assess vast amounts of distributed literature highly relevant to the investigation.

Following the text mining of OMIM records, 16 genes achieved a category score of >=3. GLUD2 was associated with 3 categories consisting of blindness, retina and retinal and present in the MitoSVM database achieving an SVM score of 0.9. TIMM8A also had a category score of 3 consisting of vision, visual loss and blindness achieving a Maestro score of 19. These are very strong LHON disease candidates as they are predicted to be mitochondrial and associated with eye-related dysfunction. All candidates mined against UniProt were not present in either MitoSVM or MitoCarta.

Unique orthologues

Following the mining of orthologues within the mouse, rat and chimpanzee several genes achieved high category scores whereby most had also been determined in the human through gene ontologies or OMIM records. Interestingly, certain genes expressing eye-related phenotypes proved to have no association currently identified in human. The following mouse orthologues had category scores of 2 or more but no associations in the corresponding human homologue: Nlgn2, Pnck and Tfdp1. These genes stand out as novel candidates for investigating LHON as no correlation has been identified in humans.

The mouse gene Nlgn2 produces the protein neuroligin-3, a neuronal cell surface protein that binds to beta-neuraxins to form intercellular junctions. The human orthologue has been associated with causing X-linked autism (AUTSX1) and X-linked Asperger syndrome (ASPGX1) resulting from a defect of synaptogenesis. However, no eye-related phenotype is reported in the human orthologue. The mouse orthologue is expressed in the retinal astrocytes during the developmental stage and is associated with visual learning (Gilbert *et al.*, 2001). Pnck produces Calcium/calmodulin-dependent protein kinase type 1B which plays a role in a calcium-triggered signalling cascade. During development in the mouse, Pnck is expressed in the brain, spinal cord and retina (Ueda *et al.*, 1999). No evidence of expression in the retina in the human orthologue has been reported. Tfdp1 codes for Transcription factor Dp-1 which stimulates E2F-dependent transcription. The E2F-1/DP complex is believed to mediate cell proliferation and apoptosis. In the mouse orthologue Tfdp1 is expressed in the developing retina, specifically in the retinoblast and ganglion cell layers (Dagnino *et al.*, 1997). Again, no such association has been found in the human orthologue. These three mouse orthologues should be classed as high priority for sequencing as they are unique candidates that have shown a phenotypic relation to an eye disorder or function in the mouse model. No unique orthologues were found in the chimpanzee or rat.

3.4.2 Technical discussion

Workflows can easily be modified to incorporate more species of interest and incorporate a larger vocabulary for text mining. These attributes can be tailored to any disease of interest and shared via workflow sharing schemas such as myExperiment. This provides a powerful mechanism for intelligently mining large amounts of biological literature that would be extremely labour intensive if performed manually. The sharing and reuse of workflows allows for the possibility to enhance and increase the quality of these methods. These techniques have rarely been applied to the candidate gene analysis of LHON and this technology is important to further our understanding of this elusive disease.

Chapter 4

Identification of Nuclear mitochondrial DNA sequences

Abstract

Background

Fragments of mitochondrial DNA have been transferred to the nucleus over evolutionary time. The majority of the sequences (NUMTs) reside in the nuclear genome as pseudogenes. However, previous research has revealed how NUMTs are transposed into the nuclear genome but have failed to highlight the mechanisms governing mtDNA loss from the mitochondria. This study aims to reveal information regarding this by performing flanking sequence analysis of the surrounding regions of these insertions and comparing these fragments to a database of 263 known mtDNA deletions.

Results

Mitochondrial DNA deletions do not appear to integrate into the nuclear genome. NUMTs appear to originate from all areas of the mitochondrial genome displaying no increase area of active loss. In contrast, mtDNA deletions are predominantly found originating in the major arc of the mitochondrial genome. There is little evidence to support the transposition of NUMTs throughout the genome following analysis of the flanking regions of the integration sites. NUMTs appear to reflect independent insertions over evolutionary time. Additionally, these fragments integate into areas of low GC content being more abundant in the intergenic, non-coding areas of the nuclear genome. Flanking regions also revealed a lack of nuclear-mitochondrial genes providing evidence to suggest NUMT integration sites are not enriched areas of mitochondrial activity.

Conclusion

NUMT integration appears to be randomly distributed throughout the genome and predominantly in areas that do not affect gene function. The mechanism for gene loss from the mitochondrial genome is still misunderstood but does not appear to be related to mtDNA deletions. Developments in sequencing technology may reveal more specific information especially on an individual level as to the mechanisms of NUMT formation and integration.

4.1 Introduction

4.1.1 Nuclear mitochondrial DNA insertions

Fragments of mitochondrial DNA (mtDNA) have been frequently transferred to the nucleus over evolutionary time resulting in nuclear-mitochondrial DNA sequences (NUMTs) (Richly and Leister, 2004). NUMTs reside in the nuclear genome as pseudogenes because despite their significant sequence homology they are not transcribed or translated into functional proteins (Woischnik and Moraes, 2002). NUMTs have been detected in excess of 100 eukaryotic genomes to date (Lopez *et al.*, 1994; Ricchetti *et al.*, 1999; Bensasson *et al.*, 2001; Pereira and Baker, 2004). The integration of mtDNA into the nuclear genome is therefore a common phenomenon amongst many species reflecting high abundance in some and complete absence in others. For example, the honeybee (*Apis mellifera*) has evidence for in excess of 1500 NUMTs which is believed to be the highest in any animal studied (Pamilo *et al.*, 2007). In the yellow fever mosquito (*Aedes aegypti*), 233 NUMTs have been detected that consist of >110Kb representing a density of 0.080bp NUMTs/Kb. This number is second to the honeybee consisting of >1.0bp NUMTs/Kb (Black Iv and Bernhardt, 2009). However, various species that have been investigated for NUMTs have displayed no evidence of mtDNA integration. These examples include the African malarial mosquito (*Anopheles gambiae*) and the pufferfish (*Takifugu rubripes*) (Richly and Leister, 2004; Venkatesh *et al.*, 2006). Hazkani-Covo *et al.* (2010) provide a large comprehensive list of determined NUMTs across 85 different species. Previous research suggested the *T. rubripes* nuclear genome contained mitochondrial pseudogenes but closer inspection through a follow-up investigation revealed these to be shotgun sequences from mitochondria that had been misassembled with the nuclear sequences (Antunes and Ramos, 2005; Venkatesh *et al.*, 2006). This highlighted the importance of genome coverage as the follow-up research used version 4.0 of the pufferfish genome assembly for their investigation. In whole genome shotgun sequencing, the efficiency and contiguity of the generated assemblies depend greatly on fold coverage of the specific genome. If the contigs are longer this results in a higher representation of the genome (Weber and Myers, 1997). Richly and Leister (2004); Pereira and Baker (2004) report no obvious correlation between the abundance of NUMTs and the size of nuclear genomes or mitochondrial genomes, or gene density within the nuclear genome. However, more recent research has found evidence of a strong correlation between NUMT content and genome size suggesting that larger genomes experience higher frequencies of double-strand breaks (DSBs) (Hazkani-Covo *et al.*, 2010). This correlation is displayed in Figure 4.1.

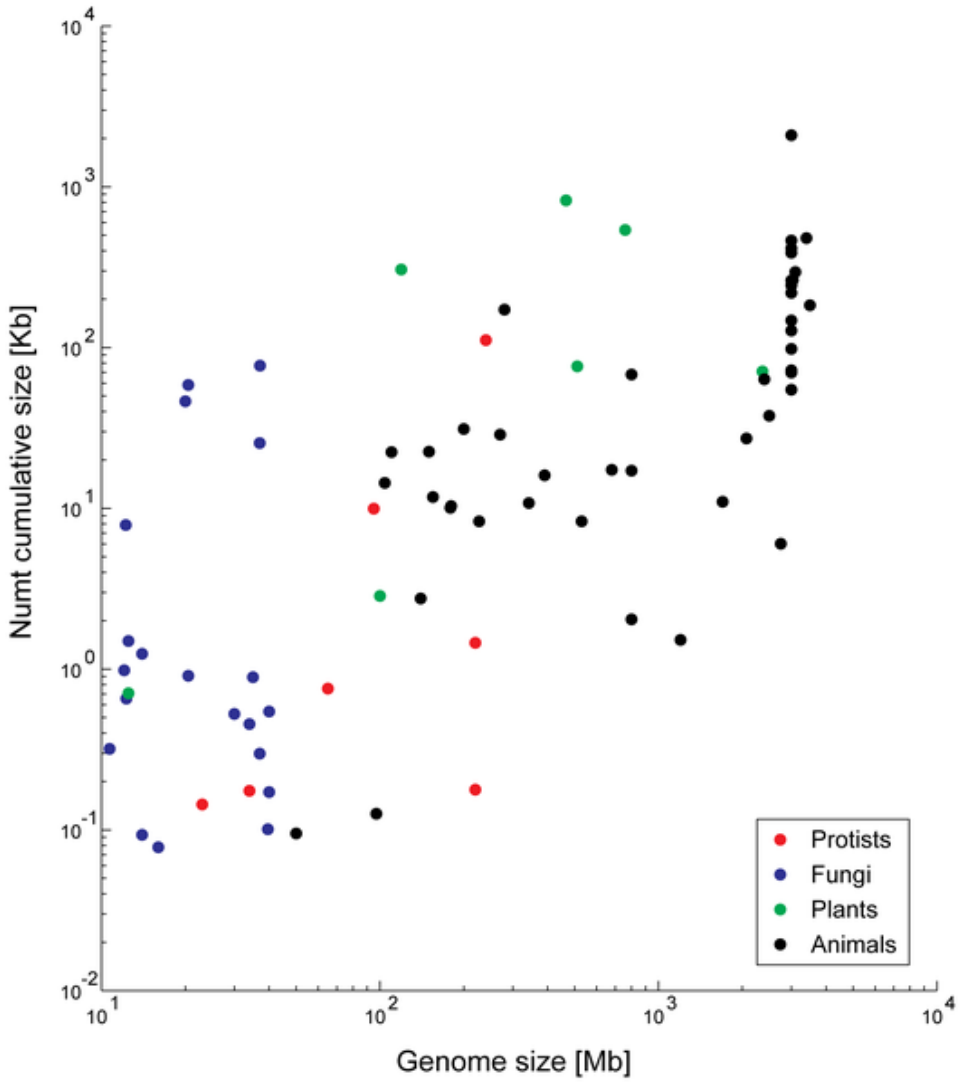

Figure 4.1: NUMT content correlating to genome size taken from Hazkani-Covo *et al.* (2010). A log-log scale graph displaying the dependency between NUMT content in genomes and genome size.

4.1.2 Mechanisms of mtDNA integration

Various mechanisms for the formation of mtDNA sequences and their subsequent insertion into the nuclear genome have been investigated (Ricchetti *et al.*, 2004; Lieber *et al.*, 2004; Blanchard and Schmidt, 1996; Honma *et al.*, 2007). Previous research suggested the transfer of genes from organelle to nucleus involved reverse transcription of an edited RNA intermediate (Nugent and Palmer, 1991). More recent evidence has shown the migrating factor to be predominantly DNA-mediated as analysis of human NUMTs have produced no evidence of splicing or polyadenylation of organellar nucleic acids prior to insertion (Woischnik and Moraes, 2002). DNA is believed to escape from mitochondria due to membrane disruption during autophagy, mitochondrial fusion or fission, and cell stress making the mtDNA available for nuclear import (Thorsness and Weber, 1996; Campbell and Thorsness, 1998). Patterns of terminal microidentities have been found in flanking regions adjacent to NUMT integration sites alluding to a mechanism of non-homologous end-joining (NHEJ) repair of DSBs believed to be a common mechanism in all eukaryotes (Ricchetti *et al.*, 2004). Figure 4.2 illustrates a model of nuclear insertion of organelle DNA.

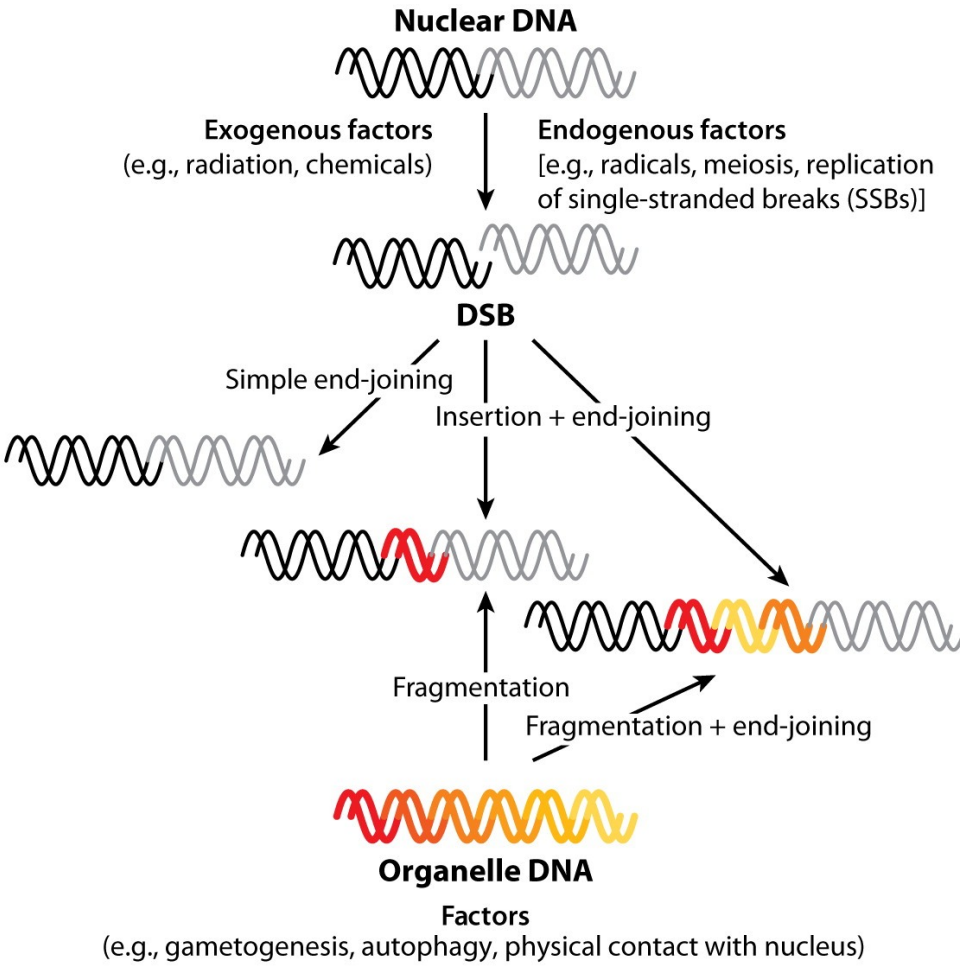

Figure 4.2: Generation of nuclear insertions from organelle DNA taken from Kleine *et al.* (2009). Double-stranded breaks (DSBs) are induced by exogenous and endogenous sources as listed. This model would imply that these mechanisms are stress-related and an increase in DSBs would result in an elevated rate of nuclear uptake of foreign DNA.

4.1.3 Mitochondrial DNA deletion formation

The most common causal variants of mitochondrial disease are mtDNA deletions and point mutations (Shoffner *et al.*, 1989; Wallace *et al.*, 1988). Mitochondrial DNA deletions are molecules that have lost a large section of the mitochondrial genome and manifest in several scenarios. 1) Single mtDNA deletions are detected in all cells within an affected tissue, 2) In a large group of affected individuals there are multiple mtDNA deletions in affected tissues namely muscle and the central nervous system involving defective nuclear genes, 3) Several reports have found mtDNA deletions accumulate with age in postmitotic tissues and neurodegenerative disorders (Bua *et al.*, 2006). The ratio of wild-type to mutated mtDNA govern the onset of disease requiring >60% of cells containing a deletion before causing a biochemical defect (Sciacco *et al.*, 1994). The majority of mtDNA deletions are located within the major arc and are flanked by two direct homologous repeat sequences (Bua *et al.*, 2006; Samuels *et al.*, 2004). Replication is thought to be the most common cause of mtDNA deletion formation but the exact nature is still largely misunderstood. Krishnan *et al.* (2008) propose that mtDNA deletions are initiated by single-stranded regions of mtDNA generated through exonuclease activity at DSBs (Figure 4.3).

4.1.4 Pseudomitochondrial genome and human genetic disease

The accidental amplification of NUMTs can pose serious problems when investigating mitochondrial diseases. Mitochondrial genome disease-associated biomarkers must be rigorously authenticated to eradicate any contamination with paralogous nuclear pseudogenes (Parr *et al.*, 2006; Yao *et al.*, 2008). An example of this involved a 5842bp NUMT on chromosome 1 originally classified as a novel mitochondrial mutation associated with low sperm motility and cystic fibrosis (Thangaraj *et al.*, 2003; Yao *et al.*, 2008). This was recorded in the HapMap database as a mitochondrial variation rather than a nuclear DNA variation (Biswas *et al.*, 2007). NUMTs have caused problems in evolutionary analysis resulting in gross misidentifications. Zischler *et al.* (1995) reported the discovery of 80 million year old DNA from dinosaur bones that on further inspection was a mitochondrial pseudogene from contaminant human nuclear DNA.

NUMTs have been implicated in several diseases in very rare cases involving the integration of mtDNA into genes. A disruption in the gene responsible for the production of plasma factor VII was caused by a 251bp NUMT insertion leading to severe factor VII deficiency (Borensztajn *et al.*, 2002). A *de novo* mutation was also responsible for a rare sporadic form of Pallister-Hall syndrome caused by a 72bp NUMT insertion into exon 14 of the GL13 gene (Turner *et al.*, 2003). This created a premature stop codon

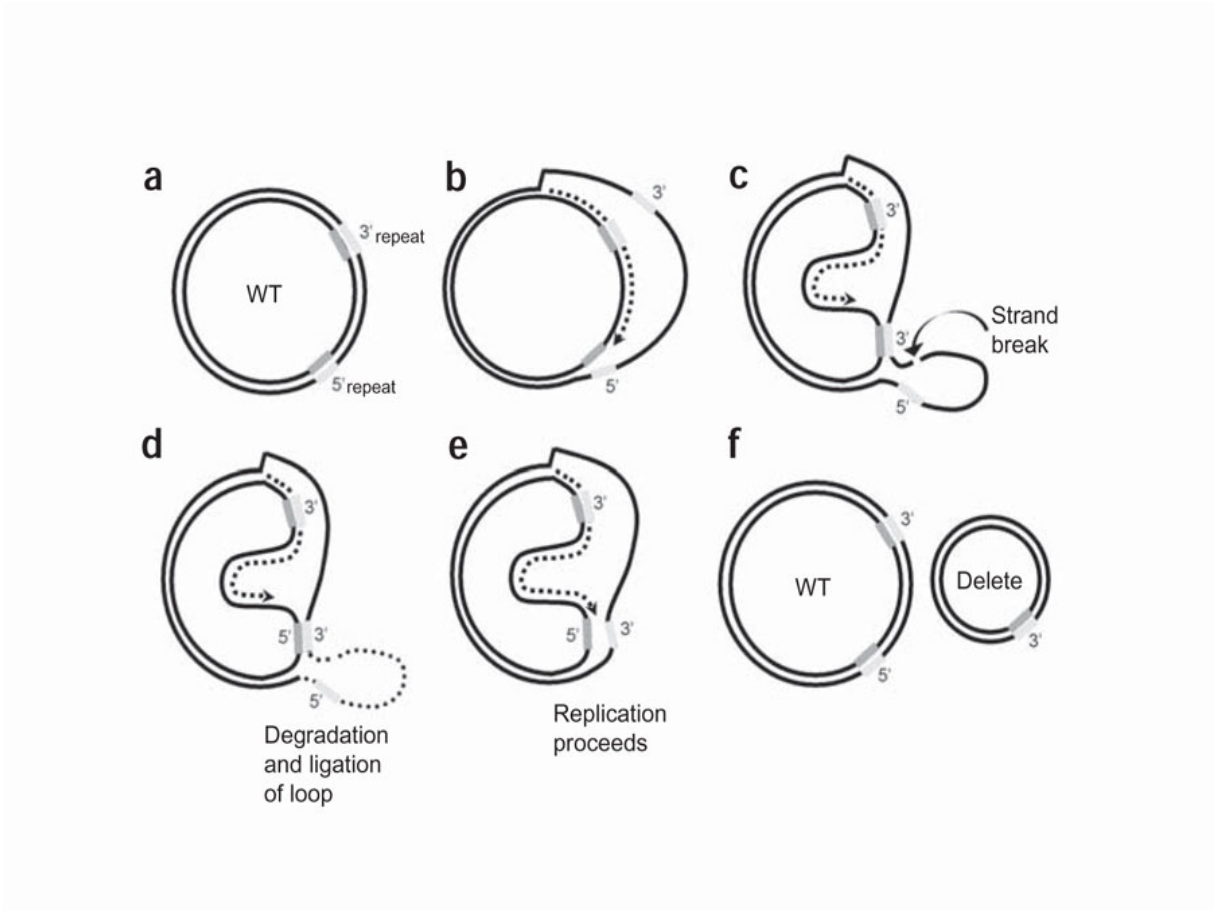

Figure 4.3: Proposed formation of a mtDNA deletion through a slipped strand model of replication taken from Krishnan *et al.* (2008). (a) mtDNA molecule and the presence of two direct repeats labeled 5' and 3'. (b) mtDNA replication begins in the D loop from OH, displacing the light strand from the heavy strand. (c) The single-stranded 3' repeat of the light strand misanneals with the newly exposed single-stranded 5' heavy-strand repeat, a downstream loop of the light strand is generated. This loop is prone to strand breaks. (d) The damaged loop is degraded until reaching the double-strand regions. Ligation of the free ends of the heavy strand occurs. (e) Replication is resumed. (f) A wild type and a deleted mtDNA molecule are produced.

resulting in a truncated protein product. Turner *et al.* (2003) state that the incident occurred in the Chernobyl region with high radioactive contamination but attribute this occurrence to coincidence. Other examples include a 93bp fragment inserted into the gene MCOLN1 eliminating correct gene splicing leading to a case of mucolipidosis IV (Goldin *et al.*, 2004) and a 36bp insertion in exon 9 of the USH1C gene associated with Usher Syndrome IC via trans-replication slippage (Chen *et al.*, 2005).

4.1.5 Sequence analysis of flanking regions

In addition to determining the location of NUMTs within the nuclear genome previous research has investigated the immediate flanking regions surrounding the NUMTs for evidence of transposable elements. Mishmar *et al.* (2004) analysed the flanking regions of 247 human NUMTs and found 59% were within 150bp of repetitive elements and the association was highly non-random ($p>0.0001$). The flanking sequences (500bp) of each NUMT were screened with RepeatMasker and aimed to identify transposons defined by DNA or RNA mediated mechanisms and not including simple repeats or free satellite elements (Mishmar *et al.*, 2004). Specific NUMTs were classed as adjacent to a repeated sequence when the element was present <150bp away. This research suggests the vicinity of transposable elements influences the ongoing integration of NUMTs and their duplication within the nuclear genome and could be facilitated by open chromatin formation, regions that are prone to chromosomal breakage. This highlights a possible correlation of foreign mtDNA integration with chromosomal structure (Martínez-López *et al.*, 2001). Mishmar *et al.* (2004) also investigated the GC content of 100kb flanking regions surrounding each NUMT finding an association with low-to-moderate GC content isochores (L1-H1) and almost completely absent from high GC content isochores (H2,H3). NUMTs appear to commonly integrate into AT-rich isochores. However, in contrast to these findings Gherman *et al.* (2007) found an initial deficit of repeats within the flanking sequences of NUMTs. This investigation compared 1kb (500bp either side of the mtDNA insert) of flanking sequence surrounding 266 NUMTs, determined earlier in the study, with the entire human genome returning to expected genome-wide levels around 500-600bp away from the insert (Figure 4.4). This research concluded that the human genome had acquired a minimum of several hundred NUMTs arising from a common ancestor as independent insertions in a process that is still active and can affect gene function (Gherman *et al.*, 2007; Turner *et al.*, 2003).

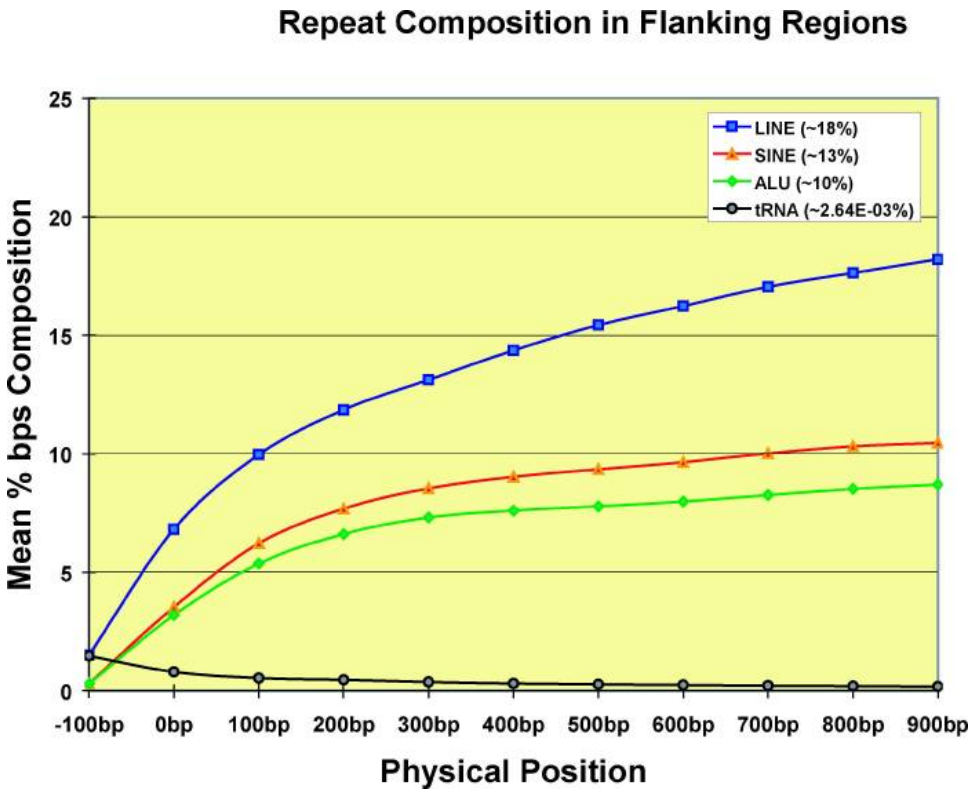

Figure 4.4: Repeat composition of flanking regions taken from Gherman *et al.* (2007). Plot comparing average repeat position of 266 independent NUMTs with 50,000 random sequence fragments of equivalent length. The x axis displays the distance from the integration site. The legend displays the different repeat elements with the average content of the human genome shown in parentheses.

Recent research has confirmed these findings claiming human-specific NUMTs appear to have integrated into regions displaying a significant deficit of transposable elements, an opposite result to chimpanzees (Jensen-Seaman *et al.*, 2009). Jensen-Seaman *et al.* (2009) compared the flanking regions of 37 human-specific NUMTs to 10,000 randomly generated flanks in 100bp windows (Figure 4.5).

4.1.6 The mitochondrial Cambridge reference sequence

The mitochondrial genome sequence was revised and fully corrected by Andrews *et al.* (1999) and can be downloaded from MITOMAP (http://www.mitomap.org) and GenBank (Accession number NC_012920, GI:251831106). The Cambridge reference sequence (rCRS) has 18 corrections or confirmations from the original sequence determined by (Anderson *et al.*, 1981). Eleven nucleotides were corrected due to instances of sequencing errors and contamination of human placental DNA with bovine or HeLa samples. However, GenBank contains a mitochondrial genome sequence (NC_001807) from an African (Yoruba) individual that contains over 40 variant nucleotides from the rCRS. Use of this variant sequence will produce spurious results and is occasionally used in error instead of the rCRS.

4.1.7 BioMart

As the deluge of biological data generated from high-throughput experiments reaches an unprecedented level, the requirement for integrated querying systems involving distributed data sources is crucial. BioMart is an integrated query data management system developed by the EBI and Ontario Institute for Cancer Research (OICR) to allow biologists to perform data mining procedures. Various databases can be interrogated via the BioMart web interface and a full list is displayed in Table 4.1. Numerous bioinformatics experiments can be conducted through implementation of BioMart including SNP (Single Nucleotide Polymorphism) selection for candidate gene analysis, microarray annotation, cross-species analysis and disease association studies (Smedley *et al.*, 2009).

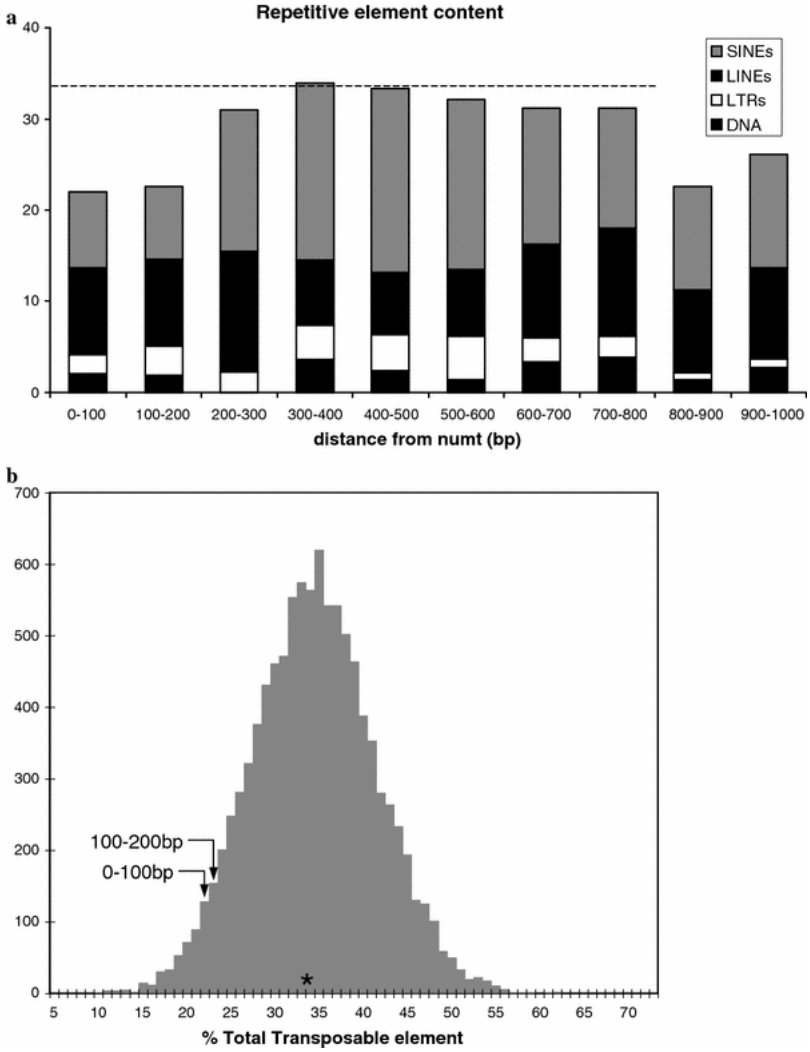

Figure 4.5: Repetitive element content taken from Jensen-Seaman *et al.* (2009). a) Transposable element (TE) content in 100bp windows flanking human-specific NUMTs. The major classes of TEs are displayed in a stacked bargraph. The dashed line represents the average (33.8%) of the total TE content found in 10,000 randomly generated datasets. b) TE distribution of all 10,000 randomly generated data sets with the first flanking windows highlighted. * = Average distribution.

Name of BioMart	Description of contents
Ensembl Genes	Automated annotation of over 40 eukaryotic genomes
Ensembl Homology	Ensembl Compara orthologues and paralogues
Ensembl Variation	Ensembl Variation data from dbSNP and other sources
Ensembl Genomic Features	Ensembl Markers, clones and contigs data
Vega	Manually curated human, mouse and zebrafish genes
HTGT	High throughput gene targeting/trapping to produce mouse knock-outs
Gramene	Comparative Grass Genomics
Reactome	Curated database of biological pathways
Wormbase	C. elegans and C. briggsae genome database
Dictybase	Dictyostelium discoideum genome database
RGD	Rat model organism database
PRIDE	Proteomic data repository
EURATMart	Rat tissue expression compendium
MSD	Protein structures
UniProt	Protein sequence and function repository
Pancreatic Expression Database	Pancreatic cancer expression database
PepSeeker	Peptide mass spectrometer data for proteomics
ArrayExpress	Microarray data repository
GermOnLine	Cross species knowledgebase of genes relevant for sexual reproduction
DroSpeGe	Annotation of 12 Drosophila genomes
HapMap	Catalogue of common human variations in a range of populations
VectorBase	Invertebrate vectors of human pathogens
Paramecium DB	Paramecium tetraurelia model organism database
Eurexpress	Mouse in situ expression data
Europhenome	Mouse phenotype data from high throughput standardized screens

Table 4.1: List of available BioMart databases taken from Smedley *et al.* (2009)

Major Class	Superfamilies
DNA transposons	Mariner, hAT, MuDR, EnSpm, piggyback, P, Merlin, Harbinger, Transib, Novosib, Mirage, Helitron, Polinton, Rehavkus
LTR retrotransposons	Gypsy, Copia, DIRS, BEL
Endogenous retroviruses	ERV1, ERV2, ERV3
Non-LTR retrotransposons	LINE1 (L1), RTE-1, CRE, CR1 (LINE3), I, Jockey, NeSL, R2, R4, Rex1, RandI, Penelope
Caulimoviridae	
Simple repeat	Satellites (SAT, MSAT)

Table 4.2: Repbase schema for transposable element classification taken from Kohany *et al.* (2006). Over 40 superfamilies are contained within the database and consists of a relational database schema that allows for simple addition using the Repbase submitter.

4.1.8 Censor - Repetitive element detection

Censor, developed by Kohany *et al.* (2006) is a computational tool that screens for repetitive elements by comparing to a database of known repeats. Censor implements WU-BLAST as the alignment algorithm chosen for speed and sensitivity. The program analyses DNA/RNA or protein sequences in a range of formats including FASTA, GenBank and GCG. Sequences are compared against the RepBase database of annotated repetitive elements developed by Jurka *et al.* (2005). The Repbase schema is displayed in Table 4.2. The relative abundance of each transposable element in the human genome is can be seen in Table 4.3.

4.1.9 EMBOSS command line applications

EMBOSS (The European Molecular Biology Open Software Suite) is an open source UNIX-based software library for molecular biology consisting of a variety of analysis tools. The multitude of software can accept varying data formats for performing many bioinformatics tasks including sequence analysis, protein motif detection, CpG island analysis and rapid database searching (Rice *et al.*, 2000). The EMBOSS package is available within the Taverna Workbench for constructing analysis pipelines and also as a standalone package for local installation to allow rapid execution of the available analysis programs (Hull *et al.*, 2006).

Major Class	Relative abundance in human genome (%)
DNA Transposons	3% (Pace and Feschotte, 2007)
ERVs	5-8% (Belshaw *et al.*, 2004)
LINES	18% (Gherman *et al.*, 2007)
LTR Retrotransposons	8% (McCarthy and McDonald, 2004)
Mariner	<1% (Pace and Feschotte, 2007)
SINES	13% (Gherman *et al.*, 2007)

Table 4.3: Relative abundance of transposable elements in the human genome.

4.1.10 R script execution in Taverna

Taverna provides the functionality to produce scripts in Java allowing scientists familiar with this language the ability to generate useful code within their workflows. However, many scientists prefer other languages such as the statistical programming language R for which Taverna previously supplied limited support (Wassink *et al.*, 2009). Following the development of an R plugin known as RShell for Taverna, R scripts can now be incorporated into workflows. RShell consists of a client-server structure requiring a local or remote installation of the R-interpreter alongside the installation of the Rserve library (Urbanek, 2003). The Rserve library converts the R-interpreter into a server enabling the communication of other applications through a socket connection. Ultimately, this architecture allows the execution of R scripts through the RShell processor and is fully compatible with the most recent version of R (Wassink *et al.*, 2009). Li *et al.* (2008) provide an example for using RShell in Taverna involving the statistical identification of differentially expressed genes extracted from microarray data followed by the annotation of their relationships to cellular processes. The design of RShell is displayed in Figure 4.6.

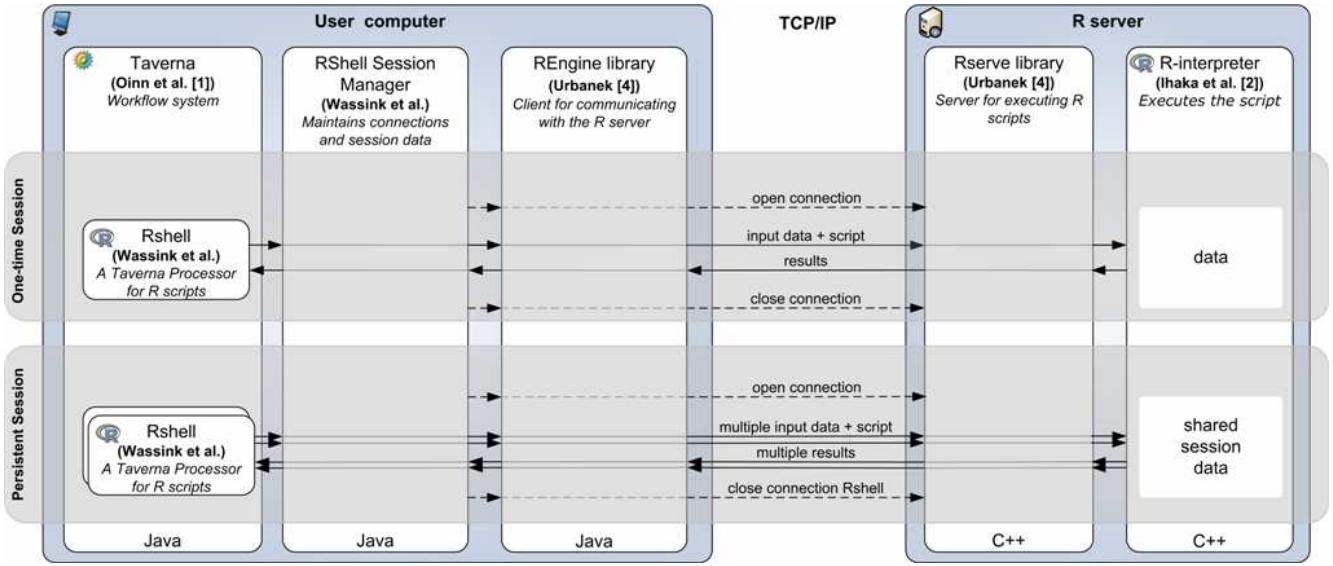

Figure 4.6: Design of the RShell processor taken from Wassink *et al.* (2009). Each processor communicates with the R-interpreter through the RShell session manager. The session manager initiates and maintains communication between the R-interpreter and Rserve library.

4.1.11 Proposed Approach

NUMTs have been determined in a wide range of organisms and various bioinformatics methods have been employed to detect these mitochondrial fragments within the nuclear genome. However, previous studies have explained how a gene is transposed into the nucleus, but fail to explain the mechanism controlling gene loss from mtDNA. An alternative hypothesis is that mtDNA deletions form fragments of double or single stranded mtDNA which are subsequently incorporated into the nuclear genome. It is possible to test this hypothesis by comparing a large existing database of known mtDNA deletions to the size and sequence of thousands of mtDNA fragments present within the cell nucleus. A close correlation would imply a related mechanism, which could be tested experimentally. The size and number of NUMTs varies greatly among species, being highly abundant in plant genomes and completely absent in others. Examples of species displaying no evidence for NUMTs include fish belonging to the order *Tetraodon* and the African malarial mosquito (*Anopheles gambiae*). The aim of this study is to quantify the number of human NUMTs and conduct various analyses implementing bioinformatics workflows. Following the determination of the nuclear location of the mitochondrial insertions, flanking regions surrounding these areas will be analysed for repetitive sequences, GC content and gene content. The analysis of these regions may reveal any areas of the genome that contain higher percentages of mitochondrial DNA. In addition, the origin of the fragments from the mitochondrial genome will be investigated and compared to an existing collection of 263 mitochondrial DNA deletions. This part of the investigation aims to reveal any related mechanisms between mtDNA deletion formation and NUMT formation.

4.2 Methods

4.2.1 Identification of Human NUMTs

In order to assess where fragments of the mitochondrial genome had been incorporated into the nuclear genome over evolutionary time, sequence analysis was required. To achieve this the mitochondrial genome was downloaded from MITOMAP (see section 4.1.6) as this contained the Cambridge Reference Sequence (rCRS). The rCRS was compared to the RefSeq genomic database using NCBI's BLASTN (nucleotide BLAST found at http://blast.ncbi.nlm.nih.gov/Blast.cgi). The RefSeq genomic database was selected as this returned real genomic positions from the nuclear genome as opposed to contig positions. Specific filters were applied to generate relevant results for the analysis. This involved using the Entrez query facility that allowed specific information from the RefSeq genomic database to be returned. This was an investigation to assess the number of human NUMTs and therefore required filters to remove unwanted BLASTN results. As a query using the human mitochondrial genome would produce vast numbers of hits to not only other mitochondrial genomes belonging to related species but their nuclear genomes also, the following filters were required to extract the unique human nuclear genome hits:

Entrez query: "*Homo sapiens* [organism] NOT alternate assembly NOT mitochondrial".

In order to reduce duplicate hits which would result in an overestimation of unique NUMTs, the filter was designed to remove any hits that were labelled alternate assembly as only reference assembly was of interest. As hits to the human nuclear genome were the only hits of interest, the final part of the filter removed any hits to the mitochondrial genome. Without this in place the majority of top scoring hits were mitochondrial related including hits to the recently sequenced Neanderthal mitochondrial genome (Accession number NC_011137, GI:196123578) and were of no relevance to this analysis. In addition, using the BLASTN parameters, the e-value was set to $e< 10^{-4}$ as this removed less significant hits and could be compared to previous research that used the same threshold and was believed to reflect a threshold of biological significance. The configured BLASTN page is illustrated in Figure 4.7. These results were then downloaded in the form of a BLAST hit table suitable for further analysis.

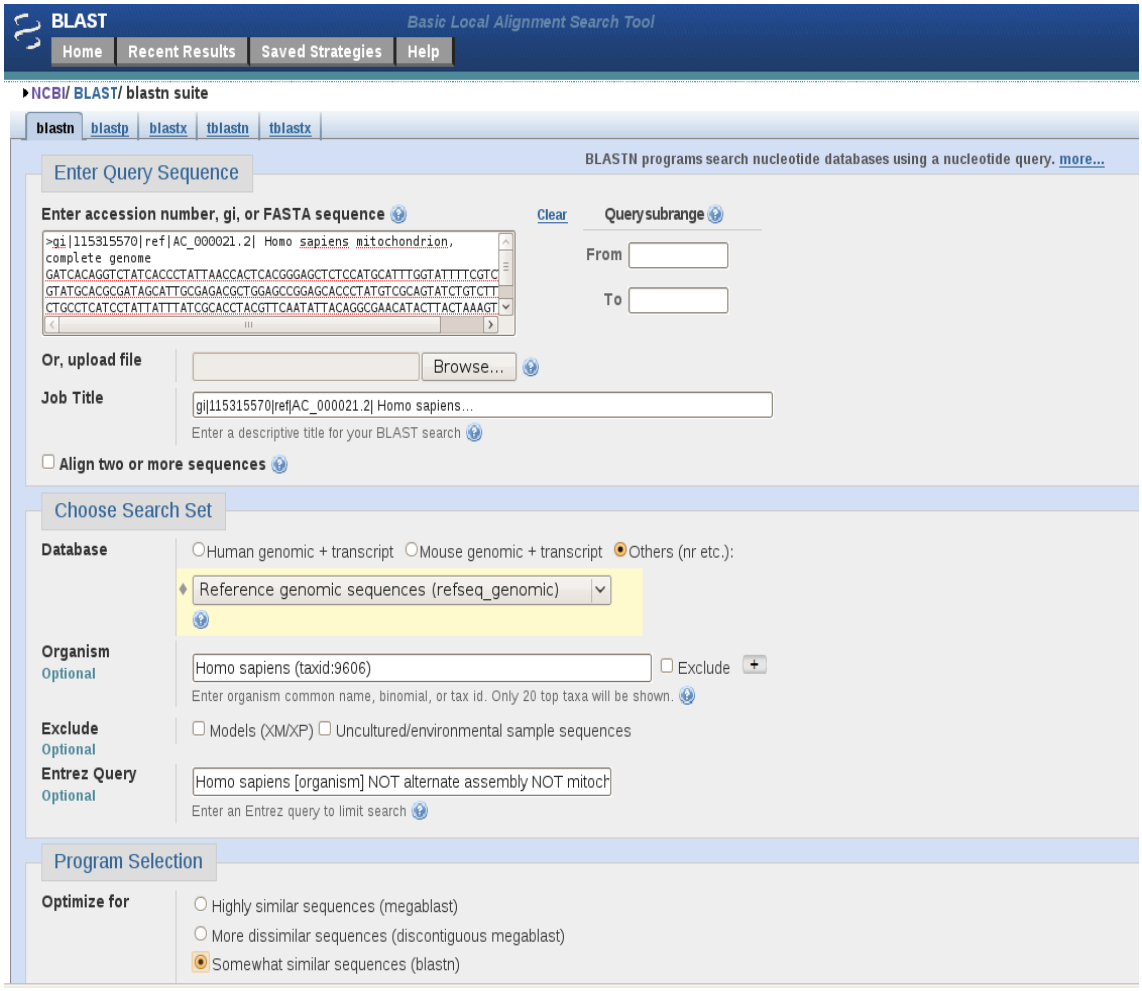

Figure 4.7: Configuration of the BLASTN parameters based at the NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) specifying specific filters in order to produce relevant hits.

4.2.2 BLASTN analysis

Following the BLASTN analysis the results were downloaded in the form of a BLAST hit table that contained all the data from the comparison of the human mitochondrial genome to the human nuclear genome. This contained columnised data that is represented in the truncated BLAST hit table displayed in Figure 4.8. This data was then imported into an excel spreadsheet and organised in preparation for further analysis. The first step involved determining which chromosome the the individual hits belonged to. Part of the subject id contained this information in the accession number. The numbers preceded by 'NC_' and a string of zeroes contained the chromosome number at the end. For example 'NC_000005' reflected chromosome 5 and 'NC_000022' reflected chromosome 22. Chromosomes X and Y were represented numerically by accession numbers 'NC_000023' and 'NC_000024', respectively. These were filtered out leaving only the chromosome numbers for this column. Following the determination of the chromosomes the hits were organised by chromosome and then by their position in relation to the mitochondrial genome in ascending order (referring to the columns q. start and q. end) as these reflect the genomic positions of the mitochondrial fragments. Finally, they were organised by bit score which represents the overall score of the individual hit that takes into account the *e*-value, mismatches in the sequence and the length of the sequence match. Unique NUMTs were determined that appeared to have high scoring hits to the nuclear genome. Any overlapping fragments were considered to be the same NUMT if they resided at the same nuclear genomic positions. If fragments appeared at the same location in duplicate, their individual length was considered and bit score whereby the highest scoring hit or longest regarding nucleotide bases would be considered as the unique NUMT. Once the unique NUMTs were determined the next step was to investigate the number of NUMTs per chromosome. This was achieved simply by summing the number of unique NUMTs in the spreadsheet per individual chromosome.

```
# BLASTN 2.2.20+
# Iteration: 0
# Query:  gi|115315570|ref|AC_000021.2| Homo sapiens mitochondrion, complete genome
# RID: W1TN7U13011
# Database: refseq_genomic
# Fields: query id, subject ids, % identity, alignment length, mismatches, gap opens, q. start, q. end,
# s. start, s. end, evalue, bit score
# 786 hits found
gi|115315570|ref|AC_000021.2| gi|51511721|ref|NC_000005.8|NC_000005 88.36 9118 1000 10 6117 15183 99418648 99409541
0.0 1.167e+04
gi|115315570|ref|AC_000021.2| gi|51511721|ref|NC_000005.8|NC_000005 94.06 5219 310 0 10269 15487 134292116 134286898
0.0 8015
gi|115315570|ref|AC_000021.2| gi|51511721|ref|NC_000005.8|NC_000005 87.27 3465 437 3 12662 16124 93928917 93932379 0.0
4251
gi|115315570|ref|AC_000021.2| gi|51511721|ref|NC_000005.8|NC_000005 94.27 2358 123 7 341 2697 79983943 79981597 0.0
3624
gi|115315570|ref|AC_000021.2| gi|51511721|ref|NC_000005.8|NC_000005 77.40 2093 451 10 5916 7992 97775384 97773298 0.0
1617
gi|115315570|ref|AC_000021.2| gi|51511721|ref|NC_000005.8|NC_000005 71.23 2461 653 27 14146 16569 8671689 8674131 0.0
1175
gi|115315570|ref|AC_000021.2| gi|51511721|ref|NC_000005.8|NC_000005 84.58 973 133 11 574 1537 123124395 123125359 0.0
1045
gi|115315570|ref|AC_000021.2| gi|51511721|ref|NC_000005.8|NC_000005 78.39 1004 212 2 6954 7956 5450081 5449082 0.0
829
gi|115315570|ref|AC_000021.2| gi|51511721|ref|NC_000005.8|NC_000005 71.54 1634 434 19 14415 16033 97040470 97038853
0.0  792
```

Figure 4.8: Output of the BLASTN analysis required for determining NUMTs

NUMT	q.start	q.end
1	1	458
2	11	241
3	11	241
4	11	76
5	115	426
6	128	455
7	335	457
8	341	2697
9	346	459
10	383	708

Table 4.4: First 10 examples for the identified NUMTs and their relevant mitochondrial positional coordinates that comprise the user query for the NUMT plotting workflow of the fragments in relation to the mitochondrial genome.

4.2.3 Distribution of NUMT origin across the mitochondrial genome

Using the results from the BLASTN analysis an important assessment was the determination of the where the NUMTs had originated from in relation to the mitochondrial genome. This would reveal any distribution pattern and could be compared to mtDNA deletion distribution in order to highlight any correlation. An automated workflow was created using Taverna1.7 that required the positions of each fragment and plotted these against the mitochondrial genome. The workflow incorporated an R script using the Rserve facility in Taverna to automatically produce the image once the workflow had completed. The workflow architecture is illustrated in Figure 4.9.

Workflow user query

From the spreadsheet containing the NUMT BLASTN results the 620 NUMTs were organised into a separate list ordered by ascending mitochondrial base start position and then numbered 1 to 620. This list could then be used as the query for the NUMT plotting workflow as seen in the example displayed in Table 4.4. The workflow then splits the positions of each identified NUMT into a list of separate objects for iteration using the local processor Split_positions. Following this the NUMT number and related positions are separated out into individual inputs by the beanshell Separate_positions in preparation for the next process.

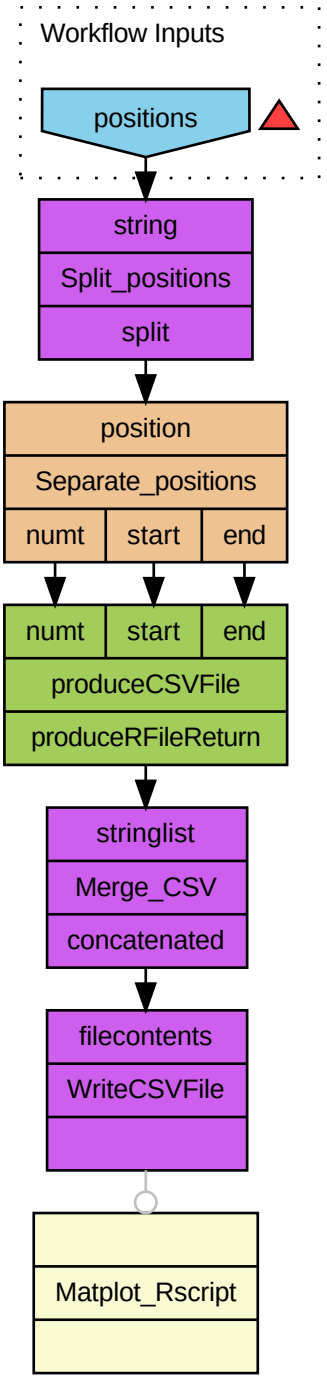

Figure 4.9: Workflow that produces a visual plot of the physical positions of NUMTs across the mitochondrial genome.

```
,,,,1,1,1,1,1,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,(..n=16569)
,,,,,,,,,2,2,2,2,2,2,2,2,2,2,2,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,(..n=16569)
,,,,,,,,,,,,,3,3,3,3,3,3,3,3,3,3,3,3,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,(..n=16569)
,,,,,,,,,,,,,,,,,,,,,,,4,4,4,4,4,4,4,4,4,4,4,4,,,,,,,,,,,,,,,,,,,,,,(..n=16569)
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,5,5,5,5,5,5,5,5,5,5,5,(..n=16569)
6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,(..n=16569)
```

Figure 4.10: File architecture required by the Matplot package in R for plotting the positions of the mitochondrial fragments across the mitochondrial genome.

Process 1: File preparation for Matplot

Using the positions for each NUMT, the next processor produceCSVFile is a Java program developed to produce a line of empty comma separated values in concordance with the size of the human mitochondrial genome. Where the NUMT is present along this line a relevant value is indicated reflecting the NUMT's previously assigned number. An example output of this program can be seen in Figure 4.10. The entire list is then merged into a single document by Merge_CSV and written to a file specified by the processor WriteCSVFile in preparation for analysis via an R script.

Process 2: Matplot invocation using RShell

RShell is utilised to execute an R script within the Taverna workflow to produce the desired plot. This is conditionally linked to the WriteCSVFile processor as this needs to have completed and prepared the file before the R script is invoked. The script initially reads in the created CSV file and then transposes the file into columns relevant for the R package Matplot. Following this the output for the resulting image is specified. Matplot is then executed producing a linear display of all the fragments in relation to their physical position on the mitochondrial genome. The workflow plots the NUMTs linearly in relation to their position in the mitochondrial genome. These are converted to polar plots to represent the circular structure of the mitochondrial genome.

Matplot R script

```
dloopdist <- read.csv(file="/home/kieren/R/NUMTs/dloopdist.csv",
 header=TRUE)
tdloopdist <- t(dloopdist)
png(filename="/home/kieren/R/NUMTs/dloopdist.png",
 height=800, width=1000)
par(font.lab=2, font.axis=2, font.main=2, cex=1.0)
matplot(tdloopdist, yaxt="n", ylab="NUMTs", xlab="bp position",
 pch=22, type="l", lwd=2, lty=1, col="black")
dev.off();
```

4.2.4 Gene mining of flanking regions surrounding NUMTs

Following the identification of the NUMT insertion sites another area of investigation was sequence analysis of the flanking regions in the nuclear genome surrounding the NUMTs to analyse gene content. This aimed to reveal if there was a high abundance of mitochondrial-related genes in close proximity and reveal potential evolutionary conservative mechanisms for mitochondrial DNA integration. To analyse these regions a workflow was constructed to analyse the flanking sequences surrounding the NUMT integration sites and the gene content (Figure 4.11). Biomart was implemented using the *Homo sapiens* genes GRCh37 database to extract relevant gene information.

Workflow user query

Nuclear genomic positions of the NUMTs were extracted from the BLASTN analysis spreadsheet and could be inserted into the gene content analysis workflow as a list of chromosomal coordinates (see Table 4.5). The individual coordinates pertaining to each unique NUMT insertion site was split into separate queries using a local Taverna processor named Split_positions. Each individual set of coordinates are then passed to a nested workflow to allow each one to be analysed one at a time.

Process 1: Nested workflow

The nested workflow (light blue processor) allows each set of coordinates to be analysed in full before the next set is consumed. A beanshell script SeparateCoordinates splits the coordinates into three separate outputs for chromosome, start bp and end bp. The chromosome number is passed straight to the Biomart query filter.

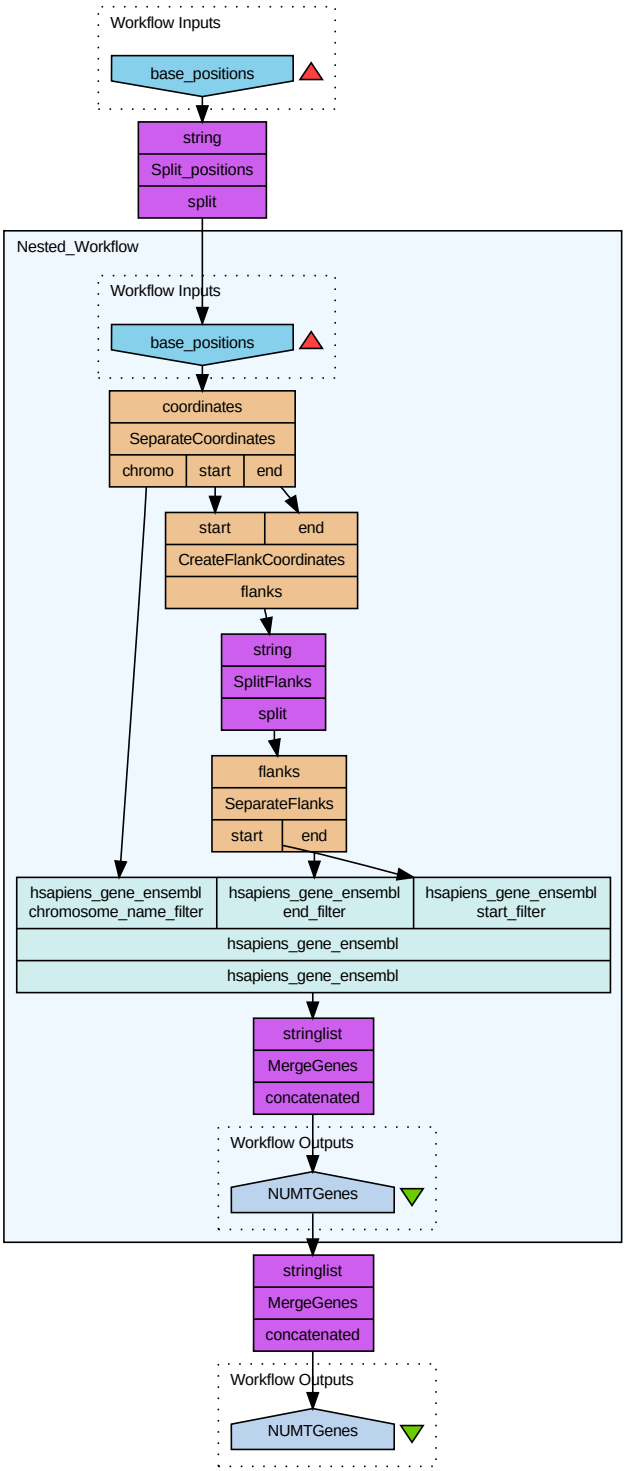

Figure 4.11: Workflow that requires the NUMT positions as input and alters the query to locate flanking regions surrounding the NUMTs using Biomart. These regions are then searched for gene content.

Chromosome	s. start	s. end
1	554324	560167
1	5537393	5537524
1	5832905	5833115
1	8892389	8892554
1	9557274	9557474
1	11407906	11408416
2	9804293	9804445
2	15578742	15578857
2	22393428	22393625
2	33846042	33846097
2	40865601	40865761
2	49310271	49310542
3	12181341	12181477
3	25483999	25484037
3	28351381	28351435
X	3987625	3987886
X	5096878	5098813
X	9733688	9733853
X	15655775	15655965
X	23984302	23984498
Y	4272822	4272892
Y	8291603	8293329

Table 4.5: An example list of nuclear chromosomal coordinates required for the gene mining workflow of NUMT flanking sequences.

Process 2: Extend flanking regions

The start and end base positions are sent to another beanshell script CreateFlankCoordinates to extend the base positions to incorporate the specified flanking regions surrounding the integration site within the nuclear genome. The script to perform the extension is shown below:

CreateFlankCoordinates example

```
input start: 5000000
input end:   5005000

i = Integer.parseInt(start);
j = Integer.parseInt(end);

#Adding 1MB flanks
flanks = (i - 1000000) + " " + i + "\n" + j + " " + (j + 1000000);

output list: 4900000 5000000
             5005000 5105000
```

Following this process the separate flanks are split into a pair of queries by the SplitFlanks local processor representing the newly generated coordinates of each flank. These flanks are separated into start bp and end bp queries in preparation for the Biomart processor by the beanshell SeparateFlanks. Each individual NUMT query from the start of the workflow now becomes two separate queries covering both flanks either side. These can then be sent in succession to the Human Biomart processor.

Process 3: Biomart gene retrieval

Taverna provides the functionality to query the Biomart database from within a workflow. This allows queries to be sent to a species-specific processor and queried using filters such as chromosomal coordinates, gene names, gene ids and genetic markers. A Biomart processor was selected from the Ensembl 56 Genes (Sanger UK) database specifying the *Homo sapiens* genes ensembl (GRCh37) subset for incorporation into the workflow. The Human Biomart was configured to return all genes using the chromosomal coordinates of the specified flanking regions (Figure 4.12).

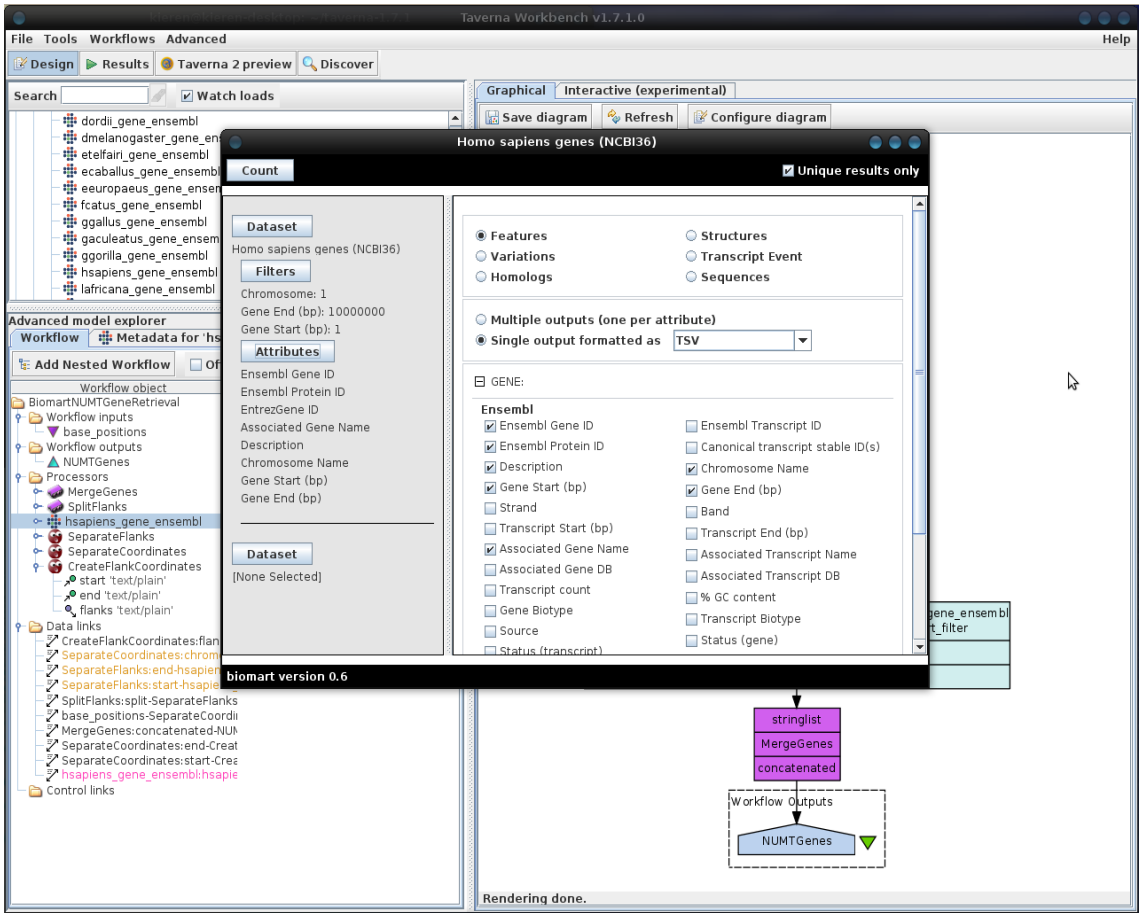

Figure 4.12: Biomart filter and attribute configuration for gene mining queries.

Using the filter page the inputs were configured within the Biomart by selecting the REGION section and specifying filters for chromosome and base pair positions. In the attributes, the outputs were configured within the Features section under the GENE and EXTERNAL headings. Specific attributes were then selected that consisted of the following:

Biomart GENE attributes configuration
Ensembl Gene ID
Ensembl Protein ID
Associated Gene Name
Chromosome Name
Gene Start (bp)
Gene End (bp)

Biomart EXTERNAL attributes configuration
Entrez Gene ID

These attributes were considered to be important when analysing the genes that were retrieved. The Gene IDs were important for cross-referencing within the local databases. MitoCarta referenced genes using the Entrez ID therefore, this was requested as an external ID to allow querying the genes within the MitoCarta database of predicted mitochondrial genes. Other attributes consisted of gene name, description and the gene's specific chromosomal coordinates. In order to correlate the chromosomal coordinates entering the Biomart query processor an iteration strategy was configured consisting of a cross product with dot product nodes. This is due to the chromosome name coming from the SeparateCoordinates beanshell and requires being compared to each set of flanking coordinates. These coordinates are produced by the SeparateFlanks beanshell and need to be correlated as pairs using the dot product within the iteration strategy. This is achieved by configuring the iteration strategy to ensure the chromosome is compared to every flanking coordinates pair whilst maintaining the pairs are correlated with each other separately. This strategy is illustrated in Figure 4.13. Finally, the output consists of 2 lists of genes representing each flank for every NUMT queried resulting in 1240 potential lists. The entire workflow result is then merged into a single document by the MergeGenes processor to allow for further analysis.

The workflow requires the positions of the insertion sites and was run five times to assess flanking regions consisting of 50Kb, 100Kb, 150Kb, 200Kb and 250Kb flanks to either side altering the base positions within the query with the NUMT excised.

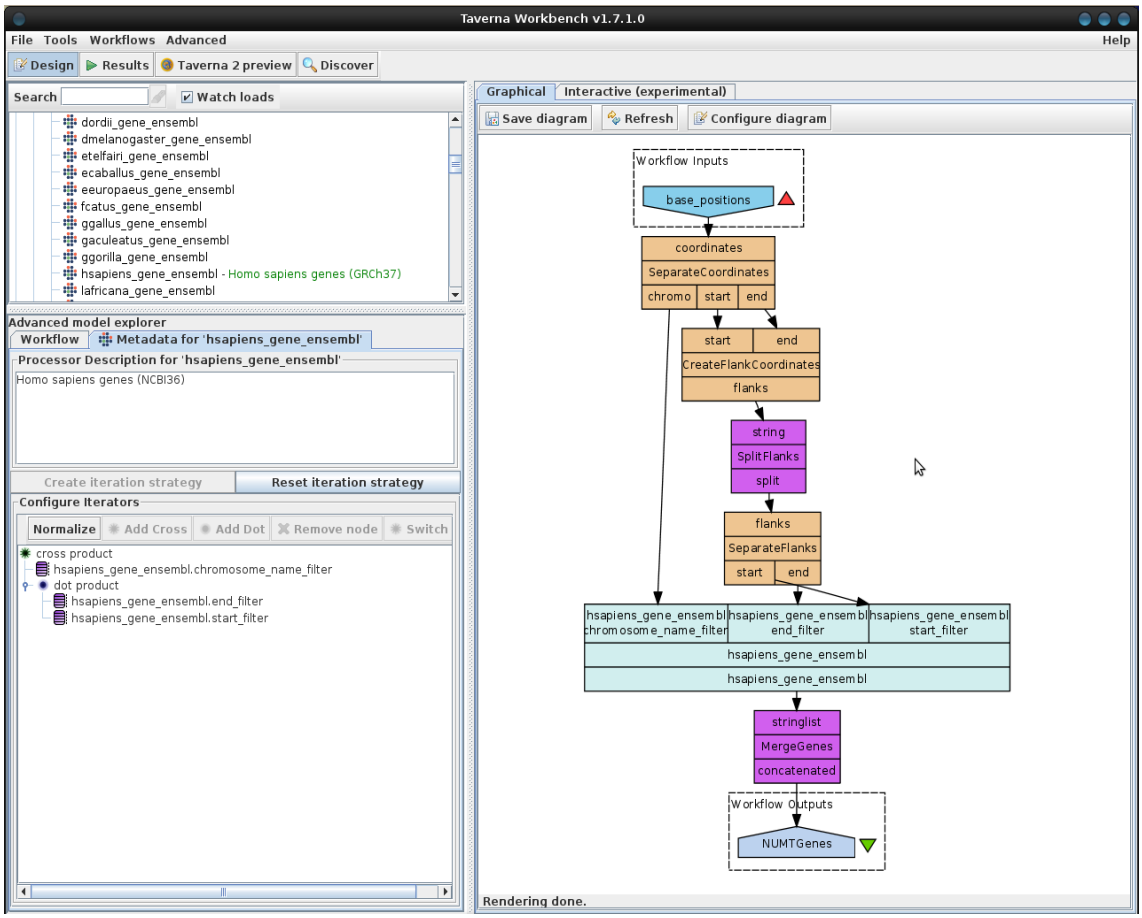

Figure 4.13: Iteration strategy configuration where the user can specify either cross product or dot product nodes.

Each gene is then assessed for mitochondrial involvement using previous established prediction techniques. Using the most successful combination of prediction methods (MitoSVM) from the previous chapter, each flanking region can be analysed for mitochondrial genes. This can be cross-referenced with MitoCarta (Pagliarini *et al.*, 2008) to reveal other potential mitochondrial genes.

4.2.5 Transposon and GC content analysis of flanking sequences

In addition to gene content analysis the flanking regions were investigated regarding transposable elements and GC content. However, this required extracting raw DNA sequence using the relevant insertion site coordinates from the human genome.

Reference DNA sequence extraction

Reference DNA sequences could not be retrieved using Biomart as this returned specific gene, protein or transcript-centric content and not raw unannotated DNA reference sequence. Querying specific coordinates required explicit filters in Biomart to return annotated information. Therefore Biomart could not been utilised in this procedure. To achieve the extraction of raw human DNA sequence, firstly each human chromosome was downloaded in FASTA format from the NCBI and stored in flat files. These could then be queried using the specific base pair coordinates using the EMBOSS (http://emboss.sourceforge.net/) command line tool extractseq. Extractseq enables the extraction of specific regions from a given sequence consisting of either an EMBL id or a specific filepath to the query sequence. Querying stored sequence files on the local machine allows for a much more rapid execution for sequence extraction. This is essential as the sequences reflect the entire human genome represented across 24 separate files. A simple implementation of extractseq can be seen below:

```
extractseq /home/kieren/NUMTs/HumanGenome/ch22.fasta -reg "10-20"
```

This query would return base pair positions 10-20 from human chromosome 22. However, due to each chromosome having multiple NUMTs this resulted in the need for batched queries to be implemented when querying the individual chromosome FASTA files. Extractseq was able to take batched queries and return a multiple FASTA file containing all the individual genomic regions.

An example of a batched query was executed as follows:

```
extractseq /home/kieren/NUMTs/HumanGenome/ch22.fasta
-reg "34899432-34900668, 34905058-34905543, 34611665-34611711,
4537-4766, 31620869-31621209, 15737114-15737254, 40490907-40490975,
48866720-48866949, 22679236-22679718, 45244827-45244894"
-separate
-outseq /home/kieren/NUMTs/HumanGenome/regions/ch22regions.fasta
```

The arguments required included the path to the relevant human chromosome FASTA file, in this case chromosome 22. The -reg command specified all the regions of interest separated by a comma. The -separate command simply separates the resulting FASTA sequences into a multiple FASTA file and labels each by the specific query using the chromosomal coordinates.

4.2.6 DNA extraction workflow for human genome

The development of a fully automated procedure to extract all the regions of raw DNA sequence comprising the insertion site flanks was required. This procedure was automated in a workflow that extracted all the regions from the entire human genome as batched queries in iteration (Figure 4.14). This was achieved by extracting the chromosomal coordinates for all of the NUMTs from the BLASTN analysis data.

Workflow user query

The DNA extraction workflow required two separate input lists consisting of all the human chromosomes in ascending order and all the specific regions in the same order whereby each newline referred to a separate chromosome. An example is shown below:

Input lists: Chromosome
1
2
3

NUMT base pair positions
10-20, 40-80, 400-670
35-55, 900-1200, 3000-3440
4550-6000, 8900-10000, 12000-12500

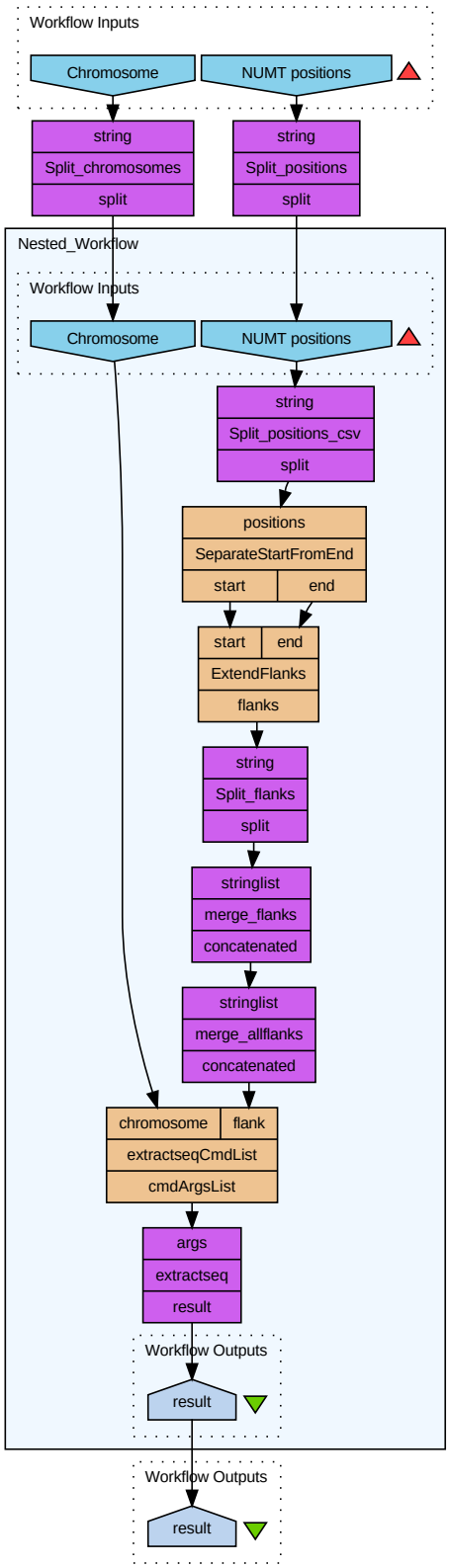

Figure 4.14: Workflow that requires the chromosome numbers and associated NUMT positional coordinates as input and extracts raw DNA sequences from the human genome in preparation for computational sequence analysis.

This enabled the workflow to assess each chromosome and its specific regions of insertion in isolation. This required a nested workflow to allow this to function correctly. Both lists were split into separate individual queries by carriage returns using local processors (purple) and sent to the next stage of the workflow.

Process 1: Nested workflow

A nested workflow (light blue processor) was constructed to consume each separate query from both lists with an iteration strategy in place. As implemented in previous workflows the strategy used here was dot product. This ensured chromosome 1 was correlated with the regions belonging to chromosome 1 and chromosome 2 correlated with regions belonging to chromosome 2 and so on, eventually extracting sequences from all chromosomes in relation to their insertion site regions.

Process 2: Flanking region extension

The first stage of the nested workflow is to alter the base pair coordinates in order to extend the flanking regions of interest in preparation for transposon and GC content analysis. The specific regions are separated into a list of queries by the processor Split_positions_csv using commas as the separator. This allows the next processor to deal with each individual region one at a time. This output enters the beanshell processor (beige) named SeparateStartFromEnd which generates two outputs start and end. An example of these procedures can be seen below:

Split_positions_csv

```
input: 250-350, 400-580, 600-750
output list: 250-350
             400-580
             600-750

#SeparateStartFromEnd
first input: 250-350
output start: 250
output end:   350
```

The separation of the start and end positions allows the next processor to extend these depending on the size of the required flanking regions. This processor is another beanshell script named ExtendFlanks that consumes the two inputs start and end and extends these regions. This script along with its inputs and outputs can be seen in Figure 4.15. The generated output are the flanking regions surrounding a specific integration site incorporating the extended positions specified in the script, not including the actual NUMT itself. This could be modified for each separate run as the flanking regions were increased regarding flank size in preparation for sequence analysis. The following script was developed to extend the flanks displaying an example for extending the flanks by 100bp:

ExtendFlanks example
input start: 250
input end: 350

```
i = Integer.parseInt(start);
j = Integer.parseInt(end);

flanks = i - 100 + "-" + i + "," + j + "-" + (j + 100);
```

output: 150-250, 350-450

Following the generation of the two new flanking region coordinates these are merged into a single string of text by the local processor merge_allflanks. This output can then be sent to the next processor in the workflow extractseqCmdList which constructs the specific query required by the extractseq command line program in the format explained in section 4.2.5.

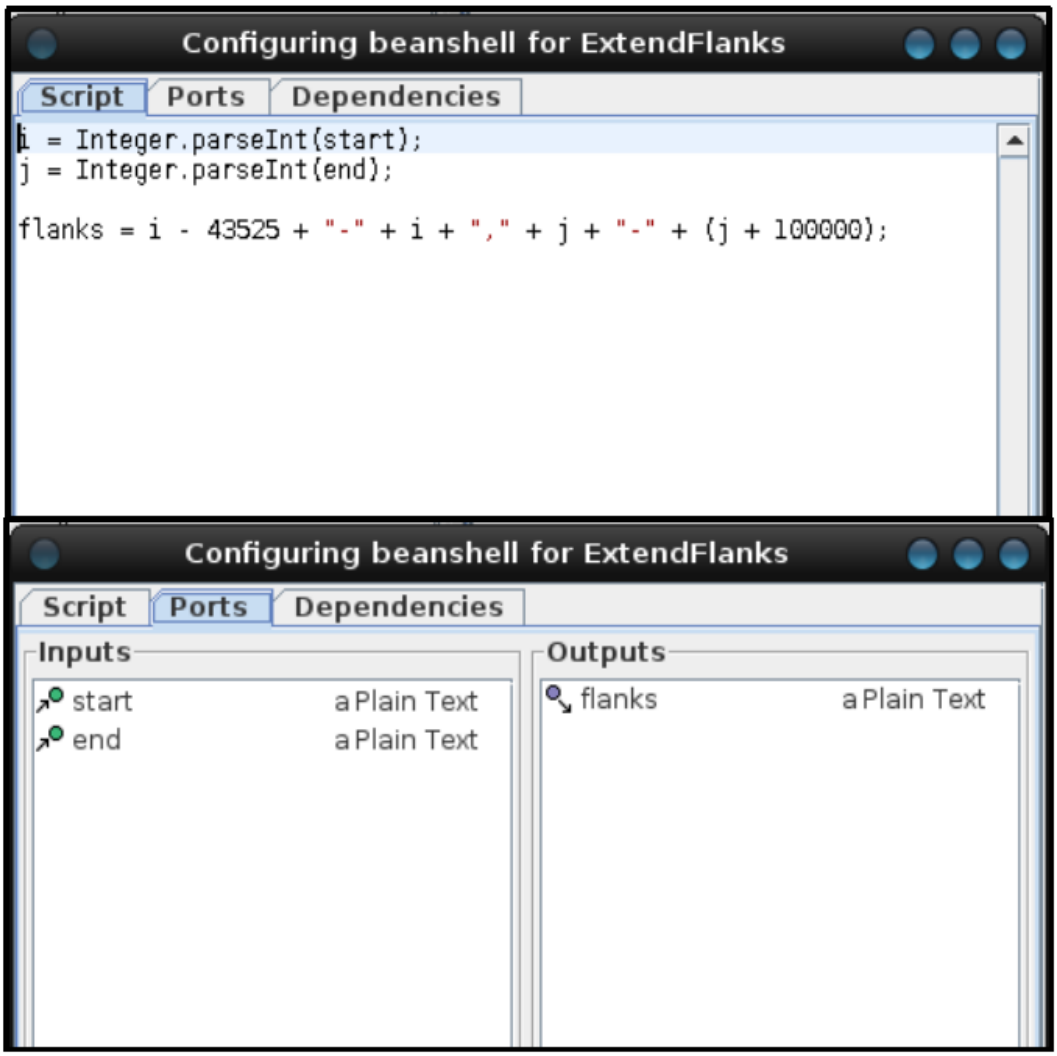

Figure 4.15: Beanshell script that requires the base pair start and end positions as input and extends these to incorporate specific flanking regions

Process 3: Extractseq batched query preparation

The next processor extractseqCmdList constructs the actual list of commands to be sent to the command line to execute the extractseq program. This list of commands can be seen below illustrating the construction of the arguments required for the extractseq command line application:

extractseqCmdList arguments

```
List cmdArgsList = new ArrayList();
cmdArgsList.add("/home/kieren/NUMTs/HumanGenome/ch"
+ chromosome + ".fasta");
cmdArgsList.add("-reg");\\
cmdArgsList.add("\"" + flank + "\"");
cmdArgsList.add("-separate");
cmdArgsList.add("-outseq");
cmdArgsList.add("/home/kieren/NUMTs/HumanGenome/regions/Final/
Flanking100KB/small/ch" + chromosome + "flank.fasta");
```

This processor requires two inputs flank and chromosome whereby flank consists of the extended regions referring to the NUMT insertion sites produced by the Extend-Flanks processor and the second input being the specific chromosome they belong to. The first argument specifies the filepath to the particular stored chromosome file that the sequences will be extracted from. Following this the -reg command consumes the output from ExtendFlanks processor referring to the flanking regions. All the flanking sequences are separated into separate FASTA sequences by the -separate command. Finally, the -outseq command specifies the filepath for where the multiple FASTA is to be stored. This requires the name of the chromosome implemented earlier in the process to specify the output filename. Therefore, the output generated represents the format illustrated earlier (section 4.2.5) as a batched query.

Process 4: Extractseq command line execution

The final part of the workflow is the execution of extractseq on the command line (purple processor). This requires two inputs, the first being a permanent command extractseq which is required by the command line to execute the application and the second being the previously constructed list of arguments. The results for each chromosome is automatically written to flat files contained in the relevant directory using the filepath specified in the arguments. Finally, the overall outcome is 24 flat files representing each chromosome. Each chromosome contains a list of FASTA DNA sequences that have the

necessary flanking regions surrounding each specific NUMT. An example output can be seen below for a user query consisting of chromosome 1 specifying positions 250-350, 400-580, 600-750:

Chromosome 1 flanks example

```
>NC_000001_150_250 Homo sapiens chromosome 1,
reference assembly, complete sequence
CTAACCCTAACCCTAACCCTAACCCT
AACCTAACCCTAACCCTAACCCTAAC
CCTAACCCTAACCCTAACCCTAACCC
TAACCCCTAACCCTAACCCTAAA
>NC_000001_350_450 Homo sapiens chromosome 1,
reference assembly, complete sequence
CCTACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCCTAACCCCT
AACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCCTAACCCT
>NC_000001_300_400 Homo sapiens chromosome 1,
reference assembly, complete sequence
CCCCAACCCCAACCCCAACCCCAACC
CTAACCCCTAACCCTAACCCTAACCC
TACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCCTAACCCC
>NC_000001_580_680 Homo sapiens chromosome 1,
reference assembly, complete sequence
GGTGCTCTCCGGGTCTGTGCTGAGGA
GAACGCAACTCCGCCGGCGCAGGCGC
AGAGAGGCGCGCCGCGCCGGCGCAGG
CGCAGACACATGCTAGCGCGTCG
>NC_000001_500_600 Homo sapiens chromosome 1,
reference assembly, complete sequence
GTCTGACCTGAGGAGAACTGTGCTCC
GCCTTCAGAGTACCACCGAAATCTGT
GCAGAGGACAACGCAGCTCCGCCCTC
GCGGTGCTCTCCGGGTCTGTGCT
```

Each individual NUMT and its flanking regions are generated into a FASTA formatted sequence. Each sequence FASTA header includes the NCBI chromosome accession number (e.g. NC_000001) followed by the extended start and end base positions. Finally the text refers to the specific assembly the sequence has been extracted from. Following completion of the sequence extraction workflow the resulting FASTA files were concatenated into one large file containing all the 1240 (620 NUMTs x 2) NUMT flanks. These sequences were the uploaded and analysed with the program Censor based at the EBI, in order to detect repeat sequences and transposons. GC content analysis was performed using the EMBOSS program geecee.

4.2.7 Transposon analysis using Censor

Transposon detection was performed using the EBI web service Censor which screens query sequences against a reference collection of repeats and censors (masks) portions that are homologous. The service produces an overall summary of all repeats found within the query sequences and generates a summary table reporting the number of fragments and their total combined length in bases. Several parameters are available in Censor including MODE and MASK PSEUDOGENES however, all parameters were left as default settings. All flanking sequences were contained in a concatenated file generated following completion of the DNA extraction workflow and contained all 1240 flanks. This file was uploaded and analysed with the Censor web service (Figure 4.16).

4.2.8 GC content analysis using geecee

Fractional GC content for all the flanking sequences was calculated using the EMBOSS command line application geecee. This program calculates the G+C bases contained within the input nucleotide sequence and produces a fractional output between 0.0 to 1.0 for the entire length of the sequence. Using the command line geecee required the name of the input file containing all the flanking sequences of 100Kb flanks and generated an output file listing all the flanks with their associated GC fraction. This was stored in a spreadsheet to allow further analysis to determine isochore groups, mean GC content of the flanks and the standard deviation.

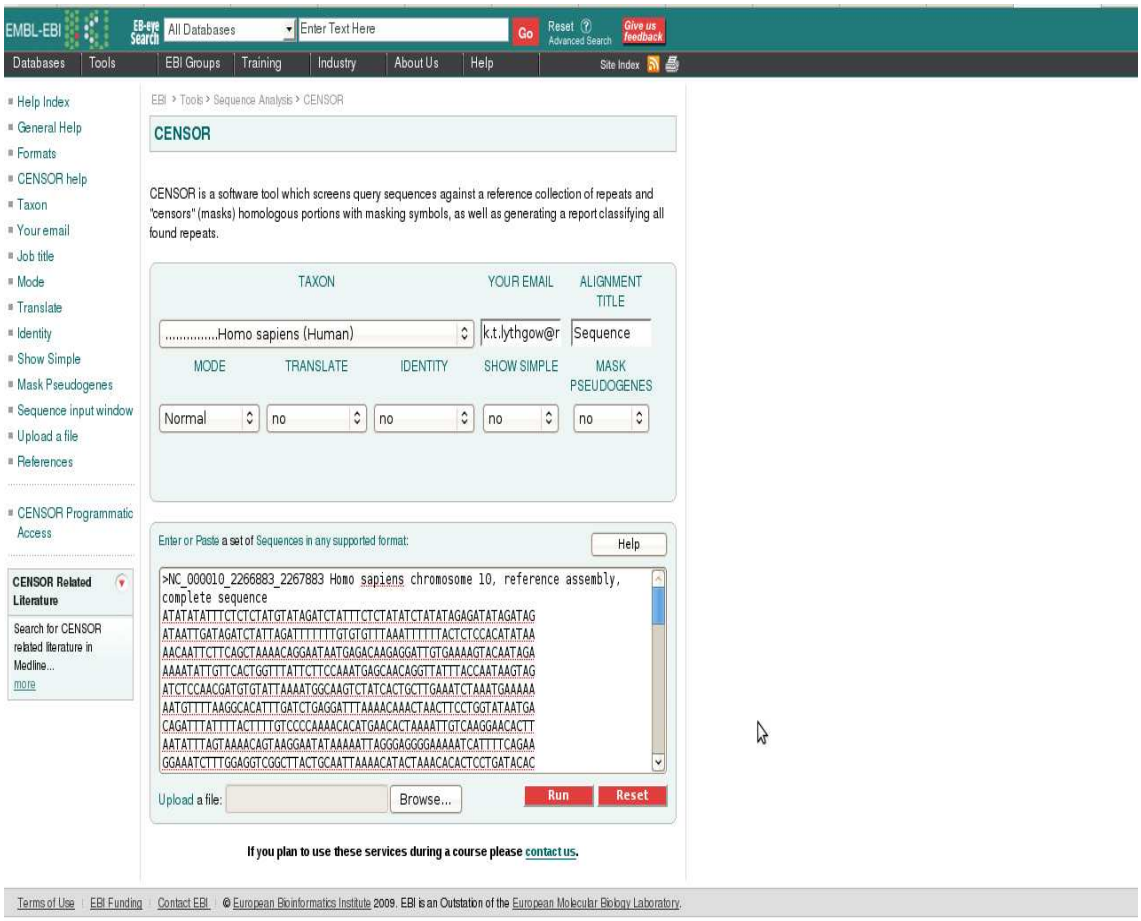

Figure 4.16: EBI Censor web service which screens query sequences against a reference database of repeat elements.

4.3 Results

4.3.1 Identification of NUMTs across the human nuclear genome

Following the BLASTN analysis of the human mitochondrial genome sequence against the human nuclear genome a total of 620 NUMTs were identified. The distribution of NUMTs across the human genome is illustrated in Figure 4.17. Chromosome 2 harbours the most NUMTs containing 62 fragments whereas chromosome 1 contains 36. The lowest number of NUMTs recorded were on chromosomes 14 and 18 both containing 6 NUMTs each. An overall trend of largest to smallest chromosome correlates with NUMT abundance.

4.3.2 Distribution of NUMT origin across the mitochondrial genome

Following the execution of a workflow in Taverna to analyse where the NUMTs had originated from in relation to their position in the mitochondrial genome a graphical plot was generated. This produced a physical plot of all the fragments reflecting their size alongside their relative position in the mitochondrial genome. These positions were determined using the results from the BLASTN analysis as these contained the mitochondrial base positions. Figure 4.18 is a linear plot whereby the x axis reflects the mitochondrial genome base positions (1-16569) and the y axis represents the stacking of each individual NUMT fragment. This also displays a linear mitochondrial genome with the 13 protein coding genes, the 2 rRNAs (12s rRNA and 16s rRNA) and the d-loop. A polarised plot was also constructed to display the NUMT fragments across a circular represented mitochondrial genome highlighting the minor arc (blue) and major arc (red) (Figure 4.19).

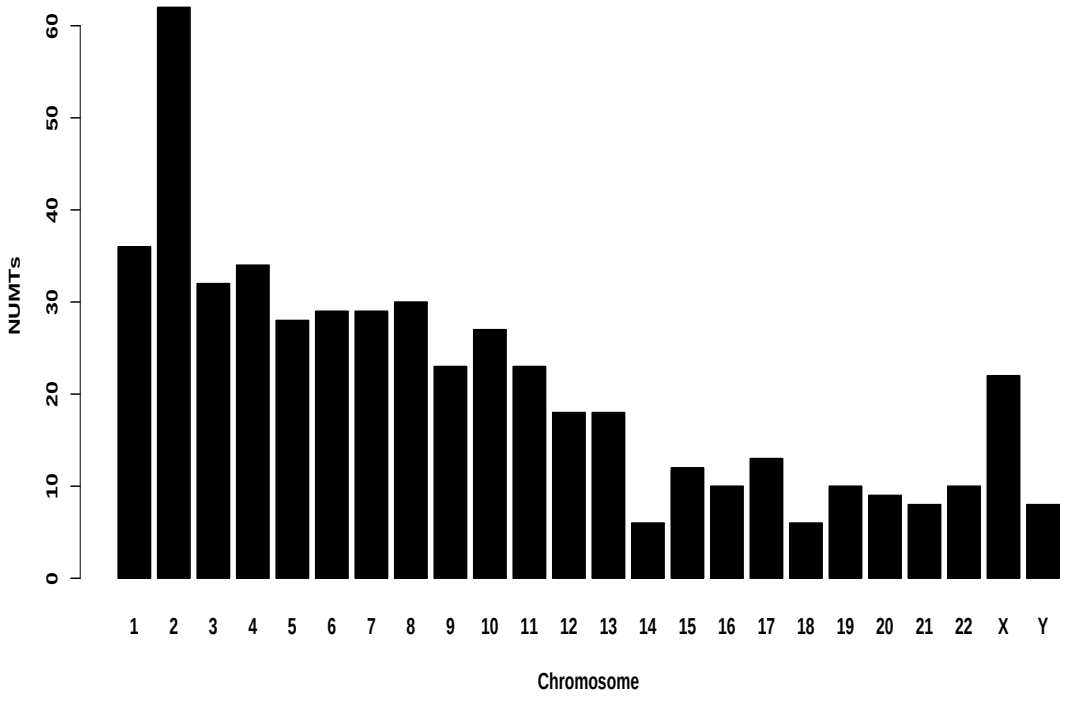

Figure 4.17: Distribution of NUMTs detected per chromosome throughout the human genome.

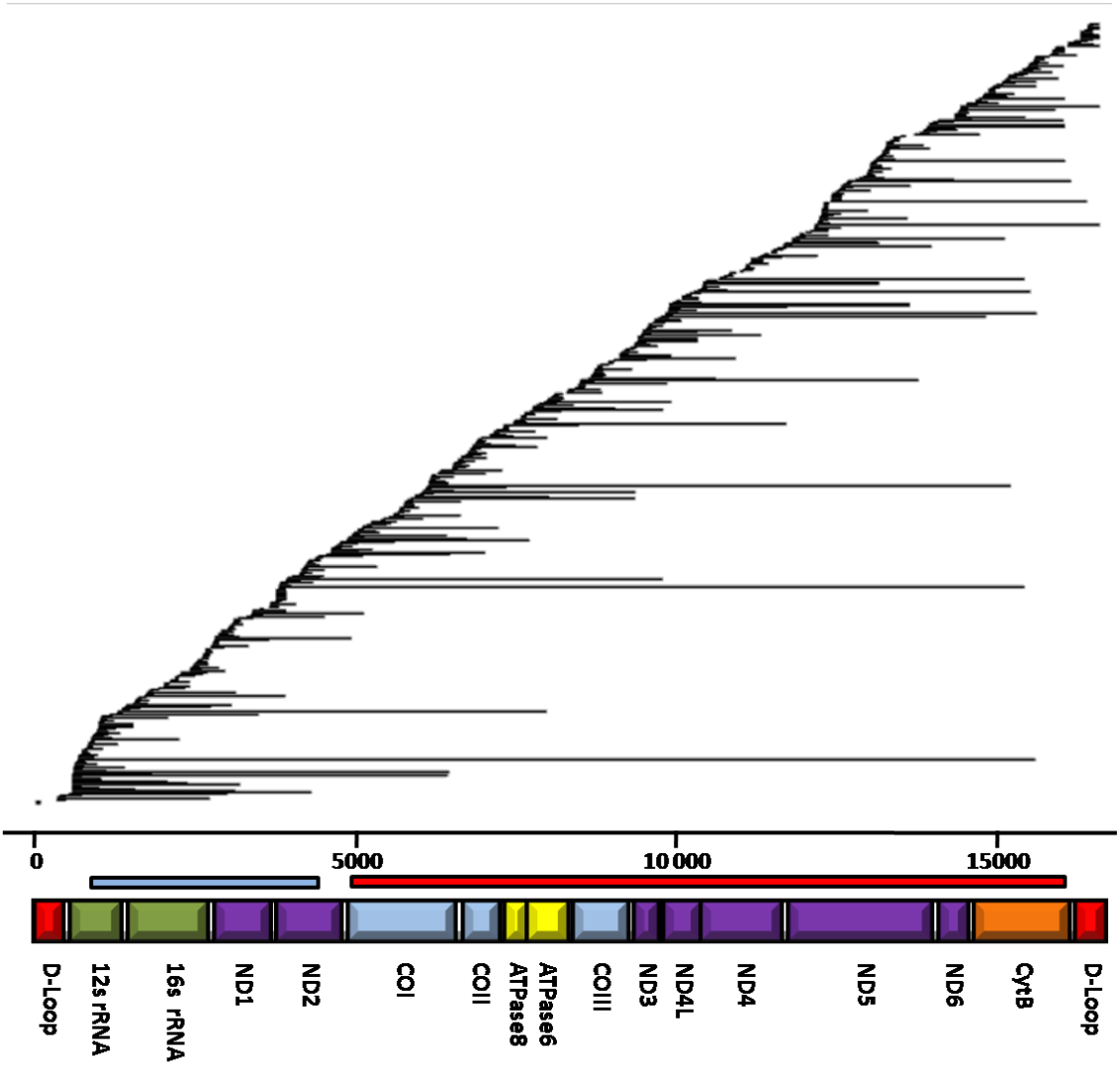

Figure 4.18: A linear plot displaying the size and location of each individual NUMT's origin across the mitochondrial genome. The mitochondrial genome is graphically represented in relation to the fragments with the 13 protein coding genes, 2 rRNAs (12s RNA and 16s rRNA) and the d-loop region labelled. Coloured bars represent the minor arc (blue) and the major arc (red).

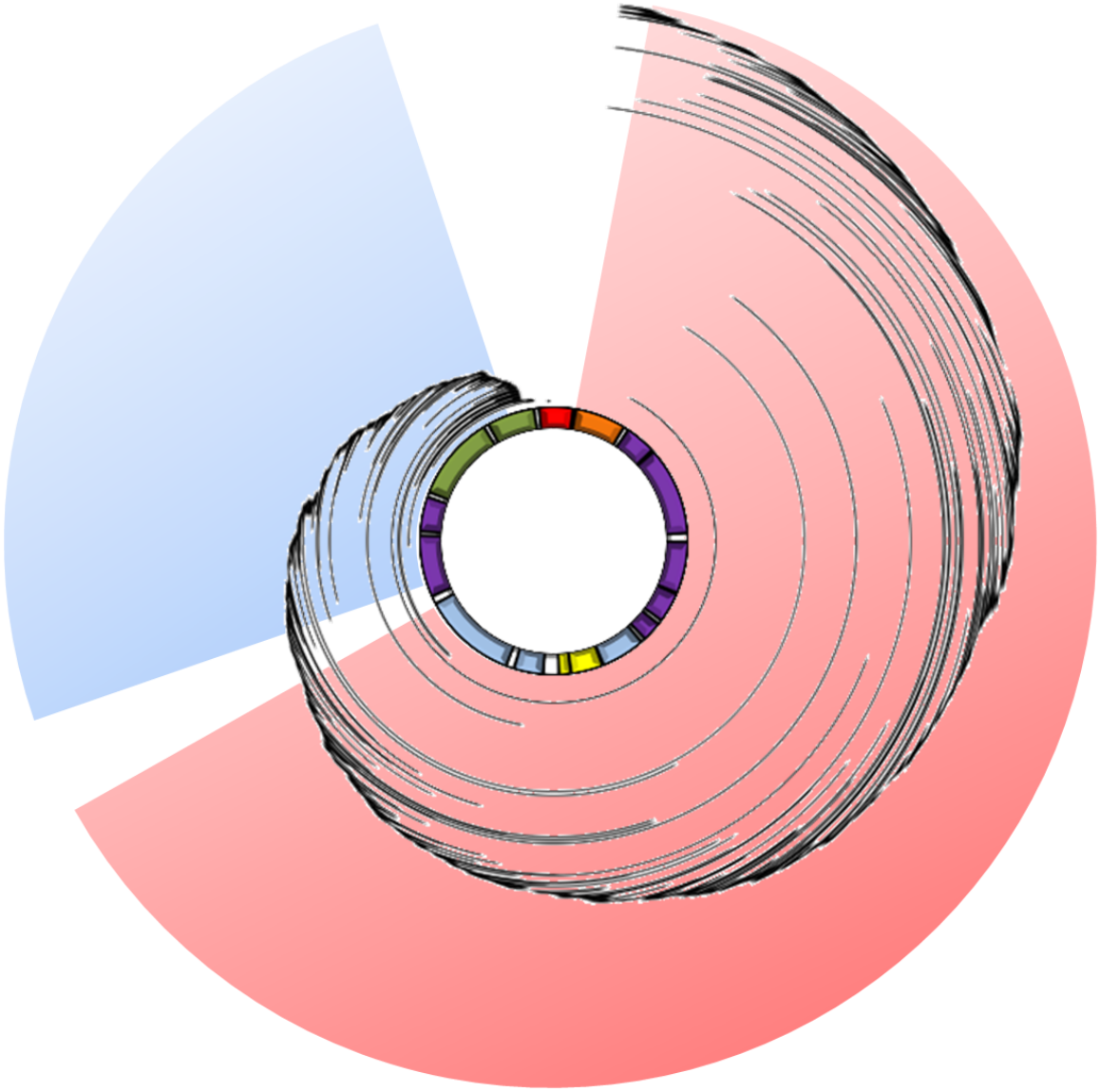

Figure 4.19: A polarised plot displaying the size and location of each individual NUMT's origin across a circular mitochondrial genome. Coloured sections represent the minor arc (blue) and the major arc (red).

4.3.3 Distribution of mtDNA deletions across the mitochondrial genome

In order to assess any potential relation to mtDNA deletion formation and NUMT formation a database of 263 previously identified mtDNA deletions was analysed. Using the same method for investigating the origination of NUMTs in relation to their known positions in the mitochondrial genome, the positions of the known 263 mtDNA deletions was physically plotted to visualise them against the mitochondrial genome. The linear plot can be seen in Figure 4.20 and the polarised plot in Figure 4.21. Both polar plots for the NUMTs and mtDNA deletions for comparative purposes are combined in Figure 4.22 to illustrate the differences of their locations and size.

4.3.4 NUMTs & mtDNA deletions per base position across the mitochondrial genome

Further to the analysis generating the physical plotting of the NUMT fragments across the mitochondrial genome another plot was generated to display the abundance of NUMTs arising per base position. This is displayed in Figure 4.23. The location of each of the 13 mitochondrial protein coding genes is highlighted in order to assess genes that may have a higher incidence of NUMT formation.

In addition to the plotting of NUMT abundance per base position in the mitochondrial genome a plot was generated for the same data regarding the abundance of mtDNA deletions per base position. This would potentially highlight any significant similarities or differences regarding areas of the mitochondrial genome more active for either phenomenon. This plot is displayed in Figure 4.24.

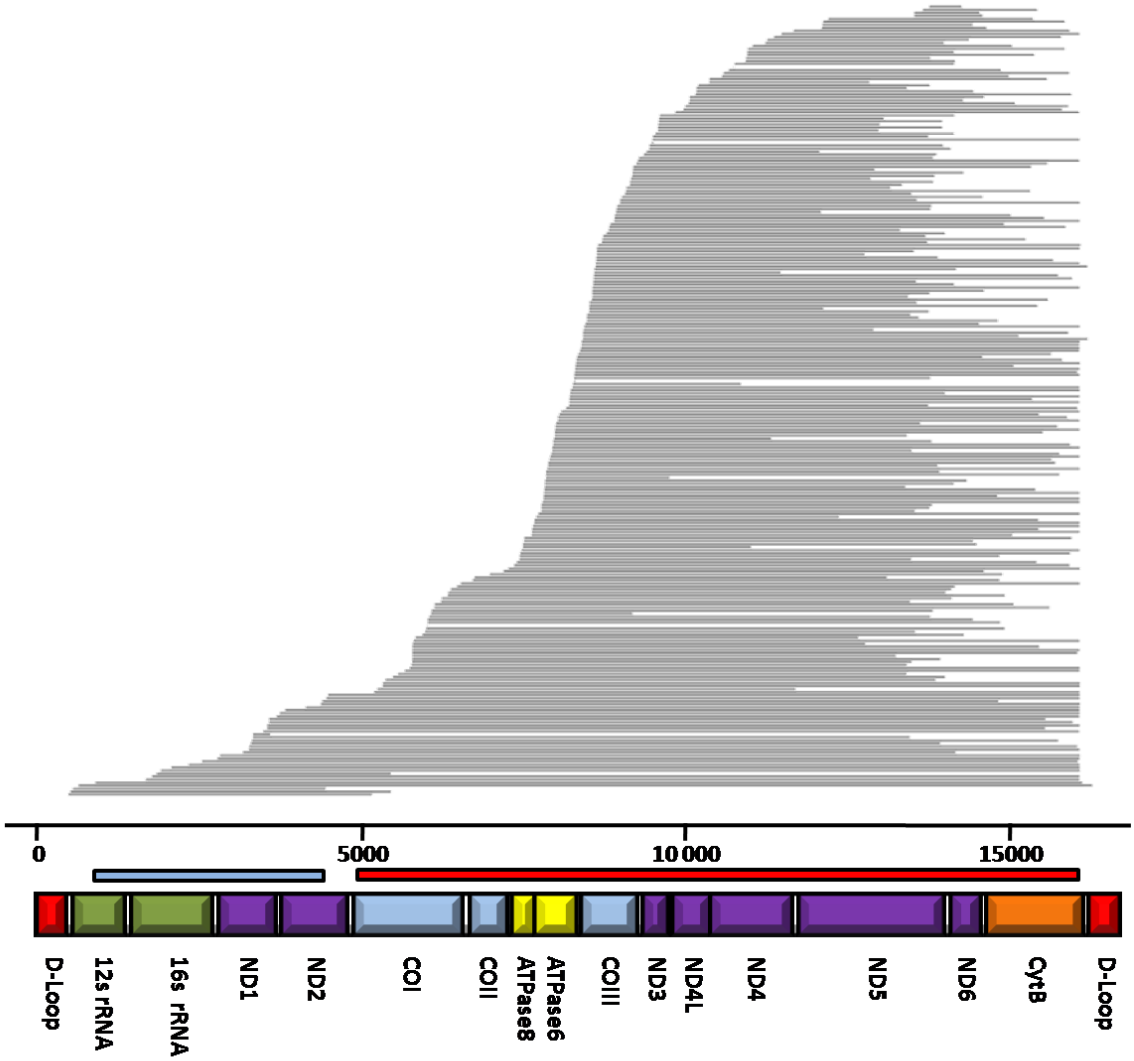

Figure 4.20: A linear plot displaying the size and location of each of the 263 individual mtDNA deletion origin across the mitochondrial genome. The mitochondrial genome is graphically represented in relation to the fragments with the 13 protein coding genes, 2 rRNAs (12s RNA and 16s rRNA) and the d-loop region labelled. Coloured bars represent the minor arc (blue) and the major arc (red).

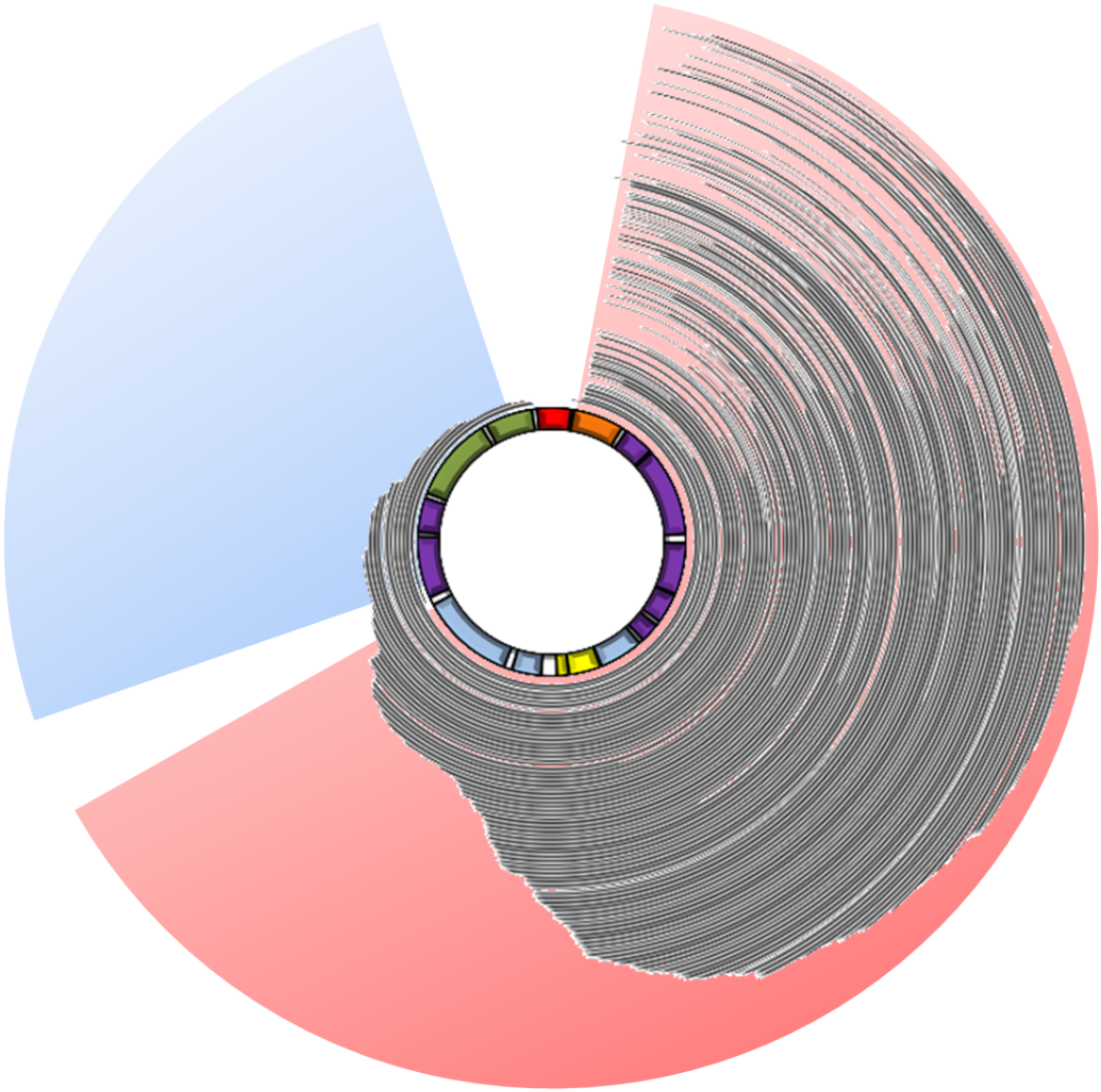

Figure 4.21: A polarised plot displaying the size and location of each individual mtDNA deletion origin across a circular mitochondrial genome. Coloured sections represent the minor arc (blue) and the major arc (red).

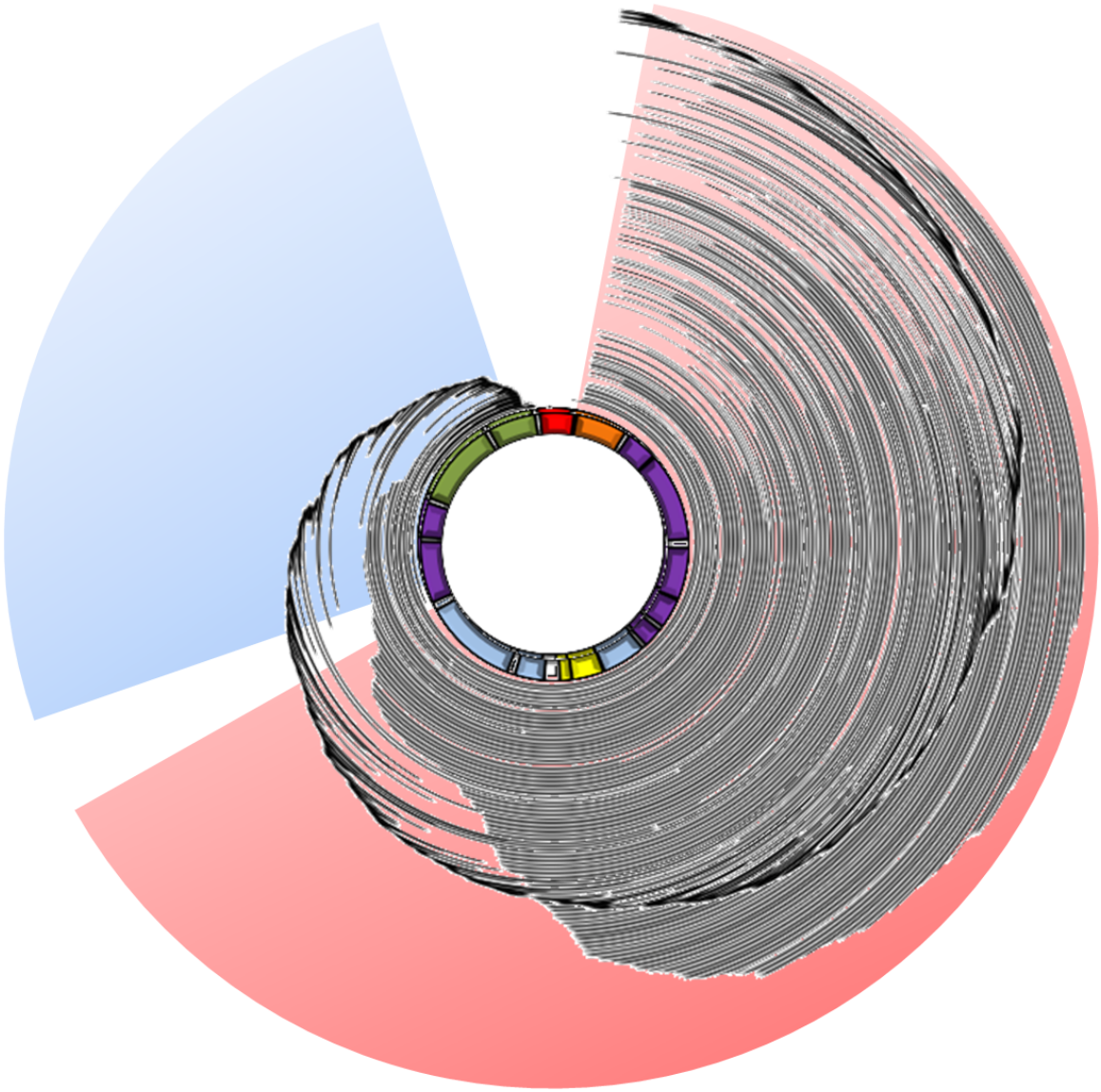

Figure 4.22: A combined polarised plot displaying the size and location of each individual NUMT and mtDNA deletion origin across a circular mitochondrial genome. Coloured sections represent the minor arc (blue) and the major arc (red).

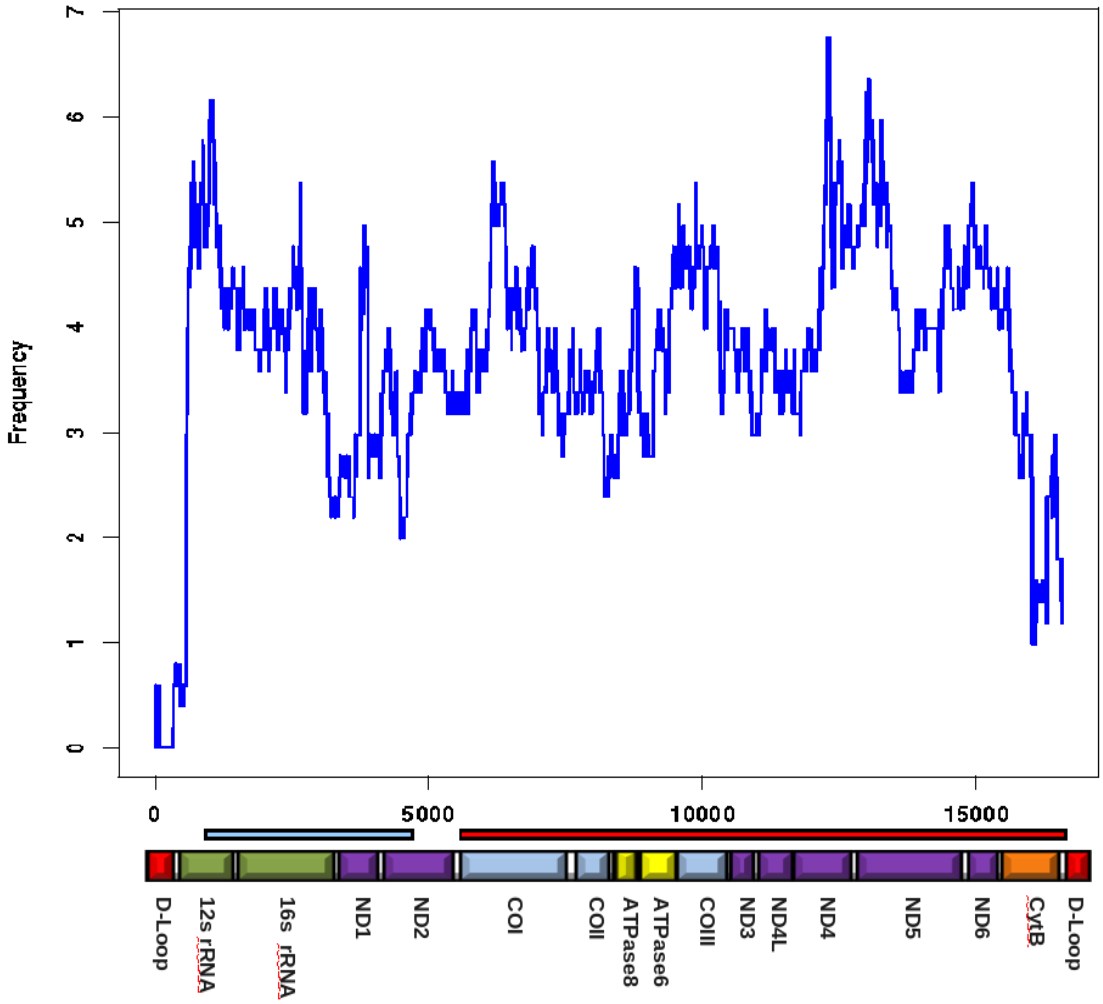

Figure 4.23: Abundance of NUMTs per base position across the mitochondrial genome displaying the location of each of the 13 protein coding genes. Coloured bars represent the minor arc (blue) and major arc (red).

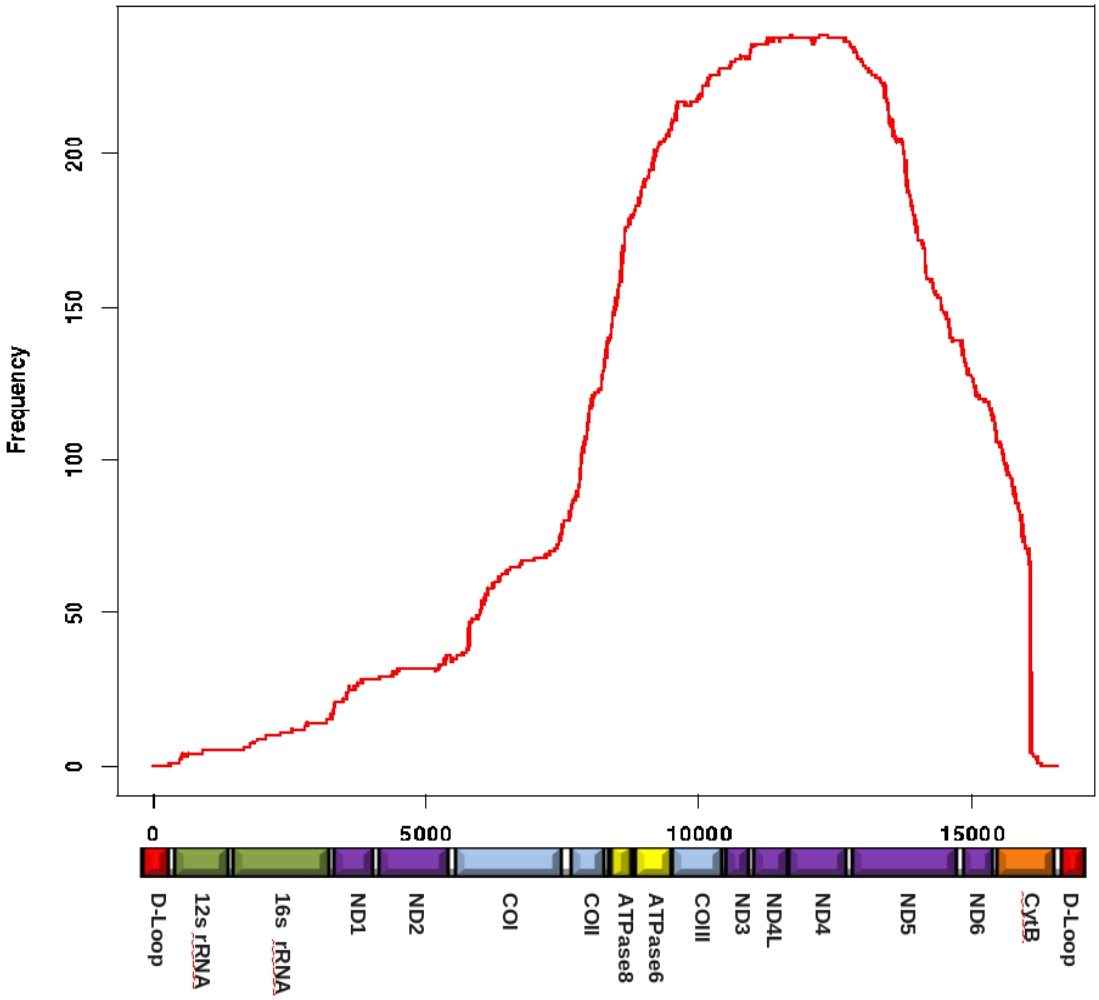

Figure 4.24: Abundance of mtDNA deletions per base position across the mitochondrial genome highlighting the positions of each protein coding gene. Coloured bars represent the minor arc (blue) and major arc (red).

4.3.5 Gene content of flanking regions

The flanking sequence positions surrounding the NUMT insertion sites were analysed for mitochondrial gene content to potentially highlight evidence of mitochondrial gene clustering around these specific areas of the nuclear genome. Following the data extraction using BioMart within Taverna an assessment of the increasing flanking regions was implemented consisting of the windows 50MB, 100MB, 150MB, 200MB and 250MB. BioMart retrieved all the important gene information including chromosomal coordinates, gene name and descriptions. Each flanking window was then analysed for genes annotated as mitochondrial. All the retrieved genes contained their Ensembl Gene Id, Ensembl Protein Id and Entrez Id if available. Figure 4.25 displays the percentage of genes annotated as mitochondrial reflecting the difference between Ensembl and Entrez regarding the number of genes they contain. This displays the percentage of mitochondrial genes present within each 50MB window moving outwards from the NUMT insertion sites that are annotated in the public databases as being mitochondrial. The expected percentage of mitochondrial genes genome wide can be approximated by calculating 100/25000 x 1000. This takes into account the estimates for the number of mitochondrial genes and genes within the nuclear genome. Therefore, the percentage of mitochondrial genes is approximately 4%. Using these values as a control, mitochondrial prediction methods MitoCarta and MitoSVM can be inferred to highlight the percentage of predicted mitochondrial genes present among the flanking regions. Figure 4.26 displays the percentage of predicted genes implementing MitoCarta for assessing the regions for mitochondrial genes overlayed with the previous figure reflecting the annotated mitochondrial genes. In addition, the same was applied using MitoSVM and is displayed in Figure 4.27.

4.3.6 Transposon analysis of flanking regions

Sequence analysis was performed on the resulting FASTA file generated from the DNA extraction workflow for each size group (100bp, 200bp,...1000bp). These files were analysed with the repeat sequence analysis web service EBI Censor. For each group a summary table was produced listing the different classes of repeat sequences found (e.g DNA transposons, LINES and SINES) along with the number of fragments and their overall length in bases. Each group had the percentage abundance calculated for each class of repeat. Figure 4.28 illustrates the percentage abundance for each type of transposable element in ascending order of group base size.

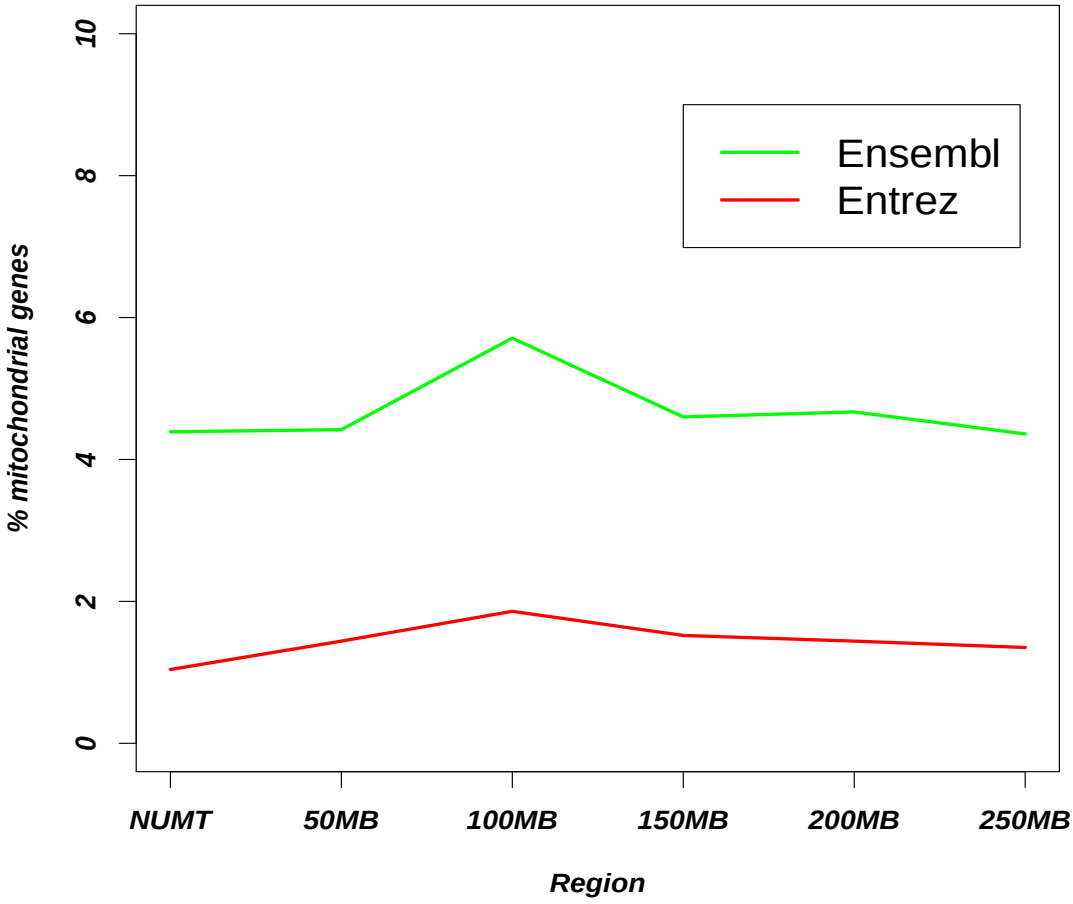

Figure 4.25: Percentage of genes detected within the specific flanking regions (50MB windows) annotated as mitochondrial by the Ensembl and Entrez databases.

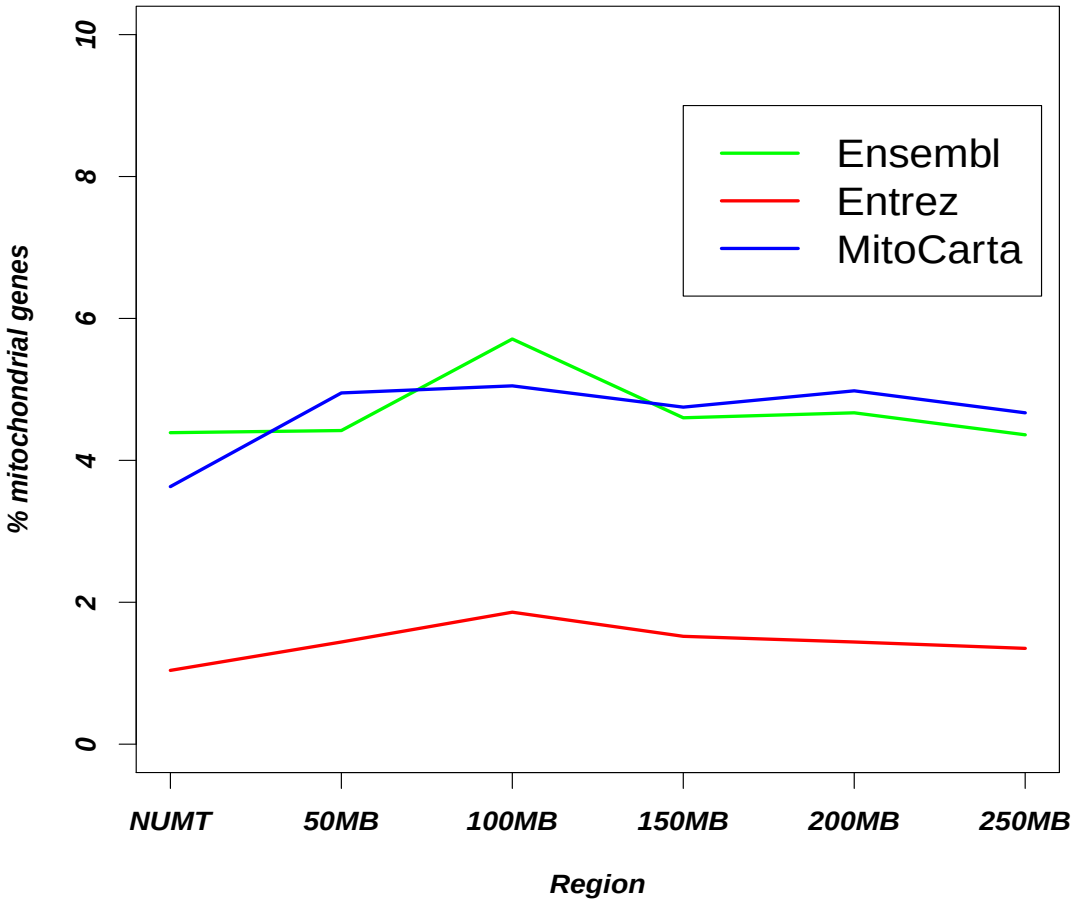

Figure 4.26: Percentage of mitochondrial genes predicted by Mitocarta within the specific flanking regions (50MB windows). The genes annotated as mitochondrial by the Ensembl and Entrez databases are displayed for comparison.

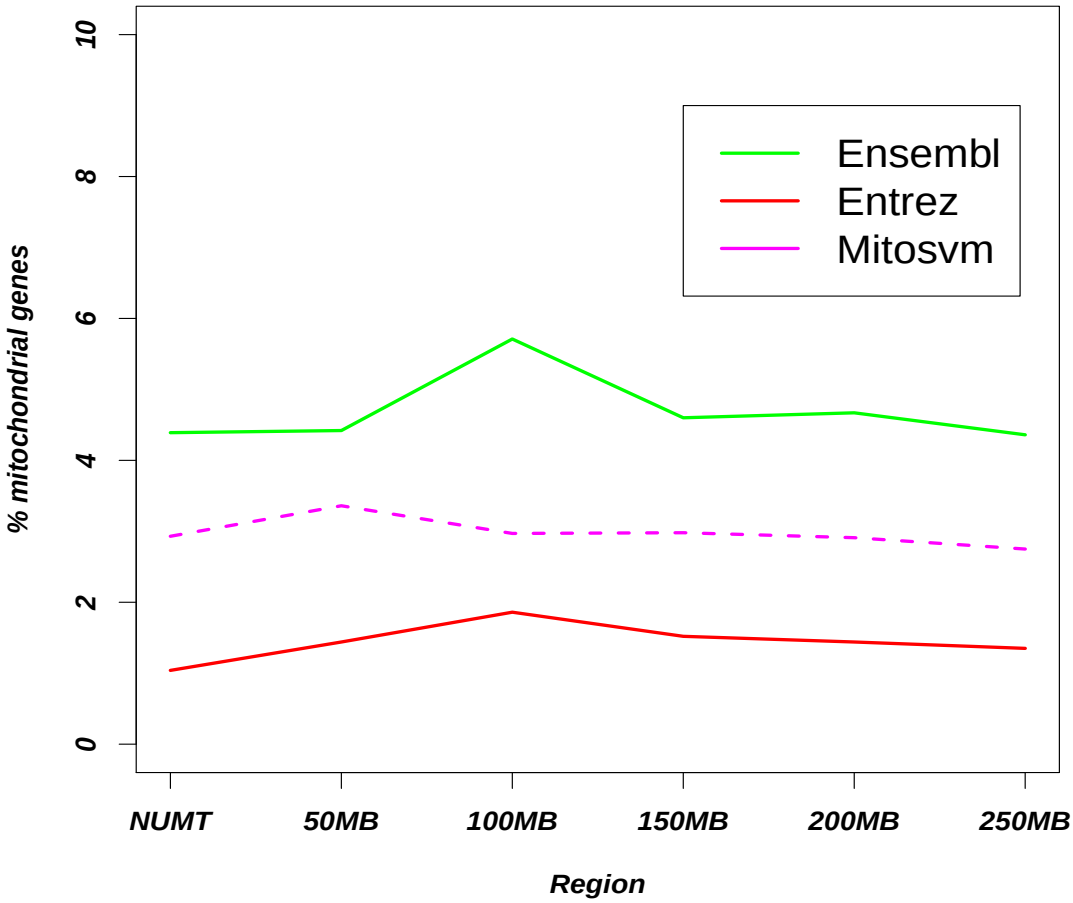

Figure 4.27: Percentage of mitochondrial genes predicted by MitoSVM within the specific flanking regions (50MB windows). The genes annotated as mitochondrial by the Ensembl and Entrez databases are displayed for comparison.

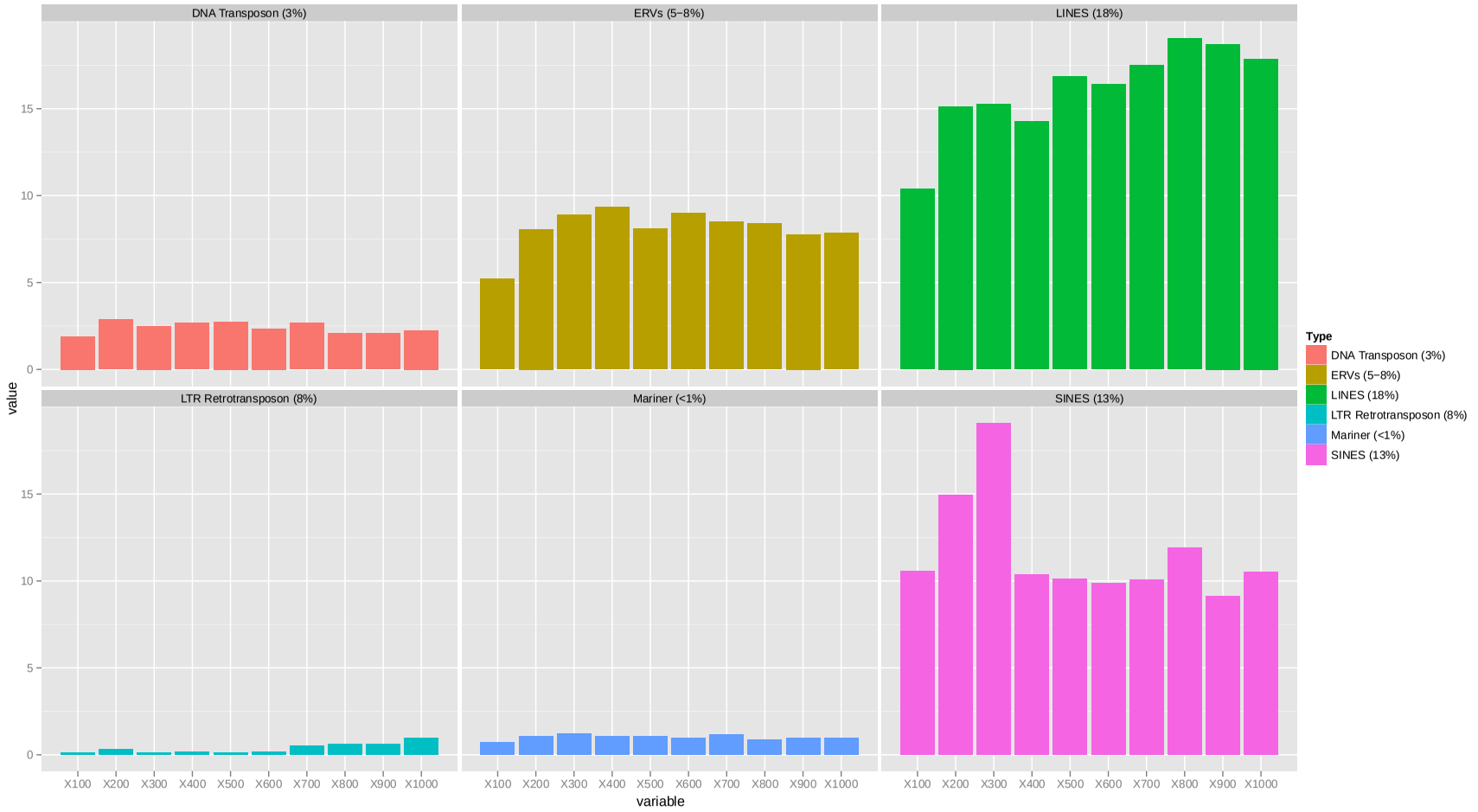

Figure 4.28: Percentage abundance of the different transposable elements detected within the specific flanking regions surrounding the 620 NUMTs using EBI CENSOR. The x axis of each individual graph represents a moving window of 100bp incrementing away from the NUMT insertion sites. Each graph represents the specific repeatable elements detected and contains the genome-wide average (%) in brackets within the key.

Isochore	GC Content
L1	$< 37\%$
L2	37-41%
H1	41-46%
H2	46-52%
H3	$> 52\%$

Table 4.6: Percentage GC content for each of the isochore families relating to the human genome

4.3.7 GC content of flanking regions

The GC content was analysed using the EMBOSS program geecee for 100Kb flanking regions for each NUMT insertion site. These were analysed in the form of a multiple FASTA file generating an output file listing each unique flank and the GC content for all 1240 (620 NUMTs x 2) flanking sequences. A spreadsheet was created and the results were ordered by descending GC percentage content. Following this the list was grouped into the five isochore families L1, L2, H1, H2 and H3. Table 4.6 displays each isochore family and its associated percentage GC content in the human genome. Using these values the percentage abundance of each isochore family regarding all flanking sequences could be determined and is displayed in Figure 4.29. The most abundant isochores were L2 (n=444, 35.81%) and H1 (n=375, 30.24%) followed by L1 (n=222, 17.9%) whereby the least abundant were H2 (n=175, 14.11%) and H3 (n=24, 1.94%).

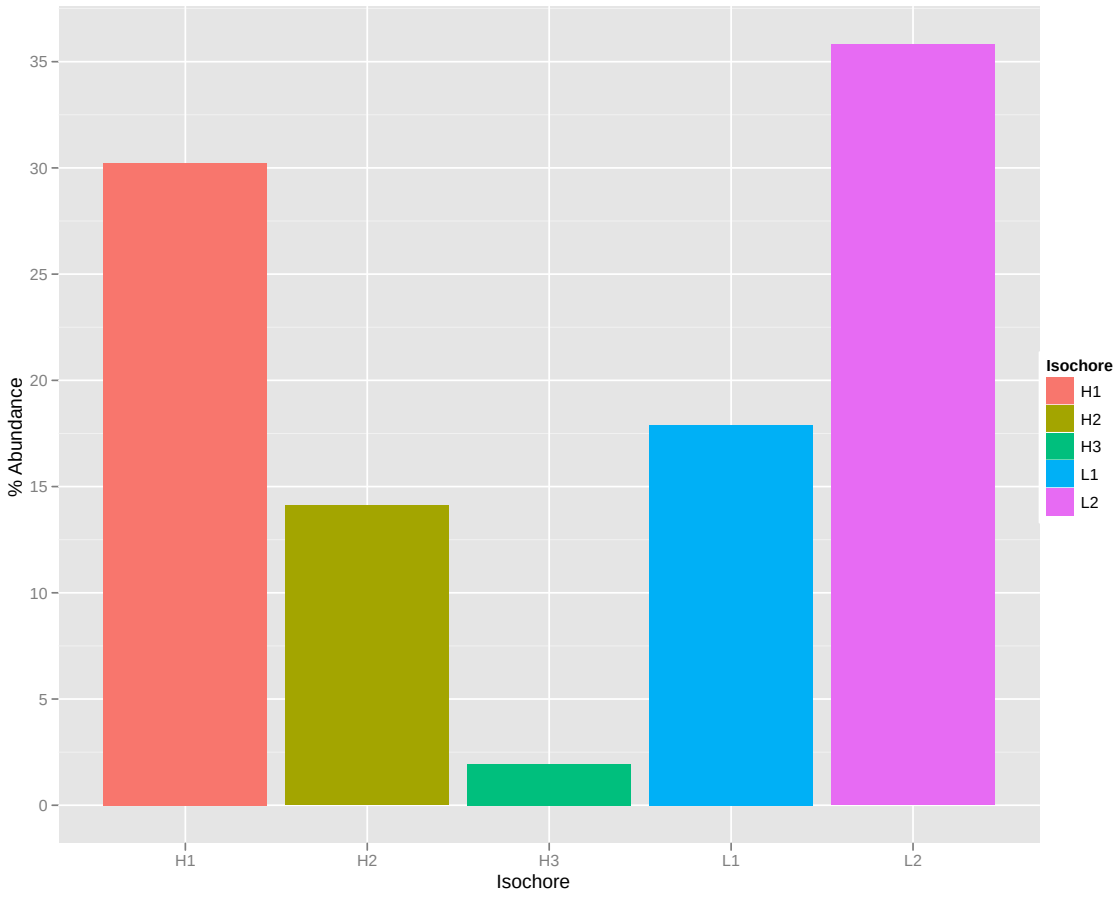

Figure 4.29: Percentage abundance of the five different isochore families (H1, H2, H3, L1 and L2) for all the flanking sequences surrounding the 620 detected NUMTs.

4.4 Discussion

4.4.1 Biological discussion

Numerous studies involving the detection of mtDNA insertions into the nuclear genome have been investigated. Various bioinformatic methods have been employed resulting in slight differences regarding the number of NUMTs for a particular genome and the computational methods implemented. This is especially relevant to studies involving human NUMT sequences as a significant range in results across the separate studies has been evident. This is clearly due to differences in methodology and sequence analysis in the identification of NUMT fragments. The protocol for determining NUMTs in this analysis ensured that any duplicate NUMTs were highlighted and only treated as separate insertions if their distance from each other in the nuclear genome was significant (i.e greater than 16,569bps, the length of the mitochondrial genome). Any overlapping fragments in the nuclear positions were treated as the same NUMT.

NUMT abundance across the human genome appears to be linear with chromosome size with minor exceptions, most significantly chromosome 2 having a considerably higher number of fragments than chromosome 1. This is based on evidence suggesting intronic and intergenic regions being larger in chromosome 2 (Sakharkar *et al.*, 2004). Chromosome 1 is only 6MB larger than chromosome 2 but has nearly 1000 additional genes. Chromosome 2 harbours larger intronic and intergenic space than chromosome 1 which could explain the higher frequency of NUMT integration in chromosome 2 when compared to chromosome 1. NUMTs appear to commonly integrate into intergenic regions which may reflect the possibility of a deleterious effect when integration occurs within intragenic regions of the genome (Woischnik and Moraes, 2002).

NUMT distribution

NUMT distribution is spread across the mitochondrial genome with no apparent preference for either the minor or major arc. This is a stark contrast to the distribution of the 263 mtDNA deletions that predominantly cluster within the major arc. Specific mitochondrial genes appear that have the highest NUMT formation are ND4, ND5, COI and COIII. Also the chicken genome appears to have preference for genes between ND4, CytB and the D loop (Pereira and Baker, 2004). Metabolism appears to have a positive correlation with small genomes exemplified by avian NUMT studies (Tiersch and Wachtel, 1991; Gregory, 2002) as birds with strong flight attributes tend to have smaller more streamlined genomes (Pereira and Baker, 2004). Pereira and Baker (2004) claims that plant and rice genomes are smaller than those of humans and other

mammals yet contain a higher level of NUMTs as they harbour the ability for bidirectional transfer of DNA between the nucleus, mitochondria and chloroplasts. However, the size of the mitochondrial genomes of (*A. thaliana* and *O. sativa*) are much larger representing sizes 30x and 22x bigger than the human mitochondrial genome, respectively (Richly and Leister, 2004). The human mitochondria alongside other similar sized genomes are evidence of a streamlining over evolutionary time where the majority of genes are encoded by the nucleus. In plants and rice the mitochondria are much larger and evidently more prone to bidirectional transfer of DNA reflecting an active symbiotic exchange of material (Kleine *et al.*, 2009). The mtDNA deletions are high for all genes apart from ND1 and ND2 that reside in the minor arc. The major arc contains two direct 13bp repeats that are believed to be associated with the majority of mtDNA deletions and are located at nucleotide positions 8470 and 13,447 (Samuels *et al.*, 2004).

Flanking region analysis

NUMTs were investigated for positional preference regarding areas of the nuclear genome for insertion. Flanking regions adjacent to the NUMT insertion sites were interrogated for annotated and predicted mitochondrial genes using the Ensembl and Entrez databases. Predicted mitochondrial genes were assessed using MitoCarta and MitoSVM. The flanking regions were separated into 50MB windows moving further away from the site of integration. When considering the Ensembl database there is a peak of 5.71% of mitochondrial genes at 100MB flanking the NUMTs significantly higher than the predicted genome-wide average. This considered both annotated mitochondrial genes in Ensembl and Entrez and predicted mitochondrial genes implementing MitoCarta and MitoSVM. However, this only serves as a small sample of the genome and would need a more holistic view to infer any strong correlation. A number of nuclear mitochondrial genes may have originated as mtDNA insertions as evidence suggests recent NUMT insertions modify exon-intron patterns in human genes (Noutsos *et al.*, 2007; Ricchetti *et al.*, 2004).

The dispersal of NUMTs post-insertion via a mode of transposition can not be assumed. It was therefore necessary to analyse the flanking regions for transposable elements to determine if this mechanism was apparent. NUMT flanking regions were assessed for repeat content including DNA transposons, ERVs, LINEs, LTR retrotransposons, mariners and SINES. These were evaluated against the repeat abundance across the entire human genome. The percentage of the different types of transposable element were quantified and compared against the average genome-wide level to highlight any elevation in concordance with the NUMT integration regions. No significant eleva-

tion of repeat content was found in any class of repeat with the exception of SINES. In flanking regions of 200bp and 300bp there was a higher incidence of SINES compared to the genome wide average. A deficit of LINES was evident up to 800bp from the NUMT insertion site. This suggests the majority of NUMTs are independent events and not duplicates via transposition but through chromosomal rearrangements. Evidence appears to corroborate findings by Gherman *et al.* (2007) and Jensen-Seaman *et al.* (2009) which both claimed a deficit of repeats in 500bp flanking sequences of NUMTs. Jensen-Seaman *et al.* (2009) did not use the Cambridge reference sequence for the BLASTN analysis therefore these results are potentially inaccurate. In contrast to these findings, Mishmar *et al.* (2004) found 59% of NUMTs were within 150bp of repetitive elements suggesting the vicinity of transposable elements influences the ongoing integration of mtDNA sequences and their duplication within the nuclear genome.

Following GC content analysis of the flanking sequences, isochores were determined for 100Kb flanking regions of the mtDNA insertion sites. This revealed that the gene-enriched isochore H3 has a very low percentage of NUMTs (1.94%) whereas H2 and L1 are significantly higher (14.11% and 17.9%, respectively). High abundance of NUMTs are found in L2 and H1 expressing much higher percentages of 35.81% and 30.24%, respectively. Areas of high GC content (H2 & H3) are gene enriched therefore suggesting NUMTs only successfully integrate into low GC areas where gene content is low. However, there appears to be a higher density of NUMTs in H1 as opposed to L1 suggesting an optimum GC percentage for NUMT integration. This may be due to selective constraints preventing foreign mitochondrial DNA disrupting functional genes (Woischnik and Moraes, 2002). Rare occurrences of NUMTs causing disease have been reported but most genic insertions are within introns (Lascaro *et al.*, 2008).

The majority of research investigating human NUMTs used the mitochondrial Cambridge reference sequence for their analyses. However, certain examples used the African Yoruba sequence that contains over 40 variant nucleotides which leads to inaccuracies. Jensen-Seaman *et al.* (2009) used this sequence in their BLASTN analysis generating precarious results. Results obtained from Jensen-Seaman *et al.* (2009) further confirmed findings by Gherman *et al.* (2007) providing supporting evidence for the lack of transposable elements flanking the NUMT insertion sites.

The honeybee and yellow fever mosquito contain the highest NUMT density per bp with the honeybee having the highest abundance. The *Apis mellifera* genome is high in AT content as opposed to gene-rich GC areas reflecting a potential correlation between low GC and mtDNA insertions. The yellow fever mosquito has high density in direct

contrast to the malarial mosquito as this has no evidence of NUMTs. The pufferfish has a relatively small and compact genome potentially with less intronic DNA which may be the reason for the lack of mtDNA insertions.

The migration of NUMTs may have peaked during a geological period that applied stressful conditions. Kleine *et al.* (2009) suggests the formation of DSBs is stress related and may result in an increase of mtDNA uptake. This may correlate with an evolutionary time period reflecting high NUMT occurrence in humans.

4.4.2 Technical discussion

Use of workflow technology and automation for performing large iterative sequence analysis is rapid, efficient and removes human error. R scripts can be invoked automatically to produce figures generated from large amounts of data. This is a unique and efficient method that can be repeated for any NUMT analysis allowing for extrapolation to any genome of interest. As genome sequencing increases due to high-throughput technology the application of rapid analysis pipelines to perform comparative assessments regarding NUMTs is needed. Workflows enable the automated extraction of sequences from the human genome and can exact further procedures assessing their flanking regions for transposable elements, GC content and mitochondrial-related gene abundance. BioMart can be interrogated as a service within Taverna thus enabling the rapid extraction of distributed biological data allowing *in silico* analysis to be performed.

Chapter 5

General Discussion

5.1 General Discussion

5.1.1 Systematic evaluation of mitochondrial protein prediction methods

Systematically identifying mitochondrial proteins has proved to be an arduous task. The wealth of bioinformatics approaches have been interrogated to maximise the efficiency in determining the mitochondrial proteome. Previous research has implied that increasing the number of independent datasets for a combined prediction increases sensitivity and specificity. However, this investigation provides strong evidence to rebuke this idea. Mitodomain proved to be the strongest prediction method among the 11 methods investigated whereas ancestry was shown to be the weakest. Specific combinations performed better than others containing more datasets increasing sensitivity without compromising specificity. However, an optimum was reached at 7 classifiers achieving the highest average sensitivity. The average sensitivity decreased when more classifiers were added. Particular bioinformatic methods were more successful than others, specifically ones based on neural networks, amino acid composition and pre-sequence determination. This research has revealed the importance of rigorously testing bioinformatic prediction methods and their performance in unison with other techniques. Implementing a support vector machine allowed each combination of datasets to be rigorously tested 100 times, each time using a different test not present in the training data. The standard deviation was consistently around 5% reflecting the variation a specific combination produced when subjected to multiple testing. Previous research conducted by Shen and Burger (2007); Pagliarini *et al.* (2008) performed 10-fold cross validation on their datasets (using 90% for training and reserving 10% for testing). However, Shen and Burger (2007) only performed the cross-validation procedure 10 times and Pagliarini *et al.* (2008) appeared to only perform this once. The MitoSVM workflow enabled the automation of large scale analysis in repetition strengthening the outcome of the sensitivities and specificities produced allowing for accurate standard deviation determination. Previous studies have failed to provide any evidence for the standard deviations produced when performing multiple testing of mitochondrial protein prediction methods. The variability reflected in the multiple testing shown by the standard deviation clearly shows the wide range of sensitivity and specificity values that are produced. Determining the most complementary combination of classifiers allows for the most accurate assessment of what proteins are contained within the mitochondrial proteome. This provides an invaluable process for biologists aiming to locate mitochondrial disease genes.

5.1.2 Identification of nuclear-mitochondrial genes involved in LHON

Following the completion of the MitoSVM database, this was applied alongside MitoCarta to produce a list of candidates genes involved in LHON. These databases were interrogated by ordering them by the chromosomal coordinates using the regions identified following linkage analysis. These both produced results containing several of the same genes with a number of differences. In addition, candidate gene analysis was applied to the entire X chromosome whereby all OMIM and gene ontology records were text mined for phenotypic correlations potentially related to LHON. This produced several candidates scoring several hits to specific keywords and phrases. 9 gene candidates were revealed that occurred in both MitoCarta and MitoSVM proving to be strong candidates for further investigation. Text mining was also applied to orthologues of mouse, rat and chimpanzee in an attempt to reveal novel disease candidates not annotated as such in humans. The mouse revealed a number of unique candidates but none were found in the rat or chimpanzee. Mouse orthologues Nlgn2, Pnck and Tfdp1 were revealed to have associations with eye-related phenotypes and crucially, the corresponding human genes had no evidence of any eye-related association. These candidates are therefore unique and require further investigation for investigations in LHON. Using a workflow to automate the text mining program against all X chromosome OMIM and UniProt records proved to be a very powerful method for mining literature. A manual approach would have been an unfeasible alternative. These candidates can been investigated further through sequencing techniques to assess their potential involvement with LHON. These methods provide new techniques in determining mitochondrial disease related genes that can be prioritised into an ordered list. This will allow biologists to focus on genes of interest based on various data pertaining to their predicted involvement in a specific disease.

5.1.3 Identification of Nuclear mitochondrial DNA sequences

NUMTs appear to commonly integrate into intergenic areas of low GC content and only very rarely disrupt genes. This was apparent from the low percentage of NUMTs in high GC% isochores that are gene rich areas of the genome. There was a high abundance of NUMTs in the low GC% isochores that are known to be gene poor regions. Little evidence was present to suggest transposable elements are responsible for the distribution of NUMT fragments within the human genome, with a small exception of SINES. However, further analysis would be required to confirm this. NUMT abundance varies wildly across species and even across closely related species exemplified by the African malar-

ial mosquito (*Anopheles gambiae*) and yellow fever mosquito (*Aedes aegypti*) ranging from none to hundreds, respectively. The mechanism of non-homologous end-joining of double-strand breaks is strongly supported by numerous research. This study aimed to compare mtDNA deletions and their point of origin within the mitochondrial genome to the formation of mitochondrial fragments that are subsequently incorporated into the nuclear genome. This potentially alludes to the same mechanism. Krishnan *et al.* (2008) claim that the majority of mtDNA deletions are caused by mitochondrial DNA repair of double-strand breaks in the mitochondrial genome, resulting in a wild type mitochondrial genome alongside a deleted genome. Mitochondrial fragments are believed to perform repair of double-strand breaks in the nuclear genome by a similar process of non-homologous end-joining. Areas of the nuclear genome prone to more double-strand breaks are likely to harbour NUMTs. Although the mechanisms are similar, mtDNA deletions occur frequently in individual and across different tissues within the same individual. However, the majority of human NUMTs occurred 55mya and likely reflect a major environmental change (Gherman *et al.*, 2007). In addition to searching for transposable elements the flanking regions surrounding the NUMTs were searched for a higher incidence of mitochondrial genes when compare to the genome-wide level. The evidence showed a slight increase regarding the abundance of mitochondrial genes when considering Ensembl genes in comparison to the overall expected abundance across the whole nuclear genome. However, this would require further analysis to confirm as significant. Utilising RShell within the Taverna workbench to incorporate R scripts in the analysis workflows enabled the automatic organisation, transposition and graphical display of mtDNA fragments. These were generated in relation to their point of origin in the mitochondrial genome in an efficient and timely experimental protocol. These images allowed for a very unique view of the distribution of both NUMTs and mtDNA deletions that would otherwise not have been possible implementing manual methods. These findings suggest NUMTs contribute to mostly to nuclear intergenic DNA and has considerable variation across species. They do not appear to reside in the neighbourhood of nuclear mitochondrial genes but more evidence is needed on a genome-wide scale. NUMT determination does not appear to be a reliable technique for finding novel nuclear mitochondrial genes.

5.2 Conclusions

MitoSVM provides a database of predicted mitochondrial proteins that has been rigorously and systematically determined by interrogating all combinations of specific datasets and validated using a high number of tests in comparison to previous methods. Standard deviations highlighted the variation between each result pertaining to a specific combination when tested multiple times. This analysis provides evidence to support the strength of prediction methods and the weaknesses harboured by others. This provides bioinformaticians with the insight to ensure the careful selection of computational methods when performing mitochondrial protein predictions.

The text mining procedure generated very interesting candidates that were not present in the mitochondrial prediction lists. Specific orthologues in the mouse proved to be strong candidates. These genes can been investigated further for their potential implications in LHON and the text mining workflow can be modified further to include a more extensive vocabulary. This technique is important for highlighting genes that may not be directly predicted as mitochondrial but indirectly associated through interaction networks.

Analysis of NUMTs can be greatly facilitated by the developments in whole genome sequencing with the rapid advancement of next generation sequencing technology. This can be applied to a much richer diversity of sequenced genomes with significantly higher depth of coverage. Following the completion of the Neanderthal nuclear genome, NUMT analysis can be applied potentially revealing any unique environmental stresses experienced by this group of ancient humans.

Various techniques have been applied to determine the mitochondrial proteome. Once this proteome is more defined, mitochondrial genetics can focus on specific diseases that are currently misunderstood. This will allow clinicians and mitochondrial geneticists to improve the lives of patients suffering from these rare diseases.

Bibliography

F Achard, G Vaysseix, and E Barillot. Xml, bioinformatics and data integration. *Bioinformatics*, 17(2):115–25, February 2001. ISSN 1367-4803.

S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, October 1990. ISSN 0022-2836.

S Anderson, A T Bankier, B G Barrell, M H de Bruijn, A R Coulson, J Drouin, I C Eperon, D P Nierlich, B A Roe, F Sanger, P H Schreier, A J Smith, R Staden, and I G Young. Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806):457–65, April 1981. ISSN 0028-0836.

R M Andrews, I Kubacka, P F Chinnery, R N Lightowlers, D M Turnbull, and N Howell. Reanalysis and revision of the cambridge reference sequence for human mitochondrial dna. *Nat Genet*, 23(2):147, October 1999. ISSN 1061-4036.

Agostinho Antunes and Maria João Ramos. Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes. *Genomics*, 86(6):708–17, December 2005. ISSN 0888-7543.

M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, May 2000. ISSN 1061-4036.

Khalid Belhajjame, Carole Goble, Franck Tanoh, Jiten Bhagat, Katherine Wolstencroft, Robert Stevens, Eric Nzuobontane, Hamish McWilliam, Thomas Laurent, and Rodrigo Lopez. Biocatalogue: A curated web service registry for the life science community. In *Microsoft eScience conference*, 2008.

Robert Belshaw, Vini Pereira, Aris Katzourakis, Gillian Talbot, Jan Paces, Austin Burt, and Michael Tristem. Long-term reinfection of the human genome by endogenous

retroviruses. *Proc Natl Acad Sci U S A*, 101(14):4894–9, April 2004. ISSN 0027-8424.

D Bensasson, D Zhang, D Hartl, and G Hewitt. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol Evol*, 16(6):314–321, June 2001. ISSN 0169-5347.

Nidhan K Biswas, Badal Dey, and Partha P Majumder. Using hapmap data: a cautionary note. *Eur J Hum Genet*, 15(2):246–9, February 2007. ISSN 1018-4813.

W C Black Iv and S A Bernhardt. Abundant nuclear copies of mitochondrial origin (numts) in the aedes aegypti genome. *Insect Mol Biol*, 18(6):705–13, November 2009. ISSN 1365-2583.

J L Blanchard and G W Schmidt. Mitochondrial dna migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. *Mol Biol Evol*, 13(3):537–48, March 1996. ISSN 0737-4038.

Mikael Bodén and John Hawkins. Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics*, 21(10):2279–86, May 2005. ISSN 1367-4803.

Keren Borensztajn, Ouerdia Chafa, Martine Alhenc-Gelas, Siham Salha, Abderrezak Reghis, Anne-Marie Fischer, and Jacqueline Tapon-Bretaudière. Characterization of two novel splice site mutations in human factor vii gene causing severe plasma factor vii deficiency and bleeding diathesis. *Br J Haematol*, 117(1):168–71, April 2002. ISSN 0007-1048.

James R Bradford and David R Westhead. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8):1487–94, April 2005. ISSN 1367-4803.

Marty C Brandon, Marie T Lott, Kevin Cuong Nguyen, Syawal Spolim, Shamkant B Navathe, Pierre Baldi, and Douglas C Wallace. Mitomap: a human mitochondrial genome database–2004 update. *Nucleic Acids Res*, 33(Database issue):D611–3, January 2005. ISSN 1362-4962.

M P Brown, W N Grundy, D Lin, N Cristianini, C W Sugnet, T S Furey, M Ares, and D Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*, 97(1):262–7, January 2000. ISSN 0027-8424.

Entela Bua, Jody Johnson, Allen Herbst, Bridget Delong, Debbie McKenzie, Shahriar Salamat, and Judd M Aiken. Mitochondrial dna-deletion mutations accumulate intracellularly to detrimental levels in aged human skeletal muscle fibers. *Am J Hum Genet*, 79(3):469–80, September 2006. ISSN 0002-9297.

Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998. URL `http://www.umiacs.umd.edu/ joseph/support-vector-machines4.pdf`.

Sarah Calvo, Mohit Jain, Xiaohui Xie, Sunil A Sheth, Betty Chang, Olga A Goldberger, Antonella Spinazzola, Massimo Zeviani, Steven A Carr, and Vamsi K Mootha. Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat Genet*, 38(5):576–82, May 2006. ISSN 1061-4036.

C L Campbell and P E Thorsness. Escape of mitochondrial dna to the nucleus in yme1 yeast is mediated by vacuolar-dependent turnover of abnormal mitochondrial compartments. *J Cell Sci*, 111 (Pt 16):2455–64, August 1998. ISSN 0021-9533.

Yiqun Cao, Tao Jiang, and Thomas Girke. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics*, 24(13): i366–74, July 2008. ISSN 1460-2059.

E Capriotti, R Calabrese, and R Casadio. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, 22(22):2729–34, November 2006. ISSN 1460-2059.

Jian-Min Chen, Nadia Chuzhanova, Peter D Stenson, Claude Férec, and David N Cooper. Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. *Hum Mutat*, 25(2):207–21, February 2005. ISSN 1098-1004.

Patrick F Chinnery. Searching for nuclear-mitochondrial genes. *Trends Genet*, 19(2): 60–2, February 2003. ISSN 0168-9525.

M G Claros and P Vincens. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur J Biochem*, 241(3):779–86, November 1996. ISSN 0014-2956.

Dawn Cotter, Purnima Guda, Eoin Fahy, and Shankar Subramaniam. Mitoproteome: mitochondrial protein sequence database and annotation system. *Nucleic Acids Res*, 32(Database issue):D463–7, January 2004. ISSN 1362-4962.

L Dagnino, C J Fry, S M Bartley, P Farnham, B L Gallie, and R A Phillips. Expression patterns of the e2f family of transcription factors during mouse nervous system development. *Mech Dev*, 66(1-2):13–25, August 1997. ISSN 0925-4773.

David De Roure and Carole Goble. Lessons from myexperiment: Research objects for data intensive research. In *eScience Workshop 2009*, Pittsburgh, US, October 2009. URL `http://eprints.ecs.soton.ac.uk/17744/`.

David De Roure, Carole Goble, Sergejs Aleksejevs, Sean Bechhofer, Jiten Bhagat, Don Cruickshank, Danius Michaelides, and David Newman. The myexperiment open repository for scientific workflows. In *Open Repositories 2009*, Atlanta, Georgia, USA, May 2009. URL `http://eprints.ecs.soton.ac.uk/17131/`.

S DiMauro and E A Schon. Mitochondrial dna mutations in human disease. *Am J Med Genet*, 106(1):18–26, 2001. ISSN 0148-7299.

Salvatore Dimauro and Guido Davidzon. Mitochondrial dna and disease. *Ann Med*, 37 (3):222–32, 2005. ISSN 0785-3890.

Andreas Doms and Michael Schroeder. Gopubmed: exploring pubmed with the gene ontology. *Nucleic Acids Res*, 33(Web Server issue):W783–6, July 2005. ISSN 1362-4962.

M Elstner, C Andreoli, T Klopstock, T Meitinger, and H Prokisch. The mitochondrial proteome database: Mitop2. *Methods Enzymol*, 457:3–20, 2009. ISSN 1557-7988.

O Emanuelsson, H Nielsen, S Brunak, and G von Heijne. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *J Mol Biol*, 300 (4):1005–16, July 2000. ISSN 0022-2836.

Paul Fisher, Harry Noyes, Stephen Kemp, Robert Stevens, and Andrew Brass. A systematic strategy for the discovery of candidate genes responsible for phenotypic variation. *Methods Mol Biol*, 573:329–45, 2009. ISSN 1940-6029.

Toni Gabaldón and Martijn A Huynen. Shaping the mitochondrial proteome. *Biochim Biophys Acta*, 1659(2-3):212–20, December 2004. ISSN 0006-3002.

Adrian Gherman, Peter E Chen, Tanya M Teslovich, Pawel Stankiewicz, Marjorie Withers, Carl S Kashuk, Aravinda Chakravarti, James R Lupski, David J Cutler, and Nicholas Katsanis. Population bottlenecks as a potential major shaping force of human genome architecture. *PLoS Genet*, 3(7):e119, July 2007. ISSN 1553-7404.

M Gilbert, J Smith, A J Roskams, and V J Auld. Neuroligin 3 is a vertebrate gliotactin expressed in the olfactory ensheathing glia, a growth-promoting class of macroglia. *Glia*, 34(3):151–64, May 2001. ISSN 0894-1491.

Carole Goble and Robert Stevens. State of the nation in data integration for bioinformatics. *J Biomed Inform*, 41(5):687–93, October 2008. ISSN 1532-0480.

Ehud Goldin, Stefanie Stahl, Adele M Cooney, Christine R Kaneski, Surya Gupta, Roscoe O Brady, James R Ellis, and Raphael Schiffmann. Transfer of a mitochondrial dna fragment to mcoln1 causes an inherited case of mucolipidosis iv. *Hum Mutat*, 24 (6):460–5, December 2004. ISSN 1098-1004.

T Ryan Gregory. A bird's-eye view of the c-value enigma: genome size, cell size, and metabolic rate in the class aves. *Evolution*, 56(1):121–30, January 2002. ISSN 0014-3820.

Chittibabu Guda. ptarget: a web server for predicting protein subcellular localization. *Nucleic Acids Res*, 34(Web Server issue):W210–3, July 2006. ISSN 1362-4962.

Chittibabu Guda, Eoin Fahy, and Shankar Subramaniam. Mitopred: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics*, 20 (11):1785–94, July 2004. ISSN 1367-4803.

Einat Hazkani-Covo, Raymond M Zeller, and William Martin. Molecular poltergeists: mitochondrial dna copies (numts) in sequenced nuclear genomes. *PLoS Genet*, 6(2): e1000834, 2010. ISSN 1553-7404.

T Hirokawa, S Boon-Chieng, and S Mitaku. Sosui: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, 14(4):378–9, 1998. ISSN 1367-4803.

Masamitsu Honma, Mayumi Sakuraba, Tomoko Koizumi, Yoshio Takashima, Hiroko Sakamoto, and Makoto Hayashi. Non-homologous end-joining for repairing i-scei-induced dna double strand breaks in human cells. *DNA Repair (Amst)*, 6(6):781–8, June 2007. ISSN 1568-7864.

N Howell. Leber hereditary optic neuropathy: respiratory chain dysfunction and degeneration of the optic nerve. *Vision Res*, 38(10):1495–504, May 1998. ISSN 0042-6989.

S Hua and Z Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–8, August 2001. ISSN 1367-4803.

Gavin Hudson, Sharon Keers, Patrick Yu Wai Man, Philip Griffiths, Kirsi Huoponen, Marja-Liisa Savontaus, Eeva Nikoskelainen, Massimo Zeviani, Franco Carrara, Rita Horvath, Veronika Karcagi, Liesbeth Spruijt, I F M de Coo, Hubert J M Smeets, and Patrick F Chinnery. Identification of an x-chromosomal locus and haplotype modulating the phenotype of a mitochondrial dna disorder. *Am J Hum Genet*, 77 (6):1086–91, December 2005. ISSN 0002-9297.

Duncan Hull, Katy Wolstencroft, Robert Stevens, Carole Goble, Mathew R Pocock, Peter Li, and Tom Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res*, 34(Web Server issue):W729–32, July 2006. ISSN 1362-4962.

M I Jensen-Seaman, J H Wildschutte, I D Soto-Calderón, and N M Anthony. A comparative approach shows differences in patterns of numt insertion during hominoid evolution. *J Mol Evol*, 68(6):688–99, June 2009. ISSN 1432-1432.

J Jurka, V V Kapitonov, A Pavlicek, P Klonowski, O Kohany, and J Walichiewicz. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, 110(1-4):462–7, 2005. ISSN 1424-859X.

Lukas Käll, Anders Krogh, and Erik L L Sonnhammer. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*, 338(5):1027–36, May 2004. ISSN 0022-2836.

Matthew Anthony Kirkman, Alex Korsten, Miriam Leonhardt, Konstantin Dimitriadis, Ireneaus F De Coo, Thomas Klopstock, Philip G Griffiths, Gavin Hudson, Patrick F Chinnery, and Patrick Yu-Wai-Man. Quality of life in patients with leber hereditary optic neuropathy. *Invest Ophthalmol Vis Sci*, 50(7):3112–5, July 2009a. ISSN 1552-5783.

Matthew Anthony Kirkman, Patrick Yu-Wai-Man, Alex Korsten, Miriam Leonhardt, Konstantin Dimitriadis, Ireneaus F De Coo, Thomas Klopstock, and Patrick Francis Chinnery. Gene-environment interactions in leber hereditary optic neuropathy. *Brain*, 132(Pt 9):2317–26, September 2009b. ISSN 1460-2156.

Tatjana Kleine, Uwe G Maier, and Dario Leister. Dna transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annu Rev Plant Biol*, 60:115–38, 2009. ISSN 1545-2123.

Oleksiy Kohany, Andrew J Gentles, Lukasz Hankus, and Jerzy Jurka. Annotation, submission and screening of repetitive elements in repbase: Repbasesubmitter and censor. *BMC Bioinformatics*, 7:474, 2006. ISSN 1471-2105.

Jacob Köhler, Jan Baumbach, Jan Taubert, Michael Specht, Andre Skusa, Alexander Rüegg, Chris Rawlings, Paul Verrier, and Stephan Philippi. Graph-based analysis and visualization of experimental results with ondex. *Bioinformatics*, 22(11):1383–90, June 2006. ISSN 1367-4803.

Eugene V Koonin. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, 39:309–38, 2005. ISSN 0066-4197.

Kim J Krishnan, Amy K Reeve, David C Samuels, Patrick F Chinnery, John K Blackwood, Robert W Taylor, Sjoerd Wanrooij, Johannes N Spelbrink, Robert N Lightowlers, and Doug M Turnbull. What causes mitochondrial dna deletions in human cells? *Nat Genet*, 40(3):275–9, March 2008. ISSN 1546-1718.

A Krogh, B Larsson, G von Heijne, and E L Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol*, 305(3):567–80, January 2001. ISSN 0022-2836.

C G Kurland and S G Andersson. Origin and evolution of the mitochondrial proteome. *Microbiol Mol Biol Rev*, 64(4):786–820, December 2000. ISSN 1092-2172.

Pedro Larrañaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Iñaki Inza, José A Lozano, Rubén Armañanzas, Guzmán Santafé, Aritz Pérez, and Victor Robles. Machine learning in bioinformatics. *Brief Bioinform*, 7(1):86–112, March 2006. ISSN 1467-5463.

Daniela Lascaro, Stefano Castellana, Giuseppe Gasparre, Giovanni Romeo, Cecilia Saccone, and Marcella Attimonelli. The rhnumts compilation: features and bioinformatics approaches to locate and quantify human numts. *BMC Genomics*, 9:267, 2008. ISSN 1471-2164.

Leonard A Levin. Mechanisms of retinal ganglion specific-cell death in leber hereditary optic neuropathy. *Trans Am Ophthalmol Soc*, 105:379–91, 2007. ISSN 1545-6110.

Peter Li, Claire Jennings, Kate Owen, Thomas Oinn, Robert Stevens, Simon Pearce, and Anil Wipat. Association of variations in i kappa b-epsilon with graves' disease using classical and mygrid methodologies. In *All Hands Meeting*, September 2003.

Peter Li, Juan I Castrillo, Giles Velarde, Ingo Wassink, Stian Soiland-Reyes, Stuart Owen, David Withers, Tom Oinn, Matthew R Pocock, Carole A Goble, Stephen G Oliver, and Douglas B Kell. Performing statistical analyses on quantitative data in taverna workflows: an example using r and maxdbrowse to identify differentially-expressed genes from microarray data. *BMC Bioinformatics*, 9:334, 2008. ISSN 1471-2105.

Michael R Lieber, Yunmei Ma, Ulrich Pannicke, and Klaus Schwarz. The mechanism of vertebrate nonhomologous dna end joining and its role in v(d)j recombination. *DNA Repair (Amst)*, 3(8-9):817–26, 2004. ISSN 1568-7864.

C.X. Ling, W.S. Noble, and Qiang Yang. Guest editors' introduction to the special issue: Machine learning for bioinformatics-part 1. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 2(2):81–82, april-june 2005. ISSN 1545-5963. doi: 10.1109/TCBB.2005.25.

J V Lopez, N Yuhki, R Masuda, W Modi, and S J O'Brien. Numt, a recent transfer and tandem amplification of mitochondrial dna to the nuclear genome of the domestic cat. *J Mol Evol*, 39(2):174–90, August 1994. ISSN 0022-2844.

Z Lu, D Szafron, R Greiner, P Lu, D S Wishart, B Poulin, J Anvik, C Macdonell, and R Eisner. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, 20(4):547–56, March 2004. ISSN 1367-4803.

P Y W Man, D M Turnbull, and P F Chinnery. Leber hereditary optic neuropathy. *J Med Genet*, 39(3):162–9, March 2002. ISSN 1468-6244.

W Martínez-López, G A Folle, G Obe, and P Jeppesen. Chromosome regions enriched in hyperacetylated histone h4 are preferred sites for endonuclease- and radiation-induced breakpoints. *Chromosome Res*, 9(1):69–75, 2001. ISSN 0967-3849.

Eugene M McCarthy and John F McDonald. Long terminal repeat retrotransposons of mus musculus. *Genome Biol*, 5(3):R14, 2004. ISSN 1465-6914.

Hamish McWilliam, Franck Valentin, Mickael Goujon, Weizhong Li, Menaka Narayanasamy, Jenny Martin, Teresa Miyar, and Rodrigo Lopez. Web services at the european bioinformatics institute-2009. *Nucleic Acids Res*, 37(Web Server issue): W6–10, July 2009. ISSN 1362-4962.

Iain Melvin, Eugene Ie, Rui Kuang, Jason Weston, William Noble Stafford, and Christina Leslie. Svm-fold: a tool for discriminative multi-class protein fold and superfamily recognition. *BMC Bioinformatics*, 8 Suppl 4:S2, 2007. ISSN 1471-2105.

Dan Mishmar, Eduardo Ruiz-Pesini, Martin Brandon, and Douglas C Wallace. Mitochondrial dna-like sequences in the nucleus (numts): insights into our african origins and the mechanism of foreign dna integration. *Hum Mutat*, 23(2):125–33, February 2004. ISSN 1098-1004.

K Nakai and P Horton. Psort: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*, 24(1):34–6, January 1999. ISSN 0968-0004.

William S Noble. What is a support vector machine? *Nat Biotechnol*, 24(12):1565–7, December 2006. ISSN 1087-0156.

Christos Noutsos, Tatjana Kleine, Ute Armbruster, Giovanni DalCorso, and Dario Leister. Nuclear insertions of organellar dna can create novel patches of functional exon sequences. *Trends Genet*, 23(12):597–601, December 2007. ISSN 0168-9525.

J M Nugent and J D Palmer. Rna-mediated transfer of the gene coxii from the mitochondrion to the nucleus during flowering plant evolution. *Cell*, 66(3):473–81, August 1991. ISSN 0092-8674.

T. Oinn, M. J. Addis, J. Ferris, D. J. Marvin, M. Greenwood, T. Carver, A. Wipat, and P. Li. Taverna, lessons in creating a workflow environment for the life sciences, 2004a. URL `http://www.citebase.org/abstract?id=oai:eprints.ecs.soton.ac.uk:9250`.

Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R Pocock, Anil Wipat, and Peter Li. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–54, November 2004b. ISSN 1367-4803.

Arzucan Ozgür, Thuy Vu, Günes Erkan, and Dragomir R Radev. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24(13):i277–85, July 2008. ISSN 1460-2059.

John K Pace and Cédric Feschotte. The evolutionary history of human dna transposons: evidence for intense activity in the primate lineage. *Genome Res*, 17(4):422–32, April 2007. ISSN 1088-9051.

David J Pagliarini, Sarah E Calvo, Betty Chang, Sunil A Sheth, Scott B Vafai, Shao-En Ong, Geoffrey A Walford, Canny Sugiana, Avihu Boneh, William K Chen, David E Hill, Marc Vidal, James G Evans, David R Thorburn, Steven A Carr, and Vamsi K Mootha. A mitochondrial protein compendium elucidates complex i disease biology. *Cell*, 134(1):112–23, July 2008. ISSN 1097-4172.

Pekka Pamilo, Lumi Viljakainen, and Anu Vihavainen. Exceptionally high density of numts in the honeybee genome. *Mol Biol Evol*, 24(6):1340–6, June 2007. ISSN 0737-4038.

B Parfait, A Percheron, D Chretien, P Rustin, A Munnich, and A Rötig. No mitochondrial cytochrome oxidase (cox) gene mutations in 18 cases of cox deficiency. *Hum Genet*, 101(2):247–50, December 1997. ISSN 0340-6717.

Ryan L Parr, Jennifer Maki, Brian Reguly, Gabriel D Dakubo, Andrea Aguirre, Roy Wittock, Kerry Robinson, John P Jakupciak, and Robert E Thayer. The pseudomitochondrial genome influences mistakes in heteroplasmy interpretation. *BMC Genomics*, 7:185, 2006. ISSN 1471-2164.

Sérgio L Pereira and Allan J Baker. Low number of mitochondrial pseudogenes in the chicken (gallus gallus) nuclear genome: implications for molecular inference of population history and phylogenetics. *BMC Evol Biol*, 4:17, June 2004. ISSN 1471-2148.

Vittoria Petruzzella and Sergio Papa. Mutations in human nuclear genes encoding for subunits of mitochondrial respiratory complex i: the ndufs4 gene. *Gene*, 286(1): 149–54, March 2002. ISSN 0378-1119.

S Pettifer, D Thorne, P McDermott, T Attwood, J Baran, J C Bryne, T Hupponen, D Mowbray, and G Vriend. An active registry for bioinformatics web services. *Bioinformatics*, 25(16):2090–1, August 2009. ISSN 1367-4811.

Anu Puomila, Petra Hämäläinen, Sanna Kivioja, Marja-Liisa Savontaus, Satu Koivumäki, Kirsi Huoponen, and Eeva Nikoskelainen. Epidemiology and penetrance of leber hereditary optic neuropathy in finland. *Eur J Hum Genet*, 15(10):1079–89, October 2007. ISSN 1018-4813.

M Ricchetti, C Fairhead, and B Dujon. Mitochondrial dna repairs double-strand breaks in yeast chromosomes. *Nature*, 402(6757):96–100, November 1999. ISSN 0028-0836.

Miria Ricchetti, Fredj Tekaia, and Bernard Dujon. Continued colonization of the human genome by mitochondrial dna. *PLoS Biol*, 2(9):E273, September 2004. ISSN 1545-7885.

P Rice, I Longden, and A Bleasby. Emboss: the european molecular biology open software suite. *Trends Genet*, 16(6):276–7, June 2000. ISSN 0168-9525.

Erik Richly and Dario Leister. Numts in sequenced eukaryotic genomes. *Mol Biol Evol*, 21(6):1081–4, June 2004. ISSN 0737-4038.

Meena Kishore Sakharkar, Vincent T K Chow, and Pandjassarame Kangueane. Distributions of exons and introns in the human genome. *In Silico Biol*, 4(4):387–93, 2004. ISSN 1386-6338.

David C Samuels, Eric A Schon, and Patrick F Chinnery. Two direct repeats cause most human mtdna deletions. *Trends Genet*, 20(9):393–8, September 2004. ISSN 0168-9525.

Andrew M Schaefer, Robert W Taylor, Douglass M Turnbull, and Patrick F Chinnery. The epidemiology of mitochondrial disorders–past, present and future. *Biochim Biophys Acta*, 1659(2-3):115–20, December 2004. ISSN 0006-3002.

Anthony H V Schapira. Mitochondrial disease. *Lancet*, 368(9529):70–82, July 2006. ISSN 1474-547X.

G Schatz. The protein import system of mitochondria. *J Biol Chem*, 271(50):31763–6, December 1996. ISSN 0021-9258.

M Sciacco, E Bonilla, E A Schon, S DiMauro, and C T Moraes. Distribution of wild-type and common deletion forms of mtdna in normal and respiration-deficient muscle fibers from patients with mitochondrial myopathy. *Hum Mol Genet*, 3(1):13–9, January 1994. ISSN 0964-6906.

Martin Senger, Peter Rice, and Thomas Oinn. Soaplab - a unified sesame door to analysis tools. In *All Hands Meeting*, September 2003. URL `http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/115.pdf`.

Suma P Shankar, John H Fingert, Valerio Carelli, Maria L Valentino, Terri M King, Stephen P Daiger, Solange R Salomao, Adriana Berezovsky, Rubens Belfort, Terri A Braun, Val C Sheffield, Alfredo A Sadun, and Edwin M Stone. Evidence for a novel x-linked modifier locus for leber hereditary optic neuropathy. *Ophthalmic Genet*, 29 (1):17–24, March 2008. ISSN 1744-5094.

Hagit Shatkay, Annette Höglund, Scott Brady, Torsten Blum, Pierre Dönnes, and Oliver Kohlbacher. Sherloc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics*, 23(11):1410–7, June 2007. ISSN 1460-2059.

Yao Qing Shen and Gertraud Burger. 'unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools. *BMC Bioinformatics*, 8:420, 2007. ISSN 1471-2105.

J M Shoffner, M T Lott, A S Voljavec, S A Soueidan, D A Costigan, and D C Wallace. Spontaneous kearns-sayre/chronic external ophthalmoplegia plus syndrome associated with a mitochondrial dna deletion: a slip-replication model and metabolic therapy. *Proc Natl Acad Sci U S A*, 86(20):7952–6, October 1989. ISSN 0027-8424.

Ian Small, Nemo Peeters, Fabrice Legeai, and Claire Lurin. Predotar: A tool for rapidly screening proteomes for n-terminal targeting sequences. *Proteomics*, 4(6):1581–90, June 2004. ISSN 1615-9853.

Damian Smedley, Syed Haider, Benoit Ballester, Richard Holland, Darin London, Gudmundur Thorisson, and Arek Kasprzyk. Biomart–biological queries made easy. *BMC Genomics*, 10:22, 2009. ISSN 1471-2164.

Sören Sonnenburg, Alexander Zien, Petra Philips, and Gunnar Rätsch. Poims: positional oligomer importance matrices–understanding support vector machine-based signal detectors. *Bioinformatics*, 24(13):i6–14, July 2008. ISSN 1460-2059.

R D Stevens, H J Tipney, C J Wroe, T M Oinn, M Senger, P W Lord, C A Goble, A Brass, and M Tassabehji. Exploring williams-beuren syndrome using mygrid. *Bioinformatics*, 20 Suppl 1:i303–10, August 2004. ISSN 1460-2059.

Robert D Stevens, Alan J Robinson, and Carole A Goble. mygrid: personalised bioinformatics on the information grid. *Bioinformatics*, 19 Suppl 1:i302–4, 2003. ISSN 1367-4803.

Heinz Stockinger, Teresa Attwood, Shahid Nadeem Chohan, Richard Cot'e, Philippe Cudr'e-Mauroux, Laurent Falquet, Pedro Fernandes, Robert D Finn, Taavi Hupponen, Eija Korpelainen, Alberto Labarga, Aurelie Laugraud, Tania Lima, Evangelos Pafilis, Marco Pagni, Steve Pettifer, Isabelle Phan, and Nazim Rahman. Experience using web services for biological sequence analysis. *Brief Bioinform*, 9(6):493–505, November 2008. ISSN 1477-4054.

Martin Szomszor, Terry Payne, and Luc Moreau. Using semantic web technology to automate data integration in grid and web service architectures. In *IEEE International Symposium on Cluster Computing and the Grid*, pages 189–195, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0780390741. doi: 10.1109/CCGRID.2005.1558553.

Kumarasamy Thangaraj, Manjunath B Joshi, Alla G Reddy, Avinash A Rasalkar, and Lalji Singh. Sperm mitochondrial mutations as a cause of low sperm motility. *J Androl*, 24(3):388–92, 2003. ISSN 0196-3635.

P E Thorsness and E R Weber. Escape and migration of nucleic acids between chloroplasts, mitochondria, and the nucleus. *Int Rev Cytol*, 165:207–34, 1996. ISSN 0074-7696.

T R Tiersch and S S Wachtel. On the evolution of genome size of birds. *J Hered*, 82 (5):363–8, 1991. ISSN 0022-1503.

V Tiranti, M Munaro, D Sandonà, E Lamantea, M Rimoldi, S DiDonato, R Bisson, and M Zeviani. Nuclear dna origin of cytochrome c oxidase deficiency in leigh's syndrome: genetic evidence based on patient's-derived rho degrees transformants. *Hum Mol Genet*, 4(11):2017–23, November 1995. ISSN 0964-6906.

Clesson Turner, Christina Killoran, Nick S T Thomas, Marjorie Rosenberg, Nadia A Chuzhanova, Jennifer Johnston, Yelena Kemel, David N Cooper, and Leslie G Biesecker. Human genetic disease caused by de novo mitochondrial-nuclear dna transfer. *Hum Genet*, 112(3):303–9, March 2003. ISSN 0340-6717.

G E Tusnády and I Simon. The hmmtop transmembrane topology prediction server. *Bioinformatics*, 17(9):849–50, September 2001. ISSN 1367-4803.

T Ueda, H Sakagami, K Abe, I Oishi, A Maruo, H Kondo, T Terashima, M Ichihashi, H Yamamura, and Y Minami. Distribution and intracellular localization of a mouse homologue of ca2+/calmodulin-dependent protein kinase ibeta2 in the nervous system. *J Neurochem*, 73(5):2119–29, November 1999. ISSN 0022-3042.

S Urbanek. Rserve – a fast way to provide r functionality to applications. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Vienna, Austria, May 2003.

L van den Heuvel and J Smeitink. The oxidative phosphorylation (oxphos) system: nuclear genes and human genetic diseases. *Bioessays*, 23(6):518–25, June 2001. ISSN 0265-9247.

V N Vapnik. An overview of statistical learning theory. *IEEE Trans Neural Netw*, 10 (5):988–99, 1999. ISSN 1045-9227.

Byrappa Venkatesh, Nidhi Dandona, and Sydney Brenner. Fugu genome does not contain mitochondrial pseudogenes. *Genomics*, 87(2):307–10, February 2006. ISSN 0888-7543.

J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea,

A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. *Science*, 291(5507):1304–51, February 2001. ISSN 0036-8075.

D C Wallace, X X Zheng, M T Lott, J M Shoffner, J A Hodge, R I Kelley, C M Epstein, and L C Hopkins. Familial mitochondrial encephalomyopathy (merrf): genetic, pathophysiological, and biochemical characterization of a mitochondrial dna disease. *Cell*, 55(4):601–10, November 1988. ISSN 0092-8674.

Ingo Wassink, Han Rauwerda, Pieter Bt Neerincx, Paul E van der Vet, Timo M Breit,

Jack Am Leunissen, and Anton Nijholt. Using r in taverna: Rshell v1.2. *BMC Res Notes*, 2:138, 2009. ISSN 1756-0500.

J L Weber and E W Myers. Human whole-genome shotgun sequencing. *Genome Res*, 7(5):401–9, May 1997. ISSN 1088-9051.

Markus Woischnik and Carlos T Moraes. Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome Res*, 12(6):885–93, June 2002. ISSN 1088-9051.

K Wolstencroft, P Alper, D Hull, C Wroe, P W Lord, R D Stevens, and C A Goble. The (my)grid ontology: bioinformatics service discovery. *Int J Bioinform Res Appl*, 3(3):303–25, 2007. ISSN 1744-5485.

Chris Wroe, Carole Goble, Mark Greenwood, Phillip Lord, Simon Miles, Juri Papay, Terry Payne, and Luc Moreau. Automating experiments using semantic data on a bioinformatics grid. *GRID.IEEE INTELLIGENT SYSTEMS*, 19:1, 2004. URL `http://www.citebase.org/abstract?id=oai:eprints.ecs.soton.ac.uk:9482`.

Y-G Yao, Q-P Kong, A Salas, and H-J Bandelt. Pseudomitochondrial genome haunts disease studies. *J Med Genet*, 45(12):769–72, December 2008. ISSN 1468-6244.

Chin-Sheng Yu, Yu-Ching Chen, Chih-Hao Lu, and Jenn-Kang Hwang. Prediction of protein subcellular localization. *Proteins*, 64(3):643–51, August 2006. ISSN 1097-0134.

P Yu-Wai-Man, P G Griffiths, G Hudson, and P F Chinnery. Inherited mitochondrial optic neuropathies. *J Med Genet*, 46(3):145–58, March 2009. ISSN 1468-6244.

M Zeviani. The expanding spectrum of nuclear gene mutations in mitochondrial disorders. *Semin Cell Dev Biol*, 12(6):407–16, December 2001. ISSN 1084-9521.

Massimo Zeviani, Antonella Spinazzola, and Valerio Carelli. Nuclear genes in mitochondrial disorders. *Curr Opin Genet Dev*, 13(3):262–70, June 2003. ISSN 0959-437X.

H Zischler, M Höss, O Handt, A von Haeseler, A C van der Kuyl, and J Goudsmit. Detecting dinosaur dna. *Science*, 268(5214):1192–3; author reply 1194, May 1995. ISSN 0036-8075.

Appendix

The Appendix containing all the analysis spreadsheets, data collection, Java and R code are all available on the accompanying CD.