Content Evaluation and Tool Development for Knowledge Management Systems

By
Kabita Shakya
January 2013

A dissertation submitted in fulfilment of the requirement for the degree of

Masters of Science by Research

Dublin City University
Dublin 9
Ireland

Supervisors:
Prof. Heather J. Ruskin
Dr. Mary J. O'Connell

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of M. Sc. is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: ___________

(Candidate) ID No.: ___________

Date: _______

Abstract

Genetic and epigenetic mechanisms play vital roles in the initiation and progression of cancer. The motivation of the work reported here is thus to support research in this area, by investigating genetic and epigenetic mechanisms and the inter-relationships between them through provision of a platform (in-house biomedical resource (*StatEpigen*)) for data collation and analysis. *StatEpigen* is targeted initially to collating information on colon cancer and the basic aim of this project is to enhance, evaluate and ensure *robustness* of this resource.

Elements involved in building towards a more comprehensive 'picture of needs' to date include: a comparative study of available epigenetic/epigenomic biomedical resources, manual augmentation of *StatEpigen* database resource and an in-depth analysis of a set of *germline* mutated colon cancer genes from the phylogenetic perspective, to link resource provision to the experimental base and address key bioinformatics questions.

Comparative study has confirmed the current importance of epigenetic studies and provided information on resources that may offer integration potential for *StatEpigen*. Manual data augmentation (15% contribution to the current datasource, URL: http://statepigen.sci-sym.dcu.ie/) permitted assessment of the data curation process itself, and also motivation and planning for some degree of future automation. The in-depth genetic analysis addressed a specific-research question relating to the suitability of the murine model as a reference organism for *colon cancer* in humans. Analysis of the mouse parallel (following 180 *MY* of independent evolution) revealed that some genes can not be used as suitable cancer model for humans. This finding provided stimulus for developed analyses (*e.g.* through *StatEpigen*) of related epigenetic characteristics and genetic-epigenetic interactions that are influential in the initiation and progression of the disease.

A future focus for *StatEpigen* includes exploitation of the data already gathered, as well as tool development for automation of the data augmentation process.

Acknowledgements

I wish to give my foremost thanks to my supervisors Prof. Heather J. Ruskin and Dr. Mary J. O'Connell who provided guidance throughout each stage of this project.

I would also like to thank DCU for its award of an O'Hare scholarship to provide financial support for this project and the Irish Centre for High-End Computing (ICHEC) for processor time and technical support. The Sci-Sym Centre (ModSci and Molecular Evolution groups) is also warmly acknowledged for the supportive working environment and technical facilities.

Kabita Shakya
7th January 2013

Table of Contents

List of Tables

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Figures

Chapter2

Chapter 4

Chapter5

Acronyms

BEB,	- Bayes empirical bayes
DDBJ	- DNA Data Bank of Japan
dN,	- Nonsynonymous substitutions per nonsynonymous site
dS,	- Synonymous substitutions per synonymous sites
EBI	- European Bioinformatics Institute
EMBL	- European Molecular Biology Laboratory
ENCODE	- Encyclopaedia of DNA Elements
EPITRON	- EPIgenetic TReatment of Neoplastic Disease
FAP	- Familial Adenomatous Polyposis
HEP	- Human Epigenome Project
HEROIC	- High-Throughput Epigenetic Regulatory Organization In Chromatin
HGP	- Human Genome Project
HNPCC	- Hereditary nonpolyposis colorectal cancer
ICGC	- International Cancer Genome Consortium
IHEC	- International Human Epigenome Consortium
LRT	- Likelihood ratio test
ML	- Maximum likelihood
MRCA	- Most Recent Common Ancester
MY	- Millions of years
NCBI	- National Center for Biotechnology Information
Ne	- Effective population size
NEB	- Naïve empirical Bayes
NHGRI	- National Human Genome Research Institute
PDB	- Protein Data Bank
PP	- Posterior probability
UHN	- University Health Network

Chapter 1

Introduction

1.1 Background to Epigenetics

In any assessment of data resource requirements, a clear overview of current availability is required. This is true also for the field of 'epigenetics', which has attracted considerable attention over the last decade (especially after the completion of Human Genome Project) (Human Genome Management Information System, 2013).

Epigenetics is concerned with the study of heritable and reversible changes in gene expression or cellular phenotype that occur without corresponding alteration in the DNA sequence – "*epi*" from the Greek meaning over (–genetics). Epigenetic changes occur due to mechanisms such as DNA methylation, histone modification (methylation/ demethylation, acetylation/ deacetylation, phosphorylation, ubiquitylation and sumoylation) and small regulatory RNA changes (Goldberg *et al.* 2007, Fernberg 2004). Gene silencing, genomic imprinting, X-chromosome inactivation are processes that utilize epigenetic processes.

DNA methylation is addition of methyl group (–CH_3) to the 5 position of the cytosine pyrimidine ring or the number 6 nitrogen of the adenine purine ring (cytosine and adenine are two of the four bases of DNA). In most of the cases, addition of methyl marks leads to repression of gene activity. In mammals, methylation is found sparsely but globally, distributed in definite CpG sequences throughout the entire genome, with the exception of CpG islands (CpG-rich regions of around 1kb that represent approximately 1% of genome, and often located around the promoters of housekeeping genes). DNA methylation patterns that extend beyond CpG methylation are also observed in different animals (Bird 2002, Lister *et al.* 2009).

Histone (protein components of chromatin around which DNA can wind for compaction and gene regulation) *modification* occurs when the binding of epigenetic factors to histone ‘tails’ alter the extent of wrapping of DNA around histones and thus allows the gene/s (in the DNA) to be switched on or off. Molecules that bind to histones include methyl, acetyl, phosphate groups and ubiquitin and sumo proteins (Esteller 2011). The attachment of these molecules to histones alters DNA activity in different ways. DNA methylation and histone modifications (explained in more detail in Chapter 3) are two of the best understood mechanisms of epigenetic modifications to date. *Genomic imprinting* is also an important epigenetic phenomenon whereby inherited genes are ‘imprinted’ due to one copy of the gene being epigenetically marked or imprinted in either the egg or the sperm. Thus the allelic expression of an imprinted gene depends on whether it is inherited maternally or paternally. Imprinted expression can also vary between tissues, developmental stages and species (Reik and Walter 2001). *Gene silencing* is the “switching off” of a gene that can be triggered by any of the epigenetic events mentioned earlier and is an epigenetic process of gene regulation.

To summarize, epigenetic events have the potential to reprogramme genomes without any genetic modification. The term “epigenetics” was coined by Conrad Waddington in early 1940s to describe ‘environmental effects on the phenotype’ but the molecular comprehension was delayed for around half a century (House 2010). The role of epigenetics in diseases is highlighted by genetically identical (monozygotic) twins that can develop distinctly different disease profiles and life projections later in their lives (also called twin discordance phenomenon, Fraga *et al.* 2005, Baranzin *et al.* 2010). In recent years, ‘epigenetics’ has grown into one of the most dynamic areas of biological research.

Epigenetic studies are also new and core to the investigation of human development, moving focus from the way in which genes and gene products bring the phenotype into being, towards a focus on the mechanism through which cells become committed to a particular structure or function and the way in which these features are transmitted in cell lineages. Epigenetics is thus important not only for its practical applications, *e.g.*, in cancer treatment, but also for gaining a greater understanding of the way in which heredity and evolution occur, since non- DNA inheritance systems may be involved

(Jablonka and Lamb 2002). In terms of involvement in the etiology[1] of human disease, the role of epigenetics is increasingly recognised (*e.g.* Jiang *et al.* 2004, Feinberg *et al.* (2004 and 2006) and more recently, Rodenheiser and Mann (2006), Steinmann *et al.* 2009 and many others), these are described in detail in the following section.

Epigenetics in Human Diseases

Many studies have been conducted to explore the role of epigenetics in the occurrence of different diseases or conditions. Genetic, epigenetic and environmental factors are likely to contribute to the pathogenesis of the diseases. Some diseases or conditions for which knowledge of epigenetic factors are acquired are: cancer (Esteller 2011), auto immune disorders (Quintero-Ronderos and Montoya-Ortiz 2012), schizophrenia (Deng 2010), autism (Grafodatskaya 2010) and aging (Susana 2010), where DNA methylation and histone modification are the predominant epigenetic mechanisms involved. Similarly, Prader Willi Syndrome (when genes derived from father) /Angelman Syndrome (when same genes derived from mother) (Adams 2008); Beckwith-Wiedeman syndrome (BWS: typically characterised by excessive growth caused by *under-expression* of a growth-suppressing gene and *over-expression* of a growth-promoting gene, both found on chromosome 11) are found to involve genomic imprinting phenomenon (Reik and Walter 2001). Additionally, other diseases/conditions, *e.g.* diabetes, neurobiological disorders and cardiovascular diseases, stress-related disorders (Post traumatic stress disorder), muscular degeneration and many more are also studied currently for possible epigenetic links.

Epigenetic Inheritance

Since epigenetic events are heritable, transgenerational epigenetic inheritance can be studied (similar to Lamarck's original evolutionary theories, Lamarck 1914). Indeed, studies have recently emerged in epigenetics research that examine (i) epigenetic inheritance in descendants (even up to a number of generations) resulting from epigenetic changes in parents (Eaxinger and Whitelaw 2010), as well as (ii) testing the evolutionary implications of epigenetic inheritance (Jablonka and Lamb 1995) and (iii) the analyses of environmental factors and their impact on epigenetic events. The thrifty phenotype hypothesis (Hales and Barker 1992) that says that the reduced fetal growth in

[1] Study of causes of disease (Footnotes occur as superscripted numbers throughout the thesis)

limited supply of nutrients is associated with a number of chronic conditions later in life has motivated thrifty epigenotype hypothesis (Stöger 2008).

1.2 Animal Reference Models

Animal models are in use for different human diseases, to understand the etiology and pathogenesis of diseases – they provide a vital cornerstone for modern medicine. Currently animal models are also used for epigenetic studies and transgenerational epigenetic study, as described in more detail in the following section. The diseases in mouse and human might not be entirely equivocal, at the phenotypic, epigenetic or genotypic level. It would be desirable to test for the existence of measurable differences between model organisms before using them as the 'disease models'.

Animal Reference Models for diseases

For different diseases (*e.g.* cancer, COPD, pneumonia, asthma, pulmonary hypertension, lung disease, rheumatoid arthritis) animals are used as references. The most common animal used in human disease studies is the mouse. The major advantages of using mouse are that, it is small in size (ease of handling), has a rapid reproductive rate/ short life span (possibility of studying several generations over a fairly short period of time), and has a high level of genetic similarity to human (78% identity at the protein coding level, Waterston *et al.* (2002)). While the disadvantages include the differences between mouse strains - studies suggest that more than 1 strain of mouse is used per study to control for strain variation (Rivera and Tessarollo 2008). Some diseases in mouse offer a completely different phenotype than human *e.g.* Asthma and fibrosis (Rivera and Tessarollo 2008). Other animals used as models for human include chimpanzee for modelling hepatitis C, monkeys for modelling polio and guinea pig for modelling diabetes and tuberculosis (Rivera and Tessarollo 2008).

Animal Reference Models for epigenetic studies

For epigenetic studies (particularly maternal nutrition based epigenetic studies), rodents, and specifically mouse, are commonly used. For environmental epigenetic studies and analyses of the phenotypic outcome, (*e.g.* how maternal diet can alter epigenetic and accompanying phenotypic response in the offspring), Agouti viable yellow (A^{vy}) and axin 1 fused ($Axin1^{Fu}$) mice are used most notably. These mice have a well-characterized locus whose methylation pattern governs dramatic phenotypic outcomes which makes them suitable for such study (Rosenfeld 2010). However, other strains of rats and mice, sheep, and Japanese macaques are also used to study the epigenetic effect of maternal diet on offspring. Similarly, for transgenerational animal studies, again Agouti viable yellow (A^{vy}) and axin 1 fused ($Axin1^{Fu}$) mice, rats and other strains of mice are also used. For methyl-deficient dietary animal studies, different strains of rats and mice are used (Rosenfeld 2010).

1.3 Cancer Focus

Cancer as a set of disease types - classification

Cancer (or neoplasm) is a generic term for group of over 100 diseases, characterized by abnormal, uncontrolled cell growth. It is the leading cause of death worldwide, resulting in over seven million fatalities each year (Ferlay *et al.* 2010). At advanced stage, cancer cells can spread to other parts of body/organs through blood and lymph systems. This process is termed ‘metastasis’ (and is a fatal stage).

Cancers are named for the organ or type of cell in which they start, *e.g.* cancer beginning in colon is called colon cancer. However, cancer can be broadly classified[2] as:

Carcinoma: Cancer that begins in skin or in tissues that line/cover internal organs.

Sarcoma: Cancer that begins in bone, cartilage, fat, muscle, blood vessels, or other connective or supportive tissue.

Leukaemia: Cancer that starts in blood-forming tissue such as the bone marrow and causes large numbers of abnormal blood cells to be produced and enter the blood stream.

Lymphoma and myeloma: Cancers that begin in the cells of the immune system.

Central nervous system cancers: Cancers that begin in the tissues of the brain and spinal cord.

The word tumour is used as a synonym for cancer sometimes, but benign tumours are not cancerous, this is true only for malignant tumours. Lung, breast, prostrate, colon, stomach, liver cancers are the most common forms of cancers. Regardless of the tissue of origin, cancer essentially arises from a single cell. The transformation from a single normal cell into a tumour cell is a multistage process, typically a progression from a

[2] Source of information : www.cancer.gov

pre-cancerous lesion to malignant tumours. The cause can be attributed to interaction between genetic factors and different carcinogenic factors, namely physical carcinogens (*e.g.* ultraviolet and ionizing radiation), chemical carcinogens (*e.g.* asbestos, components of tobacco smoke, aflatoxin[3], arsenic), and biological carcinogens (*e.g.* infection from certain pathogens). The carcinogens might have some role to play in alterations of the genetic and epigenetics variety. Thus it can be said that, at a molecular level, the genetic and epigenetic factors and the interaction between them can cause cancer. Many current studies back the genetic-epigenetic interaction for the etiology of cancer (Sawan *et al.* 2008, You and Jones 2012). Aging is also considered as one of the causes of cancer. More detail on the interaction between genetic and epigenetic factors is given in Chapter 3.

Colon cancer focus

Colon cancer (also known as colorectal cancer, CRC or bowel cancer, characterized by, rectal bleeding and anaemia that sometimes is associated with weight loss and changes in bowel habits), is the third most common cancer in males and second in females (Ferley *et al.* 2008). Because of its high incidence, there is a lot of research on this form of cancer and there is ample data available.

To study the incidence and pathogenesis of cancer from available data resources, colon cancer was selected as the starting point and is the focus of this thesis. From this work future research can expand into additional cancer types. The aim of this work is to use the genetic (mutation events) and epigenetic data and the interactions between them from existing literature. Specifically, the statistical relationship between the genetic and epigenetic events by curating the information, and collecting it in a suitable data resource (*StatEpigen Knowledge Management System* (Barat and Ruskin 2010)) so that it can be fed into different modelling software being developed in our research group. This will allow the modelling software to use real patient data to check the validity of the model itself. The database can also be used by other researchers as it is made available via the web portal http://statepigen.sci-sym.dcu.ie/, one example of how this data can be used is outlined in Chapter 4 and has been published (Morgan *et al.* 2012). *StatEpigen Knowledge Management System* is our in-house biomedical database

[3] A food contaminant

resource and is currently specific to colon cancer genetic and its associated epigenetic factors. This database intends to incorporate other cancer types in future.

1.4 Knowledge Mining

Since knowledge mining (data mining) is an important part of the knowledge discovery process, the genetic and epigenetic information are nowadays made available in database resources. More details on this are in Chapter 2. For in-house *StatEpigen* database as well, in addition of using the *StatEpigen* data in modelling software, it can also be used for data mining and valuable information can be extracted from the process. The results from this analysis will however be more accurate, with more information/data in the database. More is given on this and is in Chapter 5.

Parallel genetic resource development

The database can benefit from the links to /incorporation of addition of genetic resources (*e.g.* COSMIC, Wellcome trust). Information gathered from parallel research efforts on genetics, which can be linked to genetic and epigenetic events and their interactions, can be added continuously. Additionally, genetic and epigenetic data gathered for animal models of human diseases (*e.g.* cancer) can also be drawn upon.

Key Requirements

The efficient augmentation of the *StatEpigen* database is the key requirement for now. This will enable improved accuracy for data mining analyses based on the data as well as enable precise evaluation of cancer models. The manual augmentation process is currently implemented which is time consuming, although accurate. This needs to be automated / semi- automated in the near future for increased efficiency.

1.5 Thesis scope – Aims and Achievements

This thesis comprises five chapters. Chapter 1 provides the background information in the form of a general introduction to the field while Chapter 3 is focused on cancer and its causative agents, namely genetic and epigenetic factors, at a molecular level.

In general, the objective of this research is to understand the role of genetic and epigenetic factors and the interactions between them. To be specific, the aim is to streamline (via data curation of scientific literature and augment it manually in

StatEpigen database) the knowledge that we have at present in scattered form (in the form of literature from various fields), and make it usable to extract valuable information (via data mining). The database is made available to the scientific community via webportal http://statepigen.sci-sym.dcu.ie/ and is currently also being used to feed different modelling software being developed in the computing lab (which enable testing of the software with real patient data).

The data curation process currently is manual and during this project, 15% of the current data content (genes 728, mutations 150, and epigenetic events 318) was contributed to the database. As the project aims for an upgrade of the current system in future, some integrable database resources are also explored along with automatic source provision for *StatEpigen KMS,* (Chapter 2). Furthermore, the current availability of epigenetic resources (published in Shakya *et al.* 2012) in addition to genetic resources are also examined while assessing the data resource requirement. A framework for upgrading the current system is also proposed (explained with a schematic diagram and a flowchart), which is available in Chapter 5. Some results from an exploratory data analysis are also provided in this chapter. Besides this, genetic analysis of colon cancer genes is also carried out from phylogenetic perspective (in Chapter 4). This analysis identified differences between human and mouse and presented us with 17 genes that are positively selected (with respect to site and lineage specific analyses, Morgan *et al.* 2012) which can be further explored for epigenetic events occurring at specific sites with *StatEpigen* data. Given the importance in cancer of genetic - epigenetic interactions and their synergies, such analyses contribute both to current knowledge and future discovery.

Chapter 2

Biomedical, Epigenetic Resources and *StatEpigen*

2.1 Introduction

In this chapter, we present a comparative survey of available biomedical resources including in particular, those recording epigenetic information. Cancer-related epigenetic resources are investigated, including specialized options, which include an in-house specific database resource *StatEpigen* (Barat and Ruskin 2010), focussed on colon cancer in its initial inception. Some content of this chapter, (Section 2.4), has appeared as a recent publication, (Shakya *et al.* 2012)[4].

Completed in 2003, the Human Genome Project (HGP, a 13 year project), led to the identification of more than 20,000 genes and determined the 3 billion chemical base pairs of human DNA. In the past decade, the tremendous advances in medical technologies, corresponding development in computer power, storage capacity, inter-connectivity and cost effectiveness, has led to an explosive growth in generation and collection of all aspects of biomedical data, allied with the importance of Bioinformatics as a field (Choudhary *et al.* 2011). Data warehousing (Marakas 2003), as a way of dealing with large dataset size, combines databases across an entire enterprise, whereas independent or federated systems seek to integrate multiple autonomous databases into a single federation, with constituent databases interconnected via a network and often geographically decentralised, (Heimbigner and McLeod 1985, Sheth and Larson 1990, Devlin 2012). An example includes many bioinformatics data sources linked by the *Entrez Life Sciences* search engine.[5]

[4] In Epigenetics Volume 7, Issue 9.

[5] Managed by the U.S. National Library of Medicine, NIH, Maryland, USA .

From patient records to information from pharmaceutical studies, specific disease research, and different 'omics' studies, biomedical data cover a wide range of data. Well-known examples of the latter include genomics, proteomics and transcriptomics. Resource types can be classified by two key features; firstly, the means or method by which access is provided to entities; secondly, the nature of the entities themselves. The repository or web service that provides access to these data is a vital component of biomedical data resourcing (Tenenbaum *et al.* 2011). In general, resource providers, (such as *PubMeth* and *MutationDB*) review research papers from the domain and mine these for information relevant to the scientific audience. Typically, non-profit research institutes, such as the Sanger Institute, University of California Santa Cruz (UCSC), National Center for Biotechnology Information (NCBI) – National Institute of Health (NIH), European Molecular Biology Laboratory (EMBL), European Bioinformatics Institute (EBI) and so on, make such data publicly available over the internet for analysis.

Biological/biomedical databases themselves are a *primary* biomedical resource, providing data directly on biological quantities, and these co-exist alongside *secondary* and *composite* resources. Examples of *primary* databases include those containing information on sequence or structure alone, *e.g.* SwissProt, PIR (protein sequences), Genbank and DDBJ (genome sequences). *Secondary* resources contain derived information from *primary* sources; examples include eMOTIF (Stanford), SCOP (Cambridge) while *composite* resources typically draw information from a variety of different databases, such as those of the NCBI genome browser. Here, an overview of the principal biomedical resources is provided according to the data mining approach, the resource type and scale.

As outlined in Chapter 1 and, in particular, following completion of the Human Genome Project, increased attention has been paid to those processes that lead to heritable changes in gene expression, (during development or across generations), without altering the nucleotide sequence within the DNA. Both *epigenetics* and *epigenomics*, (the genome-wide distribution of epigenetic changes), have become major areas of research focus. Principal epigenetic phenomena encompass DNA methylation, histone modification (acetylation/deacetylation, methylation/demethylation, phosphorylation, ubiquitylation and sumoylation) while processes utilizing epigenetics include gene silencing, genomic imprinting, X-chromosome inactivation (also called lyonization), bookmarking. Recently-launched large-scale initiatives include IHEC (the International Human Epigenome Consortium)[6], with plans to map up to 1,000 reference epigenomes within a decade, and the Human Epigenome Project (HEP)[7], that aims to identify, catalogue and interpret genome-wide DNA methylation patterns of all human genes in all major tissues (Novik *et al.* 2002), amongst others. In summarising publicly-available biomedical resources here, the particular focus is thus on those that deal with epigenomic/epigenetic research. In addition, we assess the data mining capabilities, intrinsic to or accessed by these resources, and comment on their adequacy.

2.2 Knowledge Discovery and Data Mining

Data mining (broadly the application of algorithms for extracting patterns from data), is a major step in the so-called Knowledge Discovery from Data (KDD) process. Pre and post-processing of data and pattern extraction are typical steps (Fayyad *et al.* 1996). The KDD process is interactive and iterative, involving numerous steps as illustrated in Figure 2.1. The challenges include both social and technical factors, (different hardware, software platforms and automation), (Dasu and Johnson 2003), together with acquisition of metadata, and establishment of quality metrics.

[6] Paris, France (February 2010)

[7] Multinational project (Wellcome Trust Sanger: UK, Epigenomics AG: Germany/ USA, The Centre National de Genotypage: France), http://www.epigenome.org/

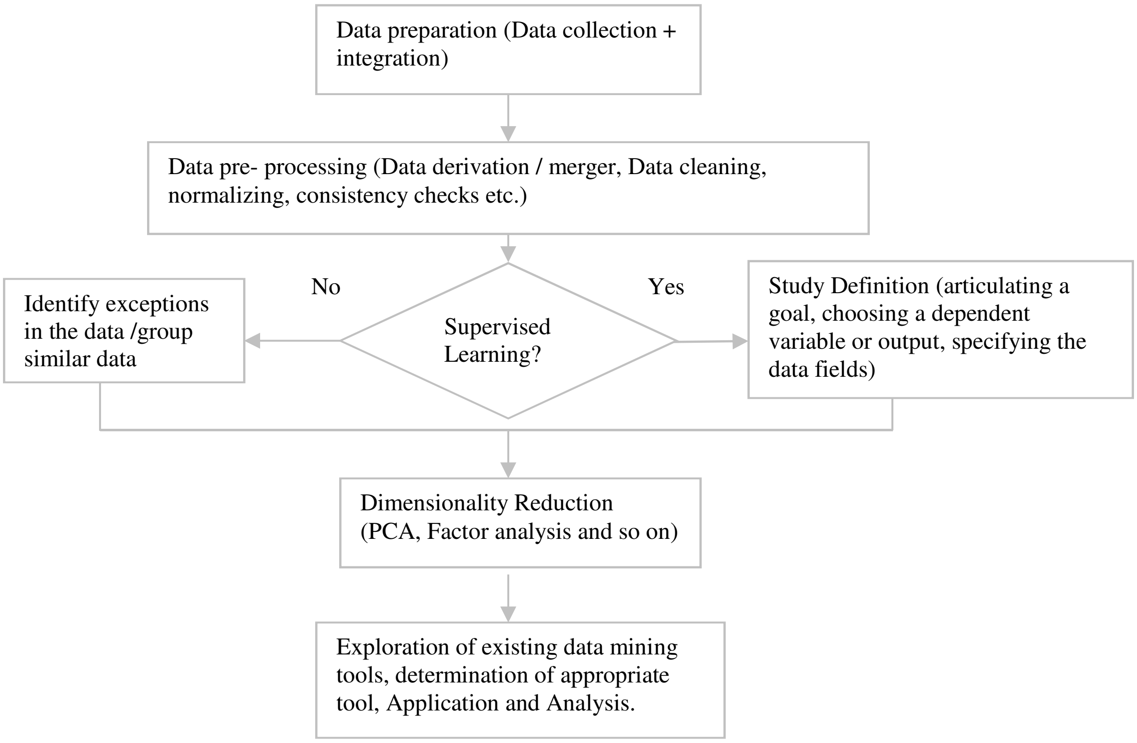
Data preparation (Data collection + integration)
Data pre- processing (Data derivation / merger, Data cleaning, normalizing, consistency checks etc.)
No
Yes
Supervised Learning?
Identify exceptions in the data /group similar data
Study Definition (articulating a goal, choosing a dependent variable or output, specifying the data fields)
Dimensionality Reduction (PCA, Factor analysis and so on)
Exploration of existing data mining tools, determination of appropriate tool, Application and Analysis.

Figure 2.1: Flowchart illustrating the knowledge discovery process

Data mining can be categorised into predictive (supervised learning), and descriptive (unsupervised learning), methods, covering a range of subtypes (Breiman *et al.* 1984, Groth and Lewis 1998 and Kerr *et al.* 2008). Semi-supervised learning incorporates a combination of both and is an important current direction for research in this area.

2.2.1 Biomedical Resource Groups: Data Mining

Attempting to present an overview of the spectrum of biomedical resources and corresponding analysis capabilities is far from trivial. As illustration, PubMed is the U.S. National Library of Medicine’s (NLM) web-based interface to MEDLINE, the premier bibliographic index to journal articles in the Life Sciences. MEDLINE includes citations from over 4600 of the worlds leading biomedical journals from 1966 to present, with even older citations (back to 1951, OLDMEDLINE). A typical classification of biomedical resources is to identify parameters or measures targeted and the analyses offered, but as new resources or extended provision come online, this needs to be continually revisited. Broad categories are highlighted below.

Text Mining

Text mining is heavily dependent on diverse knowledge sources, in part because of the complex nomenclature of the biomedical sciences. The two most important controlled vocabularies currently are MeSH[8] (Medical Subject Headings) and GO (Gene Ontology)[9]. MeSH, a list of standardised vocabulary used by NLM[10], is one of the vocabularies in the UMLS[11] Metathesaurus, a large, multipurpose and multi lingual vocabulary database that contains information on biomedical concepts. MeSH contains a hierarchical set of controlled vocabulary items taken from medicine, chemistry and genomics and now also includes the names of substances with pharmacological action. There are close to 300,000 items in MeSH and annotations also provide a rich set of metadata, describing articles, as well as a map of the entire biomedical field. GO consists of three sub-vocabularies and describes gene products in terms of their associated biological processes, cellular components, and molecular functions. PubMed articles are indexed using these vocabulary items, which also helps to find articles, relevant to the user's search, but using different terms for the same concepts.

Data Integration and Web Mining

Another major category of resource provision looks at integration and web mining of data. The UMLS (Unified Medical Language System)[12] is a set of files and software that bring together many health and biomedical vocabularies and standards to enable *interoperability* between computer systems. UMLS can be used to enhance or develop applications, such as electronic health records, classification tools, dictionaries and language translators. The purpose is the development of computer systems, which 'understand' the meaning of biomedical language (Lindberg *et al.* 1993 and more recently FactSheet, UMLS). The three tools used by UMLS are (i) Metathesaurus, (ii) Semantic Network and (iii) SPECIALIST Lexicon.

A number of other recent initiatives in data integration also exist. One is the semantic data integration environment that is part of the Biomedical Informatics Research Network (BIRN) project (Astakhov *et al.* 2005). Here, each participating institution maintains storage of their experimental or computationally-derived data while a

[8] http://www.nlm.nih.gov/mesh/MBrowser.html
[9] http://www.geneontology.org/
[10] National Library of Medicine (U.S.)
[11] Unified Medical Language System
[12] http://www.nlm.nih.gov/research/umls/

mediator-based data integration system performs semantic integration across the databases. This enables researchers to perform analysis based on larger and broader datasets than would be available from any single institution's data. The aim is a *cyber infrastructure* for biomedical research that supports advance data acquisition, storage, management, integration, mining visualisation and other computing and information processing services over the Internet. Healthgrid technology also is an emerging trend, where these grid infrastructures comprise applications, services or middleware components that deal with the specific data processing problems. This technology is an integration of bio-medical knowledge, imaging, computational tools, and other technologies in diagnosis and treatment (Healthgrid 2004).

Improvements can also be made in web searching for biomedical information. Currently, although the web provides easier and more comprehensive access to information than do physical environments, it cannot provide an accurate and efficient response to user requirements, since machines are not able to interpret and contextualise information in the form of natural language. *Web mining* is expected to improve efficient access to information, substituting traditional methods of content-based and collaborative recommendation systems, with the new trend in using technologies such as intelligent software agents, (Ferber 1999), fuzzy linguistic techniques (Herrera and Herrera-Viedma 1998, Morales-del-Castillo *et al.* 2010), and semantic web technologies (Berners-Lee *et al.* 2001, Lassila and Hendler 2007). A filtering and recommendation system that applies a fuzzy linguistic approach, based on semantic web technologies, to identify key user information has been reported in (Morales-del-Castillo *et al.* 2010), and facilitates retrieval of information from large repositories as required.

In addition to the diverse set of biomedical resources, generated by the research community itself, there are now an increasing number of shared electronic resources. The Biomedical Resource Ontology (BRO) has been developed to enable semantic resource annotation in the context of discovery on the Internet, (Tenenbaum *et al.* 2011). Bioportal, a web repository for biomedical ontologies and data resources offers researchers and clinicians access to the web for important biomedical ontologies, (Noy *et al.* 2009), while biomedical/scientific web communities are also being built by different groups to share targeted resources. Examples include Alzforum[13], (which has

[13] www.alzforum.org

over 4600 networked researchers seeking the cure for Alzheimer's), and the Schizophrenia Forum[14], amongst others. The caBIG® (Cancer Biomedical Information Grid)[15], is a similar information network, enabling all constituencies in the cancer community - researchers, physicians, and patients - to share data and knowledge. The aim of caBIG is to provide a framework for creating, communicating, and sharing bioinformatics tools, data and research results, while using shared applications, data standards and data models. Similarly, a number of biological resources are available now as part of the W3C Resource Description Framework (RDF) triples. Gene ontology (GOs), ChEBI (Chemical Entities of Biological Interest) and SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) are well-known ontology examples.

Semantic Wikis (Anmueller 2005, Oren 2006), represent another type of effort to develop collaborative annotation and knowledge management systems. Gene Wiki, (Huss *et al.* 2008), WikiProteins (Mons *et al.* 2008), and BOWiki are just some of the examples. The Science Collaboration Framework (SCF), (Das *et al.* 2008), (based on a semantically-aware content management system), aims to leverage existing knowledge repositories, using annotated information available on the semantic web.

2.3 Genetic/ Genomic Resources in Brief

These resources are categorized into gateways, genetic databases and mutation databases for ease of classification as follows. For our genetic analysis work, (Chapter 4), Ensembl genomes (Section 2.3.1) and UniProtKB/Swiss-Prot (protein sequence) databases were extensively used. The link between hereditary (for *e.g.* HNPCC, Chapter 4) as well as sporadic diseases (*e.g.* sporadic cancer caused by mutation, triggered by carcinogen, *i.e.* epigenetic event in molecular form) to genetics and epigenetics as evident by current researches (Chapter 3), links all of these, at some level.

2.3.1 Gateways

[14] www.schizophreniaforum.org

[15] https://cabig.nci.nih.gov/

Currently, the most widely-used genome browsers are: Ensembl, NCBI Map Viewer and UCSC[16] although others are also available. Ensembl is a European initiative while NCBI and UCSC are U.S. based. These genome browsers act as gateways to metadatabase connections.

Ensembl Genomes[17], automatically annotates the genome, integrates this annotation with other available biological data and makes this publicly available. The database is for vertebrates and other eukaryotic species. The latest (12th) release[18] brings the total genomes available to 335. Ensembl is a major joint project between EMBL - EBI and the Wellcome Trust Sanger Institute, targeting development of a software system to produce and maintain automatic annotation of selected eukaryotic genomes. The latest release also continues to display a joint gene set based on the merger achieved between the automatic annotation from Ensembl and the manually curated annotation from HAVANA (Human and Vertebrate Analysis and Annotation) group at Sanger. The CCDS (Consensus Coding Sequence, Pruitt *et al.* 2009) is a consensus project between the European Bioinformatics Institute (EBI), the Wellcome Trust Sanger Institute (WTSI), the National Center for Biotechnology Information (NCBI), and the University of California at Santa Cruz (UCSC) and tracks identical protein annotations on the reference mouse and human genomes with a stable identifier (CCDS ID), ensuring that these are consistently represented on the NCBI, Ensembl, and UCSC Genome Browsers[19]. Ensembl is powered by Biomart[20], a query-oriented data management system (Hubbard *et al.* 2005). Biomart can be used with any data type and is particularly suited for providing ‘data-mining’ type searches of complex descriptive data. It has powered many other databases, including those of ICGC (International Cancer Genome Consortium), HGNC (HUGO Gene Nomenclature Committee), the Gene Expression Atlas, COSMIC, Intogen (Integrative OncoGenomics), amongst others. In addition, the Biomart plug-in has been used by many third party software including Bioclipse, BioExtract, Galaxy, Cytoscape, and biomaRt-BioConductor. It should be noted that, Cytoscape is also used as a graphical data visualisation tool in *StatEpigen.*

[16] http://www.ensembl.org/index.html, http://www.ncbi.nlm.nih.gov/, http://genome.ucsc.edu/
[17] http://www.ensembl.org/index.html
[18] Version 65, Dec. 2011
[19] If both CCDS, Vega / HAVANA entry and Ensembl entry for a gene agree, it is said to have a golden transcript.
[20] http://www.biomart.org/

Entrez Gene is NCBI's repository for gene-specific information. NCBI provides search engine forms to query the data in Entrez along with eUtils (Entrez Programming Utilities)[21], for more direct access to query results. The E-Utils use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search and retrieve requested data. The E-Utils are therefore the structured interface to the Entrez system, which currently incorporates 38 databases storing data, which includes nucleotide and protein sequences, gene records, three-dimensional molecular structures, and biomedical literature.

The UCSC Genome Browser contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to the ENCODE (Encyclopaedia of DNA Elements) and Neanderthal projects[22].

2.3.2 Genetic Databases

Global efforts for gene annotation include HGNC (HUGO Gene Nomenclature Committee)[23], the only worldwide authority that assigns standardised nomenclature to human genes. This has assigned unique gene symbols and names to more than 31,000 human genes, of which over 19,000 are related to protein coding. Genenames.org is a curated online repository of HGNC-approved gene nomenclature and associated resources including genomic, proteomic and phenotypic information links, as well as dedicated gene family pages.

OMIM (Online Mendelian Inheritance in Man)[24] is a database of human genes and inherited traits and disorders[25] and provides a comprehensive and authoritative, compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known Mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype and NCBI/ Entrez links are provided from all cited references within each OMIM entry.

[21] http://www.ncbi.nlm.nih.gov/books/NBK25497/#chapter2.Introduction

[22] Project to sequence the Neanderthal genome.

[23] Jointly funded by US NHGRI(National Human Genome Research Institute) and the Wellcome Trust (UK)

[24] Housed/maintained at NCBI

[25] Maintained by Johns Hopkins University and collaborators

Genecards[26], is an integrated database of human genes that provides concise genome related information on all known and predicted human genes. It extracts and integrates a carefully selected subset of gene related transcriptomic, genetic, proteomic, functional and disease information, from dozens of relevant sources. The information is automatically mined and integrated from a variety of data sources, resulting in a web-based card for each of the 7000 human genes that currently have an approved gene symbol published by the HUGO/GDB nomenclature committee. The aim is to provide immediate current knowledge on a given gene. Source databases, mined for this information, include SWISS-PROT, OMIM, Gene Atlas and GDB. This *composite* database[27] aims to integrate information fragments, scattered over a variety of specialised databases into a coherent picture. Genecards is a freely accessible web resource that offers one hypertext card for each of the genes in the database and the recent version features both novel infrastructure and an improved search engine, which uses a persistent object/relational model, powerful tools, such as GeneAlaCart and GeneDecks. GeneALaCart is provided to produce tabulated annotations for the gene set and enable complex analysis.

2.3.3 Mutation DBs

A resource group with specific focus on mutation includes Mutation View (Minoshima *et al.* 2001) a multi-server/ client database system, which allows the users to systematically integrate information including genetic, molecular, biological and clinical findings of each disease. This enables chromosome ideograms to be generated for diseases, with OMIM document links for the mapped regions.

UMD (Universal Mutation Database), (Beroud *et al.* 2005), is a locus- specific database (LSDB) of mutations and their association with clinical and biological data. The database includes most genes and provides a large set of new analysis tools. Currently, there are new features to integrate non-coding sequences, clinical data, pictures, monoclonal antibodies and polymorphic markers (SNPs). UMD was developed as generic software to create locus specific databases (LSDBs) with the 4(th) Dimension (R) package from ACI[28], rather than the more popular MySQL and PHP, but has

[26] http://www.genecards.org/

[27] Developed and maintained by Crown Human Genome Centre ,Weizmann Institute of Science, Israel

[28] www.4D.com

comparable capability. Other mutation-related biomedical resources are tabulated, in brief, Table 2.1.

Table 2.1: Genomic Mutation Resources

Name	Resource Content	Additional Comments
IDBases[29]	Locus specific database for immunodeficiency-causing mutations	Currently 122 public databases are maintained and 3 under construction; aim is to establish database for every immunodeficiency or provide links to those maintained by others.
KinMutBase[30]	Disease-causing mutations in protein kinase domains	Database contains 582 mutations in 20 tyrosine kinase domains and 13 serine/threonine kinase domains. Database refers 1790 cases from 1322 families.
SH2Base[31]	Pathogenic SH2 (Src Homology 2) domain mutations	Genome wide search for disease-causing-mutations in the SH2 domains revealed 8 genes, mutations of which cause 9 distinct phenotypes.
Mutation DB[32]	Locus specific mutation for different genes.	Human genome variation society (HGVS) maintained. Provides HGNC gene symbol, OMIM number and the corresponding URL.
HGMD [33](Human Gene Mutation Database)	Comprehensive data on human inherited disease mutations	Provided by Biobase[34], compilation enables quick access to both single mutation queries and advanced search applications, applicable in human genetics research, diagnostics and personal genomics applications.

2.4 Epigenetic Resources and *StatEpigen*

The Human Epigenetics Project marks a further major step, following the Human Genome Project[35], driving much current ongoing research. Ongoing epigenetics and epigenomics research ranges from small-scale laboratory work to many large-scale project initiatives, with corresponding data made available via different epigenetic databases.

[29] http://bioinf.uta.fi/base_root/
[30] http://bioinf.uta.fi/KinMutBase/
[31] http://bioinf.uta.fi/SH2base/
[32] http://www.hgvs.org/dblist/glsdb.html
[33] http://www.hgmd.org/
[34] http://www.biobase-international.com/
[35] http://www.genome.gov/12011238

As outlined in Chapter 1, epigenetic abnormalities have been found to be causative factors of cancer, genetic disorders and paediatric syndromes, as well as contributory factors of autoimmune diseases and ageing (Rodenhiser and Mann 2006). The recent intensive research on cancer-epigenetics (Sharma *et al.* 2010, Esteller 2011) has also led to discovery of many epigenetic markers that play an important role in disease initiation. As a consequence, cancer-related epigenetic resources still dominate the landscape at the current time, compared to those for other diseases/conditions (autoimmune diseases, psychotic diseases, Chapter 1).

Two of the large-scale project initiatives for cancer research include ICGC (International Cancer Genome Consortium)[36] and TCGA (The Cancer Genome Atlas)[37]. Genomic and epigenomic changes that occur in various types of cancer are being investigated by the ICGC. The goal is to obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumour types and/or subtypes. Many samples from one tumour type or subtype would be analyzed in detail to provide crucial insights on genetic-epigenetic links. With respect to TCGA comprehensive sequencing, characterization and analysis of the genomic changes in various cancers has been achieved and the intent is to chart the genomic changes involved in more than 20 types of cancers.

The non-genetic inheritance phenomenon (epigenetic changes, Chapter 1; molecular details, Chapter 3), occur due to mechanisms such as DNA methylation, histone modifications and regulatory RNA changes while gene silencing, genomic imprinting, X-chromosome inactivation utilize epigenetics. Many epigenetic resources incorporating different epigenetic signatures have been observed in the survey, although unsurprisingly, less comprehensive than genetic/genomic resources. A brief summary of each is presented below. Other available epigenetic resources and available tools for epigenetic research gathered from the survey are tabulated in Appendix I, Table A and Table B respectively.

[36] http://www.icgc.org/
[37] http://cancergenome.nih.gov/

2.4.1 Methylation signatures[38]

The significant databases observed in our survey, with methylation signatures are *PubMeth*[39] (a cancer methylation database), *MethDB* (experimentally confirmed DNA methylation data, Grunau *et al.* 2001), *MethPrimerDB*[40], *MethyCancer* (He *et al.* 2008) and *Methylogix*[41] and are maintained as of the writing of this thesis.

PubMeth, provides a sorted, annotated and summarized overview of genes, reported to be methylated in various cancers, with user query based on gene or cancer type. This database draws on text-mining of Medline/PubMed abstracts, combined with manual annotation of pre-selected abstracts. The text mining approach in *PubMeth* is fast and intelligent, enabling search of multiple aliases and textual variants of these aliases, and querying of multiple key wordlists simultaneously. It also provides the facility to browse a pre-computed gene list, without having to query the database directly.

MethDB is general, more sample-oriented and not optimized to cancer-related queries. The database is designed to store and annotate information on the occurrence of methylated cytosines in DNA. It currently contains 19,905 methylation content data items and 5,382 methylation patterns or profiles for 48 species, 1,511 individuals, 198 tissues and cell lines and 79 phenotypes. It also has a public online submission system available. The resource forms part of an integrated network of biological databases through DAS (Distributed Annotation System), enabling the epigenetic data to be viewed as a layer in the human genome, and is also connected to Ensembl (for DNA sequences with available *MethDB* data aligned to NCBI Refseq).

A subset resource, *MethPrimerDB*, is a database of primer sequences used in PCR-based methylation methods. It relies on submissions by users and administrators that guarantee the required quality of the database but not necessarily its completeness. To date, there are 29 primer sets. In 2006, the MethBLAST feature was added to *MethPrimerDB* oligonucleotide sequences. Further updates since 2006, however, are not found for this resource.

[38] DNA methylation, addition of a methyl group to the cytosine base
[39] http://matrix.ugent.be/pubmeth/
[40] http://medgen.ugent.be/methprimerdb/
[41] http://www.methylogix.com/genetics/database.shtml.htm

MethyCancer is a disease-oriented database, specifically of human DNA methylation and cancer that aims to integrate methylation databases and has developed a meta-data format for data standardization, with manual curation still being used for noisy data. Four main types of data are included in *MethyCancer*, namely, (i) CGI clones and global CGI predictions, (ii) DNA methylation data, (iii) Cancer information, genes and mutations and (iv) Correlations of DNA methylation, gene expression and cancer. MethyView, a visualization tool from *MethyCancer*, is used to facilitate the browsing of methylation data in the context of existing human genome annotations. A search engine to query different data types and interactions from the *MethyCancer* database provides simple keyword search and also offers advanced options namely, "methylation," "gene," "cancer," "clone" and "repeat" searches. For example, Methylation search enables the user to specify and combine query options, such as methylation type (pattern, profile, content and domain), data source (BIG/ UHN, MethDB, HEP, Columbia University), experimental methods, sample information (tissue, sex, age and phenotype) and chromosomal positions.

Methylogix provides a high density DNA methylation database of human chromosomes 21 and 22, a CpG island DNA methylation database for male germ cells (enabling comprehensive analysis of DNA methylation variation between and within the germ lines of normal males), and a targeted DNA methylation database of late-onset Alzheimer disease. Similarly, Methtools is a collection of software tools for handling and analysis of DNA methylation data, generated by the bisulfite genomic sequencing method.

2.4.2 Genomic Imprinting

Genomic imprinting (described in Section 1.1.1) related significant database resources are the *Geneimprint database*[42], *imprinted gene and parent-of-origin effect database* (Glaser *et al.* 2006) and *mouse gene imprinting database*[43] (with genomic imprinting information for the model organism: mouse)

The *Geneimprint database* includes genes and related information on genomic imprinting for different animals including humans and gathered from NCBI. Genes are

[42] http://www.geneimprint.com/site/genes-by-species

[43] http://www.mousebook.org/about.php

listed by species and sorted by chromosomal location, name and imprinting status and are provided through the web-interface. Similarly, an *imprinted gene and parent-of-origin effect database* presents imprinted genes and related effects. This consists of two sections: (i) catalogue of current literature on imprinted genes in humans and animals and (ii) catalogue of reports of parental origin of de novo mutations in humans alone. The addition of (ii), showing a parent-of-origin effect, expands the scope of the database and provides a useful tool for examining parental origin trends for different types of spontaneous mutations. This second section currently includes more than 1,700 mutations, found in 59 different disorders. The 85 imprinted genes are described in 152 entries from several mammalian species. In addition, more than 300 other entries describe a range of reported parent-of-origin effects in animals. A further resource, containing information on *mouse gene imprinting*, also includes an imprinting catalogue, as well as chromosome anomalies on mutant mouse lines. This represents integration of curated information from the MRC Harwell stock resource and other Harwell databases, with additional information from external data resources such as IMSR (International Mouse Strain Resource).

2.4.3 Histone and chromatin-related resources

The most significant example reported is the *Histone database,* which is an inclusive resource for the analysis of chromatin structure and function. Nucleosomes (through various core histone post-translational modifications and incorporation of diverse histone variants), can serve as epigenetic markers to control processes such as gene expression and recombination. The *Histone Sequence Database* is a curated collection, assembled from major public databases, of sequences and structures of histones and non-histone proteins containing histone folds. A substantial increase in the number of sequences and taxonomic coverage for histone and histone fold-containing proteins is available. The database also provides comprehensive multiple sequence alignments for each of the four core histones (*H2A, H2B, H3* and *H4*), the linker histones (*H1/ H5*) and the archaeal histones. Also included is current information on solved histone fold-containing structures. This resource is maintained by the National Human Genome Research Institute.

Chromatin.us[44] (another webportal) includes information on chromatin proteins, histones and nucleosome structures and non-histone chromatin protein structures, and provides links to the *Protein Data Bank* (*PDB*) site (which provides further details on these). *ReplicationDomain*[45] is an online database for storing, sharing and visualizing DNA replication timing and transcription data, along with other numerical epigenetic data types. Data are typically obtained from DNA microarrays or DNA sequencing.

2.4.4 Gene silencing

Gene silencing is an epigenetic process of gene regulation that generally is used to describe the "switching off" of a gene by mechanism other than genetic modification (Chapter 1). This phenomenon has also been well-reported in the literature. Collected papers are available on *Bio-Tech Info-Net*[46]. Similarly, RNA induced epigenetics related papers on imprinting by noncoding RNAs are collated[47].

2.4.5 Other Epigenetic Database Resources

Evolution of epigenetic resources is still in its early stages, with provision associated with several specific research efforts and groups. Nevertheless, in line with genetic/genomic data examples, efforts are being made to connect information, even as new targets are emerging. Some other biomedical resources that relate to epigenetic phenomena are also noted. The *Epigenetics Database*[48] (includes all known epigenetics genes/proteins discovered to date), the *Epigenie*[49], the *Epigenetics Antibody Database* [50], *Unigene* [51] are some other examples.

The Epigenetics Database is arranged in hierarchical format, based upon gene ontology. While still in its developmental (ß) phase, it is expected that future developments will include user-submitted meta-data, which will be freely available for use in database and flat file format. Some sites, *e.g.*, *Epigenie*, also provide bioinformatics tools (*e.g.*, CpG Viewer, CpG and GC Plotter and tools for CpG Island detection). NCBI supported efforts include the *Epigenetics Antibody Database*, providing antibody information for researchers working in the field of epigenetics/epigenomics, and *Unigene*, containing

[44] http://www.chromatin.us/chrom.html
[45] http://www.replicationdomain.com/
[46] http://www.biotech-info.net/gene_silencing.html#silencing
[47] http://www.euchromatin.com/RNAepi.htm
[48] http://www.epidna.com/database.php
[49] http://www.epigenie.com/Epigenetics-Research-Products.html
[50] http://www.antibodyresource.com/antibody-database.html
[51] http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene

same locus-of-origin transcription sequences, protein similarities, gene expression, cDNA clone reagents, genomic location and associated epigenetic information. *NARNA*, supported by Newcastle University, incorporates relationships between epigenetic events, DNA methylation, gene imprinting and X-chromosome inactivation with natural antisense RNAs.

2.4.6 Large Scale Epigenetic Project Initiatives

European project initiatives including HEP

A number of European initiatives exist for centralized projects on DNA methylation. The *Human Epigenome Project* (*HEP*[52]), will provide an epigenetic resource of chromosomal DNA methylation reference profiles in human tissues and cell lines. Other initiatives include chromatin profiling (*HEROIC*, High-Throughput Epigenetic Regulatory Organization in Chromatin), treatment of neoplastic disease (*EPITRON, EPIgenetic TReatment Of Neoplastic Disease*[53]), and the *SMARTER* initiative[54], which aim to develop small inhibitors of chromatin-modifying enzymes. Another effort to provide structure to the epigenetic research landscape in Europe is that of the *Epigenetic Network of Excellence*, now known as *Epigenesys*, which aims to advance epigenetics toward Systems Biology[55].

Roadmap epigenomics program

The Roadmap Epigenomics Program (also known as Epigenomics Roadmap initiative), launched by NIH (2008), seeks to create a series of epigenome maps to study epigenetic mechanisms, develop new epigenetic analytics, generate a repository and long-term data archive, standardize procedures and practices in epigenomics and support new technologies for these. As part of the $190 million, five-year initiative, the *Roadmap Epigenomics Mapping Consortium*[56] was formed to provide a public database for human epigenomics data, the *Human Epigenome Atlas*[57]. The current release, Epigenome Atlas Release 7, includes human reference epigenomes and the results of their integrative and comparative analyses.

[52] http://www.epigenome.org/
[53] http://www.epitron.eu/
[54] http://www.smarter-chromatin.eu/
[55] http://www.epigenesys.eu/
[56] http://www.roadmapepigenomics.org/
[57] http://www.genboree.org/epigenomeatlas/index.rhtml

The NIH Roadmap Epigenomics Program has also established *IHEC*, (the *International Human Epigenome Consortium*)[58], which aims to coordinate epigenome mapping and characterization worldwide, in order to ensure high data quality standards, coordination of data storage, management and analysis and free access to the epigenomes produced. To attain substantial coverage of the human epigenome, *IHEC* aims to decipher at least 1,000 epigenomes within the next 7–10 years. Officially launched in Paris (Jan 2010), with an initial (first phase) budget target of $130 million, *IHEC* intends to coordinate the mapping of epigenomes from not only the NIH's Epigenomics Mapping Consortium but also from international efforts such as the European Epigenome Network of Excellence, the Danish National Research Foundation Centre for Epigenetics, and the Australian Epigenetic Alliance. The *IHEC* web portal provides links to databases, such as *GEO,* Array Express and *DDBJ,* where epigenetic sequencing data will be made available.

Another significant large-scale program in epigenetics is the Encyclopaedia of DNA Elements (*ENCODE*)[59], which is supported by the *ENCODE* Consortium, an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). This initiative aims to identify all functional elements, both at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active, in the human genome sequence.

ICGC

Genomic changes that occur in various types of cancer are being investigated by the International Cancer Genome Consortium (*ICGC*)[60]. The goal is to obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumour types and/or subtypes. Many samples from one tumour type or subtype will be analyzed in detail so that this initiative promises to provide crucial insights on genetic-epigenetic links.

2.4.7 *StatEpigen* Database Resource

StatEpigen is an in-house developed epigenetic database resource (Barat, Ruskin 2010). This database currently incorporates information on cancer epigenetic data, curated

[58] http://www.ihec-epigenomes.org/
[59] http://genome.ucsc.edu/ENCODE/
[60] http://icgc.org/

from published scientific literature. Initially, the database has focussed on curation of colon cancer epigenetic data, although the aim is to add data in other types of cancer in future. The database also gives correlation between different epigenetic determinants of cancer. Statistical measures and correlations or associations between a particular epigenetic event, found in a particular cancer phenotype are related to other molecular events which comprise known cancer signatures.

Since the primary aim of this project focuses on further development of *StatEpigen* database resource, the literature survey assisted in finding other database resources that can be integrated with *StatEpigen* database (Table 2.2). Following is the comparative study of some epigenetic resources that incorporate methylation signatures and can be integrated to in-house database, *StatEpigen* in future. Besides these, *COSMIC* database from Sanger Institute which provides data on *somatic* and *germline* mutation of cancer prone genes can also be integrated (Wellcome Trust Sanger Institute).

Table 2.2: Epigenetic resources with methylation signatures (integrable to *StatEpigen* database)

S.N.	Databases	Database of:	Comments:
1	*PubMeth*	Cancer methylation database	Only cancer related DNA methylation considered.
2	*MethDB*	Experimentally confirmed DNA methylation data	Aims to store all data about DNA methylation, not specific to cancer only.
3	*MethPrimerDB*	Primer sequences used in PCR based methylation data	Method of methylation focussed database, has data on mouse, human and rat, methylation data.
4	*MethyCancer*	Cancer related DNA methylation	Human DNA methylation and cancer related database –Chinese initiative. Also integrated to HEP, MethDB, Columbia and BIG/UHN.
5	*Methylogix*	DNA methylation database of human chromosomes 21 and 22 + germcells.	High density human chromosomes 21 and 22 and germcells only considered.

2.5 Discussion and Conclusion

This chapter presented a summary landscape review of the biomedical literature, with the major focus on epigenetic resources. The intention was to check for developments in epigenetic sector, whether it has paralleled the developments in genetic sector, since the former became a target area of research after the human genome project. The genetic and epigenetic biomedical resources, surveyed here, are numerous and range from small- to large-scale. There is considerable ongoing integration and new links are still being forged. As with many newly identified research targets, early-stage resources are often found to be very specific and are supported locally, and this is still the case for much of epigenetic data. Many such databases and their software tools are accessible publicly from academic/ research institutions, while some others are commercially available, Appendix I, Table A. Quality assurance, effective annotation and overall management are still major issues, but appropriate analysis must also keep pace and currently is typically uneven, Appendix I, Table B. Clearly, the generation of a centralised repository for epigenetics-related data is desirable, and currently lacking, but new technologies offer increased potential for processing solutions down the line.

Some major initiatives to ensure quality and standards for genetic and epigenetic research do exist, such as *IHEC, HEP*[61] as described in this literature review. These will also lead to development of suitable data mining tools, with improved technology, where those currently available for epigenetic/epigenomic analyses are limited, and predominantly sequence oriented, ranging from sequence identification, through PCR and initial pattern matching, Appendix I, Table B.

In general, survey results indicate that relational database management systems are still preferred and support most available data mining options, with MySQL as open source and data query achieved via form submission. Biomart, as noted earlier, is also widely-used and relies on a query-oriented data management system. These preferences are likely to change, with hybrid and object-oriented systems becoming much more general.

[61] http://www.epigenome.org/

Chapter 3

Cancer, Genetic and Epigenetic Factors, and *StatEpigen*

3.1 Introduction

This Chapter explores cancer and its etiology at molecular level, starting with cancer staging which determines the level of cancer progression in a cancer patient. Cancer Staging describes the extent or severity of a cancer diagnosis based on information about the tumour[62] (www.cancer.gov). Staging helps plan the treatment, identify a clinical trial suitable to the patient and can also be used to estimate the prognosis (likely outcome or course of the disease). The common elements considered in most staging systems are:

- Location of the primary tumour,
- Tumour size and number of tumours,
- Spread of cancer into lymph nodes,
- Cell type and tumour grade (how closely the cancer cells resemble normal tissue)
- Presence or absence of metastasis (the spread of the cancer).

The TNM system is one of the most commonly used staging systems. It is based on the extent of the tumour (T), the extent of spread to the lymph nodes (N), and the presence of metastasis (M). A number is added to each letter to indicate the size or extent of the tumour and the extent of spread. Table 3.1 below shows the indications used for TNM staging.

[62] Abnormal growth of tissue resulting from uncontrolled, progressive multiplication of cells and serving no physiological function.

Table 3.1: Tables showing the indications used for TNM staging (Source: www.cancer.gov, National Cancer Institute website)

Primary Tumour (T)	
TX	Primary tumour can not be evaluated
T0	No evidence of primary tumour
Tis	Carcinoma in situ (early cancer that has not spread to neighbouring tissue)
T1, T2, T3, T4	Size and/or extent of the primary tumour

Regional Lymph Nodes (N)	
NX	Regional lymph nodes can not be evaluated
N0	No regional lymph node involvement (no cancer found in the lymph nodes)
N1, N2, N3	Involvement of regional lymph nodes (number and/or extent of spread)

Distant Metastasis (M)	
MX	Distant metastasis can not be evaluated
M0	No distant metastasis (cancer has not spread to other parts of the body)
M1	Distant metastasis (cancer has spread to distant parts of the body)

For example, prostate cancer T2 N0 M0 means that the tumour is located only in the prostate and has not spread to the lymph nodes or any other part of the body. Similarly, breast cancer T3 N2 M0 refers to a large tumour that has spread outside the breast to nearby lymph nodes, but not to other parts of the body.

For many cancers, TNM combinations correspond to one of five stages. Criteria for stages differ for different types of cancer. For example, colon cancer T3 N0 M0 is stage II; however, bladder cancer T3 N0 M0 is stage III. Table 3.2 below shows what each stage means.

Table 3.2: Cancer stages and definitions (Source: www.cancer.gov, National Cancer Institute website)

Stage	**Definition**
Stage 0	Carcinoma in situ (early cancer that is present only in the layer of cells in which it began).
Stage I, II, and III	Higher numbers indicate more extensive disease: greater tumour size, and/or spread of the cancer to nearby lymph nodes and/or organs adjacent to the primary tumour.
Stage IV	The cancer has spread to another organ.

Cancers of the blood or bone marrow and most types of leukaemia, do not have a clear-cut staging system yet.

The summary staging system is used for all types of cancer. The categories are:

1. ***In situ*** - early cancer that is present only in the layer of cells in which it began.
2. **Localized** - cancer that is limited to the organ in which it began, without evidence of spread.
3. **Regional** - cancer that has spread beyond the original (primary) site to nearby lymph nodes or organs and tissues.
4. **Distant** - cancer that has spread from the primary site to distant organs or distant lymph nodes.
5. **Unknown** - cases for which there is not enough information to indicate a clearly described stage.

3.2 Genetic and Epigenetic Factors in Cancers

Cancer arises through the accumulation of multiple genetic and epigenetic changes (Feinberg 2004). Environmental components such as growth factors and hormones (Coffey 2001, Michels 2005, Giovannucci 2003) and changes in diet and life history that result in mismatch between ancestral and current environments are also found to increase the risk of cancers, (Jemal *et al.* 2011). This could be due to the impact of environmental components, diet and life history, in the genetic and epigenetic events.

3.2.1 Key Genetic Effects

Genetic changes or mutations are any permanent change/damage of DNA sequence of the genome such that the genetic message carried by that gene is altered. Genetic changes could be *somatic* or *germline* (hereditary) mutations of specific genes. *Somatic* mutations (changes in DNA sequence) are sporadic, that occur in cells other than *germ* cells, whereas *germline* mutations are mutations that are heritable in the *lineage* of *germ* cells. Mutations can cause (genetic) diseases, changes in enzyme activity, nutritional requirements, antibiotic susceptibility, morphology, antigenicity and other properties of cells. The most common type of mutation is a single base alteration or point mutation that can be either a transition[63] or transversion[64]. Other mutation types are classified according to the effect they have on the structure of the genes' protein product. These include: silent (where new codon specifies same amino acid), missense (new codon specifies different amino acid), nonsense (new codon is stop codon), frameshift (deletion or addition of a base), large segment deletion (unequal crossover in meiosis), splice donor or acceptor and triplet repeat expansion (Fischer and Reichert 2009).

Alterations in 3 types of genes, proto-oncogenes, tumour-suppressor genes and stability genes are responsible for tumorigenesis. Of the estimated 20-25,000 genes in human genome, currently more than 1% of human genes are known to play a role in the development of cancer (Cancer Gene Census, Sanger Institute). A single gene cannot cause cancer as there are multiple mechanisms for protection. Cancer is caused only when several genes are defective. Therefore, mutated cancer genes can be thought of as contributing to rather than causing cancer (Vogelstein and Kinzler 2004).

Oncogenes are mutated forms of normal genes called proto-oncogenes. Oncogenes cause the normal cells to grow out of control and become cancerous (Adamson 1987, Weinstein and Joe 2006). Often, proto-oncogenes encode proteins that function to stimulate cell division, inhibit cell differentiation, and stop cell death (all of these processes being important for normal human development and for the maintenance of tissues and organs). However, oncogenes typically exhibit *increased production* of these proteins, thus leading to increased cell division, decreased cell differentiation, and inhibition of cell death; taken together, these phenotypes define cancer cells. Oncogene activations can result from gene amplifications, chromosomal translocations or from subtle intragenic mutations affecting crucial residues which regulate the activity of gene product (Vogelstein and Kinzler 2004). Some studies have suggested that cancer cells may rely heavily on certain oncogenic mutations than others for their growth, proliferation and survival. This concept is termed as "oncogene addiction" (Weinstein and Joe 2006). Therefore, by aiming at those specific oncogenes, it is believed that it might be possible to target cancer cells, in spite of other mutations. These oncogenes are thus currently a major molecular target for anti-cancer drug design.

[63] Point mutation that changes a purine nucleotide to another purine (A-G) or pyrimidine nucleotide to another pyrimidine (C-T).

[64] A base-pair substitution mutation in which a purine is replaced by pyrimidine (e.g. A → C), and vice versa (e.g. T → G)

Tumour suppressor genes encode proteins that normally slow down cell division, repair DNA damage and trigger apoptosis. Mutations in tumour suppressor genes reduce the gene activity which leads to uncontrolled growth of cells, leading to cancer. Gene inactivation is caused from missense mutations at residues that are essential for its activity, from mutations that result in a truncated protein, from different deletions or insertions or from epigenetic silencing.

Both the tumour-suppressor gene and oncogene mutations ultimately result in driving the neoplastic process by increasing tumour cell number through the stimulation of cell birth, inhibition of cell death or cell cycle arrest (*i.e.* at the physiological level, the mutations result in increase of tumour cells). Stability genes (caretaker genes) on the other hand, when mutated, promote tumorigenesis in a different manner. These genes when inactivated trigger other genes to be mutated at a higher rate (Friedberg 2003). Stability genes include the mismatch repair (MMR), nucleotide-excision repair (NER) and base-excision repair (BER) genes that are responsible for repairing minor mistakes during normal DNA replication or are induced by exposure to mutagenic agents. Stability genes also control processes involving large portions of chromosomes, such as those responsible for mitotic recombination and chromosomal segregation (*e.g. BRCA1*, *ATM*) (Vogelstein and Kinzler 2004). Mutations in these 3 types of genes lead to *hereditary predisposition* to cancer if it occurs in *germline* cells or to sporadic tumours if occurs in *somatic* cells.

3.2.2 Pathways

Vogelstein and Kinzler, (2004) suggest focussing on pathways, the 3 groups of gene follow rather than the genes themselves, the reasons being:

1. Fewer numbers of pathways than the genes themselves. Studies have shown that mutations within a pathway obey an ‘exclusivity principle’; that is, only one of many genes is generally mutated in any single tumour, exactly as predicted if the functional effect of each mutation was similar (*e.g.* Rb pathway with mutations in any one of *Rb, p16, CDK4* and *Cyclin D1* genes (Sherr 2000, Ichimura 2000, Ortega *et al.* 2002, and Classon and Harlow 2002)).
2. There are more than one ways to disrupt a pathway, all of which have similar effect/consequence (*e.g. p53* pathway- which can be disrupted by point mutation

of *p53* gene; the same effect can be achieved with amplification of the *MDM2* gene and infection with DNA tumour virus whose products bind to *p53* and functionally inactivate it).

These instances suggest that, pathways rather than individual genes govern the course of tumorigenesis and the targeted therapeutics can be effective against a broad range of cancers in future (Vogelstein and Kinzler 2004) .

3.2.3 Epigenetic Factors

As per the current knowledge on epigenetic involvement in cancer and suggested complementary role of genetics and epigenetic in the etiology of the same, (Feinberg 2004, Sawan *et al.* 2008, You and Jones 2012) it would be sensible to look into the epigenetic mechanism for some of the major epigenetic events at this stage.

DNA Methylation

As defined in chapter 1, DNA methylation refers to the modification of DNA by addition of a methyl group to the cytosine base (C). It is found to be the most stable, heritable and well conserved epigenetic change. Introduced and maintained by a family of enzymes called *DNA Methyl Transferases* (*DNMT*) (Doerfler *et al.* 1990), DNA methylation can induce "*epigenetic silencing*" or the loss of expression of tumour suppressor genes, causing normal cells to be transformed into cancer cells; it is the first and most common epigenetic alteration to be observed during cancer initiation (Egger *et al.* 2004, Feinberg and Tycko 2004). Epigenetic alterations involving DNA methylation can lead to cancer by various mechanisms. *Hypomethylation* of DNA can result in genome instability (Qu *et al.* 1999), *hypermethylation* in gene promoters causes heritable silencing and therefore inactivation of tumour suppressor genes. Similarly, methylated CpG sites are “hotspots” for C->T transition mutations caused by spontaneous hydrolytic deamination (Rideout *et al.* 1990). Methylation of CpG sites also increases the binding of some chemical carcinogens to DNA and increases the rate of UV induced mutations. Methylation as a cause of mutation is also reported by Mazin (1994) and the relationship between these two events is investigated by many, including Xia (2012).

DNA methylation, in general, occurs at CpG dinucleotide sites, and a majority of CpG cytosines are methylated (in mammals). However, there are clusters of DNA near promoter regions that have higher concentrations of CpG sites (also known as CpG islands) that are unmethylated in normal cells. These CpG islands become highly methylated in cancer cells, thereby causing unexpected or unlikely genes to be silenced or to be switched off (transcriptional silencing). This abnormality is the trademark epigenetic change that occurs in tumours and takes place early in cancer development (Jones & Baylin, 2002, Egger *et al.*, 2004). Subsequently, it has been found that, hypermethylation of CpG islands can cause tumours by shutting off tumour-suppressor genes (*e.g. MGMT*, *p16INK4a*, *hMLH1*, *p14ARF*), forging the genetic-epigenetic link again (Esteller 2007). In fact, CpG island hypermethylation may be more common in human cancer than DNA sequence mutations.

Histone Modification

Histones are alkaline proteins (found in eukaryotic cell nuclei), around which DNA wraps to form nucleosomes. There are 5 major families of histones: *H1/H5, H2A, H2B, H3* and *H4*. Histones *H2A, H2B, H3* and *H4* are called core histones and two of these each, when wrapped by DNA double helix, forms a single nucleosome unit. The *H1* and *H5* are linker histones that bind the nucleosomes (*H1 in humans*), (Youngson 2006). Histone modifications lead to gene expression or inactivation (Kouzarides 2007). A direct link also exists between DNA methylation and histone modification, since a number of proteins involved in DNA methylation (*e.g. DNMTs* and *MBDs (Methyl CpG-Binding Domain proteins*)) directly interact with histone modifying enzymes, such as *Histone Methyltransferases (HMTs*) and *Histone Deacetylases (HDACs*) (Fuks 2005).

Other Epigenetic Phenomena

Genomic Imprinting is an inheritance process that is independent of classical Mendelian inheritance and is a genetic phenomenon by which genes are expressed in the parent-of-origin-specific manner. *X-chromosome inactivation* is a process in which, one out of two copies of X-chromosomes available in female mammals is inactivated. This process is also called *lyonization*. All of these processes are epigenetic processes, and thought to be contributing to the cancer initiation and progression.

3.2.4 Genetic- Epigenetic mechanisms in Cancer

Cancer was traditionally considered as a genetic disorder, however, after completion of human genome project, the role of epigenetics in cancer came to light (Feinberg 2004, Sharma *et al.* 2009). Current researches show cancer as an outcome of genetic-epigenetic interactions and there is increasing evidence that these two mechanisms are complementary (Mazin 1994, Duncan *et al.* 2012). The cross-talk between genome and epigenome is shown to be involved in cancer (Sawan *et al.* 2008, You and Jones 2012).

Feinberg (2004) suggested his hypothesis on genetic –epigenetic interplay in the initiation of cancer. He suggested, with regard to the role of epigenetics in cancer initiation and progression, it is possible that genetic mutation initiates the cancer and epigenetic change promotes its progression. It is also possible that epigenetic processes may be linked directly to initiation or may be responsible for 'priming' cells to the next mutagenic event that is involved in cellular transformation after this. Hence both genetic and epigenetic events may be co-involved in cancer progression. In fact, some literature (*e.g.* Ogino *et al.* 2006) suggest that, some genetic/epigenetic factors have a synergistic effect on the occurrence of colon cancer. Some notable publications are Rideout *et al.* (1990), Mazin (1994), Denissenko *et al.* (1997), which suggest that methylation events trigger mutation events. On the contrary, there are some other publications which suggest the opposite, *i.e.* mutation induces methylation, (Poole *et al.* 2012, Duncan *et al.* 2012). Thus genetic events can not be treated in isolation to epigenetic events without exploring possible relations, and related or co-occurring changes.

3.3 *StatEpigen* database Resource

StatEpigen Knowledge Management System (Section 2.4.7, Barat and Ruskin 2010), is our in-house epigenetic database resource. The uniqueness of this resource is that, it provides selected information on different genetic and epigenetic factors responsible for colon cancer, curated from different literature. The aim of the database is to provide a platform to explore how epigenetic events, (such as CpG island *hyper* and *hypomethylation*, different histone modifications, loss of heterozygocity etc.), are correlated with (i)each other, (ii)other molecular events namely gene expression, (iii)different types of mutations and polymorphisms, (iv)other more complex molecular signatures such as *MSI* (Microsatellite Instability), *CIN* (Chromosomal Instability), and

CIMP (CpG Island Methylator Phenotype) (v)simultaneous molecular events, *i.e.* combinations of event types occuring simultaneously in the same samples. To address these objectives in data evaluation, the following types of datasets are targeted:

1. Molecular events and their frequency of occurrence, with the phenotype of analysed samples (*Simple molecular events)* being provided,
2. Molecular events and their frequency of occurrence in sample subsets, characterised by another molecular event (occuring in all analysed samples of a subset of given phenotype). These events are referred to as 'conditional events' and are an important feature for *StatEpigen* database resource.

3.3.1 Structure and curation of data

The database basically contains the two types of records, given in bullet points above. The data after manual curation is stored in MySQL database.

Database Structure and Tables

The database itself is a phenotype-focused resource, with genetic and epigenetic data available for an extensive phenotype range. To date, data on more than 100 colon cancer cell lines are available on *StatEpigen*. Similarly, information on 728 genes, 318 epigenetic events, 579 genetic events, and 150 mutations is also available, with 5768 simple molecular events and 2952 conditional events detailed to-date[65].

The data is basically stored in database objects called 'Tables'. The major tables in current use (there are other tables as well for future addition of events or cancer types – but not in use currently) in *StatEpigen* database are : "Reference", "Cell-State", "Event" , "Gene", "single_rel", "double_rel", "pathology ". *Reference* tables store information on the references that one uses during data curtain process, *Cell-State* table stores, histology, subhistology, Origin, Cell-Line information, *event* table stores information on different events that are in curated till now, (Note: same type of event may have different different event-IDs depending on any specific details on it, *e.g.* Hypermethylation occurring in promoter region and exon2 are recorded as two different events). Similarly, *Gene* table records information on gene (*e.g.* UniProt and Entrez ID, HGNC symbol, full name), the single and double relation tables - record information on curated single and double relation information (the frequency of occurrence of that

[65] http://statepigen.sci-sym.dcu.ie/

event, sample no. are some notable columns in the table). These tables are 'relational tables' and are linked to each other with their 'keys'.

Data Curation

Data curation process firstly involves the selection of suitable literature. It involves filtering *Pubmed* database with keywords related to genetic and epigenetic events (*e.g.* hypermethylation, mutation). The selected papers are collected, scanned manually and if suitable are curated. The process involves curation of phenotypes, genes and molecular events and simple and conditional events. The phenotypes to look for are – Histology, subhistology, dysplasia, origin, cell-lines and pathology. Similarly, for genes, if *HGNC* (*HUGO Gene Nomenclature Committee)* approved gene name appears in a paper and it is a new entry, the curator needs to add all the details (*pharmagkb* site[66] can be used for gene details). If duplicate names are given, *HGNC* name is found and details are entered. For molecular events, the curator needs to look for events as shown in Table 3.3, (where this table is truncated for ease) curate any, that are found, according to the single or double relation available in the paper and augment the data in *StatEpigen*. To understand how the process works, an example of a simple event curation is given below.

[66] http://www.pharmgkb.org/

Table 3.3: Type event table (truncated) to show different epigenetic events incorporated in *StatEpigen* database

Type_ event	**Description**	**Corresponding Name_Events**	**Description**
1	**Epigenetic events**	**+Meth_CpGisland**	CpG island hyper-methylation
		-Meth_ CpGisland	CpGisland hypo-methylation
		LOI	Loss of imprinting.
		H3Ac	Histone 3 acetylation
		H4Ac	Histone 4 acetylation
		H3PhS10	Phosphorylation of threonine 10 on histone 3
		H3MeK4	Methylation of lysine 4 on histone 3
		H3MeK9	Methylation of lysine 9 on histone 3
2	**Mutations**	**mutation**	Mutations. The Details field of the db table event will serve to annotate what kind of mutation is this: **'missense';** **'antisense'** **'frameshift'** **'silent'** Or, for example, Details=**'germline**
		polymorphism	Single Nucleotide Polymorphism - the variation of a single base pair in the DNA among the individuals of the same species.
3	**Expression, both at the levels of mRNA or proteins**	**gene_expression**	Both expression of genes and presence of proteins.
		mir_expression	Expression of miRNAs
4	**Corresponding to multiple events**	**MSI**	Microsatellite Instability. Description in the adjacent column.
		CIMP	CpG island methylator phenotype. The promoters of a group of genes are methylated in the same time.
		LOH	Allelic loss in more than one loci
5	**Combined:**	**combined**	with 2 distinct events taking place simultaneously
6	**Diverse:** events that do not correspond to neither of the other classes	**MMR_defective**	Inherited mutation in a mismatch-repair genes (MLH and MSH gene family)
		activation	Activation of a protein
		phosphorylation	Phosphorylation of a protein
		translocation	Translocation of a protein from one to another cell compartment

Supposing the literature includes the contents given in the text below:

P16 methylation of tumour and/or serum of 51 colorectal cancers and 10 adenoma patients, and 10 healthy volunteers was detected with conventional MSP or IP-MSP. IP-MSP detected p16 methylation from 0.5pg/mul of the cell lysate. ***The sensitivity of IP-MSP for detecting serum p16 methylation in 27 patients with tumours characterized by p16 methylation was significantly higher than that with conventional method (81% versus 59%), particularly in Stage II patients (91% versus 45%). IP-MSP detected no p16 hypermethylation in sera of adenoma patients and volunteers*** (Sakamoto *et al.* 2010).

After the curation of phenotypes and the molecular events from this extract, use of this information to form a Single Relation (or Simple Event record) is shown below. Four potential Single Relations are found in the text:

1. p16 methylation was detected in 81% of the 27 serum samples from patients with tumours.
2. p16 methylation was detected in 91% of the serum samples from patients with tumours in Stage II.
3. p16 methylation was not detected in sera of adenoma patients.
4. p16 methylation was not detected in sera of volunteers.

In the second potential Single Relation, there is no indication on how many samples from tumours in stage II were available. For this reason, the information is incomplete and this will not be used further to create a Single Relation record.

For the remaining three statements, the ID_CELL_STATES of sera from tumours, adenomas and healthy volunteers are queried into the database to find, 75, 62 and 88 respectively. Further, the ID_EVENT for p16 methylation was queried as ID_EVENT =17. In addition, the article from which this text is extracted was given ID_REF=257.

The extract gives the frequencies of sample **positive** for p16 methylation hence Level='YES' in all the 3 cases.

The SQL code to add data is (with comments in red and double hyphen):

```
INSERT INTO single_rel (REF_ID_REF, EVENT_ID_EVENT,
CELL_STATE_ID_CELL_STATE, clinpath, Percent_Tumors, Nb_Tumors,
Quantified, Unit_ID_UNIT, Level) VALUES

(257, 17, 75, NULL, 0.81, 27, NULL, NULL, ‘YES’),
-- the sera from carcinoma patients
(257, 17, 62, NULL, 0,    10, NULL, NULL, ‘YES’),
-- no samples positive for p16 methylation in the sera from adenoma
patients
(257, 17, 88, NULL, 0,    10, NULL, NULL, ‘YES’),
-- no samples positive for p16 methylation in the sera from healthy
subjects
```

It is to be noted that the curator has to be careful here not to confuse Level ‘YES’, which indicates that the samples counted are those with positive outcome for the methylation, with the null frequency (Percent_Tumours =0) of positive samples in the ten samples analysed. After manual curation, these data are incorporated into the purpose –built database implemented in MySQL.

This is one of the simplest examples of data curation. As can be seen, manual data curation is a time consuming process, hence, automation/ semi-automation of this process in future is desirable.

Interface

The data is provided to the user via web-based user interface. The URL is http://statepigen.sci-sym.dcu.ie. The data can be accessed with straightforward querying and results are displayed. The user interface employs a design implemented in PHP, Javascript, HTML and CSS (Barat and Ruskin 2010).

Querying

Different querying facilities are available in *StatEpigen*. It can be queried: (a) in a gene/molecular event- centric manner, (b) in a phenotype – centric manner or both, (c) Additional options to carry out phenotype-centric *StatEpigen* querying, according to most frequent histologies and subhistologies, clinicopathological factors and cell lines.

Graphics

The graphical data visualisation facility is provided via the Cytoscape tool[67] and data integration facility via 'print statistics' tab. Full details are provided in (Barat and Ruskin 2010).

3.4 Summary

This chapter explains the role of genetic and epigenetic factors and their inter-relationship for cancer initiation and progression, as revealed by current research in the last decade. It also explains the need for database resources such as *StatEpigen* in the current context, as a specialised add-on resource and gives a brief description (and examples) of its main features, *i.e.* how it can be used and what it offers.

[67] http://www.cytoscape.org/

Chapter 4
Evolutionary Analysis of CRC genes

4.1 Introduction

A genetic study of colon cancer from an evolutionary perspective has been carried out with the molecular evolution group in Sci-Sym centre (http://sci-sym.dcu.ie/). This allowed exploration of, not only the genetic influences in initiation of the human disease, (Section 4.1.1) but also the inheritance features (Section 4.1.2) and human/animal parallels (Section 4.1.3), the inhibitory factors intrinsic to selection (Section 4.1.4 and 4.1 below), and the type of damage that can be attributable to various causes.

Traditionally, cancers have been considered as ‘selectively neutral’ from evolutionary point of view. Most cancer victims are at the post reproductive stage of their life, but juveniles may also be affected. In particular, the proposal that lethal cancers in juveniles played a significant role in the evolution of complex animals was put forward, by Graham in the Cancer Selection hypothesis (1992). He postulated that cancer killed numerous juveniles and concluded that the resulting accumulation of defences against the disease promoted the emergence of complexity and hence development of complex animals from simpler multicellular ancestors. The presence of neoplasias across the animal kingdom from molluscs to mammals highlights the ancestry of cancer as a disease (O’Connell 2010). This hypothesis thus tightly intertwines cancer with evolution and life history, and specifically proposes that cancer is a by-product of novel adaptation – a standpoint gaining considerable support from data and most recently reviewed by Zimmer (2007). It is highly possible that an anticancer adaptation (cancer selection in action) has evolved for survival in many environments (*e.g.* dark skin in

Africans, Jablonski and Chaplin (2010)). Cancer selection can be summarised as selection to prevent/ postpone deaths due to cancer and it is a key line of investigation in current cancer research.

To maintain the genetic stability and survival of an organism, certain mechanisms exist that function in co-ordination, for example DNA damage pathways, signal transduction mechanisms (*e.g.* Wnt signalling pathway), DNA mismatch repair pathways to mention but a few (O'Driscoll, 2008). These and all other pathways should work in perfect coordination for the survival of the organism and to reduce deleterious mutations. If one or more genes malfunction, uncontrolled or cancerous growth can result. Organisms that have survived must have evolved "anticancer mechanisms" or 'tumour suppressor genes through the process of adaptive evolution (O'Connell, 2010). Our objective is to analyse empirical cancer data from completed genomes for signatures of this adaptive evolution and to use this as a proxy for evidence of cancer selection. In particular, we want to contrast medically relevant species, specifically mouse and human and their ancestral lineages for evidence of lineage specific variations that may lead to better modelling of human colon cancer in rodents (Massaad *et al.* 1992, Chen *et al.* 2012).

In this chapter, we would thus like to know - if mouse CRC orthologs[68] have undergone any positive selection, as compared to humans, (mice have higher metabolic rates and have undergone 180 million years (*MY*) of independent evolution). Given the strong tendency for using murine models for human disease, a positive answer to this question means that such models for colon cancer may be less reliable. Therefore this analysis has predictive power in determining which models mimic the human disease more closely. In addition to answering this question, (which is important from an evolutionary medicine perspective in order to access instances where mouse may not be effective model for the human disease phenotype), analysis of all other lineages may be used to frame these results in the context of all mammals. In consequence, in this study, not only the human and mouse lineages, but all lineages leading to extant species in our new dataset are analysed, which allows us to gain a greater understanding of the level of lineage-specific functional shift that has occurred in CRC associated genes. Moreover, it

[68] genes in different species which originated from a common ancestor , **often** with similar function i.e. genes resulting from speciation events.

allows us to site these hereditary mutations and their relationships within the broader context of factors initiating and progressing the disease.

4.1.1 Genes, Pathways and Colon cancer

As discussed in Chapter 3, cancer occurs due to changes in genetic and epigenetic factors and the interactions between them. Genetic changes can be of two types - somatic mutations (causing sporadic cancers) or germline mutations (causing hereditary cancers) that could have been triggered due to epigenetic events (Toyota and Suzuki 2010). Advances in genome technology have now made it possible to screen the entire cancer genome for cancer specific/linked mutations with the hope of ultimately gaining a better understanding of cancer as a genetic disease. However, studies have also suggested that a large number of infrequently mutated cancer genes function within a relatively small number of signalling pathways that cooperate (by aberrant activation/deactivation) to induce disease(s). This means that, it is easier to target the signalling pathways themselves rather than the individual gene/ protein (Frank 2012). A recent finding has suggested that, Wnt signalling pathways behave aberrantly due to epigenetic deregulation of pathway inhibitors (Costa *et al.* 2010), which could be true for other signalling pathways as well. It is possible that the deregulation in pathways can then trigger mutation, which can then lead to uncontrolled cell growth.

For the evolutionary analysis carried out in this chapter, germline mutated colon cancer genes have been chosen for investigation (*StatEpigen* database also holds CRC data predominantly currently) which are described in more detail in Materials and methods section (section 4.2). The aim of this chapter is to test the hypothesis that genes involved in hereditary colorectal cancers have signatures of adaptive evolution. This results chapter has been published (Morgan *et al.* 2012).

4.1.2 Hereditary forms of CRC

Colon cancer can be hereditary (around 20% of cases) or non-hereditary (sporadic) accounting for around 80% of cases, (Fearnhead *et al.* 2002). Both HNPCC (Hereditary Non-Polyposis Colorectal Cancer, also known as Lynch Syndrome) and FAP (Familial Adenomatous Polyposis) are examples of hereditary colon cancer with HNPCC alone accounting for 3% of all colon cancer cases, (Strate and Syngal 2005). HNPCC is a hereditary predisposition for developing cancer of the colon and endometrium, together with other organs. The genes found to be linked with HNPCC are: *MLH1*, *PMS2*, *MSH2*, *MSH6*, and *PMS1*, all of which are members of the MMR DNA repair pathway (Strate and Syngal 2005). Missense mutations of the mismatch repair gene *MLH1* have been identified in patients with HNPCC. A large number of *MLH1* alterations are located in the C-terminal domain that is responsible for constitutive dimerization[69] together with the *PMS2* gene (Kosinski *et al.* 2010). Other studies have shown MSI to be the molecular fingerprint of a deficient mismatch repair system and that approximately 15% of CRCs display MSI owing to the epigenetic silencing (Section 1.1.1) of *MLH1* or a germline mutation in one of the mismatch repair genes, *MLH1*, *MSH2*, *MSH6* and *PMS2* (Vilar and Gruber 2010). The mismatch repair endonuclease *PMS2* is known to interact with *MLH1* and is a component of the postreplicative DNA mismatch repair system (MMR). *PMS2* is recruited to cleave damaged DNA, this recruitment being triggered by the binding of *MSH2* and *MSH6* proteins to dsDNA mismatches followed by the recruitment of *MLH1* (Figure 4.1). *PMS1* is also involved in the repair of DNA mismatches, and it can form heterodimers with *MLH1*.

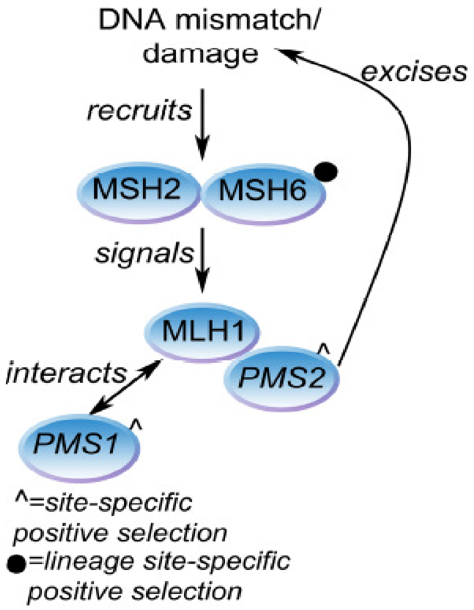
DNA mismatch/ damage
excises
recruits
MSH2
MSH6
signals
MLH1
interacts
PMS2
PMS1
^=site-specific positive selection
●=lineage site-specific positive selection

Figure 4.1: DNA mismatch repair pathway showing interaction between MMR genes. MHS6 combines with MSH2 to form the active protein complex that is capable of mismatch recognition. MLH1 acts as a heterodimer in conjunction with PMS2, PMS1 (Taken from Figure 1, Morgan *et al.* 2012).

[69] The chemical reaction joining two molecular subunits, resulting in the formation of a single dimer.

Other hereditary CRC syndromes are known, such as FAP, AAPC, which are associated with defects in DNA mismatch repair genes as well as tumour suppressor genes (Strate and Syngal 2005).

4.1.3 Mouse Modelling for Human CRC

Modelling using the mouse as reference for different human diseases and drug treatment has a considerable history (*e.g.* Hanna *et al.* 2007, Section 1.2). Mouse models of colorectal and intestinal cancers do exist and are defined as experimental systems in which mice are genetically manipulated or challenged with chemicals to develop malignancies in the gastrointestinal tract. There include mouse models used for HNPCC, FAP, and inflammation related colon cancer (Baker *et al.* 1996, Groden *et al.* 1991, Berg *et al.* 1996). There is literature that compares mouse and human with regard to several other aspects of cancer (*e.g.* Massaad *et al.* 1992 and many others), but comparison of the molecular evolution of colon cancer orthologs has, until our publication, Morgan *et al.* (2012), not been documented.

Mouse has many desirable features to be a 'disease model' (Section 1.2). At the genomic level, mouse protein coding sequences share 78.5% sequence identity with their human counterparts (Waterston *et al.* 2002). With such high levels of sequence identity, it may seem reasonable to expect that many orthologs between mouse and human would have conserved functions between the two species. However, this assumption may have failed to consider the probability for divergence due to ~180 million years (*MY*) of independent evolution, (Benton and Donoghue 2007). Another issue is the difference between the species in terms of their germline generation times, and indeed within the mouse species in terms of their natural versus artificial reproductive/germline generation times. Mice are bred in labs at faster rates than their wild cousins. It has been argued that such manipulations may have altered the evolutionary trade-off faced by mice, so that they are rewarded for investing energy in growing quickly and reproducing rapidly (Zimmer 2007). Therefore, artificial selection may be selecting negatively with respect to cancer defences (Zimmer 2007).

In particular, some cancer related genes such as Fatty Acid Synthase (*FAS*), and the entire *BRCA*/Fanconi anaemia pathway have undergone intense evolutionary changes in human, making them significantly different from their counterparts in mice, (O'Connell and McInerney 2005, O'Connell 2010). A further example of ortholog divergence between human and mouse is the *TDP1* gene, (required in Topo1-DNA complex repair with protein sequence similarity of 81%). A point mutation from an A-> G at position 1478 in human *TDP1* is linked to a disorder known as *SCAN1* (resulting in cerebellar atrophy and peripheral neuropathy). However, this mutation in mice does not result in the same condition/phenotype (Hirano *et al.* 2007). Similarly, *BRCA1* is known to be heavily associated with breast cancer in humans, with *BRCA1*$^{+/-}$ women having a 50% risk of developing breast cancer, while *BRCA*$^{+/-}$ mice do not show increased susceptibility to this cancer, (Hakem *et al.* 1996). Also, specific mutations in any of the genes *BCL10, PKLR* and *SGCA* in humans result in disease, but the same mutations in the mouse homologs do not result in phenotypic change to a disease state, (Gao and Zhang 2003). These clearly observed differences in phenotype could potentially be the result of protein functional shifts (adaptive evolution) in cancer-associated genes in either the human or mouse lineages (or indeed their ancestral lineages).

Given these observations, the specific phylogenetic study presented here aimed to address the research question as mentioned in Page 44.: "Can mouse be used as a model organism for human colon cancer, despite undergoing 180 *MY* of independent evolution?" .

4.1.4 Theory of Molecular evolution

To set the scene for the practical study, we briefly elaborate on the theory of molecular evolution and the role played by positive selection. The neutral theory of molecular evolution, (Kimura 1968) states that the vast majority of evolutionary changes at the molecular level are caused by random drift of selectively neutral mutants. It was received by some as an argument against Darwin's theory of evolution by natural selection but Kimura (1983) maintained that the two theories are compatible. It is emphasised that, the theory does not deny the role of natural selection in determining the course of adaptive evolution (Kimura 1986). However, the theory attributes a large role to genetic drift and neglects other important forms of selection that occur in genomes, *e.g.* variation of selective pressure over time or punctuated selection. The development of the nearly neutral theory to accommodate the acceptance and spread of slightly deleterious mutations in a population was an important step forward in the understanding of molecular evolutionary patterns (Ohta 1973, Ohta and Gillespie 1996). The neutralist – selectionist debate continued for almost last half century but during the last couple of decades, remarkable progress has been made in the study of molecular evolution, due to the development of new statistical methods, advances in computational technology and availability of sequence data. In particular, phylogenetic analysis of DNA or protein sequences has become a powerful tool for study of molecular evolution. This has enabled quantification of adaptive evolution that helped resolve the debate. Adaptive evolution is the modern synthesis of the process that Darwin originally described in his observations of adaptive radiations of bird morphologies (Darwin 1859).

Positive selection is generally described as retention and spread of advantageous mutations throughout a population (Yang 1998) and has long been considered synonymous with protein functional shift. Recent studies using these codon models of evolution in an ML (Maximum Likelihood) framework have combined evolutionary predictions of positive selection with biochemical verification of functional affects of these substitutions (Levasseur *et al.* 2006, Moury Simon 2011, Loughran *et al.* 2012), and thus support the link between positive selection and protein functional shift. Details on this are also given in (Morgan *et al.* 2012) which publishes this phylogenetic study. Some important driving forces for positive selection are external mechanisms, such as adaptation to different ecological niches, and response to disease, along with more internal mechanisms, such as co-evolution and compensatory mutations. All of these driving forces are relevant to analyses being carried out in fundamental CRC research (*e.g.* MacColl 2011).

Mathematically, at the molecular level, the ratio of nonsynonymous (amino-acid altering) substitutions per nonsynonymous site (*dN*) to synonymous (silent) substitutions per synonymous site (*dS*) is known as ω, and indicates the selective pressure at work in that sequence. If $\omega > 1$ it signifies positive selective pressure, $\omega = 1$ signifies neutral evolution, while $\omega < 1$ indicates purifying selective pressure, (Yang 1998).

$$i.e.\quad \omega = \frac{dN}{dS} = \frac{Nonsynonymous\ Substitutions\ per\ Nonsynonymous\ Site}{Synonymous\ Substitutions\ per\ Synonymous\ Site} \qquad \text{… Eq. 4.1}$$

$\omega > 1$	=> Positive selection / Adaptive Evolution
$\omega = 1$	=> Neutral Evolution
$\omega < 1$	=> Purifying/Negative Selective Pressure

An $\omega > 1$ means that nonsynonymous mutations offer a fitness advantages to the protein (and its carrier) and have higher fixation probabilities than synonymous mutations (synonymous mutations in this context are assumed to be a measure of the background rate of genetic drift). Previous studies (Arbiza *et al.* 2006, Kosiol *et al.* 2008) have assessed the level of positive selection present in mammal genomes under a Bayesian framework and estimated 5%-9% of genes in mammals are under positive selection, this provides us with a reference or “expected level” of positive selection for our analysis of mammal CRC genes.

In this chapter we apply a Maximum Likelihood method based on codon models of evolution to assess the selective pressures across our dataset (Yang 2007). These methods are far more robust than alternatives such as the sliding window approach (Schmid and Yang 2008). However, they do suffer from limitations and have strict criteria in terms of dataset size for statistical robustness (Bush 2001, Wong *et al.* 2004).

Another feature of sequence evolution that can negatively impact on a selective pressure analysis is recombination. In previous work by Anisimova *et al.* (2003), to evaluate the robustness of the Likelihood Ratio Tests (LRTs) simulations were performed, the results showed that type 1 error rates can be up to 90% with relatively high rates of recombination in protein coding sequences resulting in the misinterpretation of recombination as positive selection. We have incorporated a test for recombination for all genes in the dataset prior to the ML selective pressure analysis as described below.

4.2 Materials and Methods

This section explains materials/ resources used and different methods implemented for the required analyses under different sub-sections as follows.

4.2.1 Selection of germline mutated genes and their orthologs

For the investigation, 22 germline-mutated genes, known to be involved in the onset of CRC were selected from the Cancer Gene Census at the Sanger Institute (Futreal *et al.* 2004). The complete list of genes used for this analysis is provided in Table 4.1 (along with their corresponding Ensembl Identifier, the associated cancer syndrome, the tumour types observed, the pathways they are involved in and the corresponding references). Some of these genes and their association with HNPCC are described in Section 4.1.2 in more detail.

Other genes, *APC* (Syndrome: FAP, pathway: *APC*), *CDH1* (Syndrome: Familial gastric carcinoma, pathway: *APC*), *VHL* (Syndrome: Von Hippel-Lindau Syndrome, pathway: HIF1), *TP53* (Syndrome: Li_Fraumeni syndrome, pathway: p53), *STK11* (Syndrome: Peutz-Jeghers syndrome, pathway: PI3K), *PTEN* (Syndrome: Cowden syndrome, pathway: PI3K), are some examples of tumour-suppressor genes. These genes when inactivated, can cause cancer. Mutation/ alteration of stability genes on the other hand, promotes tumorigenesis by removing the maintenance of stability, namely loss of control of pathways such as MMR (mismatch repair), NER (Nucleotide-excision repair) and BER (Base excision repair). Gene *MUTYH* (causing Attenuated polyposis) is involved in the BER pathway while *MSH6* gene (causing HNPCC) is involved in the MMR (DNA mismatch repair) pathway (Table1, Vogelstein and Kinzler, 2004). Some of the pathways that are related to colorectal cancer are involved in DNA damage control pathways and the Wnt pathway (*APC* gene mutation). In general, most of the cancer genes can be grouped into 12 critical pathways including: apoptosis, DNA damage control, invasion, cell cycle signalling, KRAS signalling, and TGF-beta signalling amongst others (Vogelstein and Kinzler 2004).

Gene *TP53* is found to be involved in *P53* signalling pathway (cellular DNA damage response pathway) and acts as a hub-protein (Liu and Kulesz-Martin 2001). Some genes, such as *TP53*, *KRAS*, and *APC* are involved in cancer with great frequency, up to 100% of cancers; these are referred in certain literature to as "mountains" (Wood *et al.* 2007). Thousands of other genes are involved in cancer but are found at very low frequency, fewer than 5%; these genes are referred to as "hills" (Wood *et al.* 2007).

Collectively, however, the hills are also required for, and in some cases drive, the carcinogenesis (Wood *et al.* 2007).

Single gene orthologs were identified for these 22 genes across a set of high coverage (>6X coverage) vertebrate genomes (using Compara data from *Ensembl*) (Hubbard and Barker 2002, Hubbard and Andrews 2005). The 21 species analysed were selected based on the genome coverage. These included representatives from 3 of the 4 main lineages of Eutheria (placental mammals), namely Afrotheria (*e.g.* elephants and manatees), Euarchontoglires (includes Primates, Rodents and Glires), and Laurasiatheria (*e.g.* cows, horses, and bats), as well as outgroup species such as platypus, zebrafish, and zebra finch. Details for the complete dataset (genes and species used) are provided in Appendix II, Table A. Here, the black boxes imply that, the corresponding orthologous gene is not available in Ensembl database for that species.

Table 4.1: List of colon cancer genes used and their associated disease and genomic features

Gene	Ensembl Identifier	Taxa Number	Alignment Length	Syndrome	Tumour Types Observed	Pathway(s)	References
APC	ENSG00000134982	20	9177	Familial adenomatous Polyposis (FAP)	Colon, thyroid, stomach, intestine	APC	[Vogelstein and Kinzler 2004, Markowitz and Bertagnolli 2009]
ATM	ENSG00000149311	18	9189	Ataxia telangiectasia (A-T)	Leukaemia, lymphoma, colorectal	CIN	[Vogelstein and Kinzler 2004]
BHD	ENSG00000154803	20	1737	Birt-Hogg-Dube syndrome	Renal, colon	AMPK, mTOR, STAT	[Vogelstein and Kinzler 2004, Toro *et al.* 2008]
BMPR1A	ENSG00000107779	19	1596	Juvenile polyposis	Gastrointestinal	SMAD	[Vogelstein and Kinzler 2004]
CDH1	ENSG00000039068	15	2649	Familial gastric carcinoma	Stomach	APC	[Vogelstein and Kinzler 2004]
MADH4	ENSG00000141646	16	1656	Juvenile polyposis	Gastrointestinal	SMAD	[Vogelstein and Kinzler 2004, Markowitz and Bertagnolli 2009 (SMAD4)]
MET	ENSG00000105976	21	4146	Hereditary papillary renal cell carcinoma (HPRCC)	Kidney, colorectal	RAS, PI3K, STAT, Beta-catenin, Notch	[Vogelstein and Kinzler 2004, De Oliveira *et al.* 2009]
MLH1	ENSG00000076242	19	2274	Hereditary non-polyposis colon cancer (HNPCC)	Colon, uterus	MMR	[Vogelstein and Kinzler 2009, Markowitz and Bertagnolli 2009]
MSH2	ENSG00000095002	18	2802	Hereditary non-polyposis colon cancer (HNPCC)	Colon, uterus	MMR	[Vogelstein and Kinzler 2004, Markowitz and Bertagnolli 2009]

MSH6	ENSG00000116062	19	4101	Hereditary non-polyposis colon cancer (HNPCC)	Colon, uterus	MMR	[Vogelstein and Kinzler 2004, Markowitz and Bertagnolli 2009]
MUTYH	ENSG00000132781	21	1569	Attenuated Polyposis	Colon	BER	[Vogelstein and Kinzler 2004, Markowitz and Bertagnolli 2009]
NF1	ENSG00000196712	17	8523	Neurofibromatosis type I	Neurofibroma, colon	RTK	[Vogelstein and Kinzler 2004, Cacev *et al.* 2005]
PMS1	ENSG00000064933	20	2799	Hereditary non-polyposis colon cancer (HNPCC)	Colon, uterus	MMR	[Päivi *et al.* 2001]
PMS2	ENSG00000122512	21	2592	Hereditary non-polyposis colon cancer (HNPCC)	Colon, uterus	MMR	[Vogelstein and Kinzler 2004]
PTEN	ENSG00000171862	18	1209	Cowden syndrome	Hamartoma, glioma, uteru, colorectum	PI3K	[Vogelstein and Kinzler 2004, Markowitz and Bertagnolli 2009]
SDHB	ENSG00000117118	18	840	Hereditary paraganglioma, Carney–Stratakis	Paragangliomas, pheochromocytomas, gastrointestinal	HIF1	[Vogelstein and Kinzler 2004, Pasini *et al.* 2007]
SDHC	ENSG00000143252	16	507	Hereditary paraganglioma, Carney–Stratakis	Paragangliomas, pheochromocytomas, gastrointestinal	HIF1	[Vogelstein and Kinzler 2004, Pasini *et al.* 2007]
STK11	ENSG00000118046	18	1320	Peutz-Jeghers syndrome	Intestinal, ovarian, pancreatic, colorectal	PI3K	[Vogelstein and Kinzler 2004, Slattery *et al.* 2010]

TP53	ENSG00000141510	16	1185	Li-Fraumeni syndrome/sarcoma	Breast, sarcoma, adrenal, brain, colorectal	p53	[Vogelstein and Kinzler 2004, Markowitz and Bertagnolli 2009, Slattery *et al.* 2010]
TSC1	ENSG00000165699	18	3495	Tuberous sclerosis	Hamartoma, kidney, colorectal	PI3K	[Vogelstein and Kinzler 2004, Slattery *et al.* 2010]
TSC2	ENSG00000103197	19	5436	Tuberous sclerosis	Hamartoma, kidney, colorectal	PI3K	[Vogelstein and Kinzler 2004, Slattery *et al.* 2010]
VHL	ENSG00000134086	18	639	Von Hippel-Lindau syndrome	Kidney, colorectal	HIF1	[Vogelstein and Kinzler 2004, Giles *et al.* 2006]

[1] HGNC code

Table 4.1 Detail on each of the 22 genes analyzed, HGNC approved gene symbols and Ensembl gene unique identifiers (IDs). The total number of species analyzed for each gene, the overall length of alignment in base pairs, the syndrome, tumour type observed and pathway involved are also given. Additionally, references citing alternative gene names are also identified.

4.2.2 Multiple Sequence Alignment (MSA)

The coding DNA sequences of the single gene orthologs were translated and the resulting amino acid sequences were aligned using the default parameters in ClustalW v2.0.12 (Chenna *et al.* 2004, Larkin *et al.* 2007). Using in-house software ('Mapgap': written in perl), both the original nucleotide sequence files for the orthologous families and the translated MSA were taken and gaps added to a DNA alignment file based on the position in the amino acid alignment. All alignments were reviewed for quality and any poorly aligned regions were manually edited using Se-Al (Rambaut 1996). Se-Al does not offer any automation of the editing of alignments.

4.2.3 Selective pressure analysis using codon models of evolution

Selective pressure analyses were performed using Codeml from PAML version 4.4 (Yang 1997, Yang and Wong 2005). PAML (Phylogenetic Analysis by Maximum Likelihood) is a package of programs for phylogenetic analyses of DNA or protein sequences using maximum likelihood (maintained and distributed freely for academic use by Ziheng Yang). For our purpose, codonml[70](codon-based analysis, codeml for codons with seqtype = 1) is used, the main difference between aaml and codonml being the unit of evolution in the Markov model, referred to as a 'site' in the sequence, is a codon in case of codonml while this is an amino acid for aaml. However, both codonml and baseml use similar algorithms to fit models by ML (Yang 2007).

The aligned nucleotide file from Section 4.2.2 (in FASTA file format) is converted to (readseq.jar used) Phylip file format and made ready (with some manual editing as required by PAML) as an input file for selective pressure analysis using PAML. PAML also needs a corresponding tree for each Phylip formatted nucleotide alignment to be tested. In this case, pruned species phylogenies were used, as each gene family analysed was composed of single gene orthologs, (Benton and Donoghue 2007, Murphy and Eizirik 2001). This approach has been taken in PAML analyses in the past (O'Connell *et al.* 2010). To generate the appropriate pruned species tree for each alignment (the

[70] This software assumes that the sequences are pre-aligned codons, the sequence length is an exact multiple of 3, and the first nucleotide in the sequence is codon position 1. Introns, spacers and other noncoding regions must be removed and the coding sequences must be aligned before running the program (PAML documentation).

species present varies slightly from gene to gene), a complete tree is initially generated with 21 species involved, using nested parenthesis using treeview software[71]. This is then pruned (using Newick Utilities[72]) according to the orthologs present in each nucleotide alignment. For example, for *CDH1* gene, the pruned tree would look like:

((((((((Human,Chimpanzee),Gorilla),Orangutan),Marmoset),(((Rat,Mouse),Guinea Pig),Rabbit)), (((Cow,Pig),Horse),Dog), Elephant),Opossum),Platypus);

Criteria for Selective Pressure analysis: To increase the statistical power of the analysis performed, only single gene families containing 6 or more taxa and lengths of greater than 500 amino acids, were considered for further analysis. Studies have indicated that sequence length, taxonomic representation and divergence depths, all have an impact on the power to infer positive selection (Anisimova *et al.* 2001, Zhang *et al.* 2005). For our analysis, one of the genes, SDHD, is found to have only 3 orthologs, so this gene was not suitable for further analysis.

With the sequence data file (in Phylip file format) and treefile (in Newick format), a pipeline of analyses was performed using in-house software (under review). This software creates (for every model and every starting point on the likelihood plane) an appropriate control file, with the appropriate name of the nucleotide alignment data file, tree structure file and models parameters and specifications for the analysis, and it does this automatically for each gene family we wanted to test. The software also creates folders and subfolders for all site-specific models and lineage-site specific models (Appendix II, Table B) to be used in PAML analysis so that the output files can be retrieved directly to the respective gene family folder per model. In short, the software is used to prepare folders for all input files before analysis and to direct the processed output to appropriate folders after the selective pressure analyses are completed by PAML, and it also performs all appropriate LRTs on the resultant model output. This software was designed to reduce the scope for human error in PAML analysis, to automate an otherwise complex, involved procedure and to accommodate large-scale selective pressure analyses such as the one carried out in this chapter.

[71] http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
[72] http://cegg.unige.ch/newick_utils

Codeml (Codonml here) implements a number of codon-based models in a Maximum Likelihood framework that can be used to test alternative and nested evolutionary hypotheses. Three different types of codon model were used in this study: (i) a homogeneous model (Model 0) - a single ω-value is estimated for the entire alignment; (ii) site-heterogeneous models - the sites of the gene sequence are grouped into two or more site classes (K parameter), each with its own ω-value estimate; and (iii) lineage-specific heterogeneous models - a different ω parameter is estimated for different site classes in combination with different lineages (Yang 1997, Yang 2007, Murphy *et al.* 2001). Seven site-heterogeneous models were used, conventional annotations were retained for these models: Model 1 (Neutral), Model 2 (Selection), Model 3 Discrete (K[73]=2), Model 3 Discrete (K=3), Model 7, Model 8 and Model 8a. Two lineage-specific heterogeneous models were used: Model A and Model A Null. The complete set of models used is given in Appendix II, Table B.

The goodness-of-fit of the different models is assessed statistically using a likelihood ratio test (LRT). The LRT compares the log-likelihoods of a null model with the alternative model. For hierarchically nested models, the test statistic of an LRT approximates the χ^2 distribution with degrees of freedom equal to the number of additional free parameters in the alternative model compared to the null model. Because of this, the critical value of the test statistic can be determined from standard statistical tables. If the p-value of the test statistic exceeds that critical value (*i.e.* if the alternative model fits the data significantly better than the null model), then the null model can be rejected. For example, if the test statistic of an LRT comparing Model 1 (Neutral) with Model 2 (Selection) is greater than the critical value determined from the χ^2 distribution, Model 1 can be rejected. If $\omega_1 > 1$ under Model 2, positive selection may be inferred. The set of codon models used in this analysis and LRTs used for analysis is provided in Appendix II, Table B and Table C respectively.

The branch/lineage –specific models allow ω to vary among branches in the phylogeny and are designed for detecting positive selection acting on particular lineages (Yang 1998; Yang and Nielsen 1998). The site models allow ω to vary among sites (among codons or amino acids in the protein) (Nielsen and Yang 1998; Yang *et al.* 2000b). A

[73] Site class (Used explicitly in discrete models M2 and M3. However, For Model M0, K=1; for M1 (Neutral) and M3(Discrete K=2), it is 2, for M2(Selection) and M3(Discrete K=3), it is 3, for M7(β), K=10 and for M8 (β and ω >1) and M8a (β and ω =1), K=11 (Yang and Swanson 2002).

number of such models are implemented in codeml. The branch-site models aim to detect positive selection that affects only a few sites on pre-specified lineages (Yang and Nielsen 2002). The branches under test for positive selection are called the "foreground" branches, while all other branches on the tree are the "background" branches. The branch-site model used for our analysis is model A. In the LRT, branch-site model A is an extension of M1 (Test 1 in Zhang *et al.* 2005) and so it is compared to M1 and also to ModelA null (null model of Model A).

In cases where positive selection is inferred, the posterior probability of a site belonging to the positively selected class is estimated using two calculations: Naïve Empirical Bayes (NEB) or Bayes Empirical Bayes (BEB). If both BEB and NEB are predicted, the BEB results are used preferentially as have been shown to be more robust and account for sampling errors in the ML estimates of parameters in the model (Nielsen and Yang 1998, Yang, Wong *et al.* 2005).

PAML output files were then parsed for parameter estimates and log likelihood values and LRTs were performed. Where positively selected sites were inferred under a given model, these were mapped to the sequence(s) of interest and included in the summary file (Table 4.2). Functional annotation of sites under positive selection for each protein was obtained from UniProt (UniProt, 2011).

Recombination analysis also was carried out using GENECONV (version 1.81a) (Sawyer 1989), as described in the introduction to this chapter the presence of recombination can lead to false positive results in a codeml analysis. Additionally, an analysis of human population data was carried out to determine if the selective pressures evident at the between species level were also evident in modern human populations (Morgan *et al.* 2012), *i.e.* if these positively selected amino acids have gone to fixation in the human lineage or if there is still large amounts of variation in that position.

4.3 Results and Discussion

For recombination analysis, the results are summarized here and the complete set of results is provided in Appendix II, Table D. Only the *TP53* protein showed significant levels of recombination. The regions where recombination was present were compared to regions where positive selection was detected. If these regions overlapped - the positive selection result was deemed a false positive.

For determining selective pressure variation, both site- and lineage-specific selective pressure analyses were performed and the statistical significance of all results via LRT analyses were assessed to ascertain the codon evolutionary model of best fit. For cases with very high value, the notation, $\omega >> 1$ is used throughout the manuscript, as there is no biological significance for these extremely large ω values (the precise numbers are shown in the codeml results Tables 4.2, for positive results). As the lineage-specific analyses are more pertinent to the main question being addressed in this chapter, the lineage-specific results have been described in detail in the following section. Site-specific results are briefly summarised. The model of best fit along with associated parameter estimates are described and a summary table for all estimates for each of the 22 genes is given in Additional file 4 of Morgan *et al.* 2012.

Analysis of these orthologous datasets revealed significant levels of positive selection. A total of 17 of 22 genes were found to be under positive selection in site and/or lineage specific positive selection analyses (all positive results are summarized in Table 4.2). Of these positive selected genes, 14 are lineage-specific, 5 genes are found to be site-specific and 2 of these (*CDH1* and *MUTYH*) are both site and lineage-specific. The genes found to be positively selected in different functionally important positions include: *CDH1*, *PMS1*, *PMS2*, *MUTYH* and *TP53* (Table 4.2).

Lineage-specific models of codon evolution were assessed at multiple phylogenetic depths, (i) the extant lineages within the Euarchontoglires clade, and (ii) all ancestral lineages leading from the Euarchontoglires to modern mouse and human were also tested independently as depicted in Figure 4.2.

Analyses of the lineages within the Euarchontoglires clade resulted in significant evidence of lineage-specific positive selection, 6 genes in ancestral lineages and 12 in extant lineages (Figure 4.2 and Table 4.2). A total of 14 genes are found to be positively selected.

Lineage-specific positive selection was detected in the ancestral Euarchontoglires and Hominidae lineages for *STK11* and in the ancestral primate lineage for *CDH1*. The ancestral Muridae lineage had evidence of positive selection acting on the *TSC1* gene while the ancestral Murinae lineage showed evidence of positive selection for both *MSH6* and *SDHC*. The significant finding, therefore, is that for these two genes, *SDHC and MSH6,* the mouse and human orthologs are most likely to have different functions and therefore human cancers involving these genes may not be suitable for modelling in the mouse system.

Other evidence of positive selection in rest of the extant branches are provided in summary Table 4.2 and it is notable here that, neither of the extant human and mouse branches show any evidence of positive selection. In this table, the lineage-specific positive results for each lineage tested from the Euarchontoglires ancestor to modern lineages (see Figure 4.2) are shown in the top panel and the site-specific results are shown in the bottom panel. Parameter estimates and number of positively selected sites, BEB (Bayes Empirical Bayes) estimations, are also provided. The identified positively selected sites (final column) are separated by the posterior probability cutoffs of 0.50, 0.95 and 0.99.

The analysis is provided in context of their potential functional relevance for those genes and was carried out for all genes where functional sites and/or domains have been determined/characterised. All sites described were calculated via Bayes Empirical Bayes (BEB) analysis (unless otherwise specified). In all cases, the potential functional importance of residues based on their sequence position was assessed. There are instances where stretches of protein sequence under positive selection are identified and there is a possibility that these regions may have very different functions despite their sequence position. The corresponding alignments are provided along with the published paper in Additional file 2 (The alignments are in nexus file format, 480 pages in word, hence not included in this thesis but available at http://www.biomedcentral.com/content/pdf/1471-2148-12-114.pdf). The complete set of model estimates for the entire dataset are provided along with paper in Additional file 4 (141 pages in word, hence not included along with this thesis, but available at http://www.biomedcentral.com/content/pdf/1471-2148-12-114.pdf).

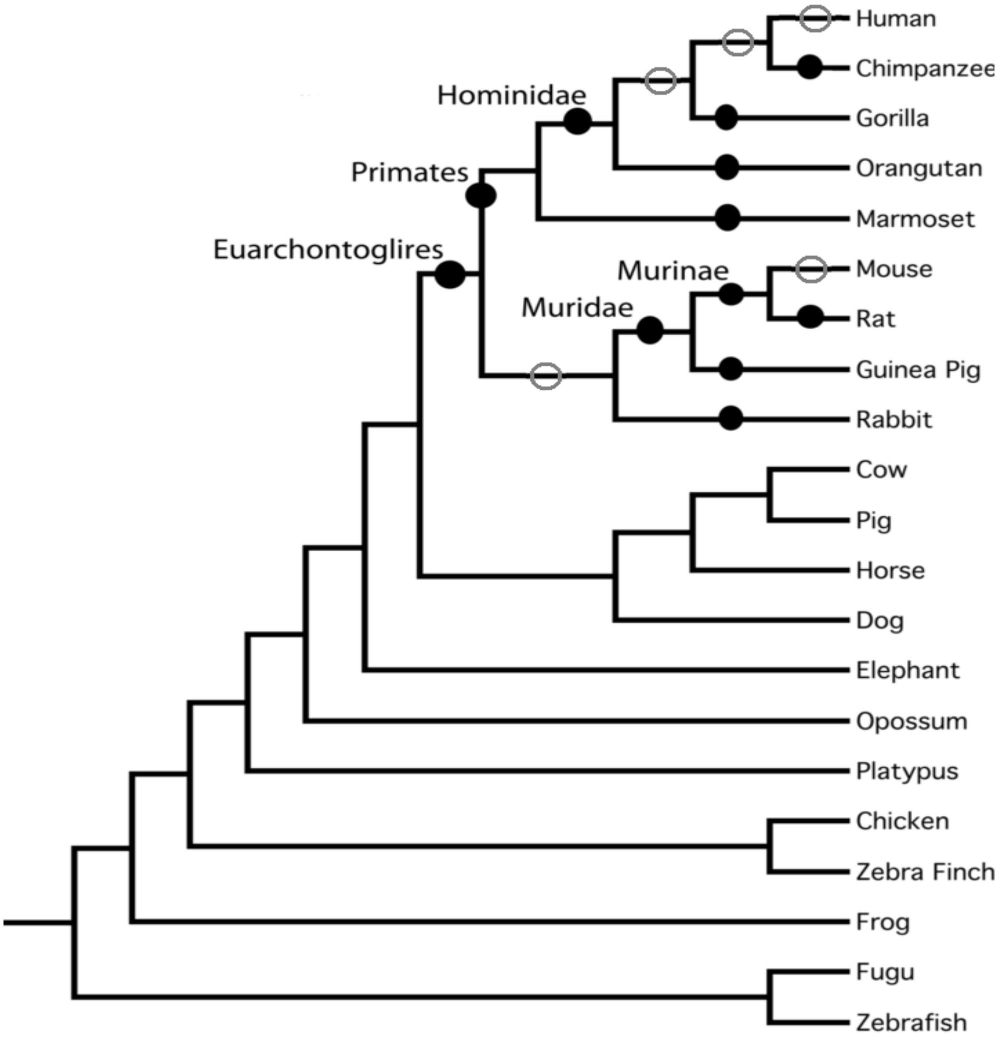

Figure 4.2: Phylogeny of animal species used. The ancestral lineages tested in the analysis are labelled with their corresponding names (as used throughout the text). Those lineages where positive selection was detected are labelled with filled circles while those lineages tested and producing no evidence of positive selection are denoted with an empty circle.

Table 4.2: Summary of parameter estimates and likelihood scores for the model of best fit showing evidence of positive selection.

Gene	Model	lnL	Parameter Estimates	Positive Selection	BEB Positively Selected Sites
			Lineage-Specific Analyses		
Euarchontoglires Ancestral Branch					
STK11	modelA	-8602.921472	p_0=0.93299, p_1=0.05633, p_2=0.01007, p_3=0.00061 ω_0=0.03346, ω_1=1.00000, ω_2=197.90897	Yes	3>0.50, 1>0.95, 0>0.99
Primate Ancestral Branch					
CDH1	modelA	-16658.03484	p_0=0.75454, p_1=0.23453, p_2=0.00834, p_3=0.00259 ω_0=0.05683, ω_1=1.00000, ω_2=10.20516	Yes	9>0.50, 1>0.95, 0>0.99
Hominidae Ancestral Branch					
STK11	modelA	-8601.056009	p_0=0.93574, p_1=0.05920, p_2=0.00476, p_3=0.00030 ω_0=0.03323, ω_1=1.00000, ω_2=44.31709	Yes	3>0.50, 2>0.95, 1>0.99
VHL	modelA	-4263.853291	p_0=0.73748, p_1=0.25109, p_2=0.00853, p_3=0.00290 ω_0=0.05985, ω_1=1.00000, ω_2=220.34533	Yes	1>0.50, 0>0.95, 0>0.99
Chimpanzee Extant Branch					
TSC2	modelA	-42659.27711	p_0=0.90352, p_1=0.09434, p_2=0.00194, p_3=0.00020 ω_0=0.04404, ω_1=1.00000, ω_2=190.09480	Yes	6>0.50, 2>0.95, 2>0.99
VHL	modelA	-4262.098043	p_0=0.73571, p_1=0.25251, p_2=0.00877, p_3=0.00301 ω_0=0.05976, ω_1=1.00000, ω_2=262.72662	Yes	3>0.50, 0>0.95, 0>0.99
Gorilla Extant Branch					
MSH2	modelA	-19485.4338	p_0=0.92233, p_1=0.06298, p_2=0.01375, p_3=0.00094 ω_0=0.06427, ω_1=1.00000, ω_2=999.00000	Yes	46>0.50, 34>0.95, 18>0.99
TSC2	modelA	-42569.22884	p_0=0.89862, p_1=0.08796, p_2=0.01222, p_3=0.00120 ω_0=0.04339, ω_1=1.00000, ω_2=999.00000	Yes	27>0.50, 14>0.95, 12>0.99
MSH6	modelA	-34009.90221	p_0=0.78382, p_1=0.18418, p_2=0.02591, p_3=0.00609 ω_0=0.06974, ω_1=1.00000, ω_2=999.00000	Yes	46>0.50, 34>0.95, 18>0.99
ATM	modelA	-69374.08393	p_0=0.80673, p_1=0.17971, p_2=0.01109, p_3=0.00247 ω_0=0.09745, ω_1=1.00000, ω_2=999.00000	Yes	48>0.50, 23>0.95, 19>0.99

Orangutan Extant Branch					
TSC1	modelA	-24068.71106	p_0=0.79963, p_1=0.18828, p_2=0.00978, p_3=0.00230 ω_0=0.08020, ω_1=1.00000, ω_2=999.00000	Yes	13>0.50, 6>0.95,5>0.99
TSC2	modelA	-42673.92339	p_0=0.90414, p_1=0.09295, p_2=0.00263, p_3=0.00027 ω_0=0.04433, ω_1=1.00000, ω_2=40.47366	Yes	9>0.50, 0>0.95, 0>0.99
Marmoset Extant Branch					
TSC2	modelA	-42616.04524	p_0=0.89841, p_1=0.09019, p_2=0.01035, p_3=0.00104 ω_0=0.04325, ω_1=1.00000, ω_2=235.10448	Yes	38>0.50, 9>0.95
MSH6	modelA	-34009.90221	p_0=0.78382, p_1=0.18418, p_2=0.02591, p_3=0.00609 ω_0=0.06974, ω_1=1.00000, ω_2=999.00000	Yes	45>0.50, 16>0.95, 12>0.99
VHL	modelA	-4262.443441	p_0=0.72045, p_1=0.22453, p_2=0.04195, p_3=0.01307 ω_0=0.05886, ω_1=1.00000, ω_2=90.26952	Yes	10>0.50, 0>0.95, 0>0.99
ATM	modelA	-69583.23068	p_0=0.81640, p_1=0.18148, p_2=0.00173, p_3=0.00038 ω_0=0.09939, ω_1=1.00000, ω_2=46.82466	Yes	2>0.50, 0>0.95, 0>0.99
Muridae Ancestral Branch					
TSC1	modelA	-24126.17894	p_0=0.80995, p_1=0.18416, p_2=0.00481, p_3=0.00109 ω_0=0.08293, ω_1=1.00000, ω_2=999.00000	Yes	1>0.59, 0>0.95, 0>0.99
Murinae Ancestral Branch					
SDHC	modelA	-3846.690164	p_0=0.87666, p_1=0.08131, p_2=0.03846, p_3=0.00357 ω_0=0.15340, ω_1=1.00000, ω_2=253.61375	Yes	9>0.50, 2>0.95, 1>0.99
MSH6	modelA	-34190.13821	p_0=0.79911, p_1=0.19671, p_2=0.00335, p_3=0.00082 ω_0=0.07057, ω_1=1.00000, ω_2=126.22513	Yes	3>0.50, 1>0.95, 0>0.99
Rat Extant Branch					
MADH4	modelA	-6092.186945	p_0=0.93360, p_1=0.01536, p_2=0.05021, p_3=0.00083 ω_0=0.01379, ω_1=1.00000, ω_2=102.33013	Yes	24>0.50, 11>0.95, 10>0.99
NF1	modelA	-37750.29866	p_0=0.96609, p_1=0.02476, p_2=0.00892, p_3=0.00023 ω_0=0.02265, ω_1=1.00000, ω_2=999.00000	Yes	39>0.50, 10>0.95, 10>0.99

Guinea pig Extant Branch					
TSC1	modelA	-24116.58577	p_0=0.80206, p_1=0.18611, p_2=0.00961, p_3=0.00223 ω_0=0.08093, ω_1=1.00000, ω_2=284.22603	Yes	9>0.50, 4>0.95, 0>0.99
NF1	modelA	-37849.50819	p_0=0.97375, p_1=0.02506, p_2=0.00116, p_3=0.00003 ω_0=0.02414, ω_1=1.00000, ω_2=171.64068	Yes	3>0.50, 1>0.95, 0>0.99
Rabbit Extant Branch					
MLH1	modelA	-19516.63525	p_0=0.80595, p_1=0.18541, p_2=0.00703, p_3=0.00162 ω_0=0.05262, ω_1=1.00000, ω_2=7.52747	Yes	5>0.05, 3>0.95, 0>0.99
MUTYH	modelA	-15911.6175	p_0=0.61027, p_1=0.37605, p_2=0.00846, p_3=0.00522 ω_0=0.07703, ω_1=1.00000, ω_2=998.99697	Yes	5>0.50, 4>0.95, 3>0.99
SDHC	modelA	-3822.683246	p_0=0.57771, p_1=0.06636, p_2=0.31926, p_3=0.03667 ω_0=0.12047, ω_1=1.00000, ω_2=3.59059	Yes	51>0.50, 10>0.95, 8>0.99
ATM	modelA	-69582.95152	p_0=0.81572, p_1=0.18045, p_2=0.00313, p_3=0.00069 ω_0=0.09930, ω_1=1.00000, ω_2=7.41594	Yes	6>0.50, 0>0.95, 0>0.99
BHD	modelA	-13523.51719	p_0=0.90728, p_1=0.05930, p_2=0.03137, p_3=0.00205 ω_0=0.02817, ω_1=1.00000, ω_2=6.50017	Yes	10>0.50, 7>0.95, 1>0.99
Site-specific Analyses					
CDH1	m8	-16589.88768	p=0.21848, p_0=0.99291, p_1=0.00709, q=0.80842 ω=4.53766	Yes	15>0.5, 1>0.95, 0>0.99
PMS1	m8	-26480.39761	p=0.61337, p_0=0.93580, p_1=0.06420, q=1.93110 ω=1.32691	Yes	25>0.50, 1>0.95, 0>0.99
PMS2	m8	-27449.3651	p=0.29104, p_0=0.91064, p_1=0.08936, q=1.31619 ω=1.28855	Yes	37>0.50, 1>0.95, 0>0.99
MUTYH	m8	-15797.6226	p=0.37255, p_0=0.97242, p_1=0.02758, q=1.00900 ω=2.44412	Yes	18>0.5, 1>0.95, 0>0.99
TP53	m8	-8688.19126	p=0.40362, p_0=0.94645, p_1=0.05355, q=1.77507 ω=1.97385	Yes	13>0.5, 3>0.95, 0>0.99

4.3.1 Positive selection in the Euarchontoglires Ancestral branch

Euarchontoglires ancestral branch (the most ancestral branch tested) resulted in a single gene with signatures of positive selection - *STK11* (Serine/Threonine-protein kinase 11). *STK11* plays an essential role in G1 cell cycle arrest and acts as a tumour suppressor. It phosphorylates and activates members of the AMPK related subfamily of protein kinases (Baas *et al.* 2003, Boudeau *et al.* 2003). Mutations in *STK11* cause Peutz-Jeghers syndrome (PJS), a rare autosomal dominant disorder characterized by multiple gastrointestinal hamartomatous polyps and an increased risk of various neoplasms including gastrointestinal cancer (Hemminki *et al.* 1998, Nakagawa *et al.*). From the literature we currently know of 17 sites across this gene that when mutated are associated with colon-cancer. The Euarchontoglires ancestral lineage has 1.1% of sites under positive selection ($\omega > > 1$). Position 206 with a PP (Posterior Probability) = 0.889 is a hydrophobic Alanine or Valine in Euarchontoglires species or a negatively charged Glutamic acid or positively charged Lysine in non-Euarchontoglires species. This residue also lies in close proximity to sporadic cancer site A205T and colorectal cancer site D208N in Human (Dong *et al.* 1998). Positively selected position 301 in Euarchontoglires (P = 0.885) is present in Euarchontoglires species as an Arginine residue and all non-Euarchontoglires as an uncharged Glutamine residue. Site 301 is close to R297K and region 303–306, both of which have been implicated in PJS (Westerman *et al.* 1999).

4.3.2 Positive selection in the Primate Ancestral branch

The branch leading from the Euarchontoglires ancestor towards the primates was analysed (termed as the ancestral Primate branch, Figure 4.2). The *CDH1* dataset consists of 15 taxa and following LRT analysis identified, lineage-specific positive selection in 1.1% of sites in the Primate Ancestor (ω=10.21). Positively selected sites were compared to human Swiss-Prot entry (P12830) that showed, position 604, with a PP of 0.549, falls in close proximity to gastric cancer variant R598Q (Berx *et al.* 1998). At position 604, Primates have a negatively charged Glutamic acid while non-primates have a polar uncharged Glutamine.

4.3.3 Positive selection in the Hominidae Ancestral branch

The Hominidae branch (Figure 4.2) also showed evidence of positive selection, in the *STK11* gene, in 0.51% of sites, or 3 positions, with ω > > 1 (Figure 4.3(a) and Table 4.2). These positions were then compared to the human Swiss-Prot sequence (Q15831). Position 347 represents a radical substitution, as the Hominidae code for an Alanine whereas the Murinae lineage encodes an Arginine at this position. For positively selected site 378, the ancestral Hominidae lineage encodes the polar residue Serine, while the closely related species studied encode Glycine. The functions of these specific sites have not been characterised so far in the literature.

A second gene showing evidence of positive selection in the Hominidae ancestral branch is the *VHL* dataset (18 taxa). The *VHL* gene encodes Von Hippel-Lindaue tumour suppressor protein. Mutations in this gene can result in von Hippel-Lindau disease (VHDL) - a dominantly inherited cancer syndrome (Latif *et al.* 1993). *VHL* exhibited weak evidence of positive selection with 1.1% of sites in the ancestral Hominidae lineage under positive selection. There was one amino acid that had low coverage in the alignment (present only in 6/18 species), so, it is not expanded further.

4.3.4 Positive selection in the Extant Primate branches

There is also evidence of positive selection in modern non-human primate lineages (Chimpanzee, Gorilla, Orangutan and Marmoset). For *VHL*, positive selection was detected in the Chimpanzee lineage with 1.2% of sites (ω > > 1), and also in the Marmoset lineage with 5.5% sites (ω > > 1). These positively selected sites were compared against human Swiss-Prot entry (P40337), however the region (1–60) was only represented by 11/18 species in the alignment and therefore we do not have sufficient confidence to explore these sites in more detail.

The *MSH6* gene (19 taxa) showed evidence of positive selection in both the Gorilla and Marmoset lineages each displaying 3.2% of sites (ω > > 1). Gorilla and Marmoset extant lineages were compared against human (P52701) Swiss-Prot entry. No relevant functional information could be extracted from positively selected sites in Gorilla, however 2/45 positively selected sites in Marmoset fall in close proximity to cancer variants. Positively selected site 803 (PP = 0.551) for Marmoset, coincides with CRC variants D803G (Kolodner *et al.* 1999) and V800A (Berends *et al.* 2002) in Human.

Position 803 in Marmoset is a Glutamic acid while in all other mammals it is Aspartic acid. Positively selected site 1099 in Marmoset (PP = 0.614) is located between human CRC variants R1095H (Kariola *et al.* 2003) and T1110C (Berends *et al.* 2002).

Mutation in *MSH2* results in HNPCC (Section 4.1.2). Lineage-specific positive selection was identified in 1.5% of sites within the extant Gorilla lineage (ω > > 1) for *MSH2*. Positively selected sites were compared to human Swiss-Prot sequence (P43246). All 15 sites occur between amino acid position 124–142 which overlaps with the region containing variants N127S, N139S and I145M associated with HNPCC1 (Ollila *et al.* 2008).

Tuberous sclerosis 2 protein (*TSC2*) interacts with *TSC1* protein, mutations in *TSC2* can cause tuberous sclerosis type 2 (Tee *et al.* 2002). Lineage-specific positive selection was observed in the following extant lineages, the percentage of sites under positive selection in each lineage being shown in brackets, in all cases ω > > 1: Chimpanzee lineage (0.2%), Gorilla (1.3%), Orangutan (0.29%), and, Marmoset (1.1%). Positively selected sites were compared against human Swiss-Prot sequence (P49815) but the functional information was not available.

ATM acts as a DNA checkpoint sensor by activating checkpoint signalling, upon double strand breaks (Kishi *et al.* 2001). Positive selection was detected in the following lineages (percentage of the alignment under positive selection shown in brackets): Gorilla (1.4%, ω > > 1), Marmoset (0.21%, ω > > 1), and Rabbit[74] (0.38%, ω = 7.42). BEB significant sites were compared to human (Q13315) and mouse (Q62388) Swiss-Prot entries. In the Gorilla lineage, positively selected site is 2067 (PP = 0.787), where if humans have a substitution of Alanine to Aspartate, it can result in Ataxia telangiectasia (AT) which causes weakened immune function and higher disposition to cancer (Kishi *et al.* 2001). No other functionally relevant information was obtained while comparing against either Marmoset or Rabbit.

The extant Orangutan lineage also showed evidence of positive selection in the *TSC1* gene for 1.2% of its alignment ω > > 1. Positively selected sites were compared against

[74] For ease, discussion of positively selected ATM gene data for rabbit extant branch is also included here.

human (Q92574) and mouse (Q9EP53) Swiss-Prot sequences but there was insufficient information to elaborate further.

4.3.4.1 Human population level analysis (using HapMap data)

Genes showing evidence of positive selection in lineages leading to *Homo sapiens*, *i.e.* the primate and Hominidae lineages (*STK11*, *CDH1* and *VHL*), were analyzed further to determine if there is evidence for any ongoing positive directional selection in modern-day human populations. The iHS (Integrated Haplotype Score) (Voight *et al.* 2006), was calculated for each SNP in *STK11*, *CDH1* and *VHL* genes across African Yorubu (Y), East Asian (A) and European (C) populations. An iHS score greater than +2 indicate that these alleles are segregating at a significant rate within their given populations. One intronic SNP in the SDK11 gene, had an iHS score of +2.0385 in European populations. For the *CDH1* gene, two intronic SNPs with iHS scores of +2.0433 and +2.5838 respectively, in the East Asian populations. For *VHL* gene, no population level directional selection was identified in modern humans.

4.3.5 Positive selection in the Ancestral Muridae branch

For ancestral Muridae (MRCA of modern mouse, rat and guinea pig species, Figure 4.2) lineage, *TSC1* gene (18 taxa) shows positive selection for 0.59% of sites in its alignment (ω > > 1). *TSC1* interacts with *TSC2* and acts as a tumour suppressor gene (Tee *et al.* 2002). Defects in *TSC1* causes tuberous sclerosis type 1 which is an autosomal dominant multi-system disorder. Positively selected sites were compared against human (Q92574) and mouse (Q9EP53) Swiss-Prot sequence, however there was insufficient information to extrapolate.

4.3.6 Positive selection in the Ancestral Murinae branch

The ancestral Murinae (MRCA of mouse and rat) lineage shows genes *MSH6* and *SDHC* under lineage specific positive selection. The *MSH6* mutation is linked to CRC (Section 4.1.2). For *MSH6* (19 taxa), 0.42% of the sites (3 residues) are under positive selection, ω > > 1 (Table 4.2). The corresponding Swiss-Prot sequence (P54276) lacked functional details for these positions, thus, potential functional effects remain unknown.

SDHC (Succinate dehydrogenase cytochrome b 560 subunit, mitochondrial), with 16 taxa, acts as a membrane-anchoring subunit for the SDH protein. Defects in this protein are reported in paragangliomas and gastric stromal sarcomas (Niemann *et al.* 2000). For *SDHC*, 4.2% of sites (9 residues) are under positive selection with ω > > 1 (Table 4.2). Comparison with the human (Q99643) and mouse (Q9CZB0) sequence from Swiss-Prot placed 8 of these sites either in transmembrane or topological domains across the gene, with the additional positively selected residue (position 128) neighbouring a metal binding site at position 127.

4.3.7 Positive selection in the Extant Rabbit branch

For extant rabbit lineage evidence of lineage-specific positive selection was observed in genes: *SDHC*, *MUTYH*, *MLH1*, *ATM* (Section 4.3.4) and *BHD*. The *SDHC* gene has 35.59% of sites under positive selection (ω = 3.59). Some 15/51 positively selected sites were identified as occurring within 10 amino acid positions of metal binding site at position 127 that is also mentioned in the ancestral Murinae analysis. While there are extremely high levels of positive selection identified in the rabbit lineage, no other relevant functional information could be gathered from the databases at this point.

The *MUTYH* gene (21 taxa), showed 1.4% of sites (ω > > 1) to be positively selected. These sites were compared to human (Q9UIF7) and mouse (Q99P21) Swiss-Prot entries, however no relevant functional information could be extrapolated. Radical substitutions occurred in all 5 BEB sites in the extant Rabbit lineage, three of which are at positions 485–487 in the Nudix hydrolase domain.

The *MLH1* gene (linked to CRC, Section 4.1.2), consists of 19 taxa and shows positive selection in 0.87% of sites (ω = 7.53). These sites were compared against human (P40692) mouse (Q9JK91) Swiss-Prot sequences. At position 120, Rabbit has a polar uncharged Serine residue while all other species tested have a hydrophobic Alanine residue. This site falls in a region dense with HNPCC2 variants at positions A111V, T116K, T117M, Y126N, A128P (Bronner *et al.* 1994, Pensotti *et al.* 1997 and Kurzawski *et al.* 2006). Positively selected residues in Rabbit: 209, 478 and 514, each fall within 8 amino acid positions of HNPCC2 variants: V213M, R474Q and V506A

(Tournier *et al.* 2008). Also, position 478 identified as under positive selection, lies in close proximity to a CRC variant R472I (Kim *et al.* 2001).

Finally, the *BHD* gene (20 taxa) shows 3.34% of sites (ω = 6.5) under positive selection. Functions of *BHD* gene are still unknown, although, it is thought that it may be a tumour suppressor and may be involved in CRC (Nickerson *et al.* 2002). BEB significant sites were compared to human (Q8NFG4) and mouse (Q8QZS3) Swiss-Prot entries. All 10 of the positively selected sites in Rabbit occur in a small region from position 61–83 and border a known human cancer variant at position 79 that when mutated from Serine to Tryptophan results in sporadic colorectal carcinoma.

4.3.8 Positive selection in the Extant Rodent (Rat and Guinea Pig)

MADH4 is the co-activator and mediator of signal transduction by TGF-beta. Its defect results in pancreatic, colorectal, juvenile polyposis syndrome, juvenile intestinal polyposis and primary pulmonary hypertension (Sayed *et al.* 2002, Sjoblom *et al.* 2006). The Rat lineage was identified as being under lineage-specific positive selection in the *MADH4* gene where 5.1% of sites are evolving with ω > > 1 (number of taxa = 16). Positively selected sites were compared to human (Q13485) and mouse (P97471) Swiss-Prot entries. The majority of positively selected residues in this protein are sequential, with 18/24 sites under positive selection in the rat lineage, within 10 amino acid positions of the natural human variant 493. When position 493 is mutated from Aspartate to Histidine pancreatic carcinoma is induced (Hahn *et al.* 1996).

NF1 is thought to be a regulator of RAS activity (Ballester *et al.* 1990). Defects in *NF1* have been shown to cause CRC and breast cancer (Sjoblom *et al.* 2006). For *NF1* (with an alignment containing 17 taxa), there was lineage-specific positive selection identified in 0.92% of sites in Rat with ω > > 1 and 0.12% of sites in guinea pig with ω > > 1. BEB significant sites were compared to human (P21359) and mouse (Q04690) Swiss-Prot sequences, however, no functionally relevant information was identified.

TSC1 also shows evidence of positive selection in the extant guinea pig lineage with 1.2% of the sites with ω> > 1. As before, the positively selected sites were compared against human (Q92574) and mouse (Q9EP53) Swiss-Prot sequences, again, no significant information was available.

4.3.9 Results of site-specific selective pressure analyses

Site-specific results have been summarized in this section as these could be prove to be beneficial for those working on rational mutagenesis and/or the identification of functionally important regions in the CRC associated genes. These results could also add to epigenetic research as links between genetic and epigenetic events have been observed (Section 3.2.4). Five genes are identified, that have signatures of site-specific positive selection, namely: *CDH1*, *MUTYH*, *PMS1*, *PMS2* and *TP53*, representing ~23% of the dataset. For each of these five genes, the model of best fit was the site-heterogeneous model “model 8” (Table 4.2).

For *CDH1* (15 taxa) gene, site-specific analysis identified 0.71% sites evolving under strong positive selection, ω= 4.54 (Table 4.2). These sites are compared to the human Swiss-Prot entry (P12830) to obtain relevant functional information, Figure 4.3 (b). The vast majority of positively selected sites (12 sites) in the protein are found within the extracellular topological domain (positions 155–709). Many of these positively selected are in close proximity to natural cancer variants *e.g.*, position 421 (positively selected) which resides within a region (418–423) known to be missing in gastric carcinoma samples (Tamura *et al.* 1996). Positions 457, 465, and 467 are under positive selection and map in close proximity to natural variant E463Q found in gastric carcinoma samples (Berx *et al.* 1998). Position 700 resides within the metalloproteinase cleavage site (700–701) of *CDH1*. Position 735 is in close proximity to a gamma-secretase/PS1 cleavage site (731–732) (Marambaud *et al.* 2002), and position 553 is in close proximity to a glycosylation site (558), essential for the posttranslational modification of proteins (Zhou *et al.* 2008). In the *CDH1* gene, the majority of species tested (8/15) have hydrophobic residues (Isoleucine, Valine, Leucine) at position 553, the glires group (mouse, rat, guinea pig and rabbit) have small residues (Alanine, Serine, Threonine), but human, gorilla, and dog have large aromatic residues (Phenylalanine) that could significantly alter the protein structure and may affect binding at the glycosylation site at position 558. Defects in the *CDH1* member of the Cadherin family are linked to hereditary diffuse gastric cancer (Vogelstein and Kinzler 2004, Yoon *et al.* 1999).

The MUTYH gene provides instructions for making an enzyme called MYH glycosylase, which is involved in the repair of DNA. This enzyme corrects particular mistakes that are made when DNA is copied (DNA replication) in preparation for cell division[75]. For *MUTYH* gene (21 taxa), site-specific analysis identified 18 sites under positive selection (ω = 2.44), representing 2.8% of the *MUTYH* protein (Table 4.2). A total of 10 unique sites are reported as natural cancer variants in human (Q9UIF7), Figure 4.3 (c). Positively selected sites 406 and 412 are in close proximity to natural cancer variants at positions 402 and 411 respectively. Positively selected sites 521, 528 and 538 also map in close proximity to natural variants, 526 and 531 respectively. Also of note are the replacement substitutions observed at Swiss-Prot positions 406 and 412 that are radical with potential effects on protein structure.

PMS1 (post meiotic segregation increased 1) gene is linked to CRC (section 4.2.1). Analysis of *PMS1* identified site-specific model of codon evolution model 8 as best fit, estimating 25 positively selected sites (6.4% of the alignment) with ω = 1.33 (Table 4.2. These sites are compared against human Swiss-Prot sequence P54277. Positively selected site 387 resides in close proximity to position 394 - a natural variant (M394T) reported in incomplete HNPCC and HNPCC3 (Wang *et al.* 1999). Due to limited functional data it was not feasible to study the remaining 24 sites. However, due to *PMS1* function in DNA MMR pathway, these positively selected sites could prove as ideal candidates for mutagenesis/epimutagenesis studies in the future.

The *PMS2* dataset contained 21 taxa and site-specific analysis identified 8.9% of sites under positive selection in this protein, ω = 1.29 (Table 4.2). Functional relevance of these sites was determined by comparison to Human Swiss-Prot sequence (P54278). The vast majority of sites (32) reside within the 430–645 region of the alignment. This region of the alignment is highly variable and could not be improved manually. Functional characterization for this region is also lacking and therefore it was not possible to assess functional relevance. Outside this region, two positively selected sites, 402 and 406 (PP = 0.632 and 0.728 respectively) flank a phosphoserine

[75] http://ghr.nlm.nih.gov/gene/MUTYH

modification site (403) (Beausoleil *et al.* 2006). Both substitutions are radical and could affect the function at position 403.

The *TP53* protein (P04637) is 393 residues in length with 343 of these sites reported as natural variants that cause/lead to cancer including but not limited to CRC and gastric cancers (UniProt, Varley *et al.* 1995, Guran *et al.* 1999). In our analysis of *TP53* we have 16 taxa. Mutations in this gene radically affect function and therefore we would expect to find evidence of strong purifying selection across sites and lineages. However, results indicate that site specific positive selection is at work with 13 sites under positive selection, ω = 1.97, Figure 4.3 (d) and Table 4.2 for detailed analysis. On inspection of the 14 sites, it is determined that 11 sites are located within the first region of the protein (positions 1–83), a region responsible for interaction with the methyltransferase HRMT1L2 and the recruiting of promoters to the *TP53* gene (An *et al.* 2004). A cluster of positively selected sites is identified, namely positions 46 and 47, along with an additional 7 sites within ten residues 39, 52, 53, 54, 55, 56, and 59 (see Additional file 4, Publication). Mutation of position 46 can abolish phosphorylation by HIPK2 and acetylation of K-382 by CREBBP (Hofmann *et al.* 2002). Region 66–110 of *TP53* is involved in interaction with WWOX protein and two sites are identified (Swiss-Prot positions: 72 and 81), under positive selection within this region. Positively selected position 129, is located within a region reported to interact with HIPK1 (100–370) and AXIN1 (116–292), and in addition is also located within a region (positions 113–236) that is required for interaction with FBX042. Positively selected residue 355 is located within the CARM1 interaction region (300–393), the HIPK2 interacting region (319–360), and the oligomerization region (325–356).

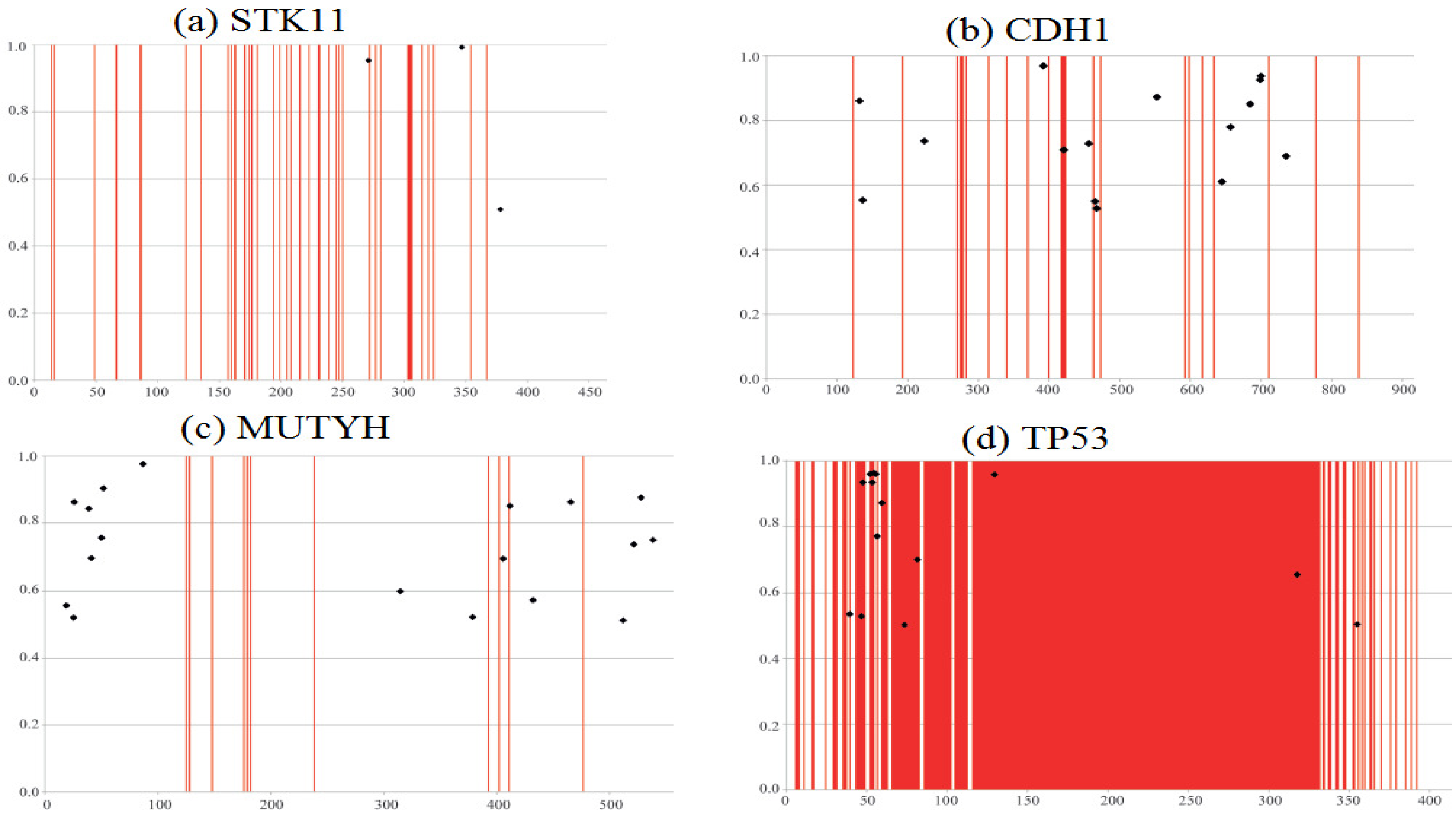

Figure 4.3 Positive selection analysis for 4 genes: (a) *STK11*, (b) *CDH1*, (c) *MUTYH*, and (d) *TP53*. The x-axis depicts the gene from start to end of alignment. The Y-axis is the posterior probability. The vertical red bars on each graph represent the known cancer causing variants from human populations. The black dots on each graph represent the positively selected sites identified in this study.

4.4 Conclusion

The results obtained are indicative of selective pressures acting in a lineage-specific manner. In this study, the positively selected sites identified, are found to reside frequently in regions of functional importance, such as glycosylation sites, protease cleavage sites, and sites known to interact with proteins involved in MMR pathways. Also of note, positively selected residues are frequently found to be located at, or in close proximity to, known cancer associated sites. However, the statistical significance of these occurrences is not conclusive (sample sizes are small). Larger sample sizes (more genes) and more complete functional information (from Swiss-prot) would be helpful in resolving whether these sites are most likely positioned at variants associated with cancer.

While comparing the functional divergence in the extant human and mouse lineages for the genes analysed, none were observed in our analysis. However, upon testing the lineages leading from the MRCA (most recent common ancestor) of mouse and human, *i.e.* Euarchontoglires, positive selection has occurred on certain branches and in specific lineages. In the ancestral lineages from the divergence of primates, rodents and glires there is evidence of positive selection in 6 of the 22 genes tested (including weak *VHL* result, as evidenced by *p*-values (from Table 4.2). Considering all, including extant lineages analyzed, lineage-specific positive selection was detected in total of 64% of the genes (*i.e.* 14/22 genes). Studies on the levels of polymorphism observed in Drosophila shows positive selection in ~25% of the genes (Eyre-Walker 2004) while studies for primates compared to rodents and in the Hominidae reveal much lower levels of positive selection, 5-9% of genes in the genome, (Arbiza *et al.* 2006, Kosiol *et al.* 2008). If these previous analyses of primates were to act as references, then the expected number of identified genes under positive selection in our dataset, would be 1. However, taking the Drosophila data as the upper bound, ~6 genes with evidence of positive selection would be expected. In our case we find 64% of our dataset under positive selection in either lineage or site specific manner. The genes taken for study here are known CRC genes, and their function and close link with cancer might have been the reason for this discrepancy – further supporting cancer as a selective pressure.

Further, grouping the CRC genes according to their involvement in functional pathways, it is found that the MMR pathway has evidence of positive selection in 3 components of the pathway – 2 of which are site-specific and one is specific to the ancestral Murinae lineage, suggesting a specific selective pressure in this clade for this process. The site-specific analyses identified a total of 5 genes that are positively selected: *CDH1*, *MUTYH*, *PMS1*, *PMS2* and *TP53*. These results are important for contributing to our understanding of fundamental functions of these proteins and have provided potential targets for rational mutagenesis. Positive selection sites in MMR pathway genes might have some connection with epigenetic silencing of these genes (these could be hypermethylation prone) as these genes are known to undergo epigenetic silencing and thereby initiating deregulation of MMR pathway.

In general, these results indicate that the function of certain proteins associated with colon cancer display distinct lineage-specific patterns of substitution indicative of positive selection in the ancestral human and mouse lineages. There are a number of selective pressures on any given protein that can contribute to patterns of substitution that are "falsely" indicative of positive selection. The necessity to continue to interact with protein partners may be strong, driving the evolution of the proteins in this study, as many form functional complexes with each other or other proteins (Fraser *et al.* 2002). Compensatory mutations may also contribute to elevated levels of *ω* (Lunzer *et al.* 2010). The effective population size (*Ne*) of the species tested in this study varies enormously (with estimations for modern human populations in the range of *Ne* = 7,500 to 3,100 (Tenesa *et al.* 2007), while estimations for modern mouse populations range from *Ne* = 58,000 to 25,000 (Salcedo *et al.* 2007), which might have also contributed to the detection of false positives.

A weak evidence for ongoing selective pressure in the human genome on the *STK11* and *CDH1* has also been detected, but these may be artefacts of the very small effective population size[76]*(Ne)* of modern humans. Smaller *Ne* values are associated with increased fixation of slightly deleterious substitutions and subsequent elevated *ω* values, (Eyre-Walker *et al.* 2002). Such slightly deleterious mutations in turn can lead to

[76] the number of breeding individuals in an idealised population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration

additional compensatory substitutions that become fixed. Distinguishing substitutions that have become fixed, due to positive selection from slightly deleterious substitutions, and small *Ne,* (Eyre-Walker, Keightley 2009) will aid to a more complete understanding of protein evolution in the future.

Chapter 5

Exploratory Data Mining and Analysis of *StatEpigen* Database

5.1 Exploratory Data Mining and Analysis

In this chapter, an exploratory data analysis on *StatEpigen* data has been carried out. An exploratory data analysis examines the data for simple correlations or consistent occurrences in the data, at the distributions or ranges. This initial analysis also helped to evaluate the database contents and checks for quality and consistency. The following Figures 5.1, 5.2 and 5.3 look at the occurrence of methylation events at different colon cancer stages.

Let's consider the analysis of *MLH1* data depicted in Figure 5.1. The methylation frequency values are, number of positive results (for methylation)/number of samples taken. Each dot in the plot represents a result of a single experiment for *MLH1*, *e.g.* if an experiment for *MLH1* has 100 samples (patients) and 10 of them are methylated at adenoma stage, represented as CC_stage = 7 in the plot, the methylation frequency value at Y-axis is 0.1. Similarly, for second experiment, for *MLH1*, for carcinoma stage (CC_stage = 10), if there were 50 samples and 20 of them are methylated, this methylation value is represented as 0.4 on the Y-axis. It is of note that, methylation frequency is not 'degree of methylation' in this context.

For the exploratory data analysis, *StatEpigen* data tables, 'cell-state' , 'single relation' and 'event' tables (in MySQL) are combined, selecting colon cancer data only (as we have some data for other forms of cancer) and methylation event only (as we have other events *e.g.* mutation), and the combined table is further processed (*e.g.* removing -

identical columns and non-methylated rows). Finally, the colon cancer stages are numbered (for ease of plotting) as in Table 5.1 as follows:

Table 5.1: Different stages of colon cancer (numbered for ease of plotting) in *StatEpigen* database

SN	Stages	SN	Stages
1	Adjacent to Polyps	8	Stool- Adenoma
2	Adjacent to adenoma	9	Premalignant mucosa
3	Adjacent to Carcinoma	10	Carcinoma
4	ACF (Aberrant Crypt Foci)	11	Stool- Carcinoma
5	Polyps, stool	12	Serum - Carcinoma
6	Polyps	13	CRC
7	Adenoma	14	Metastasis.

The resultant data then is plotted in R, to show methylation frequency (defined as the number of methylated samples within each sample used for study, *Y*-axis), for different colon cancer stages, (*X*-axis) for 3 different genes selected based on their frequency of occurence in the current database, these are *MLH1* (272 occurences), *CDKN2Ap16* (241 occruences) and *MGMT* (221 occurences), as shown in Figures 5.1, 5.2 and 5.3 respectively. It is of note, that, the numbers here (*e.g.* 272), are instances of data introduced in database for *MLH1* (the sample/patient numbers could, in each instance vary). Database content is clearly subject to updating in future (with addition of more data). Also, it is not necessarily the case that these genes occur most frequently in CRC, but could be biased due to the selection of papers (during curation) or due to the preference for these genes for experiment by various authors. However, it could also be possible that these genes are found to be undergo epigenetic changes more frequently during experiments, are hence picked up more frequently in experiments.

It is observed that most of the methylation occurs in carcinoma stage for the selected genes (*MLH1, CDKN2Ap16* and *MGMT*), the most common methylation frequency being in the range 0.1-0.4 for *MLH1* and *CDKN2Ap16,* and 0.2-0.4 for *MGMT*, indicating that in a random sample, around 10 to 40% of the former and 20-40% of the later genes are methylated. Methylation also is seen to be occuring more prominently in adenoma and carcinoma stage as compared to other stages (*e.g.* metastasis) in the genes selected. These are clearly initial observations only, given provisions as above and any subsequent analyses would require cross checking with other data.

Methylation frequency versus stages plot in different genes (single relations):

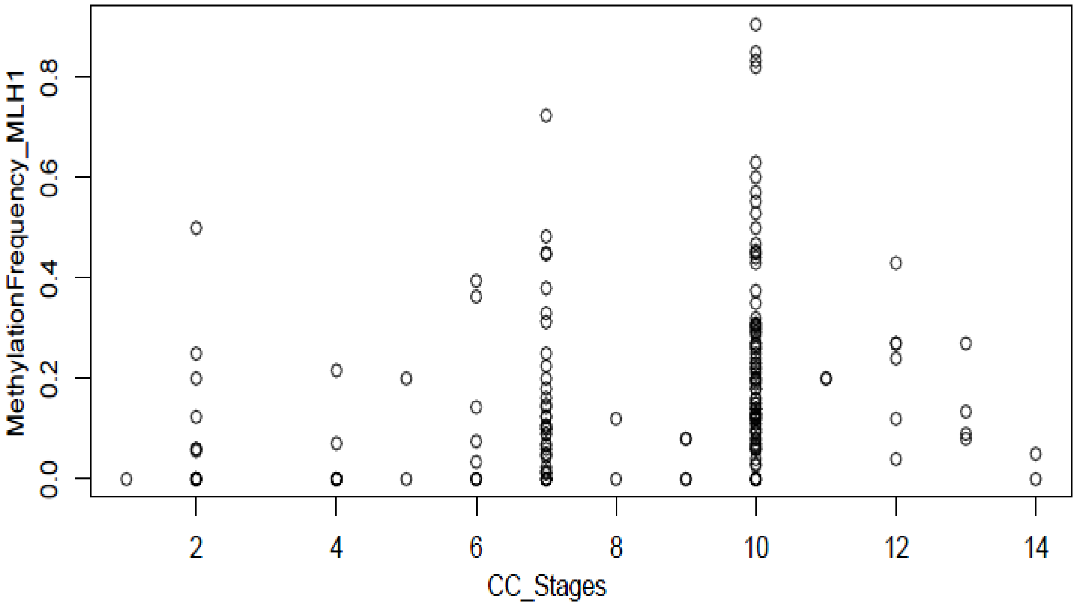

Figure 5.1: Methylation frequency in *MLH1* gene versus colon cancer stages (14 metastasis, 10- carcinoma, 7- Adenoma, 2- adjacent to adenoma).

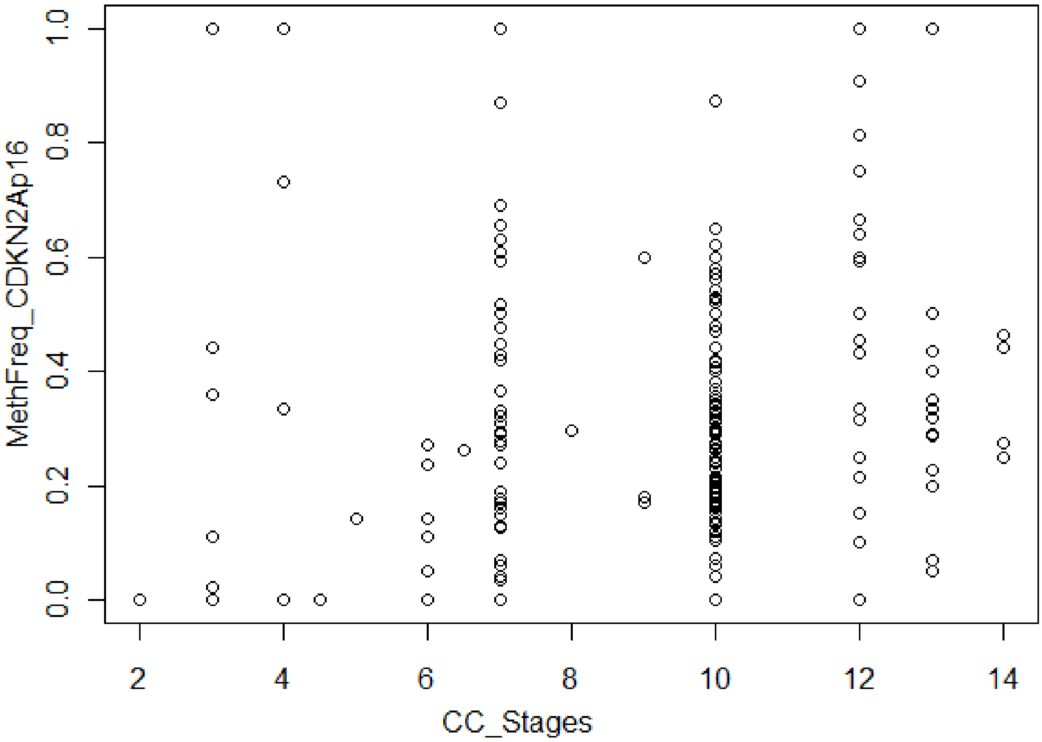

Figure 5.2: Methylation frequency in *CDKN2Ap16* versus colon cancer stages (14- metastasis, 10- carcinoma, 7- Adenoma, 2- adjacent to adenoma).

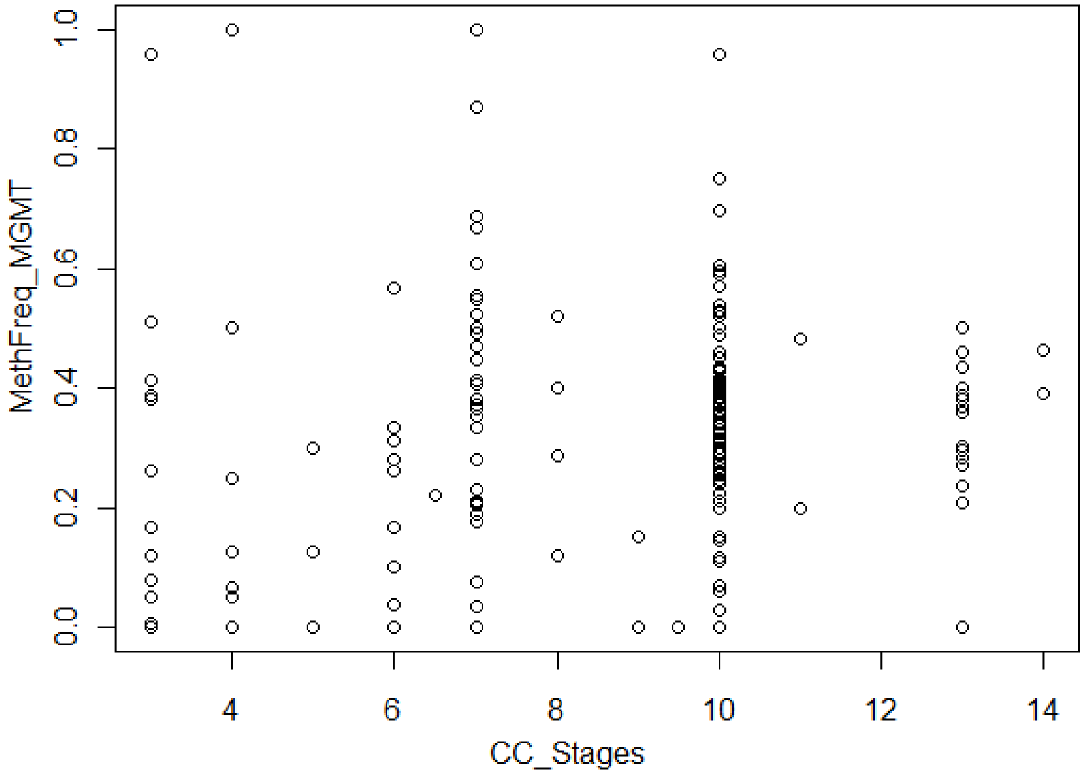

Figure 5.3: Methylation frequency in *MGMT* gene versus colon cancer stages (14- metastasis, 10- carcinoma, 7- Adenoma, 2- adjacent to adenoma).

These are experimental data and can be traced back if any outliers found. The higher variability at early stages might be due to smaller sample size.

This analysis answers the question - which genes are augmented in the database the most (as data selection could be due to mere coincidence, author preference or constraints on the experiments, or could actually be the most hypermethylated genes, as mentioned before). The second question that is addressed here is which stages are observed to be hypermethylated (adenoma and carcinoma here, again given conditions as mentioned before), according to our current database release. This analysis is one of the examples that show the type of questions that can be posed, with the *StatEpigen* database, in order to evaluate its contents.

Provided, these genes occur with high incidence, backed up by the evidence that *MLH1* is mismatch repair gene, the other two genes are also shown in the literature to be involved in the colon carcinogenesis, and given that literature selecion was not determined by any deliberate bias towards any particular gene set, methylation is seen to be occuring more prominently in adenoma stage (after carcinoma stage) as compared to other stages (*e.g.* metastasis) in the genes selected. In first instance, it implies that cancer initiation might be detectable with methylation profiling of the genes involved during adenoma stage. These probabilities evidently need to be further investigated.

Similar many other queries, genetic, epigenetic and genetic-epigenetic interactions can be dealt with, using *StatEpigen*. This analysis in brief, signifies the scope of *StatEpigen.*

5.2 Data analyses: Future work

Following is the list of data analyses that can be carried out with *StatEpigen* in future.

(1) The genes in *StatEpigen* database can be grouped into tumour suppressor genes, oncogenes and mismatch repair (*MMR)* gene groups and their methylation data can be compared.

(2) Grouping the entire genes available in *StatEpigen* into somatic and germline mutated genes (for colon cancer) and comparing the methylation data could be another way of looking into CRC research. *COSMIC* database information would help to differentiate between *somatic* and *germline* mutated genes. *COSMIC* database is designed to store and display somatic mutation information and related details and also contains information relating to human cancers.

(3) Epigenetic analysis with respect to *driver* and *passenger* genes in colon cancer and the difference in their methylation data would be another useful analysis since different *driver* and *passenger* genes provide insight on targets for colon cancer medication. Currently, few genes are known to be *driver* genes, with most of those associated with colon and other cancers, known as *passenger* genes. In fact, the key challenge is to distinguish *driver* from *passenger* genes and to discriminate between these.

(4) In addition to the above, comprehensive cluster analysis of the colon cancer genes in *StatEpigen*, is currently underway to identify the markers while analysis of gene pathways and the timeline for gene involvement is still lacking in any derived network.

5.3 Data Augmentation in *StatEpigen* database

The platform of the in-house developed, database resource, *StatEpigen* (the *knowledge management system/ KMS*) consists of a database (back-end) and a user interface (front-end). This resource correlates molecular events (gene expressions and mutations etc.) to various colon cancer phenotypes (focus- early stage phenotypes)/ epigenetic events. It presently contains around 5700 records on simple events and 2900 records on conditional events (on colon cancer). The objective of this database is to facilitate future research in cancer – especially for early diagnosis and risk assessment.

Maintenance and upgrading of the *StatEpigen* resource needs augmentation of the database on a regular basis that currently incorporates manual data curation. Manual data curation is described in Section 3.3.1 and is the gold standard, maintaining data quality, consistency and accuracy. However, it is costly in effort and very time consuming. Approximately 45 papers, 15% of current database content in *StatEpigen* database is due to work done over 12 months. One solution to this labour-intensive requirement, is to address the data curation process, such that the system is self-sustainable in future and the data entry, (curation and augmentation), in the database can be carried out in an automatic manner. While providing robustness to the resource, such a feature is non-trivial and is discussed in Section 5.4.

5.4 Future Work

A major possible upgrade in the current *Knowledge Management System (KMS)* would be to design and construct an automatic data-entry pipeline/module that would act as an interface between the (suitable) data source(s) and *KMS*. A review of the literature leads to recognition of *MIAME (Minimum Information About a Microarray Experiment)* compliant microarray data in public repositories, including *GEO* (*Gene Expression Omnibus*)[77] at *NCBI* (US), *Array Express* at the *EBI* (UK)[78] and *CIBEX* at *DDBJ* (Japan)[79]. These databases aim at storing these high-throughput data in accordance with *MGED* (*Microarray Gene Expression Data*) recommendations.

The DNA methylation arrays and gene expression arrays in these public repositories would be valuable as input source(s) to *KMS*. The raw data from these resources could be downloaded, processed (filtered and normalized) and the results fed into the *StatEpigen KMS* as relational event(s), which are then made available to the webportal. The new module could then be used as a data processing tool and as an interface initially between *KMS* and *GEO*.

[77] http://www.ncbi.nlm.nih.gov/geo/
[78] http://www.ebi.ac.uk/arrayexpress/
[79] http://trace.ddbj.nig.ac.jp/cibex/index.shtml

Data Processing: Processing of methylation array data (for example infinium DNA methylation assay considered initially to start with) might include discarding the data with *p*-value >0.05 (for statistical significance), removal of unreliable probe data and so on. The *β*-value (level of DNA methylation at each CpG sites (Chapter 1)) then needs to be normalized to eliminate batch effects.

Normalization of data, would be one of the important, as well as challenging, tasks during the upgrade process, particularly due to the different platforms (*e.g.* Affymetrix chips, Illumina chips, Beadchips). Different normalization methods are likely to be required including least square, lowess, total intensity and so on depending on the platform being used (Shakya *et al.* 2010, Schmid *et al.* 2010).

Once these data are filtered and normalized, these are subjected to series of statistical analysis so that the processed output from these arrays can be incorporated as *simple* and *conditional* (*single* and *double relational*) *events* in *StatEpigen KMS*. A schematic box diagram and a flow chart (Figures 5.4 and 5.5 respectively below) are provided to clarify the initial outline conception of the system.

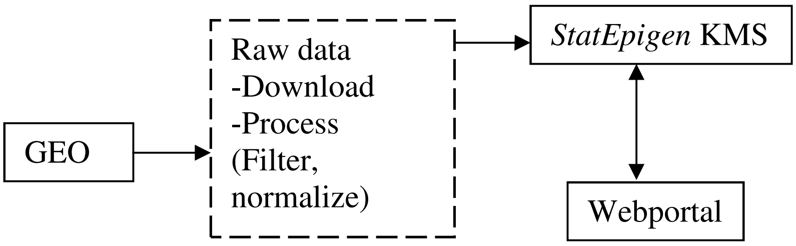

Figure 5.4: Schematic box illustration of anticipated tool

Anticipated tools to be used

The following tool base is anticipated to be required.

1. *Biobase* , *Limma* and other tools of *Bioconductor* package in R.
2. *GEOquery* - Tool to access *GEO* from *Bioconductor* (Sean and Meltzer 2007)
3. *MySQL* and php for *StatEpigen* Biomedical Resource (Currently used).

An illustration of how the processed data from new input source would appear on the web portal is explained below.

There would be a button (left hand pane) below the ‘advanced search’- ‘cell-line’ in the home page which will be designated “Search by Arrays”. If this button is clicked, all available-processed arrays will be displayed at the right hand pane. When one is selected, the respective single or double relations will be displayed in a separate single window. A ctrl-select feature will be enabled, which makes it possible for multiple relations from the multiple arrays to be displayed in the window. For the initial design phase, data on arrays can be kept in separate tables in *StatEpigen* database and will be integrated/ migrated later, once the test-phase migration is successful. A simplified flowchart on work-flow for possible upgrade on *Knowledge Management System*, *StatEpigen* is given in Figure 5.5, below.

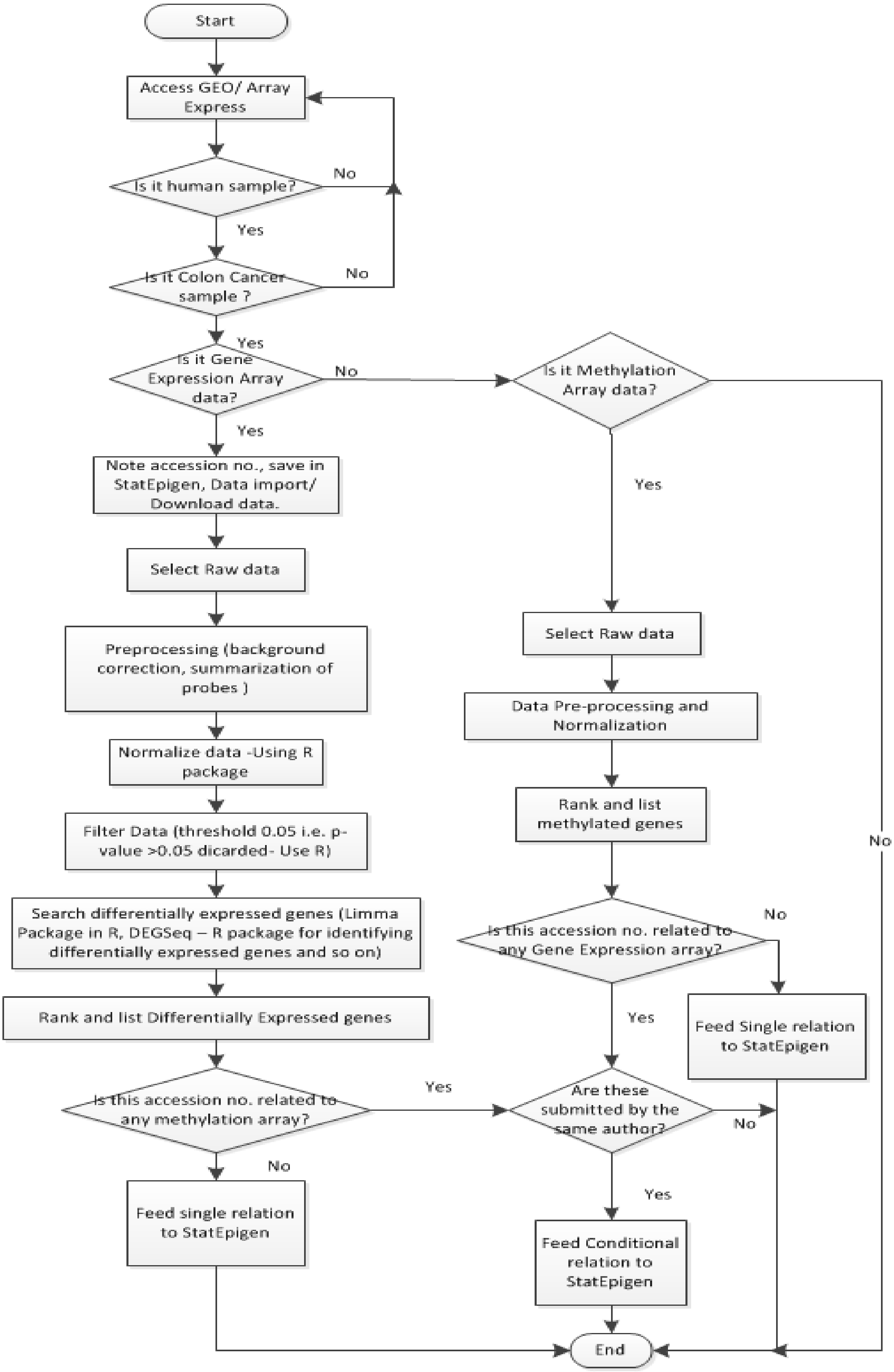
Start
Access GEO/ Array Express
Is it human sample?
No
Yes
Is it Colon Cancer sample ?
No
Yes
Is it Gene Expression Array data?
No
Is it Methylation Array data?
Yes
Note accession no., save in StatEpigen, Data import/ Download data.
Select Raw data
Preprocessing (background correction, summarization of probes)
Normalize data -Using R package
Filter Data (threshold 0.05 i.e. p-value >0.05 dicarded- Use R)
Search differentially expressed genes (Limma Package in R, DEGSeq – R package for identifying differentially expressed genes and so on)
Rank and list Differentially Expressed genes
Is this accession no. related to any methylation array?
No
Feed single relation to StatEpigen
Yes
Select Raw data
Data Pre-processing and Normalization
Rank and list methylated genes
Is this accession no. related to any Gene Expression array?
No
Feed Single relation to StatEpigen
Yes
Are these submitted by the same author?
No
Yes
Feed Conditional relation to StatEpigen
No
End

Figure 5.5: Flowchart (simplified) showing single iteration work-flow for proposed upgrade on *Knowledge Management System, StatEpigen.*

Upgrade Advantages and Future Features

Advantages include:

1. The scope of *StatEpigen* would be broadened. The array –input upgrade would give *StatEpigen* a second input source to populate the database in addition to data curation information from literature /publications only.
2. Reduction in the volume of manual curation, as most papers from which data are currently extracted have the array information uploaded into *GEO* or *Array Express* (given introduction of new standards[80]).
3. Economy in the data augmentation process.
4. Timely and regular update of data as new research on colon (and other cancers in future) becomes available.
5. Improved possibilities for data analysis and refinement
6. A degree of self-sustainability for the KMS and for inputs to modelling softwares currently being developed.

There is also a possibility of extension to other platforms, other cancer forms and other resources.

In addition to the automation feature, Integration of the *StatEpigen* data resource with other resources (*COSMIC* is a targeted example in relation to which some interaction has already taken place) can also be a flagged as future work. *COSMIC* database features (*somatic* and *germline*) mutation information. Other integrable resources are potentially, *MethyCancer, MethDB* (Section 2.4.6) and useful linkage could also include murinae epigenetic database, given the close association with human lineage, and use of these species as model organisms for human cancers. Furthermore, there is a considerable body of work available in murinae lineage for epigenetic research.

[80] http://www.ncbi.nlm.nih.gov/geo/info/MIAME.html

Chapter 6

Summary and Conclusions

The investigations reported here make use of the *StatEpigen* database resource, which incorporates data on genetic and epigenetic factors and statistical information on the inter-relationships between them. It is seen as complementary to some of the other resources, detailed in chapter 2, with potential for integration with and input-sourcing from several others. As a small resource, it has nevertheless been plumbed for data on gene lineage shift and robustness of murine models for colon cancer in humans, as well as potential epigenetic factors in cancer etiology.

In addition to genetic and epigenetic links in initiation and progression of cancer, as discussed in Section 3.2.4, some studies (Sawan *et al.* 2008, You 2012) suggest that genetic- epigenetic interactions are complementary in cancer etiology, some others suggest they have synergistic effect (*e.g.* Ogino *et al.* 2006, in case of colon cancer). Denissenko *et al.* (1997) have suggested that (cytosine) methylation determines mutation 'hotspots' in human tumour suppressor gene *P53* (indicating the possibility that majority of these in human genes are at CpGs). Some notable publications, Rideout *et al.* (1990), Mazin (1994) suggest that methylation events trigger mutation events. In contrast, other publications suggest the opposite, *i.e.* some mutation events can induce methylation (Poole *et al.* 2012, Duncan *et al.* 2012).

Since positive selection (Chapter 4) also is a form of (advantageous) mutation, positively selected genes (identified from selective pressure analysis), may be associated with methylation events (they could be methylation prone) or *vice-versa.* Such inter-relationships can be explored further through *StatEpigen experimental* *epigenetic* data that links the phylogenetic analysis to the big-picture of genetic-epigenetic interaction again.

While investigating colon cancer from phylogenetic/ evolutionary perspective, identifying positive selection in the *primate, Hominidae, Muridae* and *Murinae* lineages suggested an ancestral functional shift in these genes between the rodent and primate lineages. Our finding that, for *SDHC* and *MSH6* genes, mouse might not be an ideal model organism for human colon cancer (as these genes are positively selected in *Murinae* lineage), as well as our identification of 15 other genes that are positively selected in different functionally important sites/lineages are significant, as these genes can now be targeted for more detailed analyses. The site-specific analyses have provided potential targets for rational mutagenesis. The positively selected sites identified in this study were found to frequently reside in regions of functional importance, such as glycosylation sites, protease cleavage sites, and sites known to interact with proteins involved in DNA damage repair pathways, with indications also that, positively selected residues are frequently located at, or in close to, known cancer associated sites. On grouping cancer associated genes according to their involvement in functional pathways, it was determined that the mismatch repair (*MMR*) DNA damage response pathway shows evidence of positive selection in 3 components, 2 of which are site-specific and one of which is specific to the ancestral *Murinae* lineage, suggesting a specific selective pressure in this clade (Chapter 4).

The literature survey (Chapter 2) carried out as an initial phase of this research provided a background check on the stage of development regarding epigenetics related biomedical resources. Epigenetic databases are already found to be numerous and range from small to large-scale, with considerable ongoing integration and new links still being implemented. Major issues are typical of early-stage development, namely quality assurance, effective annotation and overall management. The generation of a centralised repository for epigenetics-related data is clearly desirable and is projected for the future with many large scale initiatives currently launched. Small scale, specialised databases are typically sustainable through integration and compatible targets for *StatEpigen* objectives are to include *COSMIC, MethDB*, *MethPrimerDB,* amongst others. The biomedical resources*, GEO, Array Express* and *CIBEX* are also noted as the potential data source for *StatEpigen* resource after its upgrade.

Chapter 3 discussed one of the simplest data curation examples in order to illustrate what is involved and highlighted the need for automation, Chapter 5 indicated what is involved for automation of the data curation process and provided a schematic diagram and initial conceptualization, for future implementation. This chapter also provided initial example exploratory data analysis. This simple analysis Section 5.1 noted that a significant number of adenoma stage ‘patients’ showed key genes to be methylated, which suggests that methylation profiling is useful as a screening measure. For early stages (*e.g.* polyp, stage 6), samples are, however, too small to determine what useful thresholds can be set.

References

1. Adams, J. 2008. Imprinting and genetic disease: Angelman, Prader-Willi and Beckwith-Weidemann syndromes. *Nature Education,* 1(1).
2. Adamson, E.D. 1987. Oncogenes in development. *Development,* 99, pp. 449–471.
3. Albert Tenesa, A., Navarro,P., Hayes, B.J., Duffy, D.L., Clarke, G.M.,Goddard, M.E.and Visscher, P.M. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res*, 17(4), pp. 520–526.
4. An, W., Kim, J. and Roeder R.G. 2004. Ordered cooperative functions of PRMT1, p300, and CARM1 in transcriptional activation by p53. *Cell,* 117(6), pp. 735–748.
5. Anisimova, M., Bielawski, J.P., and Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol*, 18(8), pp.1585–1592.
6. Anisimova, M., Nielsen, R. and Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*, 164(3), pp.1229–1236.
7. Arbiza, L., Dopazo, J. and Dopazo, H. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput Biol,* 2(4), e38, doi: 10.1371/journal.pcbi.0020038.
8. Astakhov, V., Grethe, J.S., Ross, E., Little, D., Sanders, B. and Gupta, A. 2007. Semantic Data Integration Environment for Biomedical Research (Mediator Infrastructure for Information Integration), BIRN Coordinating Center, University of California San Diego, La Jolla, CA 92093-0715, USA.
9. Aumueller, D., 2005. Proceedings of 2nd ESWC. Semantic authoring and retrieval within a Wiki, Department of Computer Science, University of Leipzig, Germany.
10. Baas, A.F., Boudeau, J., Sapkota, G.P., Smit, L., Medema, R., Morrice, N.A., *et al.* 2003. Activation of the tumour suppressor kinase LKB1 by the STE20-like pseudokinase STRAD. *EMBO J*, 22(12), pp. 3062–3072.
11. Baker, S.M., Plug, A.W., Prolla, T.A., Bronner, C.E., Harris, A.C., Yao, X., Christie, D.M., Monell, C., Arnheim, N., Bradley, A., *et al.* (1996). Involvement of mouse Mlh1 in DNA mismatch repair and meiotic crossing over. *Nat. Genet.* 13 (3), pp. 336–42.

12. Ballester, R., Marchuk, D., Boguski, M., Saulino, A., Letcher, R., Wigler, M. and Collins, F. 1990. The NF1 locus encodes a protein functionally related to mammalian GAP and yeast IRA proteins. *Cell*, 63(4), pp. 851–859.

13. Baranzini, S.E., Mudge, J., Van Velkinburgh, J.C., Khankhanian, P., Khrebtukova, I., Miller, N.A. *et al.* 2010. Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature,* 464(7293), pp.1351-1356.

14. Barat, A. and Ruskin, H.J. 2010. A Manually Curated Novel Knowledge Management System for Genetic and Epigenetic Molecular Determinants of Colon Cancer. *The Open Colorectal Cancer Journal*, 3, pp. 36-46.

15. Beausoleil, S.A., Villén, J., Gerber, S.A.,Rush, J. and Gygi, S.P. 2006. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol*, 24(10), pp. 1285– 1292.

16. Benton, M.J. and Donoghue, P.C.J. 2007. Paleontological evidence to date the tree of life. *Mol Biol Evol*, 24(1), pp. 26-53.

17. Berends, M.J., Wu, Y., Sijmons, R.H., Mensink, R.G., van der Sluis, T., Hordijk-Hos, J.M., de Vries, E.G., Hollema, H., Karrenbeld, A., Buys, C.H. *et al.*. 2002. Molecular and clinical characteristics of MSH6 variants: an analysis of 25 index carriers of a germline variant. *Am J Human Genet*, 70(1), pp. 26–37.

18. Berg, D.J., Davidson, N., Kühn, R., Müller, W., Menon, S., Holland, G., Thompson-Snipes, L., Leach, M.W. and Rennick, D.1996. Enterocolitis and colon cancer in interleukin-10-deficient mice are associated with aberrant cytokine production and CD4(+) TH1-like responses. *J. Clin. Invest.,* 98 (4), pp. 1010–1020.

19. Berners-Lee, T., Hendler, J. and Lassila, O. 2001. The Semantic Web. *The Sci Am,* 284 (5), pp. 34-43.

20. Beroud, C., Hamroun, D., Collod- Beroud, G., Boileau, C., Soussi, T. and Claustres M. 2005. UMD (Universal Mutation Database): 2005 Update, *Human Mutation*, 26 (3), pp. 184-191.

21. Berx, G., Becker, K.F., Höfler, H. and Van Roy, F. 1998. Mutations of the human E-cadherin (CDH1) gene. *Hum Mutat*, 12(4), pp. 226–237.

22. Bierne, N. and Eyre-Walker, A. 2004. The genomic rate of adaptive amino acid substitution in Drosophila. *Mol Biol Evol*, 21(7), pp. 1350–1360.

23. Bird, A. 2002. DNA methylation patterns and epigenetic memory. *Genes & Dev*., 16, pp. 6-21.

24. Blackwell, L.J., Bjornson, K.P. and Modrich P. 1998. DNA-dependent activation of the hMutSalpha ATPase. *J Biol Chem*, 273(48), pp. 32049–32054.

25. Blackwell, L.J., Martik, D., Bjornson, K.P., Bjornson, E.S. and Modrich, P. 1998. Nucleotide-promoted release of hMutSalpha from heteroduplex DNA is consistent with an ATP-dependent translocation mechanism. *J Biol Chem*, 273(48), pp. 32055–32062.

26. Boudeau, J., Baas, A.F., Deak, M., Morrice, N.A., Kieloch, A., Schutkowski, M., *et al.* 2003. MO25alpha/beta interact with STRADalpha/beta enhancing their ability to bind, activate and localize LKB1 in the cytoplasm. *EMBO J*, 22(19), pp. 5102–5114.

27. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. 1984. *Classification and Regression Trees*, Published: Wadsworth, Publication No.: 0412048418.

28. Bronner, C.E., Baker, S.M., Morrison, P.T., Warren, G., Smith, L.G., Lescoe, M.K., Kane, M., Earabino, C., Lipford, J., Lindblom, A. *et al.* 1994. Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. *Nature*, 368(6468), pp. 258–261.

29. Bush, R.M. 2001. Predicting adaptive evolution. *Nat Rev Genet*, 2(5), pp. 387–392.

30. Chen, Z., Cheng, K., Walton, Z., Wang, Y., Ebi, H., Shimamura, T., *et al.* 2012. A murine lung cancer co-clinical trial identifies genetic modifiers of therapeutic response. *Nature*, 483 (7391), pp. 613-617.

31. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., *et al.* 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, 31(13), pp. 3497-3500.

32. Choudhary, A.N., Honbo, D., Kumar, P., Ozisikyilmaz, B., Misra, S. and Memik G. 2011. Accelerating Data Mining Workloads: Current Approaches and Future Challenges in System Architecture Design, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1 (1), pp. 41-54.

33. Classon, M. and Harlow, E. 2002. The retinoblastoma tumour suppressor in development and cancer. *Nat. Rev. Cancer,* 2(12), pp. 910–917.

34. Coffey, D.S., 2001. Similarities of prostate and breast cancer: evolution, diet, and estrogens. *Urology,* 57(4 Suppl 1), pp. 31–38.

35. Copeland, N.G. and Jenkins, N.A. 2009. Deciphering the genetic landscape of cancer-from genes to pathways, *Trends Genet.*, 25(10), pp. 455-62.

36. Costa, V.L., Henrique, R., Ribeiro, F.R., Carvalho, J.R., Oliveira, J., Lobo, F., et al. 2010. Epigenetic regulation of Wnt signaling pathway in urological cancer. *Epigenetics*, 5(4), pp. 343-351.

37. Darwin, C. 1859. *The Origin of Species by Means of Natural Selection*: *Or, the Preservation of Favored Races in the Struggle for Life,* John Murray publishers, London.

38. Das, S., Girard, L., Green, T., Weitzman, L., Lewis-Bowen, A. and Clark, T. 2009. Building Biomedical web communities using a semantically aware content management system. *Briefing in Bioinformatics*, 10 (2), pp. 129-138.

39. Dasu, T. and Johnson, T. 2003. *Exploratory Data Mining and Data Cleaning*, John Wiley & Sons Inc., ISBN 0-471-26851-8.

40. Davis, S. and Meltzer, P.S. 2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, 23(14), pp. 1846-1847.

41. Daxinger, L. and Whitelaw, E. 2010. Transgenerational epigenetic inheritance: More questions than answers. *Genome Res.*, 20(12), pp. 1623-1628.

42. Deng, Z., Sobell, J.L. and Knowles, J.A. 2010. Epigenetic Alternations in Schizophrenia. *Focus*, 8, pp. 358-365.

43. Denissenko, M.F., Chen, J.X., Tang, M. and Pfeifer, G.P. 1997. Cytocine methylation determines hot spots of DNA damage in the human P53 gene. *Proc Natl Acad Sci*, 94(8), pp. 3893-3898.

44. Devlin, B. 2012. The Big data zoo- taming the beast, The need for an integrated platform for enterprise information. *9Sight Consulting,* white paper.

45. Doerfler, W., Toth, M., Kochaneka, S., Achten, S., Freisem-Rabien, U., Behn-Krappa, A., *et al.* 1990. Eukaryotic DNA methylation : facts and problems. *FEBS Lett.*, 268(2), pp. 329-333.

46. Dong, S.M., Kim, K.M., Kim, S.Y., Shin, M.S., Na, E.Y., Lee, S.H., Park, W.S., Yoo, N.J., Jang, J.J., Yoon, C.Y. *et al.* 1998. Frequent somatic mutations in serine/threonine kinase 11/Peutz-Jeghers syndrome gene in left-sided colon cancer. *Cancer Res*, 58(17), pp. 3787–3790.

47. Duncan, C.G., Barwick, B.G., Jin, G., Rago, C., Kapoor-Vazirani, P., Powell, D.R., *et.al.* 2012. A heterozygous IDH1R132H/WT mutation induces genome-wide alterations in DNA methylation. *Genome Research*, 22(12), pp. 2339-2355.

48. Egger, G., Liang, G., Aparicio, A. and Jones, P.A. 2004. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429 (6990), pp. 457-463.

49. Esteller, M. 2007. Cancer epigenomics: DNA methylomes and histone –modification maps. *Nat Rev Genet*., 8(4), pp. 286-298.

50. Esteller, M. 2011. Epigenetic changes in cancer. *F1000 Biol Reports*, 3:9 (doi: 10.3410/B3-9).

51. Eyre-Walker, A. and Keightley, P.D. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol*, 26(9), pp. 2097–2108.

52. Eyre-Walker, A., Keightley, P.D., Smith, N.G. and Gaffney, D. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. *Mol Biol Evol*, 19(12), pp. 2142–2149.

53. Fayyad, U. M., Djorgovski, S. G. and Weir, N. 1996. *Advances in knowledge discovery and data mining*, eds. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R., AAAI Press/The MIT Press, Menlo Park, CA. ISBN 978-0-262-56097-9.

54. Fearnhead, N.S., Wilding, J.L. and Bodmer, W.F. 2002. Genetics of colorectal cancer: hereditary aspects and overview of colorectal tumorigenesis. *Br Med Bull.,* 64 (1), pp. 27-43.

55. Feinberg, A.P. 2004. The epigenetics of cancer etiology. *Semin Cancer Biol,* 14(6), pp. 427-432.

56. Feinberg, A.P. and Tycko, B. 2004. The history of cancer epigenetics. *Nat Rev Cancer*, 4(2), pp. 143-153.

57. Feinberg, A.P., Ohlsson, R. and Henikoff, S. 2006. The epigenetic progenitor origin of human cancer. *Nat Rev genet.,* 7(1), pp. 21-33.

58. Ferber, J. 1999. *Multi-agent systems:An Introduction to distributed artificial intelligence*, Harlow: Addison- Wesley Longman, ISBN 0-201-36048-9.

59. Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C. and Parkin, D. M. (Editors). 2010. GLOBOCAN 2008 v 2.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10, Lyon, France: International Agency for Research on Cancer. Available from: http://globocan.iarc.fr [Accessed on 01/09/2012].

60. Fischer, C. and Reichert, S. 2009. The Genetic code, mutations and Translation, Kaplan Medical USMLE Master the Boards Step 3, pp. 43-62

61. Fraga, M.F., Ballestar, E., Paz, M.F, Ropero, S., Setien, F., Esteller, M. *et al.* 2005. Epigenetic differences arise during the lifetime of monozygotic twins. *PNAS,* 102 (30), pp. 10604- 10609.

62. Frank, D.A. (Ed.) 2012. *Signalling Pathways in Cancer Pathogenesis and Therapy.* 1st ed. Springer.

63. Fraser, H.B., Hirsh, A.E., Lars, M., Steinmetz, L.M., Scharfe, C. and Feldman, M.W. 2002. Evolutionary rate in the protein interaction network. *Science*, 296(5568), pp. 750–752.

64. Friedberg, E.C. 2003. DNA damage and repair. *Nature*, 421(6921), pp. 436-440.

65. Fuks, F. 2005. DNA methylation and histone modifications: teaming up to silence genes. *Curr Opin Genet Dev,* 15(5), pp. 490-495.

66. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., *et al*. 2004. A census of human cancer genes. *Nat Rev Cancer*, 4(3), pp. 177-183.

67. Gao, L. and Zhang, J. 2003. Why are some human disease-associated mutations fixed in mice? *TRENDS in Genetics*, 19(12), pp. 678-681.

68. Giovannucci, E. 2003. Nutrition, insulin, insulin-like growth factors and cancer. *Horm Metab Res,* 35 (11-12), pp. 694–704.

69. Glaser, R.L., Ramsay, J.P. and Morison, I.M. 2006. The imprinted gene and parent-of-origin effect database now includes parental origin of *de novo* mutations. *Nucleic Acids Research*, 34(Database issue): D29-31.

70. Grafodatskaya, D., Chung, B., Szatmari, P. and Weksberg, R. 2010. Autism spectrum disorders and epigenetics. *J Am Acad Child Adolesc Psychiatry,* 49(8), pp. 794-809.

71. Graham, J. 1992. Cancer Selection: The New Theory of Evolution, First Edition, Aculeus Press, Lexington, Virginia.

72. Groden, J., Thliveris, A., Samowitz, W., Carlson, M., Gelbert, L., Albertsen, H., Joslyn, G., Stevens, J., Spirio, L., Robertson, M., *et al.* (1991). "Identification and characterization of the familial adenomatous polyposis coli gene". *Cell,* 66 (3), pp. 589–600.

73. Groth, R. and Lewis, W. J. (Reviewer). 1997. *Data Mining: A Hands*-On *Approach for Business Professionals*. *Prentice* Hall. PTR (ECS Professional), 1997. ISBN 0-13-756412-0.

74. Grunau, C., Renault, E., Rosenthal, A. and Roizes, G. 2001. MethDB - a public database for DNA methylation data. *Nucleic Acids Research,* 29(1), pp. 270–274.

75. Guo, A., Salomoni, P., Luo, J., Shih, A., Zhong, S., Gu, W. and Pandolfi, P.P. 2000. The function of PML in p53-dependent apoptosis. *Nat Cell Biol*, 2(10), pp. 730–736.

76. Güran, S., Tunca, Y. and Imirzalioğlu, N. 1999. Hereditary TP53 codon 292 and somatic P16INK4A codon 94 mutations in a Li-Fraumeni syndrome family. *Cancer Genet Cytogenet*, 113(2), pp.145–151.

77. Hales, C. N. and Barker, D. J. 1992. Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. *Diabetologia*, 35(7), pp. 595–601.

78. Hahn, S.A., Schutte, M., Hoque, A.T., Moskaluk, C.A., da Costa, L.T., Rozenblum, E., Weinstein, C.L., Fischer, A., Yeo, C.J., Hruban, R.H., *et al.* 1996. DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1. *Science,* 271(5247), pp. 350–353.

79. Hakem, R., de la Pompa, J.L., Sirard, C., Mo, R., Woo, M., Hakem, A., *et al*. 1996. The tumor suppressor gene Brca1 is required for embryonic cellular proliferation in the mouse. *Cell*, 85(7), pp. 1009-1023.

80. He, X., Chang, S., Zhang, J., Zhao, Q., Xiang, H., Kusonmano, K., *et al.* 2008. MethyCancer: the database of human DNA methylation and cancer. *Nucleic Acids Research*, 36 (Database issue), D836–D841.

81. HealthGrid-White_Paper-Draft_v.1.1-5, 2004.

82. Heimbigner, D. and McLeod, D. 1985. A Federated Architecture for Information Management, *ACM Transactions on Office Information Systems,* 3(3).

83. Hemminki, A., Markie, D., Tomlinson, I., Avizienyte, E., Roth, S., Loukola, A., Bignell, G., Warren, W., Aminoff, M., Höglund, P. *et al.* 1998. A serine/threonine kinase gene defective in Peutz- Jeghers syndrome. *Nature*, 391(6663), pp. 184–187.

84. Herrera, F. and Herrera–Viedma, E. 2000. Linguistic Decision Analysis: Steps for Solving Decision Problems under Linguistic Information, *Fuzzy Sets Systems,* 115(1), pp. 67 -82.

85. Hirano, R., Interthal, H., Huang, C., Nakamura, T., Deguchi, K., Choi, K., *et al.* 2007. Spinocerebellar ataxia with axonal neuropathy: consequence of a Tdp1 recessive neomorphic mutation? *The EMBO Journal*, 26(22), pp. 4732-4743.

86. Hofmann, T.G., Möller, A., Sirma, H., Zentgraf, H., Taya, Y., Dröge, W., Will, H. and Schmitz ML. 2002. Regulation of p53 activity by its interaction with homeodomain-interacting protein kinase-2. *Nat Cell Biol*, 4(1), pp. 1–10.

87. House, S., Handbook of Epigenetics: The New Molecular and Medical Genetics, Editor: Trygve Tollefsbol, Publisher: Academic Press; 1 edition (21 Oct 2010), ISBN-10: 0123757096, ISBN-13: 978-0123757098.

88. Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., *et al.* 2005. Ensembl 2005. *Nucleic Acids Res*, 33(Database issue), D447–D453.

89. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark L., *et al.* 2002. The Ensembl genome database project. *Nucleic Acids Res*, 30(1), pp. 38–41.

90. Hunter B. Fraser, H.B.,Hirsh, A.E., Steinmetz, L.M., Scharfe, C. and Feldman M.W. 2002. Evolutionary rate in the protein interaction network. *Science*, 296(5568), pp. 750–752.

91. Huss, J.W., Orozco, C., Goodale, J., Wu, C., Batalov, S., Vickers, T.J., *et al.* 2008. A Gene Wiki for community annotation of gene function, *PLoS Biol*, 6(7), e175.

92. Ichimura, K., Bolinm, M.B., Goike, H.M., Schmidt, E.E., Moshref, A. and Collins, V.P. 2000. Deregulation of the p14ARF/MDM2/p53 pathway is a prerequisite for human astrocytic gliomas with G1-S transition control gene abnormalities. *Cancer Res,* 60(2), pp. 417–424.

93. Iwata, K., Matsuzaki, H., Takei, N., Manabe, T. and Mori, N. 2010. Animal models of autism: An epigenetic and environmental viewpoint. *Journal of Central Nervous System Disease*, 2, pp. 37-44.

94. Jablonka, E. and Lamb, M. 1995. Epigenetic Inheritance and Evolution: The Lamarckian Dimension. *Oxford University Press USA*, ISBN-10: 0198540639, ISBN-13: 978-019854063.

95. Jablonka, E. and Lamb, M. J. 2002. The changing concept of epigenetics. *Ann. N. Y. Acad. Sci.*, 981, pp. 82-96.

96. Jablonski N.G. and Chaplin G. 2010. Human skin pigmentation as an adaptation to UV radiation. *PNAS*, 107 (2), pp. 8962-8968.

97. Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E. and Forman, D. 2011. Global Cancer Statistics. *CA Cancer J Clin*, 61(2), pp. 69-90.

98. Jiang, Y.H., Bressler, J. and Beaudet A.L. 2004. Epigenetics and Human disease. *Annu Rev Genomics Hum Genet.,* 5, pp. 479-510.

99. Jones, P.A. and Baylin, S.B. 2002. The fundamental role of epigenetic events in cancer. *Nat Rev Genet*, 3 (6), pp. 415-428.

100. Kariola, R., Otway, R., Lönnqvist, K.E., Raevaara, T.E., Macrae, F., Vos, Y.J., Kohonen-Corish, M., Hofstra, R.M. and Nyström-Lahti, M. 2003. Two mismatch repair gene mutations found in a colon cancer patient–which one is pathogenic? *Hum Genet*, 112(2), pp. 105–109.

101. Kerr, G., Ruskin, H.J., Crane, M. and Doolan, P., 2008. *Techniques for clustering gene expression data, Computers in Biology and Medicine*, 38 (3), pp. 283-293.

102. Kim, J.C., Kim, H.C., Roh, S.A., Koo, K.H., Lee, D.H., Yu, C.S., Lee, J.H., Kim, T.W., Lee, H.L., Beck, N.E., *et al*. 2001. hMLH1 and hMSH2 mutations in families with familial clustering of gastric cancer and hereditary non-polyposis colorectal cancer. *Cancer Detect Prev*, 25(6), pp. 503–510.

103. Kimura, M. 1968. Evolutionary Rate at the Molecular Level. *Nature*, 217, pp. 624 – 626.

104. Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press.

105. Kishi, S., Zhou, X.Z., Ziv, Y., Khoo, C., Hill, D.E., Shiloh, Y. and Lu, K.P. 2001. Telomeric protein Pin2/TRF1 as an important ATM target in response to double strand DNA breaks. *J Biol Chem*, 276(31), pp. 29282–29291.

106. Kolodner, R.D., Tytell, J.D., Schmeits, J.L., Kane, M.F., Gupta, R.D., Weger, J., Wahlberg, S., Fox, E.A., Peel, D., Ziogas, A. *et al*. 1999. Germ-line msh6 mutations in colorectal cancer families. *Cancer Res*, 59(20), pp. 5068–5074.

107. Kosinski, J., Hinrichsen, I., Bujnicki, J.M., Friedhoff, P. and Plotz, G. 2010. Identification of Lynch syndrome mutations in the MLH1-PMS2 interface that disturb dimerization and mismatch repair. *Hum Mutat*, 31(8), pp. 975–982.

108. Kosiol, C., Vinař, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., *et al*. 2008. Patterns of Positive Selection in Six Mammalian Genomes. *PLoS Genet,* 4(8), e1000144.

109. Kouzarides, T. 2007. Chromatin modifications and their function. *Cell*, 128(4), pp. 693–705.

110. Kurzawski, G., Suchy, J., Lener, M., Kłujszo-Grabowska, E., Kładny, J., Safranow, K., Jakubowska, K., Jakubowska, A., Huzarski, T., Byrski, T. *et al*. 2006. Germline MSH2 and MLH1 mutational spectrum including large rearrangements in HNPCC families from Poland (update study). *Clin Genet*, 69(1), pp. 40–47.

111. Lamarck, J. 1914. *Zoological Philosophy* trans. Elliot H., p.113.

112. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., *et al*. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21), pp. 2947-2948.

113. Lassila, O. and Hendler, J. 2007. Embracing "Web 3.0", *Internet Computing, IEEE*, 11(3), pp. 90-93.

114. Latif, F., Tory, K., Gnarra, J., Yao, M., Duh, F.M., Orcutt, M.L., Stackhouse, T., Kuzmin, I., Modi, W., Geil, L. *et al*. 1993. Identification of the von Hippel-Lindau disease tumor suppressor gene. *Science*, 260(5112), pp. 1317–1320.

115. Levasseur, A., Gouret, P., Lesage-Meessen, L., Asther, M., Asther, M., Record, E., *et al*. 2006. Tracking the connection between evolutionary and functional shifts using the fungal lipase/feruloyl esterase A family. *BMC Evol Biol*, 6, pp. 92.

116. Lindberg, D.A.B., Humphreys, B.L. and McCray, A.T. 1993. The Unified Medical Language System, *Meth. Inform. Med.* 32, pp. 281-291.

117. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q. *et al.* 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462, pp. 315-322.

118. Liu, Y. and Kulesz-Martin, M. 2001. p53 protein at the hub of cellular DNA damage response pathways through sequence-specific and non-sequence-specific DNA binding. *Carcinogenesis*, 22 (6), pp. 851-860.

119. Loughran, N.B., Hinde, S., McCormick-Hill, S., Leidal, K.G., Bloomberg, S., Loughran, S.T., *et al.* 2012. Functional consequence of positive selection revealed through rational mutagenesis of human myeloperoxidase. *Mol Biol Evol.* 29(8), pp. 2039-2046.

120. Loughran, N.B., O'Connor, B., O'Fágáin, C and O'Connell, M.J. 2008. The phylogeny of the mammalian heme peroxidases and the evolution of their diverse functions. *BMC Evol Biol*, 8, pp. 101. doi: 10.1186/1471-2148-8-101.

121. Lunzer, M., Golding, G.B. and Dean A.M. 2010. Pervasive cryptic epistasis in molecular evolution. *PLoS Genet.*, 6(10), pp. e1001162, doi: 10.1371/journal.pgen.1001162.

122. MacColl, A.D. 2011. The ecological causes of evolution. *Trends Ecol Evol.*, 26(10), pp. 514–522.

123. Marakas, G.M. 2003. *Modern Data Warehousing, Mining and Visualization: Core Concepts*. First ed. Prentice Hall, ISBN 0131014595.

124. Marambaud, P., Shioi, J., Serban, G., Georgakopoulos, A., Sarner, S., Nagy, V., Baki, L., Wen, P., Efthimiopoulos, S., Shao, Z., *et al.* 2002. A presenilin-1/gamma-secretase cleavage releases the Ecadherin intracellular domain and regulates disassembly of adherens junctions. *EMBO J*, 21(8), pp. 1948–1956.

125. Massaad, L., de Waziers, I., Ribrag, V., Janot, F., Beaune, P.H., Morizet, J., Gouyette, A. and Chabot, G.G. 1992. Comparison of mouse and human colon tumors with regard to phase I and phase II drug-metabolizing enzyme systems. *Cancer Res.,* 52(23), pp. 6567-6575.

126. Mazin, A.L. 1995. Methylation of the Factor IX Gene is a Main Cause of Mutations Responsible for Hemophilia B. *Molecular Biology*, 29(1), pp. 73-92.

127. McInerney, J.O. 2006. The causes of protein evolutionary rate variation. *Trends Ecol Evol.*, 21(5), pp. 230-232.

128. Michels, K.B. 2005. The role of nutrition in cancer development and prevention. *Int. J. Cancer,* 114 (2), pp. 163–165.

129. Minoshima, S., Mitsuyama, S., Ohtsubo, M., Kawamura, T., Ito, S., Shimizu, N., *et al.* 2001. The KMDB/MutationView: a mutation database for human disease genes, *Nucleic Acids Res.*, 29 (1), pp. 327-328.

130. Mons, B., Ashburner, M., Chichester, C., Mulligen, E.V., Weeber, V., Dunnen, J. *et al.* 2008. Calling on a million minds for community annotation in WikiProteins. *Genome Biol.* 9 (5), R89.

131. Morales-del-Castillo, J.M., Peis, E., Ruiz, A.A. and Herrera-Viedma, E. 2010. Recommending Biomedical Resources: A Fuzzy Linguistic Approach Based on Semantic Web, *International Journal of Intelligent Systems*, 25 (12), pp. 1143-1157.

132. Morgan, C. C.*, Shakya, K.*, Webb, A. E., Walsh, T. A., Lynch, M., Loscher, C. E., *et al.* 2012. Colon cancer associated genes exhibit signatures of positive selection at functionally important positions. *BMC Evol Biol*, 12(114), doi:10.1186/1471-2148-12-114.

133. Moury B and Simon V. 2011. dN/dS-based methods detect positive selection linked to tradeoffs between different fitness traits in the coat protein of potato virus Y. *Mol Biol Evol.*, 28(9), pp. 2707–2717.

134. Murphy, W.J., Eizirik, E., O'Brien, S.J., Madsen, O., Scally, M., Douady, C.J., Teeling, E., Ryder, O.A., Stanhope, M.J., de Jon, W.W. *et al.* 2001. Resolution of the early

placental mammal radiation using Bayesian phylogenetics. Science, 294(5550), pp. 2348-2351.

135. Nakagawa, H., Koyama, K., Miyoshi, Y., Ando, H., Baba, S., Watatani, M., Yasutomi, M., Matsuura, N., Monden, M. and Nakamura, Y. 1998. Nine novel germline mutations of STK11 in ten families with Peutz-Jeghers syndrome. *Hum Genet*, 103(2), pp.168–172.

136. National Cancer Institute, 2010. *Factsheet, Cancer Staging* [Online]. Available from: http://www.cancer.gov/cancertopics/factsheet/detection/staging [Accessed 29 November 2012].

137. Newick Utilities 2012. Newick Utilities: A suite of shell tools to process phylogenetic trees, 2012. http://cegg.unige.ch/newick_utils

138. Nickerson, M.L., Warren, M.B., Toro, J.R., Matrosova, V., Glenn, G., Turner, M.L., Duray, P., Merino, M., Choyke, P., Pavlovich, C.P., *et al.* 2002. Mutations in a novel gene lead to kidney tumors, lung wall defects, and benign tumors of the hair follicle in patients with the Birt-Hogg- Dube syndrome. *Cancer Cell*, 2(2), pp. 157–164.

139. Nicolaides, N.C., Papadopoulos, N., Liu, B., Wei, Y.F., Carter, K.C., Ruben, S.M., Rosen, C.A., Haseltine, W.A., Fleischmann, R.D., Fraser, C.M., *et al.* 1994. Mutations of two PMS homologues in hereditary nonpolyposis colon cancer. *Nature* 1994, 371(6492), pp. 75–80.

140. Nielsen, R. and Yang, Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene, *Genetics*, 148(3), pp. 929-936.

141. Niemann, S. and Muller, U. 2000. Mutations in SDHC cause autosomal dominant paraganglioma, type 3. *Nat Genet.*, 26(3), pp. 268–270.

142. Novik, K. L., Nimmrich, I., Genc, B., Maier, S., Piepenbrock, C., Olek, A., *et al.* 2002. Epigenomics: genome-wide study of methylation phenomena. *Curr. Issues Mol. Biol.*, 4(4), pp. 111-128.

143. Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., *et al.* 2009. BioPortal: ontologies and integrated data resources at the click of a mouse, *Nucleic Acids Res*., 37(Web Server issue), W170–W173.

144. O'Connell, M.J. 2010. Selection and the cell cycle: positive Darwinian selection in a well-Known DNA damage response pathway. *J Mol Evol*., 71(5-6), pp. 444-457.

145. O'Connell, M.J. and McInerney J.O. 2005. Adaptive evolution of the human fatty acid synthase gene: support for the cancer selection and fat utilization hypotheses? *Gene*, 360(2), pp. 151-159.

146. O'Connell, M.J., Loughran, N.B., Walsh, T.A., Donoghue, M.T.A., Schmid, K.J., Spillane, C. 2010. A phylogenetic approach to test for evidence of parental conflict or

gene duplications associated with protein-encoding imprinted orthologous genes in placental mammals. *Mammalian Genome*, 21 (9-10), pp. 486-498.

147. O'Driscoll, M. and Jeggo, P.A. 2008. The role of the DNA damage response pathways in brain development and microcephaly: insight from human disorders. DNA Repair (Amst.), 7(7), pp. 1039-1050.

148. Ogino, S., Brahmandam, M., Kawasaki, T., Kirkner, G.J., Loda, M. and Fuchs, C.S. 2006. Combined analysis of COX-2 and p53 expressions reveals synergistic inverse correlations with microsatellite instability and CpG island methylator phenotype in colorectal cancer. *Neoplasia*, 8(6), pp. 458-464.

149. Ohmiya, N., Matsumoto, S., Yamamoto, H., Baranovskaya, S., Malkhosyan, S.R. and Perucho, M. 2001. Germline and somatic mutations in hMSH6 and hMSH3 in gastrointestinal cancers of the microsatellite mutator phenotype. *Gene,* 272(1–2), pp. 301–313.

150. Ohta, T. 1973. Slightly Deleterious Mutant Substitutions in Evolution. *Nature*, 246 (5428), pp. 96–98.

151. Ohta, T., and Gillespie, J.H. 1996. Development of Neutral and Nearly Neutral Theories. *Theoretical Population Biology*, 49 (2), pp. 128–142.

152. Ollila, S., Dermadi Bebek, D., Jiricny, J. and Nyström, M. 2008. Mechanisms of pathogenicity in human MSH2 missense mutants. *Hum Mutat.*, 29(11), pp. 1355–1363.

153. Oren, E., Delbru, R., Moller, R., Volkel, M. and Handschuh, S. 2006. Annotation and Navigation in Semantic Wikis, SemWiki WS at ESWC.

154. Ortega, S., Malumbres, M. and Barbacid, M. 2002. Cyclin D-dependent kinases, INK4 inhibitors and cancer. *Biochim. Biophys. Acta.,* 1602(1), pp. 73–87.

155. Pancione, M., Remo, A. and Colantuoni, V. 2012. Genetic and Epigenetic Events Generate Multiple Pathways in Colorectal Cancer Progression. *Pathology Research International*, doi:10.1155/2012/509348.

156. Pensotti, V., Radice, P., Presciuttini, S., Calistri, D., Gazzoli, I., Grimalt Perez, A., Mondini, P., Buonsanti, G., Sala, P., Rossetti, C. *et al.* 1997. Mean age of tumor onset in hereditary nonpolyposis colorectal cancer (HNPCC) families correlates with the presence of mutations in DNA mismatch repair genes. *Genes Chromosomes Cancer*, 19(3), pp. 135–142.

157. Plaschke, J., Krüger, S., Dietmaier, W., Gebert, J., Sutter, C., Mangold, E., Pagenstecher, C., Holinski-Feder, E., Schulmann, K., Möslein, G. *et al.* 2004. Eight novel MSH6 germline mutations in patients with familial and nonfamilial colorectal cancer selected by loss of protein expression in tumor tissue. *Hum Mutat*, 23(3), pp 285.

158. Poole, R.L., Leith, D.J., Docherty, L.E., Shmela, M.E., Gicquel, C., Splitt, M., *et al.* 2012. Beckwith-Wiedemann syndrome caused by maternally inherited mutation of an

OCT-binding motif in the IGF2/H19-imprinting control region, ICR1. *European journal of human genetics*, 20(2), pp. 240-243.

159. Pruitt, K.D., Harrow, J., Harte, R.A., Wallin, C., Diekhans, M., Maglott, D.R., *et al.* 2009. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes, *Genome Res.*, 19(7), pp. 1316-1323.

160. Qu, G.Z., Grundy, P.E., Narayan, A. and Ehrlich, M. 1999. Frequent hypomethylation in Wilms tumors of pericentromeric DNA in chromosomes 1 and 16. *Cancer Genet Cytogenet.*, 109(1), pp. 34-39.

161. Quintero-Ronderos, P. and Montoya-Ortiz, G. 2012. Epigenetics and Autoimmune Diseases. *Autoimmune Diseases*, doi:10.1155/2012/593720.

162. Rambaut, A. 1996. Se-AL Sequence alignment editor. Oxford.

163. Reik, W. and Walter, J. 2001. Genomic imprinting: parental influence on the genome. *Nat Rev Genetics*, 2(1), pp. 21-32.

164. Rideout, W.M., Coetzee, G.A., Olumi, A.F. and Jones, P.A. 1990. 5- Methylcytosine as an endogenous mutagen in the human LDL receptor and *p53* genes. *Science,* 249 (4974), pp. 1288-1290.

165. Rodenhiser, D. and Mann, M. 2006. Epigenetics and human disease: translating basic biology into clinical applications. *CMAJ*, 174(3), pp. 341-348.

166. Rosenfeld C.S. 2010. Animal Models to Study Environmental Epigenetics. *Biology of Reproduction,* 82(3), pp. 473-488 .

167. Roth, T.L., Lubin, F.D., Sodhi, M. and Kleinman, J.E. 2009. Epigenetic mechanisms in schizophrenia. *Biochim Biophys Acta*, 1790(9), pp. 869-877.

168. Sacho, E.J., Kadyrov, F.A., Modrich, P., Kunkel, T.A. and Erie, D.A. 2008. Direct visualization of asymmetric adenine-nucleotide induced conformational changes in MutL alpha. *Mol Cell.*, 29(1), pp. 112–121.

169. Sakamoto, J., Fujiya, M., Okamoto, K., Nata, T., Inaba, Y., Moriichi, K., Tanabe, H., Mizukami, Y., Watari, J., Ashida, T., *et al.* 2010. Immunoprecipitation of nucleosomal DNA is a novel procedure to improve the sensitivity of serum screening for the p16 hypermethylation associated with colon cancer. *Cancer Epidemiol.*, 34 (2), pp. 194-199.

170. Salcedo, T., Geraldes, A. and Nachman, M.W. 2007. Nucleotide variation in wild and inbred mice. *Genetics*, 177(4), pp. 2277–2291.

171. Sawan, C., Vaissière, T., Murr, R. and Herceg, Z. 2008. Epigenetic drivers and genetic passengers on the road to cancer. *Mutat Res*. 642(1-2), pp. 1-13.

172. Sawyer, S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol.,* 6(5), pp. 526–538.

173. Sayed, M.G., Ahmed, A.F., Ringold, J.R., Anderson, M.E., Bair, J.L., Mitros, F.A., Lynch, H.T., Tinley, S.T., Petersen, G.M., Giardiello, F.M., *et al.* 2002. Germline SMAD4 or BMPR1A mutations and phenotype of juvenile polyposis. *Ann Surg Oncol*, 9(9), pp. 901–906.

174. Schmid, K. and Yang, Z. 2008. The trouble with sliding windows and the selective pressure in BRCA1. *PLoS One*, 3(11), e3746.

175. Schmid, R., Baum, P., Ittrich, C., Fundel-Clemens, K., Huber, W., Brors, B., *et al.* 2010. Comparison of normalization methods for Illumina BeadChip HumanHT -12 v3. *BMC Genomics*, 11, 349, doi:10.1186/1471-2164-11-349.

176. Shakya, K., O'Connell, M.J. and Ruskin, H.J. 2012. The landscape for epigenetic/epigenomic biomedical resources. *Epigenetics*, 7 (9), pp. 982 - 986.

177. Shakya, K., Ruskin, H.J., Kerr, G., Crane, M. and Becker, J. 2010. Comparison of microarray pre-processing methods. Chapter 2, *Adv Exp Med Biol.*, 680 (2), pp. 139-147.

178. Sharma, S., Kelly, T.K. and Jones, P.A. 2009. Epigenetics in cancer. *Carcinogenesis*, 31(1), pp. 27-36.

179. Sherr, C.J. 2000. Cancer cell cycles revisited. *Cancer Res*., 60, pp. 3689–3695.

180. Sheth, A.P. and Larson, J.A. 1990. Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases, *ACM Computing Surveys*, 22(3), pp. 183-236.

181. Sjöblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., *et al.* 2006. The consensus coding sequences of human breast and colorectal cancers. *Science*, 314 (5797), pp. 268–274.

182. Steinmann, K., Sandner, A., Schagdarsurengin, U., *et al*. 2009. Frequent promoter hypermethylation of tumor-related genes in head and neck squamous cell carcinoma. *Oncol. Rep*., 22, pp. 1519–1526.

183. Stöger, R. 2008. The thrifty epigenotype: an acquired and heritable predisposition for obesity and diabetes?. *BioEssays,* 30 (2), pp. 156–166.

184. Strate, L.L. and Syngal, S. 2005. Hereditary colorectal cancer syndromes. *Cancer Causes Control*, 16 (3), pp. 201-213.

185. Susana Gonzalo, S. 2010. Epigenetic alterations in aging. *Journal of Applied Physiology*, 109(2), pp. 586-597.

186. Tamura, G., Sakata, K., Nishizuka, S., Maesawa, C., Suzuki, Y., Iwaya, T., Terashima, M., Saito, K. and Satodate, R. 1996, Inactivation of the E-cadherin gene in primary gastric carcinomas and gastric carcinoma cell lines. *Jpn J Cancer Res*, 87(11), pp. 1153–1159.

187. Tee, A.R., Fingar, D.C., Manning, B.D., Kwiatkowski, D.J., Cantley, L.C. and Blenis, J. 2002. Tuberous sclerosis complex-1 and −2 gene products function together to inhibit mammalian target of rapamycin (mTOR)-mediated downstream signaling. *Proc Natl Acad Sci U S A*, 99(21), pp. 13571–13576.

188. Tenenbaum, J.D., Whetzel, P.L., Anderson, K., Borromeo, C.D., Dinov, I.D., Gabriel, D., *et al*. 2011. The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research*, Journal of Biomedical Informatics*, 44(1), pp. 137-145.

189. Tenesa, A., Navarro, P., Hayes, B.J., Duffy, D.L., Clarke, G.M., Goddard, M.E., *et al.* 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res.*, 17(4), pp. 520–526.

190. Tournier, I., Vezain, M., Martins, A., Charbonnier, F., Baert-Desurmont, S., Olschwang, S., Wang, Q., Buisine, M.P., Soret, J., Tazi, J. *et al*. 2008. A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. *Hum Mutat*, 29(12), pp. 1412–1424.

191. Toyota, M. and Suzuki, H. 2010. Epigenetic drivers of genetic alterations. *Advances in Genetics*, 70, pp. 309-323.

192. Tudzarova, S., Colombo, S.L., Stoeber, K., Carcamo, S., Williams, G.H. and Moncada, S. 2011. Two ubiquitin ligases, APC/C-Cdh1 and SKP1- CUL1-F (SCF)-beta-TrCP, sequentially regulate glycolysis during the cell cycle. *Proc Natl Acad Sci,* 108(13), pp. 5278–5283.

193. UMLS 2013. Unified medical Language System [Online]. Available from: http://www.nlm.nih.gov/pubs/factsheets/umls.html [Accessed 6 Jan 2013]

194. UniProt Consortium. 2011. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research,* 39 (Database issue), D214-219.

195. Varley, J.M., McGown, G.,Thorncroft, M., Tricker, K.J.,Teare, M.D., Santibanez-Koref, M.F., Martin, J.,Birch, J.M., and Evans, D.G. 1995. An extended Li-Fraumeni kindred with gastric carcinoma and a codon 175 mutation in TP53. *J Med Genet*, 32(12), pp. 942–945.

196. Vilar, E. and Gruber, S.B. 2010. Microsatellite instability in colorectal cancer-the stable evidence. *Nat Rev Clin Oncol*, 7(3), pp. 153–162.

197. Vogelstein, B. and Kinzler, K.W. 2004. Cancer genes and the pathways they control, *Nature Medicine,* 10 (8), pp. 789-799.

198. Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. 2006. A map of recent positive selection in the human genome. *PLoS Biol*, 4(3), e72.

199. Wang, Q., Lasset, C., Desseigne, F., Saurin, J.C., Maugard, C., Navarro, C., *et al.* 1999. Prevalence of germline mutations of hMLH1, hMSH2, hPMS1, hPMS2, and hMSH6

genes in 75 French kindreds with nonpolyposis colorectal cancer. *Hum Genet.*, 105(1–2), pp. 79–85.

200. Wang, Q., Lasset, C., Desseigne, F., Saurin, J.C., Maugard, C., Navarro, C., Ruano, E., Descos, L., Trillet-Lenoir, V., Bosset, J.F., *et al.* 1999. Prevalence of germline mutations of hMLH1, hMSH2, hPMS1, hPMS2, and hMSH6 genes in 75 French kindreds with nonpolyposis colorectal cancer. *Hum Genet*, 105(1–2), pp. 79–85.

201. Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., *et al.* 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915), pp. 520-562.

202. Weinstein, I.B. and Joe, A.K. 2006. Mechanisms of disease: oncogene addiction—a rationale for molecular targeting in cancer therapy. *Nature Clinical Practice Oncology,* 3, pp. 448–457.

203. Weksberg, R., Smith, A.C., Squire, J. and Sadowski, P. 2003. Beckwith Wiedemann syndrome demonstrates a role for epigenetic control of normal development. *Human Molecular Genetic*, 12(1), pp. R61–R68.

204. Wellcome Trust Sanger Institute 2013. Catalogue of Somatic Mutations in Cancer [Online]. Available from: http://www.sanger.ac.uk/genetics/CGP/cosmic/ [Accessed 6 Jan 2013]

205. Westerman, A.M., Entius, M.M., Boor, P.P., Koole, R., de Baar, E., Offerhaus, G.J., Lubinski, J., Lindhout, D., Halley, D.J., de Rooij, F.W. *et al.* 1999, Novel mutations in the LKB1/STK11 gene in Dutch Peutz-Jeghers families. *Hum Mutat*, 13(6), pp. 476–481.

206. Wong, W.S., Yang, Z., Goldman, N. and Nielsen, R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*, 168(2), pp. 1041–1051.

207. Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjoblom, T., Leary, R.J., *et al.* 2007. The Genomic Landscapes of Human Breast and Colorectal Cancers, *Science,* 318(5853), pp. 1108-1113.

208. Wu, Y., Berends, M.J., Sijmons, R.H., Mensink, R.G., Verlind, E., Kooi, K.A., van der Sluis, T., Kempinga, C., van dDer Zee, A.G., Hollema, H. *et al.* 2001. A role for MLH3 in hereditary nonpolyposis colorectal cancer. *Nat Genet*, 29(2), pp. 137–138.

209. Xia, J., Han, L. and Zhao, Z. 2012. Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome. *BMC Genomics*, 13(Suppl 8) :S7, doi:10.1186/1471-2164-13-S8-S7.

210. Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.*, 13(5), pp. 555-556.

211. Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*, 15(5), pp. 568-573.

212. Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.*, 24(8), pp. 1586-1591.

213. Yang, Z. and Swanson, W. J. 2002. Codon-Substitution Models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol.,* 19(1), pp. 49-57.

214. Yang, Z., Wong, W.S. and Nielsen, R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol.*, 22(4), pp. 1107-1118.

215. Yoon, K.A., Ku, J.L., Yang, H.K., Kim, W.H., Park, S.Y. and Park, J.G. 1999. Germline mutations of E-cadherin gene in Korean familial gastric cancer patients. *J Human Genet.,* 44(3), pp. 177–180.

216. You, J.S. and Jones, P.A. 2012. Cancer Genetics and Epigenetics: Two Sides of the Same Coin?, *Cancer Cell*, 22(1), pp. 9-20.

217. Youngson, R. M. 2006. *Collins Dictionary of Human Biology*. Glasgow: HarperCollins. ISBN 0-00-722134-7.

218. Zhang, J., Nielsen, R. and Yang, Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol.,* 22(12), pp. 2472– 2479.

219. Zhou, F., Su, J., Fu, L., Yang, Y., Zhang, L., Wang, L., *et al.* 2008. Unglycosylation at Asn-633 made extracellular domain of Ecadherin folded incorrectly and arrested in endoplasmic reticulum, then sequentially degraded by ERAD. *Glycoconj J*, 25(8), pp. 727–740.

220. Zhou, F., Su, J., Fu, L., Yang, Y., Zhang, L., Wang, L., Zhao, H., Zhang, D., Li, Z. and Zha, X. 2008. Unglycosylation at Asn-633 made extracellular domain of Ecadherin folded incorrectly and arrested in endoplasmic reticulum, then sequentially degraded by ERAD. *Glycoconj J*, 25(8), pp. 727–740.

221. Zimmer, C. 2007. Evolved for Cancer? *Scientific American*, 296(1), pp. 68-74.

Publications

1. Shakya, K., O' Connell, M. J. and Ruskin, H. J. 2012. The Landscape for Epigenetic/Epigenomic Biomedical Resources*, Epigenetics*, 7 (9), pp. 982 - 986.
2. Morgan, C. C.*, Shakya, K.*, Webb, A. E., Walsh, T. A., Lynch, M., Loscher, C. E., Ruskin, H. J. and O'Connell, M. J. 2012. Colon cancer associated genes exhibit signatures of positive selection at functionally important positions. *BMC Evol Biol*, 12(114), doi:10.1186/1471-2148-12-114.
3. Shakya, K., Ruskin, H. J., Kerr, G., Crane, M. and Becker, J. 2010. Comparison of Microarray Pre-Processing Methods. Chapter 2, *Advances in Computational Biology*, Springer, 680 (2), pp. 139- 147, doi: 10.1007/978-1-4419-5913-3_16.

The landscape for epigenetic/epigenomic biomedical resources

Kabita Shakya,[1,2,3,*] Mary J. O'Connell[2,3] and Heather J. Ruskin[1,3]

[1]Centre for Scientific Computing and Complex Systems Modelling (SCI-SYM); Dublin City University; Dublin, Ireland;
[2]Bioinformatics and Molecular Evolution Group; School of Biotechnology; Dublin City University; Dublin, Ireland;
[3]School of Computing; Dublin City University; Dublin, Ireland

Keywords: biomedical resource, data mining, epigenetics, epigenomics, primary database, secondary database, methylation

Abbreviations: BIG, Beijing Institute of Genomics; BRO, Biomedical Resource Ontology; DDBJ, DNA Data Bank of Japan; EBI, European Bioinformatics Institute; ENCODE, Encyclopaedia of DNA Elements; EPITRON, Epigenetic Treatment of Neoplastic Disease; HEP, Human Epigenome Project; HEROIC, High-Throughput Epigenetic Regulatory Organization in Chromatin; HGP, Human Genome Project; ICGC, International Cancer Genome Consortium; IHEC, International Human Epigenome Consortium; NCBI, National Center for Biotechnology Information; NHGRI, National Human Genome Research Institute; NIH, National Institute of Health; NLM, National Library of Medicine; PDB, Protein Data Bank; SCOP, Structural Classification of Proteins; UHN, University Health Network

Recent advances in molecular biology and computational power have seen the biomedical sector enter a new era, with corresponding development of bioinformatics as a major discipline. Generation of enormous amounts of data has driven the need for more advanced storage solutions and shared access through a range of public repositories. The number of such biomedical resources is increasing constantly and mining these large and diverse data sets continues to present real challenges. This paper attempts a general overview of currently available resources, together with remarks on their data mining and analysis capabilities. Of interest here is the recent shift in focus from genetic to epigenetic/epigenomic research and the emergence and extension of resource provision to support this both at local and global scale. Biomedical text and numerical data mining are both considered, the first dealing with automated methods for analyzing research content and information extraction, and the second (broadly) with pattern recognition and prediction. Any summary and selection of resources is inherently limited, given the spectrum available, but the aim is to provide a guideline for the assessment and comparison of currently available provision, particularly as this relates to epigenetics/epigenomics.

Introduction

The Human Genome Project (HGP) in 2003 led to identification of more than 20,000 genes and determined the three billion chemical base pairs of human DNA. With the tremendous advances in medical technologies, corresponding development in computational power, storage capacity, interconnectivity and cost effectiveness, this explosive growth has resulted in the generation and collection of all aspects of biomedical data and, in the past decade, the importance of bioinformatics has been recognized. [1] Data warehousing[2] as a way of dealing with large data set size, combines databases across an entire enterprise, whereas independent or federated systems seek to integrate multiple autonomous databases into a single federation, with constituent databases interconnected via a network and often geographically decentralised.[3,4] One example is the many bioinformatics data sources linked by the Entrez life sciences search engine.[5]

Biomedical data cover a wide range, from patient records to information from pharmaceutical studies, specific disease research and different "omics" studies, including genomics, proteomics and transcriptomics. Resource types can be classified by two key features: first, the means or method by which access is provided to entities; second, the nature of the entities themselves. The repository or web service that provides access to these data are a vital component of biomedical data resourcing. [6] An example is PubMed, the NLM's web-based interface to MEDLINE, the premier bibliographic index to journal articles in the life sciences. In general, resource providers, such as PubMeth and MutationDB, review research papers from the domain and mine these for information relevant to the scientific audience. Typically, non-profit research institutes, such as the Sanger Institute, University of California Santa Cruz (UCSC), National Center for Biotechnology Information (NCBI), National Institute of Health (NIH), European Molecular Biology Laboratory (EMBL) and European Bioinformatics Institute

*Correspondence to: Kabita Shakya;
Email: kabita.shakya@gmail.com
Submitted: 04/12/12; Revised: 07/11/12;
Accepted: 07/13/12
http://dx.doi.org/10.4161/epi.21493

(EBI), among others, make such data publicly available over the internet so that these can be further analyzed/mined for knowledge discovery.

Biological/biomedical resources may be one of several types, primary, secondary or composite. Examples of primary database containing information on biological quantities themselves indicate those for sequence or structure, *e.g.*, SwissProt, PIR (protein sequences), GenBank and DDBJ (genome sequences). Secondary resources contain derived information from primary sources and examples include eMOTIF (Stanford) and SCOP (Cambridge). Composite resources typically draw information from a variety of different databases, such as those of the NCBI genome browser and Genecards.[7] The most popular genome browsers today are Ensembl, NCBI Map Viewer and UCSC, which act as gateways for access to genetic and epigenetic information.

Following completion of the Human Genome Project, increased attention has been paid to processes that lead to heritable changes in gene expression, during development or across generations, without altering the nucleotide sequence within the DNA. Both epigenetics and epigenomics, the genome-wide distribution of epigenetic changes, have become major areas of research focus. Principal epigenetic phenomena encompass DNA methylation, histone modification (methylation/demethylation, acetylation/ deacetylation, phosphorylation, ubiquitylation and sumoylation), gene silencing, genomic imprinting and X-chromosome inactivation. Recently-launched large-scale initiatives include, among others, IHEC (International Human Epigenome Consortium),[8] which plans to map up to 1,000 reference epigenomes within a decade, and the Human Epigenome Project (HEP),[9] which aims to identify, catalog and interpret genome-wide DNA methylation patterns of all human genes in all major tissues.[10]

Epigenetics, cancer and other diseases. Epigenetic abnormalities have been found to be causative factors of cancer, genetic disorders and pediatric syndromes, as well as contributory factors of autoimmune diseases and aging.[11] The recent intensive research on cancer-epigenetics has also led to the discovery of many epigenetic markers that play an important role in disease initiation. As a consequence, cancer-related epigenetic resources preponderate over others. Two of the large-scale project initiatives for cancer research include ICGC (see "ICGC" section below) and TCGA (The Cancer Genome Atlas). TCGA has achieved comprehensive sequencing, characterization and analysis of the genomic changes in various cancers and intends to chart the genomic changes involved in more than 20 types of cancers.[12] All of the epigenetic resources are outlined in the following sections, with additional assessment of their data mining capabilities, intrinsic or externally accessed, and their adequacy provided where possible.

DNA methylation can induce "epigenetic silencing" (or loss of expression) of tumor suppressor genes, causing normal cells to be transformed into cancer cells and is the first and most common epigenetic alteration to be observed.[13,14] A direct link also exists between DNA methylation and histone modification, since a number of proteins involved in DNA methylation (*e.g.*, DNMTs and MBDs) directly interact with histone modifying enzymes, such as histone methyltransferases (HMTs) and histone deacetylases (HDACs).[15] Epigenetic resources incorporating methylation signatures are described in the "Methylation" section below.

Resources for Epigenetic/Epigenomic Signatures

Epigenetic/epigenomic resources are inevitably less comprehensive to date but can be broadly categorized in terms of type of data content, tools and access, and are described below.

Methylation. Pubmeth,[16] a cancer methylation database, provides a sorted, annotated and summarized overview of genes, reported to be methylated in various cancers, with user query based on gene or cancer type. PubMeth draws on text-mining of Medline/PubMed abstracts, combined with manual annotation of pre-selected abstracts. The text mining approach in PubMeth is fast and intelligent, enabling search of multiple aliases and textual variants of these aliases, and querying of multiple keywordlists simultaneously. PubMeth also provides the facility to browse a pre-computed gene list, without having to query the database directly.

MethDB[17] is also a major source for experimentally confirmed DNA methylation data but is general, more sample-oriented and not optimized to cancer-related queries. The database is designed to store and annotate information on the occurrence of methylated cytosines in DNA. It currently contains 19,905 methylation content data items and 5,382 methylation patterns or profiles for 48 species, 1,511 individuals, 198 tissues and cell lines and 79 phenotypes. MethDB also has a public online submission system available.[18] The resource forms part of an integrated network of biological databases through DAS (Distributed Annotation System), enabling the epigenetic data to be viewed as a layer in the human genome, and is also connected to Ensembl (for DNA sequences with available MethDB data aligned to NCBI Refseq).

A subset resource, MethPrimerDB,[19] is a database of primer sequences used in PCR-based methylation methods. The database depends on submissions by users and administrators that guarantee the required quality of the database but not necessarily its completeness. To date, there are 29 primer sets. In 2006, the MethBLAST feature was added to MethPrimerDB oligonucleotide sequences. Further updates since 2006, however, are not found for this resource.

MethyCancer[20] is a disease-oriented database, specifically of human DNA methylation and cancer that aims to integrate methylation databases and has developed a meta-data format for data standardization, with manual curation still used for noisy data. Four main types of data are included in MethyCancer, namely, (1) CGI clones and global CGI predictions, (2) DNA methylation data, (3) cancer information, genes and mutations and (4) correlations of DNA methylation, gene expression and cancer. MethyView, a visualization tool from MethyCancer, is used to facilitate the browsing of methylation data in the context of existing human genome annotations. A search engine to query different data types and interactions from the MethyCancer database provides simple keyword search and also offers advanced options namely, "methylation," "gene," "cancer," "clone" and "repeat" searches. For example, Methylation search enables the user to specify and combine query options, such as methylation type (pattern, profile, content and domain), data source (BIG/ UHN, MethDB,[17] HEP,[9] Columbia University), experimental methods, sample information (tissue, sex, age and phenotype) and chromosomal positions.

On similar lines, Methylogix[21] provides a high density DNA methylation database of human chromosomes 21 and 22, a CpG island DNA methylation database for male germ cells, enabling comprehensive analysis of DNA methylation variation between and within the germ lines of normal males, and a targeted DNA methylation database of late-onset Alzheimer disease. Similarly, Methtools is a collection of software tools for handling and analysis of DNA methylation data, generated by the bisulfite genomic sequencing method.[22]

Genomic imprinting related resources. Genomic imprinting is an important epigenetic phenomenon whereby inherited genes are "imprinted" due to one copy of the gene being epigenetically marked or imprinted in either the egg or the sperm. Thus, the allelic expression of an imprinted gene depends on whether it is inherited maternally or paternally. Imprinted expression can also vary between tissues, developmental stages and species.[23] The Geneimprint database[24] includes genes and related information on genomic imprinting for different animals including humans and gathered from NCBI. Genes are listed by species and sorted by chromosomal location, name and imprinting status and are provided through the web-interface. Similarly, an imprinted gene and parent-of-origin effect database[25] presents imprinted genes and related effects. This consists of two sections: (1) catalog of current literature on imprinted genes in humans and animals and (2) catalog of reports of parental origin of de novo mutations in humans alone. The addition of (2), showing a parent-of- origin effect, expands the scope of the database and provides a useful tool for examining parental origin trends for different types of spontaneous mutations. This second section currently includes more than 1,700 mutations, found in 59 different disorders. The 85 imprinted genes are described in 152 entries from several mammalian species. In addition, more than 300 other entries describe a range of reported parent-of-origin effects in animals.[26] A further resource, containing information on mouse gene imprinting,[27] also includes an imprinting catalog, as well as chromosome anomalies on mutant mouse lines. This represents integration of curated information from the MRC Harwell stock resource and other Harwell databases, with additional information from external data resources such as IMSR (International Mouse Stain Resource).

Histone and chromatin-related resources. The Histone Database,[28] of the National Human Genome Research Institute, provides a complete set of histone protein sequences. Nucleosomes, through various core histone post-translational modifications and incorporation of diverse histone variants, can serve as epigenetic markers to control processes such as gene expression and recombination. The Histone Sequence Database is a curated collection, assembled from major public databases, of sequences and structures of histones and non-histone proteins containing histone folds. A substantial increase in the number of sequences and taxonomic coverage for histone and histone fold-containing proteins is available. The database also provides comprehensive multiple sequence alignments for each of the four core histones (H2A, H2B, H3 and H4), the linker histones (H1/ H5) and the archaeal histones. Also included is current information on solved histone fold-containing structures. The database is thus an inclusive resource for the analysis of chromatin structure and function.

Chromatin.us is another web portal that includes information on chromatin proteins, histones and nucleosome structures and non-histone chromatin protein structures, and provides links to the protein data bank (PDB) site, which provides further details.[29] ReplicationDomain[30] is an online database for storing, sharing and visualizing DNA replication timing and transcription data, along with other numerical epigenetic data types. Data are typically obtained from DNA microarrays or DNA sequencing.

Gene silencing. An important epigenetic phenomenon, gene silencing, has also attracted attention and has been well reported in the literature. Collected papers are available on Bio-

Tech Info- Net.[31] Similarly, RNA induced epigenetics related papers on imprinting by non-coding RNAs are collated.[32]

Other epigenetic biomedical resources. The evolution of epigenetic resources is still in its early stages, with provision associated with several specific research efforts and groups. Nevertheless, in line with genetic/genomic data examples, efforts are being made to connect information, even as new targets are emerging. The Epigenetics Database[33] includes all known epigenetics genes/proteins discovered to date. The database is arranged in hierarchical format, based upon gene ontology. While still in its developmental (ß) phase, it is expected that future developments will include user-submitted meta-data, which will be freely available for use in database and flat file format. Some sites, *e.g.*, Epigenie,[34] also provide bioinformatics tools (*e.g.*, CpG Viewer, CpG and GC Plotter and tools for CpG Island detection). NCBI supported efforts include the Epigenetics Antibody Database,[35] providing antibody information for researchers working in the field of epigenetics/epigenomics, and Unigene,[36] containing same locus-of-origin transcription sequences, protein similarities, gene expression, cDNA clone reagents, genomic location and associated epigenetic information. NARNA,[37] supported by Newcastle University, incorporates relationships between epigenetic events, DNA methylation, gene imprinting and X-chromosome inactivation with natural antisense RNAs. Other, locally developed or supported, current resources include StatEpigen,[38] with an initial focus on colon cancer, although incorporating some information on other pathologies for comparison. Data are provided on simple and conditional molecular events, since many genetic and epigenetic alterations are expected to be mutually correlated and synergistic, and drive model input at the micro-layer.[39] Specialized resources also exist for plant data.[40]

Large-Scale Epigenetic Project Initiatives

European project initiatives including HEP. A number of European initiatives exist for centralized projects on DNA methylation. The Human Epigenome Project (HEP**[9]**) will provide an epigenetic resource of chromosomal DNA methylation reference profiles in human tissues and cell lines. Other initiatives include chromatin profiling (HEROIC, High-Throughput Epigenetic Regulatory Organization In Chromatin), treatment of neoplastic disease (EPITRON, Epigenetic Treatment Of Neoplastic Disease[41]) and the SMARTER[42] initiative, which aims to develop small inhibitors of chromatin-modifying enzymes. Another effort to provide structure to the epigenetic research landscape in Europe is that of the Epigenetic Network of Excellence, now known as Epigenesys, which aims to advance epigenetics toward Systems Biology.[43]

Roadmap epigenomics program. The Roadmap Epigenomics Program (also known as Epigenomics Roadmap initiative), launched by NIH (2008), seeks to create a series of epigenome maps to study epigenetic mechanisms, develop new epigenetic analytics, generate a repository and long-term data archive, standardize procedures and practices in epigenomics and support new technologies for these. As part of the $190 million, five-year initiative, the Roadmap Epigenomics Mapping Consortium[44] was formed to provide a public database for human epigenomics data, the Human Epigenome Atlas.[45] The current release, Epigenome Atlas Release 7, includes human reference epigenomes and the results of their integrative and comparative analyses.

The NIH Roadmap Epigenomics Program has also established IHEC (International Human Epigenome Consortium),[8] which aims to coordinate epigenome mapping and characterization worldwide, in order to ensure high data quality standards, coordination of data storage, management and analysis and free access to the epigenomes produced. To attain substantial coverage of the human epigenome, IHEC aims to decipher at least 1,000 epigenomes within the next 7–10 years. Officially launched in Paris (Jan 2010), with an initial (first phase) budget target of $130 million, IHEC intends to coordinate the mapping of epigenomes from not only the NIH's Epigenomics Mapping Consortium but also from international efforts such as the European Epigenome Network of Excellence, the Danish National Research Foundation Centre for Epigenetics, and the Australian Epigenetic Alliance. The IHEC web portal provides links to databases, such as GEO, ARRAYEXPRESS and DDBJ, where epigenetic sequencing data will be made available.

Another significant large-scale program in epigenetics is the Encyclopedia of DNA Elements (ENCODE).[46] This is supported by the ENCODE Consortium, an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). This initiative aims to identify all functional elements, both at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active, in the human genome sequence.

ICGC. Genomic changes that occur in various types of cancer are being investigated by the International Cancer Genome Consortium (ICGC).47 The goal is to obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes. Many samples from one tumor type or subtype will be analyzed in detail so that this initiative promises to provide crucial insights on genetic-epigenetic links.

Discussion and Conclusions

The biomedical resources relating, primarily, to epigenetic data that were surveyed here are numerous and range from small to large-scale, with considerable ongoing integration and new links still being forged. In common with many newly identified research targets, early-stage resources are often very specific and are supported locally, and this is still the case for much useful epigenetic data. Many such databases and their software tools are publicly accessible from academic/research institutions, while others are commercially available (**Table S1**). Major issues remain quality assurance, effective annotation and overall management, but appropriate analysis must also keep pace and is typically uneven (**Table S2**). Clearly, the generation of a centralized repository for epigenetics-related data is desirable and currently lacking, but new technologies offer increased potential for processing solutions down the line. Notably, biomedical needs are an important focus for federated database development, health-grid technology and, of course, Cloud computing.

Major initiatives to ensure quality and standards for genetic and epigenetic research do exist and some, such as IHEC and HEP, are described in this review. With improved technology, these should lead to improved data mining tools where those currently available for epigenetic/epigenomic analyses are limited and predominantly sequence-oriented, ranging from identification, through PCR and initial pattern matching (**Table S2** presents the current summary).

Acknowledgments

The authors would like to acknowledge funding from the Daniel O'Hare Scholarship program DCU, which made it possible to carry out this study.

Supplemental Materials

Supplemental materials may be found here: www.landesbioscience.com/journals/epigenetics/article/21493

References

1. Choudhary AN, Honbo D, Kumar P, Ozisikyilmaz B, Misra S, Memik G. Accelerating Data Mining Workloads: Current Approaches and Future Challenges in System Architecture Design. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2011; 1:41-54; http://dx.doi.org/10.1002/widm.9.
2. Marakas GM. Modern Data Warehousing, mining and Visualization Core Concepts. Pearson Education, 2003.
3. Sheth AP, Larson JA. Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases. ACM Comput Surv 1990; 22;183-236. http://dx.doi.org/10.1145/96602.96604.
4. Heimbigner D, McLeod D. A Federated Architecture for Information Management. ACM Transactions on Office Information Systems 1985; 3:253-78.
5. Entrez gene (Internet). 24 February2012. http://www.ncbi.nlm.nih.gov/Entrez/
6. Tenenbaum JD, Whetzel PL, Anderson K, Borromeo CD, Dinov ID, Gabriel D, *et al.* The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research. J Biomed Inform 2011; 44:137-45; PMID:20955817; http://dx.doi.org/10.1016/j.jbi.2010.10.003.
7. Weizmann Institute of Science (Internet). Genecards. 24 February 2012. http://www.genecards.org/
8. International Human Epigenome Consortium (Internet). IHEC. 24 February 2012: http://www.ihecepigenomes.org/
9. Human Epigenome Project (Internet). HEP. 24 February 2012. http://www.epigenome.org/
10. Novik KL, Nimmrich I, Genc B, Maier S, Piepenbrock C, Olek A, *et al.* Epigenomics: genome-wide study of methylation phenomena. Curr Issues Mol Biol 2002; 4:111-28; PMID:12432963.
11. Rodenhiser D, Mann M. Epigenetics and human disease: translating basic biology into clinical applications. CMAJ 2006;174:341-8; http://dx.doi.org/10.1503/cmaj.050774; PMID:16446478.
12. The Cancer Genome Atlas (Internet). TCGA. 21 May 2012. http://cancergenome.nih.gov/
13. Egger G, Liang G, Aparicio A, Jones PA. Epigenetics in human disease and prospects for epigenetic therapy. Nature 2004; 429:457-63; PMID:15164071; http://dx.doi.org/10.1038/nature02625.
14. Feinberg AP, Tycko B. The history of cancer epigenetics. Nat Rev Cancer 2004; 4:143-53; PMID:14732866; http://dx.doi.org/10.1038/nrc1279.
15. Fuks F. DNA methylation and histone modifications: teaming up to silence genes. Curr Opin Genet Dev 2005; 15:490-5; PMID:16098738; http://dx.doi.org/10.1016/j.gde.2005.08.002.
16. Pubmeth, Reviewed methylation database in cancer (Internet). 24 February 2012. http://matrix.ugent.be/ pubmeth/
17. Grunau C, Renault E, Rosenthal A, Roizes G. MethDB- -a public database for DNA methylation data. Nucleic Acids Res 2001; 29:270-4; PMID:11125109; http://dx.doi.org/10.1093/nar/29.1.270 . 986 E pigenetics V olume 7 Issue 9
18. Negre V, Grunau C. The MethDB DAS server: adding an epigenetic information layer to the human genome. Epigenetics 2006; 1:101-5;

PMID:17965614; http:// dx.doi.org/10.4161/epi.1.2.2765.

19. MethPrimerDB, the DNA methylation analysis PCR primer database (Internet). 24 February 2012. http:// medgen.ugent.be/methprimerdb/

20. He XM, Chang SH, Zhang JJ, Zhao Q, Xiang H, Kusonmano K, *et al.* MethyCancer: the database of human DNA methylation and cancer. Nucleic Acids Res 2008; 36(Database issue):D836-41; http://dx.doi. org/10.1093/nar/gkm730; PMID:17890243.

21. Epigenetics Group (Internet). Dr. Axel Schumacher. 24 February 2012. http://www.methylogix.com/gen etics/ database.shtml.htm

22. Methtools (Internet). 27 February 2012. http:// genome.imb-jena.de/methtools/

23. Reik W, Walter J. Genomic imprinting: parental influence on the genome. Nat Rev Genet 2001; 2:21-32; PMID:11253064; http://dx.doi. org/10.1038/35047554.

24. Gene imprint, Imprinted Gene Database (Internet). 24 February 2012. http://www.geneimprint.com/sit e/ genes-by-species

25. University of Otago (Internet). New Zealand, Catalogue of Parent of Origin Effects, Parental Origins of de novo Mutations. 24 February 2012. http://igc. otago.ac.nz/home.html

26. Glaser RL, Ramsay JP, Morison IM. The imprinted gene and parent-of-origin effect database now includes parental origin of de novo mutations. Nucleic Acids Res 2006; 34(Database issue):D29-31; http://dx.doi. org/10.1093/nar/gkj101; PMID:16381868.

27. Mousebook (Internet). 24 February 2012. http://www. mousebook.org/about.php

28. National Human Genome Research Institute (Internet). Histone Database. 24 February 2012. http://genome. nhgri.nih.gov/histones/complete. shtml.

29. Chromatin, Chromatin Structure and Function Page (Internet). 24 February 2012. http://www.chromatin. us/chrom.html

30. Florida State University (Internet). Replication Domain. 24 February 2012. http://www.replicationdomain. com/

31. Gene Silencing (Internet). Technical papers. 24 February 2012. http://www.biotech-info.net/gene_ silencing.html#silencing

32. RNA Induced Epigenetics (Internet). 24 February 2012. http://www.euchromatin.com/R NAepi.htm

33. The Epigenetics Database (Internet). Beta. 24 February 2012. http://www.epidna.com/database .php

34. The EpiGenie (Internet). EpiGenie Epigenetics Product Database. 24 February 2012. http://www. epigenie.com/Epigenetics-Research-Products.html

35. National Center for Biotechnology Information (Internet). NCBI. 24 February 2012. http://www.ncbi. nlm.nih.gov/guide/all/#database s_

36. Unigene, NCBI (Internet). 24 February 2012. http:// www.ncbi.nlm.nih.gov/sites/entr ez?db=unigene

37. Platform for research on natural antisense transcripts, NARNA (Internet). 24 February 2012. http://www. narna.ncl.ac.uk

38. Statepigen, Database of colon cancer epigenetic statistics (Internet). 24 February 2012. Available from: http://statepigen.sci-sym.dcu.ie

39. Raghavan K, Ruskin HJ, Perrin D, Goasmat F, Burns J. Computational micromodel for epigenetic mechanisms. PLoS One 2010; 5:e14031; http://dx.doi.org/10.1371/ journal.pone.0014031; PMID:21152421.

40. Epigenomics of Plants International Consortium (Internet). EPIC. 24 February 2012. https://www. plant-epigenome.org/groups

41. EPIgenetic Treatment of Neoplastic disease, EPITRON (Internet). 24 June 2012. http://www.epitron.eu/.

42. SMARTER Network (Internet). 24 June 2012. http:// www.smarter-chromatin.eu/

43. EpiGenesys (Internet). 27 February 2012. http://www. epigenesys.eu/

44. NIH Roadmap Epigenomics Mapping Consortium (Internet). 24 February 2012. http://www.roadmapepigenomic s.org

45. Human Epigenome Atlas (Internet). 24 June 2012. http://www.genboree.org/epigen omeatlas/index.rhtml

46. Encyclopedia of DNA Elements (Internet). 24 February 2012. http://genome.ucsc.edu/ENCOD E/

47. ICGC Cancer Genome Projects (Internet). 24 June 2012. http://www.icgc.org/

Table S1: Table of available Epigenetic Resources

Other Epigenetic Resources:	**Details:**	**URLs:**
NCBI Epigenomic sample browser	Current most comprehensive collection.	http://www.ncbi.nlm.nih.gov/epigenomics/browse
ChromDB	Incorporates 3 types of sequences, genome-based, transcript based and NCBI refseq.	http://www.chromdb.org/
Zhao lab (NIH) data	Genome-wide mapping of histone H3 modifications in human CD4 and T cells, High-Resolution profiling of histone methylations in the human genome and combinatorial patterns of histone acetylation and methylation in the human genome.	http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/ ; http://dir.nhlbi.nih.gov/Papers/lmi/epigenomes/hgtcell.aspx ; http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellacetylation.aspx
Young Lab Data (MIT)	Genome-wide map of nucleosome acetylation and methylation in yeast.	http://web.wi.mit.edu/young/nucleosome/
REBASE (Restriction Enzyme Database)	Collection of information about restriction enzymes, methylases, the microorganisms from which they have been isolated, recognition sequences, cleavage sites, methylation specificity, the commercial availability of the enzymes, and references - both published and unpublished observations (dating back to 1952).	http://rebase.neb.com/rebase/rebase.html
Rett Syndrome DB	An interactive version of the mutation frequency and symptoms databases originally established by Dr. Brian Hendrich and Skirmantas Kriaucionis at the University of Edinburgh. Aims to collate and display mutation and symptom information from Rett Syndrome patients to allow the analysis of how symptoms correlate with MeCP2 mutation status.	http://www.mecp2.org.uk/
Web IRSF MECP2 Variation DB (RettBASE)	This is constructed by merging mutation and polymorphism data from the published literature pertaining to Rett syndrome and related clinical disorders, and by incorporating unpublished mutation and polymorphism data that have been submitted.	http://mecp2.chw.edu.au/
CREMOFAC	A web-database of chromatin remodelling factors, currently with 64 types of remodelling factors from 49 organisms reported in literature.	http://www.jncasr.ac.in/cremofac/
Antibody Validation DB	Aims to collect and share experimental results on antibodies – aiding researchers in selection and validation of antibodies. Site began with 200+ histone antibodies tested as part of the ENCODE and Roadmap Epigenomics projects. Non histone antibodies from these projects will also be added.	http://compbio.med.harvard.edu/antibodies/

Table S2: Table of available Tools for Epigenetic Research

Tools :	Details:	URLs:
CpG Island Searcher	Screens for CpG Islands that meet the criteria listed in the website.	http://www.uscnorris.com/cpgislands2/cpg.aspx
Methylator	Predicts if CpG in a DNA sequence is likely to be methylated or not.	http://bio.dfci.harvard.edu/Methylator/
MethPrimer	Online tool for designing bisulfite-conversion-based Methylation PCR Primers. Can design primers for Methylation-Specific PCR (MSP) and Bisulfite-Sequencing PCR or Bisulfite-Restriction PCR. The input sequence is DNA sequence in any format. The program returns results in both text and graphic view.	http://www.urogene.org/methprimer/
BiSearch	Provides access to a PCR primer-test and primer-design algorithm that can be used for both bisulfite converted and not modified sequences. Search can be carried out on various genomes with the designed primers by a fast ePCR method	http://bisearch.enzim.hu/
methBLAST	Allows for checking primers for bisulfite converted DNA by blasting them against the unmethylated and methylated genomic sequences of human, mouse and rat.	http://medgen.ugent.be/methBLAST/
Snake-charmer	Tool to select restriction enzymes for COBRA	http://insilico.ehu.es/restriction/two_seq/snake_charmer.html
MethTools (ver. 1.x and 2.x)	Software tools for the analysis of bisulfite treated DNA. The software does the comparison between unconverted mother sequences and deaminated sequences, generates graphical outputs of methylation patterns and methylation density, estimates the systematic error of the experiment and searches for conserved methylated nucleotide-patterns. The software can be used to generate files suitable for the submission to MethDB.	http://genome.imb-jena.de/methtools/ ; http://194.167.139.26/methtools/MethTools2_submit.html
BiQ Analyzer	Software tool for easy visualization and quality control of DNA methylation data from bisulfite sequencing (Java tool, capable of generating archive files that can be directly uploaded to MethDB)	http://biq-analyzer.bioinf.mpi-inf.mpg.de/
CpGviewer	Simple integrated tool for handling bisulphite sequencing projects. It can process plain text sequences or a variety of electropherogram formats, and allows interactive editing of the sequences, aligned to a reference sequence.	http://dna.leeds.ac.uk/cpgviewer/
CyMATE	CyMATE (Cytosine Methylation Analysis Tool for Everyone) - a software platform to perform in silico analyses of DNA methylation at cytosine sites. Suitable for analyses of sequence data obtained with bisulfite genomic sequencing and hairpin-bisulfite sequencing, *i.e.* single-strand and double-strand DNA data.	http://www.cymate.org/
BISMA	BISMA (Bisulfite Sequencing DNA Methylation Analysis) analyses the bisulfite sequencing data that are derived from sequencing of subcloned molecules of a PCR product.	http://biochem.jacobs-university.de/BDPC/BISMA/
EpiGRAPH	Software for advanced (epi)-genome analysis and prediction.	http://epigraph.mpi-inf.mpg.de/WebGRAPH/

Appendix I

Table A: Table of available Epigenetic Resources (Table reproduced from Table S1, Shakya *et al*. 2012)

Other Epigenetic Resources:	**Details:**	**URLs:**
NCBI Epigenomic sample browser	Current most comprehensive collection.	http://www.ncbi.nlm.nih.gov/epigenomics/browse
ChromDB	Incorporates 3 types of sequences, genome-based, transcript based and NCBI refseq.	http://www.chromdb.org/
Zhao lab (NIH) data	Genome-wide mapping of histone H3 modifications in human CD4 and T cells, High-Resolution profiling of histone methylations in the human genome and combinatorial patterns of histone acetylation and methylation in the human genome.	http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/ ; http://dir.nhlbi.nih.gov/Papers/lmi/epigenomes/hgtcell.aspx ; http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellacetylation.aspx
Young Lab Data (MIT)	Genome-wide map of nucleosome acetylation and methylation in yeast.	http://web.wi.mit.edu/young/nucleosome/
REBASE (Restriction Enzyme Database)	Collection of information about restriction enzymes, methylases, the microorganisms from which they have been isolated, recognition sequences, cleavage sites, methylation specificity, the commercial availability of the enzymes, and references - both published and unpublished observations (dating back to 1952).	http://rebase.neb.com/rebase/rebase.html
Rett Syndrome DB	An interactive version of the mutation frequency and symptoms databases originally established by Dr. Brian Hendrich and Skirmantas Kriaucionis at the University of Edinburgh. Aims to collate and display mutation and symptom information from Rett Syndrome patients to allow the analysis of how symptoms correlate with MeCP2 mutation status.	http://www.mecp2.org.uk/
Web IRSF MECP2 Variation DB (RettBASE)	This is constructed by merging mutation and polymorphism data from the published literature pertaining to Rett syndrome and related clinical disorders, and by incorporating unpublished mutation and polymorphism data that have been submitted.	http://mecp2.chw.edu.au/
CREMOFAC	A web-database of chromatin remodelling factors, currently with 64 types of remodelling factors from 49 organisms reported in literature.	http://www.jncasr.ac.in/cremofac/
Antibody Validation DB	Aims to collect and share experimental results on antibodies – aiding researchers in selection and validation of antibodies. Site began with 200+ histone antibodies tested as part of the ENCODE and Roadmap Epigenomics projects. Non histone antibodies from these projects will also be added.	http://compbio.med.harvard.edu/antibodies/

Table B: Table of available Tools for Epigenetic Research (Table reproduced from Table S2, Shakya *et al.* 2012)

Tools :	Details:	URLs:
CpG Island Searcher	Screens for CpG Islands that meet the criteria listed in the website.	http://www.uscnorris.com/cpgislands2/cpg.aspx
Methylator	Predicts if CpG in a DNA sequence is likely to be methylated or not.	http://bio.dfci.harvard.edu/Methylator/
MethPrimer	Online tool for designing bisulfite-conversion-based Methylation PCR Primers. Can design primers for Methylation-Specific PCR (MSP) and Bisulfite-Sequencing PCR or Bisulfite-Restriction PCR. The input sequence is DNA sequence in any format. The program returns results in both text and graphic view.	http://www.urogene.org/methprimer/
BiSearch	Provides access to a PCR primer-test and primer-design algorithm that can be used for both bisulfite converted and not modified sequences. Search can be carried out on various genomes with the designed primers by a fast ePCR method	http://bisearch.enzim.hu/
methBLAST	Allows for checking primers for bisulfite converted DNA by blasting them against the unmethylated and methylated genomic sequences of human, mouse and rat.	http://medgen.ugent.be/methBLAST/
Snake-charmer	Tool to select restriction enzymes for COBRA	http://insilico.ehu.es/restriction/two_seq/snake_charmer.html
MethTools (ver. 1.x and 2.x)	Software tools for the analysis of bisulfite treated DNA. The software does the comparison between unconverted mother sequences and deaminated sequences, generates graphical outputs of methylation patterns and methylation density, estimates the systematic error of the experiment and searches for conserved methylated nucleotide-patterns. The software can be used to generate files suitable for the submission to MethDB.	http://genome.imb-jena.de/methtools/ ; http://194.167.139.26/methtools/MethTools2_submit.html
BiQ Analyzer	Software tool for easy visualization and quality control of DNA methylation data from bisulfite sequencing (Java tool, capable of generating archive files that can be directly uploaded to MethDB)	http://biq-analyzer.bioinf.mpi-inf.mpg.de/
CpGviewer	Simple integrated tool for handling bisulphite sequencing projects. It can process plain text sequences or a variety of electropherogram formats, and allows interactive editing of the sequences, aligned to a reference sequence.	http://dna.leeds.ac.uk/cpgviewer/
CyMATE	CyMATE (Cytosine Methylation Analysis Tool for Everyone) - a software platform to perform in silico analyses of DNA methylation at cytosine sites. Suitable for analyses of sequence data obtained with bisulfite genomic sequencing and hairpin-bisulfite sequencing, *i.e.* single-strand and double-strand DNA data.	http://www.cymate.org/
BISMA	BISMA (Bisulfite Sequencing DNA Methylation Analysis) analyses the bisulfite sequencing data that are derived from sequencing of subcloned molecules of a PCR product.	http://biochem.jacobs-university.de/BDPC/BISMA/
EpiGRAPH	Software for advanced (epi)-genome analysis and prediction.	http://epigraph.mpi-inf.mpg.de/WebGRAPH/

Appendix II

Table A: Details of the data used in the analysis, the 21 species and their genome coverage.

Species	Assembly[1]	APC	ATM	BHD	BMPR1A	CDH1	MADH4	MET	MLH1	MSH2	MSH6	MUTYH	NF1	PMS1	PMS2	PTEN	SDHB	SDHC	*STK11*	TP53	TSC1	TSC2	VHL
Human	GRCh37p2																						
Chicken	WASHUC2					■	■											■		■			
Chimpanzee	CHIMP2.1																						
Cow	Btau_4.0																						
Dog	CanFam_2.0																						
Elephant	loxAfr																		■	■			
Frog	JGI4.1					■							■										
Fugu	FUGU4.0	■				■	■										■						
Gorilla	gorGor3																						
Guinea Pig	cavPor3															■							
Horse	EquCab2																						
Marmoset	culJac3								■										■				
Mouse	NCBIM37																						
Opossum	monDom5						■			■												■	
Orangutan	PPYG2																	■					
Pig	Sscrofa9		■	■					■	■	■						■	■			■		
Platypus	OANA5		■							■	■		■			■		■		■		■	■
Rabbit	oryCun2.0																■		■				■
Rat	RGSC3.4													■						■			
Zebra Finch	teaGut3.2.4				■	■	■						■					■		■	■		
Zebrafish	Zv9		■		■	■	■						■			■					■		■

Table B: Table with Models used in phylogenetic study.

Model Name	**Characteristics**	**Positive Selection (ω>1)?**
M0	One ω allowed across all sites	Allowed
Site – specific Models		
M1: Neutral	Two classes, ω_1 fixed at 0 (ω_1 <1 after version 3.14 PAML/Codeml) and ω_2 fixed at 1. Returns the proportion p_2 of sites in the second category. 1- p_2 returns $p_{1.}$	Not allowed
M2: Selection	M1 plus an additional class where ω is estimated from the data and can be larger than 1.	Allowed
M3 : Discrete K =2	Two classes of ω allowed without constraint on either value, these values are estimated as their relative proportions and ω can be larger than 1.	Allowed
M3 : Discrete K =3	As M3 (K=2) but with 3 unconstrained classes of ω.	Allowed
M7: Beta	ω is assumed to have a beta distribution between values of 0 and 1 inclusive. 10 classes of ω allowed.	Not Allowed
M8: Beta and ω >1	As model 7 but further ω category is estimated from the data and can be larger than 1.	Allowed
M8a: Beta and ω =1	As model 8 but with ω fixed to 1; is the null model of M8.	Not Allowed
Lineage – specific Models		
Model A	Lineage – specific extension of M1. Four ω classes allowed, two of which can vary between the foreground and background lineages, one which is unconstrained apart from equality for the foreground and background lineages, and one which is set ω =1.	Allowed
Model A Null	As Model A but with both the ω classes that are allowed to differ between the foreground and background lineages set to 1.	Not Allowed

Table C: Likelihood ratio tests (LRTs) performed using all evolutionary models used in selection analysis.

Comparison	***df***	***Δl***	**Critical χ^2 values**
M0 v M3k2	2	X2	≥ 5.99
M3k2 v M3k3	-	X1	≥ 1.00
M1 v M2	2	X2	≥ 5.99
M7 v M8	2	X2	≥ 5.99
M8 v M8a*	1	X2	≥ 2.71 (@5%) ≥ 5.41 (@1%)
M1 v Model A	2	X2	≥ 5.99
Model A v Model A1(denoted as model a null throughout the manuscript)	1	X2	≥ 3.84 (@5%)

* One degree of freedom for the chi square and comparison using a 50:50 mixture of point mass 0 and χ^2, so the critical χ^2 values are 2.71 at **5%** and 5.41 at **1%**, and not 10% and 2% respectively.

Table D: Full set of recombination test results on a per gene and per species basis

Seq Name	Sim *p*- value	BC KA Value	Alignment start (nuc)	Alignment End (nuc)	Len	Positively Selected Sites Removed (AA positions)
CDH1						
NA	NA	NA	NA	NA	NA	NA
PMS1						
NA	NA	NA	NA	NA	NA	NA
MLH1						
Chicken;Fugu	0.0037	0.00967	1713	1739	27	None
Horse;Platypus	0.0051	0.01317	952	985	34	None
Dog;Platypus	0.0093	0.02324	952	985	34	None
Orangutan;Platypus	0.015	0.03912	958	985	28	None
Human;Platypus	0.0189	0.04617	958	985	28	None
Gorilla;Platypus	0.0198	0.04772	958	985	28	None
Chimpanzee;Platypus	0.0209	0.04931	958	985	28	None
MSH2						
Zebra_Finch;Frog	0.034	0.10025	1435	1463	29	None
TSC2						
Orangutan;Zebrafish	0.0392	0.08831	1579	1604	26	None
MET						
Rabbit;Opossum	0.001	0.00217	2530	2606	77	None
Human;Opossum	0.0054	0.01115	2530	2606	77	None
Orangutan;Opossum	0.0054	0.01115	2530	2606	77	None
Chimpanzee;Opossum	0.0054	0.01115	2530	2606	77	None
Marmoset;Dog	0.0485	0.11156	2875	3032	158	None
BMPR1A						
Cow;Dog	0	0.00015	205	362	158	None
Rabbit;Elephant	0.0027	0.00958	226	284	59	None
Orangutan;Rabbit	0.0141	0.0455	217	285	69	None
Chimpanzee;Rabbit	0.0141	0.0455	217	285	69	None
Human;Rabbit	0.016	0.0494	217	285	69	None
Gorilla;Rabbit	0.0201	0.05818	217	285	69	None
Marmoset;Dog	0.0317	0.09429	241	332	92	None
Marmoset;Rabbit	0.0493	0.14138	226	285	60	None
MSH6						
Gorilla;Chimpanzee	0	0.00006	1066	1529	464	None
Human;Gorilla	0	0.00007	1066	1529	464	None
Zebrafish;Chicken	0.0037	0.00783	532	550	19	None
Gorilla;Orangutan	0.0073	0.01704	2632	2885	254	None
SDHB						
NA	NA	NA	NA	NA	NA	NA
STK11						
Cow;Rat	0.0009	0.00338	379	446	68	None
Cow;Horse	0.0028	0.00893	890	951	62	None
Cow;Pig	0.0128	0.0475	895	951	57	None
PMS2						
NA	NA	NA	NA	NA	NA	NA
MUTYH						
NA	NA	NA	NA	NA	NA	NA
VHL						
NA	NA	NA	NA	NA	NA	NA

APC						
Elephant;Mouse	0.0003	0.00041	1045	1280	236	None
Dog;Rat	0.0025	0.00565	994	1157	164	None
Horse;Mouse	0.0028	0.00724	1105	1307	203	None
Cow;Rat	0.0032	0.0089	994	1157	164	None
Horse;Rat	0.0052	0.01403	985	1157	173	None
Orangutan;Mouse	0.0085	0.02012	1045	1241	197	None
Cow;Mouse	0.009	0.02047	1105	1280	176	None
Human;Mouse	0.0105	0.02385	1045	1241	197	None
Pig;Rat	0.0108	0.02465	994	1157	164	None
Chimpanzee;Mouse	0.0112	0.02552	1045	1241	197	None
Cow;Pig	0.0118	0.02559	916	1289	374	None
Gorilla;Mouse	0.0123	0.0273	1045	1241	197	None
Mouse;Pig	0.0235	0.05456	1105	1280	176	None
TP53						
Cow;Guinea_Pig	0.0127	0.07289	927	970	44	317
MADH4						
Rat;Guinea_Pig	0.0157	0.07756	682	752	71	None
SDHC						
Cow;Guinea_Pig	0.0127	0.07289	927	970	44	None
ATM						
Gorilla;Chicken	0.0026	0.00616	3541	3587	47	None
Gorilla;Chimpanzee	0.0035	0.00819	3049	3793	745	None
Chimpanzee;Chicken	0.0047	0.01096	3541	3587	47	None
Human;Chicken	0.005	0.0115	3541	3587	47	None
Orangutan;Chicken	0.0052	0.01222	3541	3587	47	None
Human;Gorilla	0.0093	0.01851	4927	5641	715	None
Rabbit;Chicken	0.0173	0.03145	3547	3588	42	None
BHD						
Cow;Zebrafish	0.001	0.00282	236	254	19	None
Marmoset;Zebrafish	0.0011	0.00313	236	254	19	None
Mouse;Zebrafish	0.0011	0.00313	236	254	19	None
Opossum;Zebrafish	0.0013	0.00365	238	257	20	None
Cow;Mouse	0.0048	0.01657	208	266	59	None
Elephant;Zebrafish	0.0136	0.04138	238	251	14	None
TSC1						
Orangutan;Horse	0.0242	0.06153	1037	1136	100	None
Horse;Marmoset	0.0474	0.12863	1030	1136	107	None
PTEN						
NA	NA	NA	NA	NA	NA	NA
NF1						
Cow;Mouse	0.0218	0.05108	7390	7445	56	None
Gorilla;Opossum	0.0448	0.10141	7102	7148	47	None
Opossum;Orangutan	0.0463	0.10479	7102	7148	47	None
Chicken;Mouse	0.0464	0.10541	7406	7439	34	None
Human;Opossum	0.0472	0.10828	7102	7148	47	None
Chimpanzee;Opossum	0.048	0.11007	7102	7148	47	None

Table D: Full set of recombination test results on a per gene and per species basis. The value highlighted in yellow for TP53 represents a region where recombination was detected with reasonable confidence that also coincided with a positively selected residue (*i.e.* false positive).