UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1862

Predictive Healthcare

Cervical Cancer Screening Risk Stratification and Genetic Disease Markers

NICHOLAS BALTZER

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019

ISSN 1651-6214
ISBN 978-91-513-0768-8
urn:nbn:se:uu:diva-394293

Dissertation presented at Uppsala University to be publicly examined in Room A1:111, BMC, Husargatan 3, Uppsala, Thursday, 28 November 2019 at 09:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Mark Jit (University of Hong Kong).

Abstract

Baltzer, N. 2019. Predictive Healthcare. Cervical Cancer Screening Risk Stratification and Genetic Disease Markers. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1862. 62 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-0768-8.

The use of Machine Learning is rapidly expanding into previously uncharted waters. In the medicine fields there are vast troves of data available from hospitals, biobanks and registries that now are being explored due to the tremendous advancement in computer science and its related hardware. The progress in genomic extraction and analysis has made it possible for any individual to know their own genetic code. Genetic testing has become affordable and can be used as a tool in treatment, discovery, and prognosis of individuals in a wide variety of healthcare settings. This thesis addresses three different approaches towards predictive healthcare and disease exploration; first, the exploitation of diagnostic data in Nordic screening programmes for the purpose of identifying individuals at high risk of developing cervical cancer so that their screening schedules can be intensified in search of new disease developments. Second, the search for genomic markers that can be used either as additions to diagnostic data for risk predictions or as candidates for further functional analysis. Third, the development of a Machine Learning pipeline called ||-ROSETTA that can effectively process large datasets in the search for common patterns. Together, this provides a functional approach to predictive healthcare that allows intervention at early stages of disease development resulting in treatments with reduced health consequences at a lower financial burden.

Keywords: Bioinformatics, Cervical Cancer, Screening, Computer Science, Algorithmics, Machine Learning, Genetics, SNPs, Rough Sets

Nicholas Baltzer, Department of Cell and Molecular Biology, Computational Biology and Bioinformatics, Box 596, Uppsala University, SE-751 24 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-0768-8
urn:nbn:se:uu:diva-394293 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-394293)

To my brother Harald, the reason I
do not feel alone.

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Baltzer N., Sundström K., Nygård J. F., Dillner J., Komorowski J. (2017) Risk stratification in cervical cancer screening by complete screening history: Applying bioinformatics to a general screening population. *Int J Cancer 2017;141:200–9.*

II Cavalli M., Baltzer N., Umer H. M., Grau J., Lemnian I., Pan G., Wallerman O., Spalinskas R., Sahlén P., Grosse I., Komorowski J., Wadelius C. (2019) Allele specific chromatin signals, 3D interactions, and motif predictions for immune and B cell related diseases. *Sci Rep 2019;9:2695.*

III Baltzer N., Nygård J. F., Sundström K., Nygård M., Dillner J., Komorowski J., (2019) Risk Stratification in Cervical Cancer Screening – Validation and Generalization of a Data-driven Functional Screening Recall Model. *Manuscript*

IV Cavalli M., Baltzer N., Pan G., Bárcenas Walls J. R., Smolinska Garbulowska K., Kumar C., Skrtic S., Komorowski J., Wadelius C. (2019) Studies of liver tissue identify functional gene regulatory elements associated to gene expression, type 2 diabetes, and other metabolic diseases. *Hum Genomics* 2019;13:20.

V Baltzer N., Komorowski J. (2019) ||-ROSETTA. *Accepted manuscript, Lecture Notes on Computer Science, Transactions on Rought Sets, Springer*

Reprints were made with permission from the respective publishers.

Additional Papers

These additional papers are not included in the thesis.

I Dąbrowski, M. J., S. Bornelöv, M. Kruczyk, N. Baltzer, J. Komorowski. 'True' Null Allele Detection in Microsatellite Loci: A Comparison of Methods, Assessment of Difficulties and Survey of Possible Improvements. *Molecular Ecology Resources 15, no. 3 (May 1, 2015): 477–88.*

II Kruczyk, M., Baltzer N., Mieczkowski J., Draminski M., Koronacki J., Komorowski J. Random Reducts: A Monte Carlo Rough Set-Based Method for Feature Selection in Large Datasets. *Fundam. Inf. 127, no. 1–4 (2013): 273–88.*

Reprints were made with permission from the respective publishers

Contents

Abbreviations

PSA – Prostate specific antigen

DNA – Deoxyribonucleic acid

HPV – Human papillomavirus

RNA – Ribonucleic acid

TF – Transcription factor

SNP – Single nucleotide polymorphism

CTCF – CCCTC binding factor

3C – Chromatin conformation capture

4C – Circularized chromatin conformation capture

5C - Chromosome conformation capture carbon copy

ChIA-PET - Chromatin interaction analysis by paired end tag sequencing

GWAS – Genome wide association study

eQTL – Expression Quantitative Trait Loci

A – Adenine

C – Cytosine

G – Guanine

T – Thymine

ML – Machine Learning

Introduction

The progress of computing technology over the last 15 years has opened new areas of research in almost every field of science. In biology, chemistry, medicine and in physics, researchers have turned to the use of computing for testing hypotheses *in silico* and for making wide hypothesis-free searches for leads and answers[1].

The future of science lies to a large degree in quantitative studies, where complex systems can be explored for functional mechanisms much like an 1862 gold digger in Boise Basin would sift through his pan looking for a nugget. This thesis addresses the search for nuggets in vast amounts of data to explain which factors control an outcome, for example, the search for variants in our DNA to explain why some people have certain diseases, and the search for shared behavior in cancer screening to explain why some patients are detected early enough for effective treatment whereas some patients are detected too late.

The advancement of computing has made possible Machine Learning (ML), predictions based on statistical inference from previous observations much like a human would predict traffic congestion based on his or her previous experience. While ML has been around for quite some time in theory[2], it has not been practically applicable on a wide scale until the last twenty years or so. The sheer amount of data and computing power needed for ML to produce statistically sound results made it unfeasible. Even today we are greatly limited in terms of computing, especially in the search for combinatorial effects. As an example, the scoring schema in Paper III required 142 days. During this time, the algorithm computed over 2,508,271,955,205 possible solutions requiring 571,057,686,141,794 propagations of the data.

The availability of ML in scientific computing is a recent development, and it remains to be deployed in many areas. In biological science the importance of statistically sound analyses was quickly recognized as it is a field of research where data with a high noise level is frequently encountered. In medicine it has taken longer, as befits a field where lives are at stake in validation tests, but the use of ML has increased in recent years.

Computing has also changed the field of genetics, making possible the rapid analysis of entire genomes, corresponding to enormous amounts of quaternary data. The capacity for analysis has opened up genomics far beyond simply looking at nucleotides, allowing for exploring the genome and its related products in a systems context and observing how components interact with the genome.

The central theme of this thesis is computing in medicine and genetics. It addresses the use of ML in the context of cervical cancer screening in the search for the differences in the history of women that develop cancer and those who do not and a program has been developed to make those ML computations practical. Also, the development of a pipeline for filtering genomic variations down to only those highly relevant for the development of disease is described.

The Basics

Nucleotides

A nucleotide is an organic molecule that serves as the building block (monomer) for deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) polymers. The structure of the nucleotide is the basis for all genetic information, from that controlling the simplest virus to that of humans, and sequences of nucleotides are used to encode all the data we need to function. It consists of three parts; a nitrogenous base, a five-carbon sugar, and a phosphate group (Figure 1). Nucleotides in the genome form long chains (polymers) by forming a bond between the phosphate group of one nucleotide and the sugar molecule of another. These chains will form helices, either single (RNA) or double (DNA).

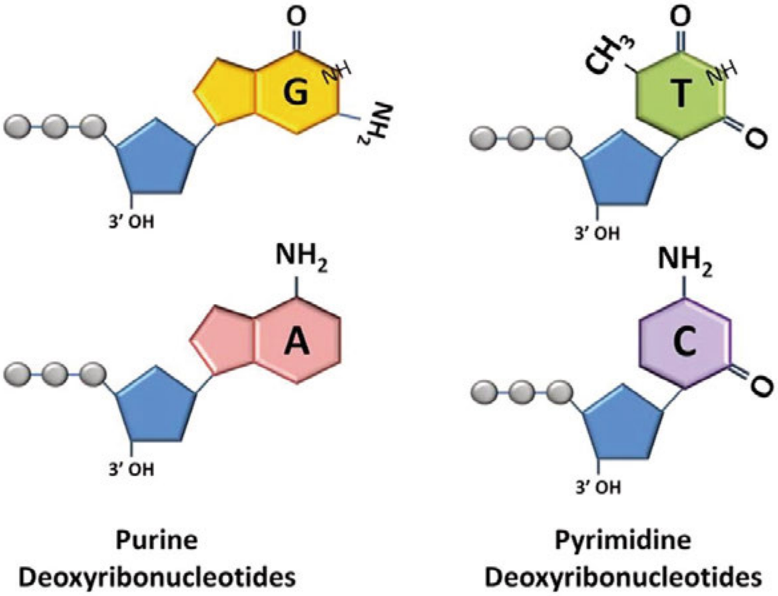


Figure 1 The four nucleotides of DNA. The pentose sugar (blue) binds to the phosphate group (grey) of another nucleotide to form polymers capable of dimerizing and forming helices. The nitrogenous bases are labelled with the character representing them in DNA (G = guanine, T = thymine, A = adenine, C = cytosine). Image courtesy of Scientific Reports[3]

DNA

All information required to develop a human being is stored in deoxyribonucleic acids (DNA). DNA is stored in the cell nucleus, in long X-shaped stretches called chromosomes. To fit in the nucleus, the DNA is wrapped tightly around circular proteins called histones. There are 23 chromosome pairs in total, each parent contributing one chromosome to the pair. Of these, 22 pairs are the same for men and women. The last pair contains the X and Y chromosomes, and these differ between male and female as women have two X chromosomes, while men have one X and one Y.

DNA has many different regions used for different purposes. A gene will contain a starting point and a stopping point so that the transcription to form proteins knows where to begin and where to end, as well as "data" regions called exons and "flow control" regions called introns. When put together, the exons contain the "code" of a gene, and introns offer control flow so that different versions of the "code" can be created. Before the starting point of a gene there is also a promoter region that is needed to activate the transcription of the gene, and somewhere there might also be an enhancer region, which greatly increases the transcription rate of the gene.

DNA is regulated by multiple systems. The availability of DNA is regulated by modifications of the histone tails, enabling the DNA to uncoil from the histones so it can be accessed by proteins. The activity of DNA is regulated by methylation, as methyl groups attached to the cytosine nucleotide prevent it from binding peptides. The gene transcription is regulated by transcription factors (TFs), peptides that bind to the promoter and enhancer of a gene to attract RNA polymerase, the enzyme that creates RNA strands. The proximity of the enhancer to the promoter is regulated by architectural TFs such as CTCF that create loops in the DNA to bring distant regions together.

DNA is a complex blueprint for proteins, the building blocks of the cell. To create a protein, the corresponding DNA sequence is unwound from the histones and replicated into single-stranded RNA. The RNA is then processed further into a final blueprint, which is read by the ribosome protein complex and the sequence of the amino acids, the protein, is based on this RNA.

The DNA of any two humans is estimated to differ by 0.6%[4], or 20 million nucleotide pairs. This variation comes from mutations. These mutations can be inherited from the parents, called germline, or they can be developed over the lifetime of the individual, called somatic. Most mutations will have no effect on the individual and will never be noticed unless the DNA is sequenced. A few mutations may have severe effects on the individual, such as causing cancer. In the case of sickle-cell anemia[5] only a single nucleotide mutation is needed in the β-globin gene. Sources of somatic mutations include exposure to ultraviolet radiation, errors in DNA replication, and certain chemicals. Mutations include a variety of specific changes to the genome, such as single nucleotide polymorphisms.

Single Nucleotide Polymorphisms

The substitution of a single nucleotide at a specific position in the genome is called Single Nucleotide Polymorphism (SNP). The possible variations are referred to as alleles for the affected gene. A SNP will change the nucleotide in one sequence only, leaving the other in its original state. This is the most common type of genetic variation, and each human carries somewhere between four to five million SNPs[4] in their DNA.

SNPs can have a wide range of effects depending on which region they are located in. If they are within a gene coding region there is a chance they will alter which amino acid is expressed at the position, though in most cases this has no effect on the function of the expressed protein, the phenotype. Sometimes this single change is enough to cause a disease and single SNPs can cause sickle-cell anemia and β-thalassemia[6]. If the SNPs fall within the non-coding region, there is a chance they might alter the specificity of regulatory TFs like CTCF, resulting in altered expression for the whole domain of the affected binding site and even adjoining domains[7].

SNPs can also have combinatorial effects, making it difficult to identify the function of any singular SNP without proper context. The study of proximal contexts have provided some insights, i.e. studying genes close to the SNP. The addition of techniques for capturing long distance interactions of DNA, such as Chromatin Conformation Capture[8] (3C) and its derivations (4C, 5C, Hi-C, ChIA-PET), have allowed for identifying distal interactions of SNPs. Even with these additions it is difficult to trace a disease back to the causative SNPs, as there are many factors involved and there can be multiple disease pathways, as for example with cancer[9], a disease driven by mutations in DNA.

Genome-wide Association Studies

Genome wide association studies (GWAS) attempt to identify disease causing SNPs using statistical analysis[10]. Observing the genetic variants of a population, the GWA study attempts to find statistical correlation between observed SNPs and observed traits (Figure 2). There have been many GWA studies to date, and some have been successful in identifying disease associated SNPs[11]. Unfortunately, statistical analysis means that a stronger correlation is related to higher frequency of occurrence. As a result, to find less common disease associated SNPs, larger and larger study populations are needed, with recent studies exceeding 1.3 million participants[12]. Even at these numbers, it can be hard to find associated SNPs with low frequencies of occurrence.

A GWA study only explores the association between SNP and a trait, it does not assert a causative relation; that part will have to be explored in a more direct study of the functional effect of the SNP. In a combinatorial setting, only a part of the set of SNPs causing the disease may be discovered, complicating the more in-depth analysis at the level where causative mechanisms are studied. Given the vast number of disease-associated SNPs and the time required to properly explore their function, finding the correct ones to evaluate further is quite important.

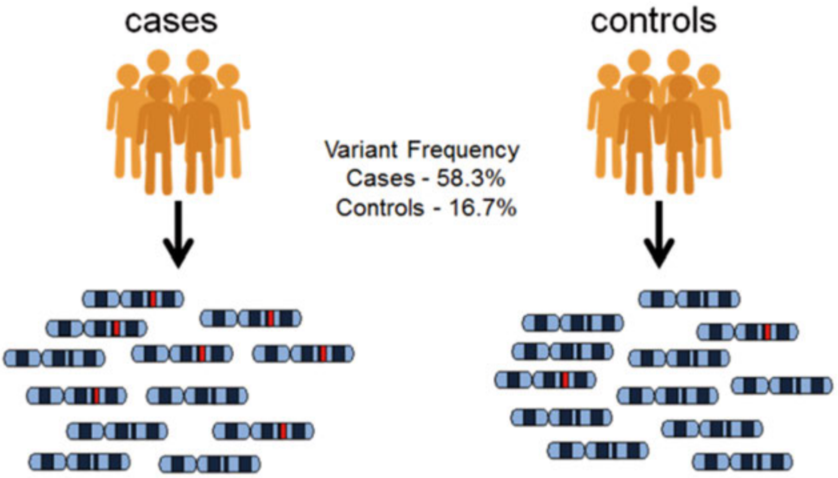


Figure 2 Example of how a genome-wide association is measured. The variant observed has a higher frequency in the case group than the control group for some disease. Image courtesy of EMBL-EBI.

DNA Sequencing and Chromatin Immunoprecipitation

DNA sequencing is the identification of the nucleotide sequence in DNA and forms the basis for research on genetic inheritance and mutation. Since its invention it has opened up new areas of research in biology and medicine[13]. Sequencing data is often combined with large-scale transcription factor binding analysis to see not only the nucleotide sequence, but also which areas of the nucleotide sequence that interact with various peptides and proteins. This large scale binding analysis is studied using Chromatin Immunoprecipitation sequencing[14], or ChIP-seq. ChIP-seq is a two-step process where DNA associated to a TF is first selected and then sequenced using high-throughput sequencing[15]. ChIP-seq must be done for a specific TF and will quantify the interactions between regions in the DNA and this protein to an accuracy of 50-100 base-pairs. The choice of TF will allow conclusions about the likely role of that DNA region. An architectural protein like CTCF can indicate the bases of DNA looping regions, which cluster long DNA sequences into a single domain of related genes. Proteins more specific for the expression of certain genes can be used to test the effect of mutations on the related expression.

In practice, multiple proteins are used to ensure that the analysis is correct. For experiments on the effect of a single SNP on gene expression, it is common to use a protein involved in the gene expression, histone modification proteins to observe the chromatin status of the region, and DNase I enzyme (through DNase-seq) to show the overall transcriptional activity of the region.

Some Biology Concepts

eQTL

Expression Quantitative Trait Loci (eQTL) are genomic loci involved in some or all of the variation in expression levels of mRNA[16]. An eQTL SNP is a polymorphism that causes the transcription rate of a gene to change, either to increase or to decrease. It can be compared to a dial on the oven, increasing or decreasing the temperature, potentially turning the oven off completely. eQTL SNPs always affect the expression levels, but the process through which that occurs can be of several kinds.

AS-SNP

Allele-specific Single Nucleotide Polymorphisms (AS-SNPs) are SNPs that have a statistically significant impact on the binding affinity of an allele[17]. This causes transcription factors (TFs) to bind preferentially to one allele over the other resulting in one of the two DNA sequences to become dominantly expressed.

TAD

A Topologically Associated Domain (TAD) is a region in the DNA that is physically associated[18]. It is a loop that is created by the DNA when architectural protein complexes bind together an anchor site pair, forcing them together, forming a shape much like a noose. This noose can be a taut circle or a loose and serpentine one. This noose-like loop effectively allows distal enhancers to fold in and connect to their associated promoters and assist in gene transcription. TADs are often co-expressed and have associated functions, making it practical to express them all at the same time.

LD

Linkage Disequilibrium (LD) in population genetics is a term for correlation of occurrence between alleles, either negative or positive[19]. That is, in any given genome they appear together more frequently than expected by chance, much like how socks frequently but not always are of matching color and size. The co-occurrence is simply too frequent to have arisen randomly.

PMM

A Parsimonious Markov Model (PMM) is a predicted motif that accounts for the spatial context of TF binding[20]. A predicted motif is a sequence of nucleotides computed *in silico* for the likelihood of binding some specific set of TFs.

Cervical Cancer and Screening

Human Papillomavirus

Papillomaviridae is an ancient family of non-enveloped DNA viruses[21]. The many types of this family have been found to infect every type of mammal investigated[22] as well as various other vertebrates[23].

The virus replicates by entering a host cell nucleus and inserting its DNA into the host cell DNA. The viral proteins can then be expressed by the same machinery as the regular cell proteins.

Most often an infection is asymptomatic, but some types may cause benign tumors, more commonly known as warts or papilloma. Certain types of Human Papillomavirus (HPV) are well known for causing cervical and anal cancer and are also implicated in oropharyngeal, penile, vaginal, vulvar[24], and Head & Neck cancers[25].

Papillomaviruses are usually host specific and rarely transmit between species. They replicate exclusively in their type-specific basal layer of surface tissue[22], such as skin or the mucosal epithelium of genitals, anus, mouth, or airways; a quality that can make it difficult for the immune system to detect the infection[26].

Human Papillomaviruses infect a variety of surfaces: HPV1 infects the soles of the feet while HPV2 infects the palms of the hands. HPV6 infects the penile, vaginal, and anal epithelial layers. The most well-known HPV types are 16 and 18, both of which can cause cervical cancer. Infections by these two types account for approximately 70% of all cervical cancer cases in the west. HPV types 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73, and 82[27], account for the remaining 30%.

Cervical Cancer

Cervical cancer is estimated to have afflicted around 570,000 women worldwide in 2018 alone[28]. It is a disease driven by HPV DNA expressing proteins that have inactivated important tumor suppressing functions within the cell, allowing it to grow cancerous and replicate without inhibition. Infection by persistent high-risk HPV, most commonly HPV16 or HPV18[29], is a requirement for developing cervical cancer. The development of cervical cancer from HPV starts with the transfection of HPV DNA into the cell

DNA. The cell's gene expression process then starts expressing virus proteins. These proteins, in particular E6 and E7, suppress the expression of tumor suppressor genes[30,31]. E6 primarily binds and initiates the degradation of the p53 tumor suppressor protein, a critical cancer inhibiting component that can kill the cell when tumor-inducing behavior occurs, and E7 acts similarly towards several proteins of the Retinoblastoma family, proteins involved in the suppression of genes required for cell cycle progression. In HPV16, only a very specific form of E7 will induce carcinogenesis, and it is possible that it develops in situ as a result of the human immune system[29].

With these important functions inactivated, the cell can replicate freely and thus create more virus particles. As a side effect of virus replication the surface of the cervix may in time develop into a cancer tumor.

The development of cervical cancer tends to be slow, and there are several clinical diagnoses for the different stages. When changes occur, but before the growth is considered a cancer, it is called a Cervical Intraepithelial Neoplasia, or CIN. This occurs in three stages defined by how abnormal the cells look under a microscope and how much of the cervical tissue is affected (Figure 3, Figure 4).

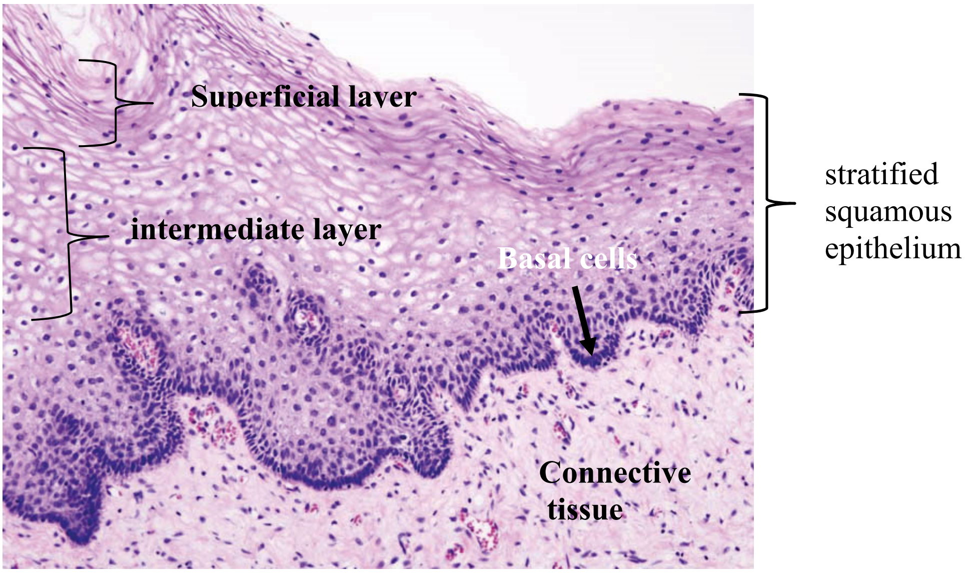


Figure 3 Normal cervical epithelium. (hematoxylin/eosin staining).

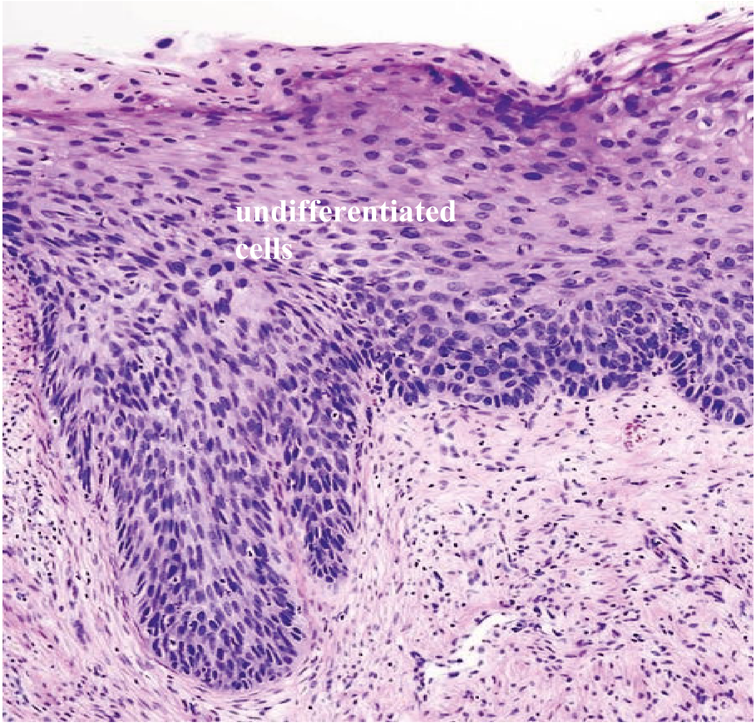


Figure 4. Cervical Intraepithelial Neoplasia Grade 3 (CIN3). There are many undifferentiated cells and they are spanning more than two thirds of the epithelium, the cells differ greatly in size, and many cells have irregular shapes.

A CIN3 is what may eventually develop into an invasive cancer. The mildest form of cancer is similar to CIN3 and can be treated the same way, with removal of the tumor via a surgical procedure. If the cancer develops further the likelihood of survival is reduced and the treatments become more severe, including radiotherapy, chemotherapy and surgical hysterectomy.

HPV Vaccination

The introduction of nationwide vaccination has been effective in reducing cervical cancer incidence, in some countries by up to 72%[32]. The first vaccine, Cervarix, protected against HPV 16 and 18, the two main culprits behind cervical cancer development. The vaccines that followed extended the protection. Gardasil protected against four different types: 6, 11, 16, 18. HPV 6 and 11 are low-risk types, but they do cause papilloma. The latest vaccine, Gardasil 9, protected against HPV types 6, 11, 16, 18, 31, 33, 45, 52 and 58, with all except 6 and 11 being high-risk HPV types. The clinical effects of Gardasil 9 will not be seen for a while yet, as it was approved in December of 2014.

Screening

Screening is the process of systematically testing a population for symptoms of a disease before it has developed. The purpose is to find dangerous conditions early on in the development stage, before they become fatal, as a medical intervention at an early stage is both safer to perform and more likely to prevent fatal disease outcomes. Cervix, breast, and prostate cancer are well known diseases to screen for, but there are many other diseases screened for, such as tuberculosis, depression, fetal abnormalities, or pneumoconiosis.

In screening, individuals from the at-risk population are invited to attend a local clinic where they can be examined for disease biomarkers. A biomarker is anything indicative of an underlying condition, such as visual changes in a cell that could eventually lead to cervical cancer, or high levels of prostate specific antigen (PSA) in the blood that could signify a potential development of prostate cancer. This examination of biomarkers is repeated at intervals, usually every three years for cervical cancer, until the individual is no longer considered part of the at-risk population. If a risk biomarker is discovered during these assessments, the individual is remitted for further examination and possible medical intervention.

Screening can be of different types. Mass screening tests a whole population regardless of status while high-risk or selective screening test only the individuals considered likely to develop a disease. Well known screening programmes, such as liquid-based cytology or mammography, commonly test a majority of the whole population, based on age, without consideration of risk factors.

Diseases should only be screened for when intervention is practical. There are considerable financial and social burdens associated with screening in the form of high costs and overdiagnosis, the latter leading to unnecessary medical interventions. Overdiagnosis refers to the discovery and treatment of disease symptoms that would not lead to a disease outcome, such as benign growths in the prostate or breast. Treating these symptoms cause unnecessary risk to the individual for little or no benefit, and further congests the clinics.

There is considerable discussion regarding the use of prostate cancer screening[33] due to the side effects of the confirmation test, where a small sample of the prostate is extracted, and it is currently unclear whether screening for prostate cancer actually reduces the mortality of the disease[34]. Breast cancer screening seems to offer only a marginal, if any, reduction in mortality as well[35]. Conversely, screening for cervical cancer has been highly successful in reducing both cancer incidence and mortality[36].

Cervical Cancer Screening Registries

Screening programmes record all examinations and store them in a registry. This information can then be used for quality control and auditing of the programme, as well as tracing problems or faults at the associated clinics. Each screening examination stored contains information about the individual, the diagnosis, the clinic, and the date.

Some Screening Concepts

There are many different aspects of screening and most countries with a cervical cancer screening programme have their own standards and protocols. There may even be regional differences within countries. To provide the best care and transparency for those involved in the screening programmes, there is a set of standards for definitions and processes. These are intended to make screening programmes comparable and to enforce minimum standards of performance and safety.

SNOMED

The Systematized Nomenclature of Medicine (SNOMED) is a computer-processable collection of medical items[37]. It is an international standardization protocol such that a screening diagnosis of cervical intraepithelial neoplasia I (M74006) in Sweden means exactly the same as mild dysplasia

(M74006) in the United States. In a medical scenario, small differences can have drastic consequences and standardization prevents these differences from causing problems. It also furthers research and communication between different countries as the population-wide results become directly comparable.

ICD

The Tenth International Statistical Classification of Diseases and Related Health Problems (ICD-10) is a set of medical diagnoses intended to clearly specify diseases[38]. The ICD standard defines diseases and health problems, such as diabetes or a personal history of breast cancer.

Auditing

During the course of a screening programme, it is necessary to test and validate the performance of the processes involved. This is done via audits. A sample of the recorded statistics from the registry is collected and compared to expected outcomes. If cancer incidence is higher in certain counties then further analysis and observation of the guidelines and practices of these regions are warranted. The level of detail of the data in Swedish registries allows for tracing irregularities and unexpected statistical outcomes down to the clinic and the responsible clinician if necessary.

Some Clinical Abbreviations

ACIS – Adenocarcinoma in Situ
AGUS – Atypical Glandular Cells of Unknown Significance
ASCUS – Atypical Squamous Cells of Unknown Significance
ASC-H – Suspected Malignant Dysplasia of the Squamous Epithelium
CIN1/2/3 – Cervical Intraepithelial Neoplasia Grade 1/2/3
HSIL – High-grade Squamous Intraepithelial Lesion
LSIL - Low-grade Squamous Intraepithelial Lesion
NILM - Negative for Intraepithelial Lesion or Malignancy

Machine Learning

Machine Learning (ML) is the use of statistical knowledge from data for the purpose of inferring knowledge about unknown data. It can be described as a statistical version of a medical doctor, diagnosing a possible disease in a patient based on the patient's symptoms, or if no disease fits the symptoms, inferring what family of disease the unknown malady belongs to. The reason the doctor can make such a diagnosis is that he has previous experience of disease symptoms, and based on that experience he guesses the most likely disease from the current symptoms.

ML comes in two formats: supervised and unsupervised. Supervised refers to objects having a known outcome, and this type of outcome is what classification will predict in objects where the outcome is unknown. This method focuses on finding differences between objects with different outcomes. An example of this would be trying to predict if a boat will perform well during certain weather conditions by looking at its performance during other weather conditions.

Unsupervised learning does not have objects with an outcome and focuses on clustering objects in an n-dimensional space based on what similarities the objects have. An example of this would be trying to cluster a sample of different cells into groups based on their observed similarities.

ML can be applied in any number of ways to create a model for classification. Support Vector Machines (SVMs) work by creating an equation that separates the objects from different outcomes in a n-dimensional space. Decision Trees take a sequential approach to classification and create a pathway to different outcomes based on whichever object features provides the most information at each point. Rough Set classifiers create minimal sets of features that can separate between some of the objects belonging to different outcomes.

Statistical Variance

Variance refers to the frequency of variation, that is, the likelihood that any datapoint in a dataset will differ between samples. In the context of the human genome, any position in the DNA can have one of four nucleotide bases: adenine, cytosine, guanine, or thymine (A, C, G, T). If a position always has the same nucleotide when looking at genomes from different people, it has no variance.

Aims

The purpose of this study was to find pragmatic approaches to predictive health, in particular the advancement of cervical cancer screening practices. This included three different aspects:

- The development of tools for finding genetic markers of future risk to increase the number of variables that could be used for prediction.
- The development of Machine Learning models for projecting future risk of developing cervical cancer.
- The development of tools for facilitating the discovery process of candidate markers to be used in the predictions.

Methods

Some statistical concepts

Object

An object in ML is a row in the dataset. The object can represent anything, the values of different tests run on a patient, the different properties of a car, the expression levels of genes from a particular cell, or the various physical characteristics of a person.

Feature

A feature is known by many names, and these usually vary between fields as well. It can be referred to as feature, attribute, variable, property, characteristic, parameter, dimension, class, vector, array, and many more. It refers to anything for which the objects have a recorded value. For instance, *color* can be a feature of cars and *blood pressure* can be a feature of patients.

Decision Class

The decision class is the outcome of an object, the purpose of the prediction. It is the “goal value” of the object that the classifier attempts to find out by using the other feature values. For predicting the risk of cervical cancer, the decision class can be the cancer status of the individual, and the classifier will try to predict the objects as either being a *cancer* case or a *control* case. For predicting the sex of a person based on physical properties, the outcome can be *male* or *female* and the classifier will be using features such as *height*, *weight*, *shoe size* and so on.

Classifier

A classifier is an algorithm that takes a dataset and attempts to classify all objects in the dataset to one of the decision classes, outcomes, using whatever patterns have been developed. It can be seen as the “prediction machine” that guesses the outcome of an object based on its experiences of other objects. It is usually the last algorithm to be run in ||-ROSETTA, and generates all the statistics about the prediction performance. In ||-ROSETTA, the classifier uses the rules to predict the most likely outcome.

Odds Ratio

The Odds Ratio (OR) is a comparison of the likelihoods of two groups reaching some outcome, for example the likelihood of developing lung cancer based on smoking or not[39]. OR requires an exposure (smoking) and a known outcome (lung cancer), creating four different categories (Table 1). It is the odds of the outcome given exposure and non-exposure.

Table 1 An example of the Odds Ratio calculation. The four values correspond to the numbers in the study population. D_e = smoker with lung cancer, H_e = smoker and healthy, D_n = non-smoker with lung cancer, H_n = non-smoker and healthy. D_e/H_e is the ratio of smokers with lung cancer to smokers who do not have cancer, i.e. a measure of the risk of having cancer if you smoke. D_n/H_n is the ratio of non-smokers with lung cancer to healthy non-smokers, i.e. the risk for cancer for those who do not smoke. The ratio of ratios describes the risk of smokers vs the risk of non-smokers for having lung cancer.

	Lung cancer	**Healthy**
Smoking	D_e	H_e
Non-smoking	D_n	H_n

From Table 1, The OR can be calculated as $R = \frac{D_e H_n}{D_n H_e}$.

The OR is useful in that it compares a group against a background. Non-smokers get lung cancer as well, so to say that smoking is the reason for lung cancer is not accurate. The OR can explain just how much the likelihood of lung cancer increases for smokers compared to non-smokers. In terms of cervical cancer, there is a population-wide incidence rate that varies depending on for instance which country the comparison is made in. To clearly see what effect certain measures can have, the comparison must therefore be made against the baseline odds of the population in question.

To properly describe the OR, it is also useful to get the confidence interval (CI) of the OR. The standard 95% CI gives a range for the OR which describes how the randomness of the data used to compute the OR might have affected the number. This is similar to how a single temperature reading in a city does not necessarily give the correct temperature, but the more readings taken at different locations in the city the more certain it is that the temperature is somewhere in the given range. The size of the CI range indicates how certain the OR value is; a large range means that the value is less precise, probably due to a limited sample size. A 95% CI means that repeating the procedure on new data from the same population an unlimited number of times will result in 95% of the confidence intervals including the real OR[40].

The CI is calculated after the OR. It is given by the formula $95\%\ CI = e^{\left(ln(OR)\pm 1.96\sqrt{\frac{1}{D_e}+\frac{1}{D_n}+\frac{1}{H_e}+\frac{1}{H_n}}\right)}$. OR is frequently used in medical settings to estimate population-wide risk for disease.

Risk Ratios

The Risk Ratio (RR) is similar in nature to the OR, but somewhat simpler. Instead of comparing an exposure to a background, the entirety of the population is used. The formula is $RR = \frac{D_e(H_eD_e)}{D_n(H_nD_n)}$. In scenarios where a disease is rare, the OR approximates the RR.

ROC

The Receiver Operating Characteristics (ROC) describe the performance of classification given two characteristics: sensitivity and specificity. Sensitivity is the fraction of how many ”positive” cases were predicted correctly, and specificity is the fraction of how many “negative” cases were predicted correctly. These are sometimes called True Positive Rate (TPR) and True Negative Rate (TNR). Sensitivity is the measure of how many sick individuals are correctly identified as sick, while specificity is the measure of how many healthy individuals are correctly identified as healthy. A high sensitivity in medicine means that many individuals with a disease are discovered. In the case of cancer, discovering it in time is crucial for successful treatment and rehabilitation. A high specificity is less important in many cases. Incorrectly predicting disease in an individual can be unpleasant, but is far less dangerous to the healthy individual than missing the presence of disease in a sick individual. This does not always hold as it can also be important with a high specificity as well. For example, the confirmation test for prostate cancer can lead to considerable complications, causing erectile dysfunction and difficulty urinating. Prostate cancer predictions therefor need a high specificity.

The ROC curve gives the total accuracy, number of correct predictions as a fraction of all predictions, as a function of both sensitivity and specificity, giving a picture of how these two perform. In almost all cases of prediction, sensitivity and specificity are contrary, and choosing a high sensitivity results in a lowered specificity and vice versa (Figure 5).

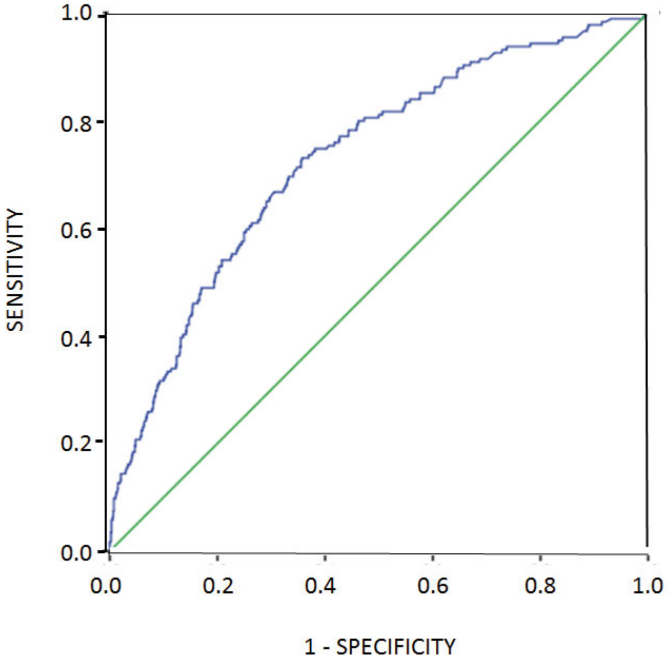


Figure 5 A ROC curve. The blue line is the performance of the classification, and the green line is the baseline, given by random guessing. The y-axis gives the sensitivity, and the x-axis gives 1 – specificity. The best prediction is in the upper left corner. The parable of the curve demonstrates that a higher sensitivity leads to a lower specificity.

The overall performance of the ROC curve is called Area Under Curve (AUC). It is the total area under the parable. The baseline for the comparison is an accuracy of 50%, which corresponds to just guessing what the outcome is and being correct half the time.

ROSETTA

ROSETTA is a program for building and running computational pipelines[41]. The ROSETTA pipeline takes a set of algorithms and a dataset, such as a set of patient medical parameters, and runs the algorithms on the dataset.

Usually the purpose of a ROSETTA pipeline is to train and test a classifier on the data in the form of a cross-validation (CV). This is done in several steps. First the dataset is divided into a number of pieces. These pieces are then grouped into a training set, for training the classifier on, and a testing set, for estimating the performance of the classifier. This is similar to dividing a deck of cards into two half-decks and using the first half-deck to practice the card game and then the second half-deck to play the card game for real. The cards will not be the same in the two half-decks but the principles will be the same. Some cards will have a different value from the training cards but the same suite, and other cards will have the same value as the training cards but a different suite. The practice half deck is very useful in learning the game, but it is not perfect.

The outputs from a pipeline like this are the rules used in the classifier and the statistics for the classifier: accuracy, ROC, OR and RR of each classifier rule. The pipeline that ROSETTA runs is highly customizable and can be designed to account for any type of data of any size as long as sufficient computational resources are available.

The pipeline consists of two parts: a training and a testing phase. The training phase revolves around preparing the data for processing and extracting the informative patterns, while the testing phase is used to evaluate the performance of the classifier created in the training phase.

Completion

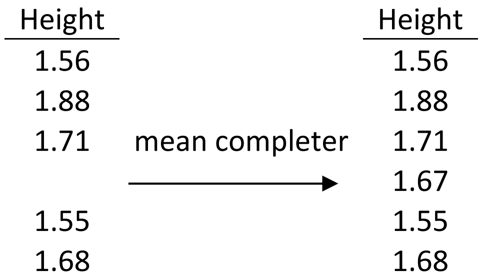
When working with incomplete datasets such as patient data sets from different hospitals, the first step of the training phase is usually completing the data, which means filling in the blanks where information is missing. Clinical data will often have missing values, and completion is a way to handle that. For example, the missing values in a patient visit record can be assigned as the mean value of those features (Table 2). If the blood pressure value is missing, then that value can be assigned as the average blood pressure from the data that is available. This is useful when the value is expected to follow rigid patterns; if the patient did not receive a blood pressure test it is unlikely that the blood pressure would deviate significantly from previously measured values.

The value can also be assigned as a zero or other unused value, to indicate simply that it is missing. This is useful when the omission of the value itself is an indicator of the outcome. If the blood pressure is missing from a patient visit record, that can be used to indicate that the patient showed no overt symptoms of any disease that would have given rise to blood pressure related anomalies.

A missing value can also be approximated from correlation with other features. A simple correlation test, such as Spearman or Fisher, can show that features have a linear, or direct, relation in values. In essence, this can be described as if feature f_a has value x then most likely feature f_b has value y. For blood pressure, if the patient has a kidney disease then it is likely that the blood pressure will be high.

There are many other ways of assigning missing values, and the method adopted should be chosen according to the purpose. The more information that is available regarding the relationship between the parameter with the missing value and other parameters, the better the estimation of the missing value will be.

Table 2 An example parameter before and after mean completion. The missing values in the parameter list are replaced with the mean value for that parameter.



Height		Height
1.56		1.56
1.88		1.88
1.71	mean completer	1.71
	→	1.67
1.55		1.55
1.68		1.68

The next step of the pipeline is discretization. This is the process of converting specific numerical values into intervals. For example, when predicting payment default on loans it is unnecessary to know the exact sum. Instead, consolidating the values into larger groups gives a much better overview of the situation. The three loan values $1,224, $1,335, and $1,687 can all be described as the same interval, [$1,000, $2,000], without impacting the accuracy of the predictions. This discretization is needed for non-linear relationships in the data, such as multimodal distributions, where a fitted regression model doesn't necessarily work well (Figure 6).

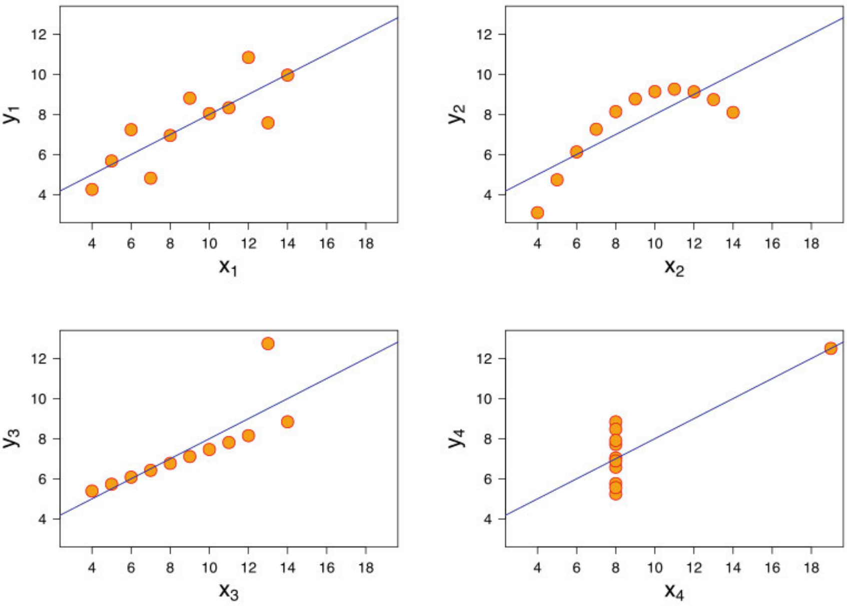


Figure 6 Anscombe's Quartet[42]. Four very different datasets all produce the same linear regression. Fitting a regression to the data requires a suitable model, but not all datasets have a suitable model available. Using discretization would instead create distinct clusters useful for classification.

Discretization

Discretization will sort, or “bin” the values of a feature into different intervals, changing a value x into an interval $[a, b]$ where $x \in [a, b]$. Doing this will avoid the need for mathematical understanding of the data, as the numerical data has been categorized instead (Table 3). Discretization thus eliminates the need to find an analytical mathematical equation that can be fitted to the data, which is most often neither possible nor desirable. Discretization will simplify a context into only the relevant parts. For instance, there is no absolute value for a healthy blood pressure. It varies from person to person, resulting in a wide interval of values to be considered. There is also no medical difference between a blood pressure of 50 and 51. No matter how you interpret this data on its own, there is no way to make a useful prediction from it. Discretization can be used to cluster the values into the categories that matter for doctors, such as low, normal, and high. This does not in itself increase the prediction value of the feature, but it does make the feature much more powerful when used in conjunction with other features. The combination of blood pressure: *low* and weight: *obese* is more predictive than blood pressure *51* and weight *139.*

There are many algorithms for discretizing data. Equal frequency binning, one of the simplest forms, sorts all the values of a feature and then divides them into a number of intervals. Applying equal frequency binning with 3 intervals to a feature of integers {1, 1, 2, 3, 4, 5, 6, 6, 21} would produce three intervals [1, 2], [3, 5], and [6, 21]. This type of algorithm is useful mostly when it is the relative changes in the data that are of interest and the purpose is to create a number of states corresponding simply to *low*, *medium*, or high values of the desired granularity, or in the case of differential expression analysis, values that are *unchanged*, *down-*, or *up-* regulated compared to previous or following data points. Treating values as relative assumes that all changes are significant, as small changes can end up in the same bin as large changes. Using the example of blood pressure, patients with increasing blood pressure would end up in the same group regardless of whether the blood pressure was low or high from the beginning, as would patients with unchanged or decreasing blood pressure. This is a very useful measure in for example evaluating the effects of new medications.

Table 3 Discretization using Equal Frequency Binning at two bins. The values are ordered from lowest to highest, and the cuts are placed to create an equal number of values into each bin.

Height		Height
1.56		[1.55, 1.67]
1.88		[1.67, 1.88]
1.71	Equal Frequency Binning	[1.67, 1.88]
1.67	→	[1.55, 1.67]
1.55		[1.55, 1.67]
1.68		[1.68, 1.88]

Manual discretization is also an option. Data-driven algorithms generate intervals from what is available in the dataset, but sometimes it is more efficient to create intervals based on knowledge from other sources. Looking at the distributions of a feature in the data by decision class, e.g. the distribution for the amount of money borrowed when looking at payment default, can yield effective cut-points for intervals and also create smaller high-impact intervals with great accuracy if desired. Relying on existing literature or expertise for relevant intervals can also be of great help, especially in interdisciplinary studies where standards may differ between fields (Figure 7).

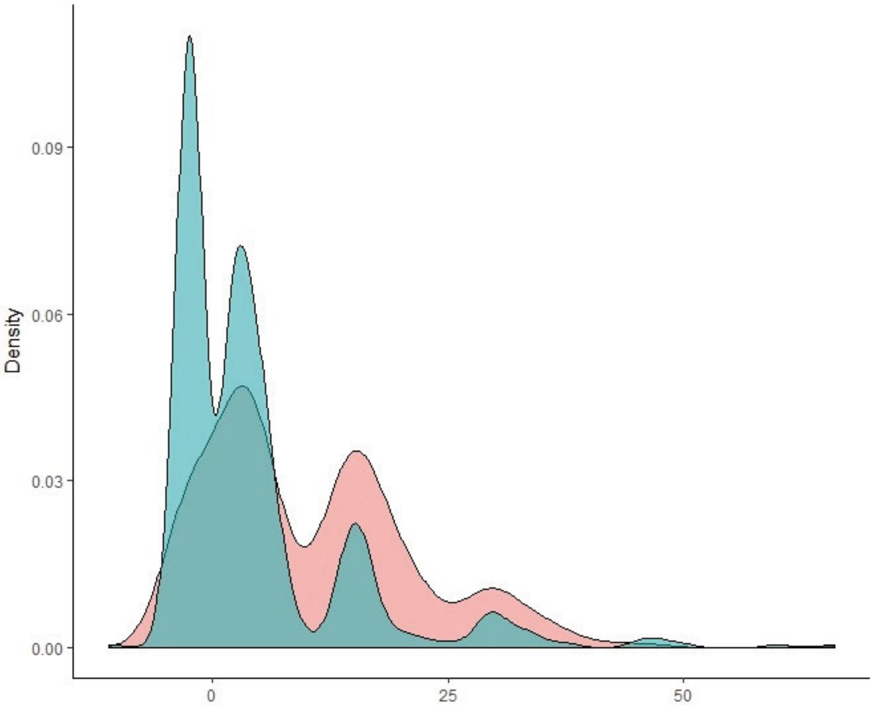


Figure 7 The distributions by class of a sample feature. There are many ways to define intervals of a feature depending on purpose. A single cut at around 10 would create two intervals within which one class would be dominant, creating a binary-value feature that can easily be combined with other features for classification. The greatest separation would yield cuts at approximately -5, 0, 5, and 45, leading to a feature with less likelihood of combination with others, smaller impact per rule, but better accuracy.

Reducts

A reduct can be described as a minimal set of features which together hold meaningful information regarding the decision value of objects. For example, when predicting if a man speaks English it is not important to know his hair color, shoe size, or mother's name, it is enough to know his nationality and educational grade in school. A likely reduct, or minimal set of informative features, from this dataset would thus be the pair *nationality* and *educational grade*. These two features are not necessarily enough on their own, but together provide enough information for a mostly accurate prediction.

Reducts in the simplest form are computed by observing which features can separate between objects in a sequential process. This is similar to looking for all the features that have different values between a man and a woman, such as height or weight. Usually multiple features are needed for this.

In the post-discretization decision system presented in Table 4 there is no singular feature that can be used to determine whether an object *o* belongs to the decision class *Sex(F)* or *Sex(M)*. By using combinations of features, the decision class can be predicted.

Table 4 An example decision system after discretization.

	Height	Weight	Hair color	Age	Sex
O_1	[1.55, 1.67]	[49,64]	Brown	[17, 39]	M
O_2	[1.68, 1.88]	[67, 90]	Brown	[40, 56]	F
O_3	[1.68, 1.88]	[67, 90]	Black	[17, 39]	M
O_4	[1.55, 1.67]	[49,64]	Black	[40, 56]	F
O_5	[1.55, 1.67]	[49,64]	Black	[17, 39]	F
O_6	[1.68, 1.88]	[67, 90]	Black	[40, 56]	M

This process is handled in two steps. First, all features are evaluated between every pair of objects in what is called a discernibility matrix, and for each pairing the features that have different values for the two objects are added to a list. For complete discernibility between objects, this is the list that will be used (Table 5). In a practical setting this list will be too long and too specific to create strong feature sets that can be used on a diverse population.

Table 5 A discernibility matrix for the objects in Table 4. The set of features that can be used to separate between objects is written in disjunctive form.

	O_1	O_2	O_3	O_4	O_5
O_1	X				
O_2	*Height* ∨ *Weight* ∨ *Age*	X			
O_3	*Height* ∨ *Age* ∨ *Hair*	*Hair* ∨ *Age*	X		
O_4	*Hair* ∨ *Age*	*Height* ∨ *Weight* ∨ *Hair*	*Height* ∨ *Weight* ∨ *Age*	X	
O_5	*Hair*	*Height* ∨ *Weight* ∨ *Hair* ∨ *Age*	*Height* ∨ *Weight*	*Age*	X
O_6	*Height* ∨ *Weight* ∨ *Hair* ∨ *Age*	*Hair*	*Age*	*Height* ∨ *Weight*	*Height* ∨ *Weight* ∨ *Age*

When all the features that can discern between each object pair have been noted, the feature sets that separate objects with the same decision class are removed. This simplifies the list and removes features that are not necessary for the classification, as there is no benefit to classification from retaining information that separates between objects with the same decision class. In the example given earlier about looking for features that have different values between a man and a woman, there is no benefit to classification from retaining features that can separate between two different men. The decision-relative discernibility matrix is used for computing the reduct (Table 6).

Table 6 The decision-relative discernibility matrix. All feature sets that discern between objects with the same decision class are removed.

	O_1	O_2	O_3	O_4	O_5
O_1	X				
O_2	*Height* ∨ *Weight* ∨ *Age*	X			
O_3	X	*Hair* ∨ *Age*	X		
O_4	*Hair* ∨ *Age*	X	*Height* ∨ *Weight* ∨ *Age*	X	
O_5	*Hair*	X	*Height* ∨ *Weight*	X	X
O_6	X	*Hair*	X	*Height* ∨ *Weight*	*Height* ∨ *Weight* ∨ *Age*

After the decision-relative matrix has been computed, the feature sets are simplified from disjunctive form to conjunctive form. This is the logical reduction that produces the smallest possible set of features. The complete form (*Height* ∨ *Weight* ∨ *Age*) ∧ (*Hair* ∨ *Age*) ∧ (*Hair*) ∧ (*Hair* ∨ *Age*) ∧ (*Hair*) ∧ (*Height* ∨ *Weight* ∨ *Age*) ∧ (*Height* ∨ *Weight*) ∧ (*Height* ∨ *Weight*) ∧ (*Height* ∨ *Weight*) ∧ (*Height* ∨ *Weight* ∨ *Age*) can be reduced to *Height* ∧ *Hair*. This minimal set of features is called a reduct, and can be used to determine, for each object in the dataset, whether it belongs to decision class *Sex(M)* or *Sex(F)*.

The reducts are used to build the rules which form the “knowledge” of the classifier.

Reduct computations in practice are more complicated as there is rarely such a clear separation between decision classes and often it is necessary to allow for some error rate.

Rules

The basis for the classification with ROSETTA consists of two parts: the classification schema or voter, which determines how to count votes, and the rules. Rules are patterns in the data that predict a decision combined with the statistical relevance of that pattern (Table 7). The rule can be seen as an ID card. Like the pattern, the name and the picture on the ID are the most important parts, and they matter to everyone that views the ID. The other notes on the card are more specialized, and matter more or less depending on the situation. A bouncer at a night club would care only about the age statistic on the card, a registrar might care only about the region of origin statistic of the card, and airport security might care only about the verification code of the card. When using the rules, the purpose and assumptions about the data determine which statistics are important.

Each rule has eight components. The first is the pattern. This is a conjunction of features in the form of an "if" statement: "IF F_a = X AND F_b = Y THEN Decision = 1". This pattern determines which objects the rule applies to. All objects *o* where $F_a(o) = X$ and $F_b(o) = Y$ will be voted on according to the rule. The second component of a rule is the left-hand side (LHS) support. LHS support is a number indicating how many objects in the data follow the pattern of the rule. LHS support is often used as a measure for how general the rule is, and a high support is a strong indicator that the rule is applicable beyond the data onto the population that the data represents. Moreover, a high support often gives the rule more importance in voting. The third component is the right-hand side support (RHS) support. This is a set of numbers of how the LHS support is split amongst the decision values. It is rare that a rule with high support only applies to objects with the same decision value, and the RHS support shows how the objects are divided. The fourth component is the accuracy. It shows, for each decision value represented by the objects of the rule, what the prediction accuracy is for that particular decision value. The further apart the prediction is between the decision values, the better the accuracy of the rule. Rules with only a single possible decision value always have an accuracy of 100%. The fifth component is LHS coverage. This value is equal to the LHS support divided by the number of objects in the dataset and represents how big a fraction of the entire dataset matches the pattern of the rule. It can be used in lieu of LHS support to determine how general a rule is, provided that the dataset is a reasonable representation of the population it is intended to emulate. The sixth component is the RHS coverage. This is a set of numbers that shows how big a fraction of each decision value is covered by the pattern of the rule. Each number is equal to RHS support for that decision value divided by the total number of objects in the dataset with that decision value. The seventh and eighth components of the rule are the Odds Ratios (ORs) and Risk Ratios (RRs) for the rule. These give the likelihood of the decision values given the pattern, with the comparison base being every object that does not fit the pattern.

Table 7 A sample rule taken from the cervical screening classifier.

IF Abnormal diagnoses < 2 AND HPV tests = 0 AND Last diagnosis = Benign AND Inconclusive tests = 0 THEN	CASE	CONTROL)
Support (LHS)	13,480 object(s)	
Support (RHS)	5,026 object(s)	8,454 object(s)
Accuracy (RHS)	0.37	0.63
Coverage (LHS)	0.35	
Coverage (RHS)	0.26	0.44
Odds Ratio	0.45 (0.43 - 0.47)	2.2 (2.1 - 2.3)
Risk Ratio	0.66 (0.64 - 0.67)	1.45 (1.42 - 1.48)

The IF => THEN pattern at the top indicates which objects are covered by the rule. The decision value has two possible values with the pattern, CASE and CONTROL. The left-hand side (LHS) support shows the number of objects covered by the pattern, and the right-hand side (RHS) support how those objects are split between the decision values. The accuracy indicates how often an object is correctly predicted using the pattern. The LHS coverage is the fraction of the dataset that can be predicted using the pattern, and the RHS coverage shows the fraction of the dataset covered when only looking at the same decision value. The OR and RR shows the likelihood of an object having the decision value when covered by the pattern compared to not being covered by the pattern.

The classification process usually generates a large number of rules. These can be similar or not, and it is common for an object to be covered by multiple rules. These rules can have differing predictions, and to resolve the classification of the object it is necessary to develop a voting system whereby each rule that fires for an object can be evaluated and given voting power commensurate to the relevance of the rule. This voting system is called a schema or voter.

Classification Schemas

Classification schemas, or voters, are a type of meta-algorithms that classify objects given a ruleset for those objects. The classification schema determines how to evaluate the relevance of each rule, how to deal with objects that have no qualified prediction, and how to assess the classification. The voting process will take each object to be classified, tally the votes from each rule that fires for the given object, and give a prediction for that object, much like a courtroom judge will do after hearing all the evidence for and against. If the object was correctly classified, the accuracy of the classification increases. The schema will look at the rules that fire. Each rule casts its vote for its most likely decision value, and the classification given is whichever decision value has the highest amount of voting power after all rules have voted.

The most common schema is that a rule is given voting power equal to its LHS support, hence general rules are valued higher than more specific rules.

This assures that the rules most likely to be applicable to the population represented by the data are the ones that dominate the voting process. This is useful for data where there is no specific pattern discernable or where there is only a single pathway for the decision outcome, such as behavioral studies on a wide demographic of humans or the pathogenicity of specific strains of avian influenza[43]. In scenarios where there are multiple pathways using only LHS support can instead obfuscate the outcome, such as rules for carcinogenesis patterns losing voting power to rules for general inflammation patterns because there are several pathways to cancer[44] but only a single general inflammation pattern. Using accuracy as the quality of a rule results in similar problems. The accuracy of a rule is always inversely proportional to the support of a rule in a relational sense. Relaxing the pattern of the rule will result in a greater support value, but also reduce the accuracy of that rule. If a single object matches a rule, the accuracy will be 100%, but it is likely that this pattern arose by chance and does not represent any pattern in the actual population. Hence, some level of support is required for the rule to have statistical significance before the accuracy can be used as a measurement of quality.

In order to classify an object, some level of certainty might be required for one or more of the decision values. For instance, predicting a potential cancer case as non-cancer should only be done with certainty, as incorrectly predicting non-cancer is far more dangerous than incorrectly predicting cancer. The schema determines what level of certainty is required for classification. Classifying an object as non-cancer might require that at least 75% of the voting power predicts non-cancer with the object being otherwise classified as cancer. The schema can also refuse to classify an object, or classify it as a separate decision entirely if the rules do not provide a strong enough vote for any decision value.

Results

Paper I

Aim

The first paper aimed to create a proof of concept stratification model for cervical cancer screening attendants such that the frequency of visits for medical tests could be modified based on the perceived risk of developing cancer. This model would be used to increase the number of tests of high-risk individuals but reduce the number of visits to the laboratory for low-risk individuals.

Methods

An updated version of an audit set[45] of all 4,137 cervical cancer cases in Sweden 2002 – 2010 with 121,339 age-matched controls was used as the study population. The data contained the entire history of SNOMED-defined[37] screening results and biopsies for the included women as well as the ICD-10[38] cancer status outcomes. The data was filtered to remove those with non-standardized diagnosis codes, those related to non-cervical cancer, those obtained too close in time to the cancer diagnosis, data collected too close in time to another, all data obtained after a cancer diagnosis had been given, data for individuals that appeared in both the case and the control populations, and finally any data in the control population that no longer had a match in the case population. The filtered data were gathered into complete histories, and these histories were then fitted with metadata such as number of missed screening opportunities and worst biopsy result. Further, each SNOMED diagnosis was assigned a value in terms of how likely it was to indicate a possible cancer diagnosis in the future by a medical expert, and these values were then used to compute a total risk score, the Cumulative Risk Score (CRS) for each history. The CRS was weighted by time, ensuring that results obtained long ago did not contribute as much in estimating the risk for cancer as more recent events.

After the histories had been processed this way, the relevant cases were selected for testing and combined in a dataset. All cases with at least four datapoints, where a datapoint represents a medical examination, and at most ten datapoints, were selected and matched with one control of equal history size drawn randomly from the matched control group of that case. This dataset was then used to train a rule-based classifier[46]. The accuracy and ROC of the classifier was used as an indicator of the overall performance of the protocol while the rules generated were used as predictors of the performance of individual features such as the computed risk score. Accurate rules were tested on the entire study population to get Odds Ratios and Risk Ratios that reflected the total population of the dataset.

Results

The accuracy of the classifier was low (64%), mainly due to the large contingent of asymptomatic histories (62% of the study population). Accuracy for high-risk subgroups in the population was significantly better, with ORs ranging from 8 to 36 for various patterns, meaning that these subgroups were up to 36 times more likely to develop cervical cancer than normal. Approximately 98% of the controls had a risk score below 10, while 11% of cases had a risk score of 10 or higher. The risk score identified groups with increasingly high risk: at CRS 15+ the OR was 20.3 (16.0–25.8), at CRS 20+ the OR was 24.4 (19.0–31.2) and at CRS 25+ the OR was 36.6 (27.3–49.2). Also, low-risk groups were identified: at CRS -3 the OR was 0.77 (0.62–0.95), at CRS -5 the OR was 0.65 (0.48 – 0.88). Testing the CRS time weights by removing all events which occurred a certain number of years ago showed a high level of consistency in the data, with most intervals showing similar predicted risk (Figure 8).

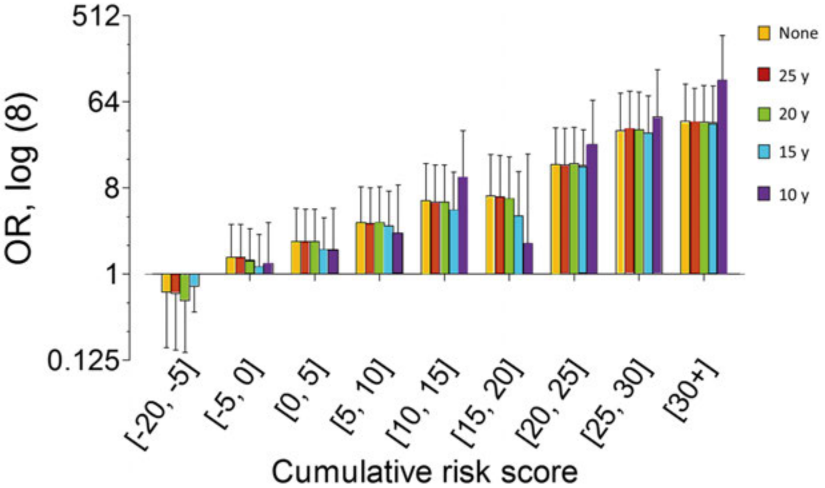


Figure 8 Log OR from CRS based on time constraints. The histories were censored at different time points and clustered in intervals. Censoring the oldest diagnoses still yielded consistent scores.

The results suggest that the model described here is a functional tool for separating low-risk from high-risk groups in the screening population for cervical cancer. The different cutoffs used to test robustness of the time-based weighting indicate that the model is stable, though the sparse amount of data available in the 10-year cutoff makes for uncertain intervals, as expected. The patterns discovered include both well known risk factors for cervical cancer such as non-attendance and pre-cancer diagnoses, and previously unknown factors not necessarily related to diagnoses, such as an over-diligence in screening attendance. The risk score performs as intended, providing a semi-linear score indicating the current risk of cancer development.

Overall, the study has successfully shown that a bioinformatical approach to risk stratification within the cervical cancer screening programme is possible and practical.

Paper II

For definitions and descriptions of terms, see Some Biology Concepts.

Aim

The aim of the paper was to create a pipeline to identify allele-specific single nucleotide polymorphisms (AS-SNPs) with possible functional significance in complex diseases with the intent that sorting through vast amounts of genomic data for relevant biological targets should be automated and accessible. This process could either serve as a tool to search large genome panels for inherited risk factors and the likelihood of cancer development or for further genetic inquests into disease progression and development.

Methods

We used ChIP-seq data from seven lymphoblastoid cell lines for three histone modifications and two architectural proteins. This data was downloaded from the Gene Expression Omnibus[47]. From each of the seven cell lines we generated two *in silico* genomes using the diploid genome reconstruction model from the ALEA toolbox[48]. These genomes were analyzed for AS-SNPs which were further filtered for significance and relevance by removing all SNPs within the signal artifact blacklisted ENCODE regions[49], centromeric and telomeric regions. The remainder were assessed for their frequencies in the population from 1000 Genomes SNP collection[4] and DHSs.

The remaining set of AS-SNPs was intersected with 1,545 unique GWAS SNPs for B-cell related traits[50] as well as 26,300 SNPs in linkage disequilibrium (LD) with them, and with 5,565 eQTL SNPs associated with gene expression in lymphoblastoid cell lines[51] as well as 66,935 in LD with them. AS-SNPs found to coincide with those in the databases were selected. AS-SNPs were also assessed with RegulomeDB scores[52] for quality and filtered by region associations with ChromHMM [53]and tfNet[54]. SNPs within the following categories were selected for further analysis: enhancer, heterochromatin, insulator, mixed, promoter, repressed, and transcribed.

The SNPs were also intersected with 130,915 anchor loci from Hi-C data of GM12878 and with TADs of GM12878[55] for a more comprehensive picture of the likely topological domain the SNPs might affect.

To assess the possible types of effect SNPs might have, we gathered all putative motifs from the HOCOMOCO database[56] as well as PMMs predicted by the de novo module of InMoDe[57] for a total of 404 PMMs. We then scanned the AS-SNP positions ± 300bp on both strands for sequences matching these PMMs.

Results

We identified 17,293 unique AS-SNPs in total, of which 1,199 were rare. 2,050 of these AS-SNPs were discovered only when pooling all ChIP-reads together. The number of AS-SNPs discovered from each cell line correlated well with the number of reads available for that cell line. To identify which AS-SNPs were likely to play a role in disease and gene expression we intersected the AS-SNPs with GWAS and eQTL top hits and those in LD with top hits. We identified 237 AS-SNPs related to traits, 18 of which showed allele-specific signals. A total number of 216 were in LD with GWAS SNPs, and 714 AS-SNPs were related to gene expression, of which 98 were allele-specific and 603 in LD with eQTL SNPs. Of the AS-SNPs in LD with eQTL SNPs, 3 were rare.

Many of the SNPs were found in the Human Leukocyte Antigen (HLA) region on chromosome 6, a highly polymorphic region difficult to explore. For example, the Type 1 Diabetes-associated GWAS SNP rs9272346 is located in the HLA region, which is confounding given its regulatory effect. We identified 26 AS-SNPs in LD with rs9272346, located in at least 10 different regulatory elements. These AS-SNPs were located in untranslated regions (UTRs) and intronic regions of the HLA alleles DQA1 and DQB1, reported as coding for HLA epitopes and most strongly associated to type 1 diabetes (T1D) susceptibility, suggesting that expression varies between alleles and may contribute to the risk of T1D. Similarly, we identified AS-SNPs with loci for several auto-inflammatory diseases including 8 alleles for celiac disease and inflammatory bowel disease and 11 for ulcerative colitis. We found AS-SNPs clustered around loci associated with multiple sclerosis (10 loci), systemic lupus erythematosus (4 loci), rheumatoid arthritis (3 loci), amyotrophic lateral sclerosis (4 loci), vitiligo (3 loci) and celiac disease (4 loci). We also identified 61 AS-SNPs associated with multiple traits, suggesting that some functional regulatory elements and genes are shared between multiple immune diseases.

Using PMMs, putatively altered binding sites were predicted for 98 and 325 AS-SNPs from the SNPs associated with GWAS and eQTL, respectively. Loss of predicted binding affinity was more prominent than gain and the average loss of predicted binding affinity was approximately 54%.

The discovery of regulatory AS-SNPs in LD with known GWAS SNPs shows the complementary power of our approach to GWAS and helps refine the candidate list for further exploration of autoimmune diseases. Predicting the level of alteration in affinity for the binding sites using PMMs provides an extra layer of filtering for candidates based on which transcription factors are involved in the gene expression and further emphasizes what effect a candidate may have on the region.

Given the results, the model described herein has proven to be an efficient filter for candidate AS-SNPs in immune diseases, complementary to existing databases such as GWAS and eQTL.

Paper III

For a list of clinical abbreviations, see Some Clinical Abbreviations.

Aim

The cervical cancer screening programmes in Sweden and Norway have successfully reduced the frequency of cervical cancer incidence but have no evaluation of or prediction for future screening needs. This means that the screening frequency for individuals can be suboptimal, increasing either the cost of the programmes or the risk of missing early stage cancer development.

The aim of this paper was to validate the proof-of-concept model based on Swedish data and further develop the framework for assessing an individual's risk of cervical cancer based on the available screening history. The earlier creation of a risk assessment score, called the Cumulative Risk Score (CRS), was to be further developed as a data-driven separation model together with multiple derived attributes.

Methods

A selection of 10,817,130 screening diagnoses was collected from the Norwegian Cancer Registry (NCR). These data contained cytology, HPV status, and histology diagnoses, accounting for 5,055 cancer case histories and 1,726,789 non-cancer histories, labelled as controls. The data were cleaned to remove any inconsistencies or unrelated discoveries such as non-cervical cancer, resulting in the removal of approximately 1% of the diagnoses. After cleaning, HPV tests were appended as a status marker to all diagnoses taken within a year of the HPV test, and the HPV test itself was subsequently removed. Any diagnosis less than one year before the actual cancer diagnosis and all diagnoses following it were removed. If diagnoses in a history were clustered together with others given within a time-span of less than a year, all but the last one were removed. Any control histories that exhibited a pattern signifying a medical intervention at a pre-cancerous stage such as CIN2/3 or ACIS followed by a recidivistic diagnosis such as Normal, NILM, ASCUS, Inconclusive, or CIN1, were relabeled as interventions instead of controls. This was to indicate that the history up to that point was more likely to be indicative of carcinogenesis than any form of benign pattern. After filtering, the study population contained 2,928 cases, 1,372,071 controls, and 53,120 interventions. These histories were fitted with derived data, such as non-attendance, total number of HPV tests etc.

The previous model for computing the perceived risk of cancer given a diagnosis history[58] was replaced with a data-driven separation model using C++ with Gecode[59] libraries. In this separation schema, all diagnoses were given a range of possible weights, a value for each history was calculated given these weights, and then the overlap of these history values between case and control histories was measured as a fitness function. The weights for the diagnoses were continuously adjusted until the overlap of history values between controls and cases was as low as possible. These weights were then used to calculate the CRS.

Excerpts from the study population, all the cancer cases and interventions matched with controls, were then evaluated in ||-ROSETTA for predictive patterns. The patterns discovered were further evaluated on the whole study population to see how well the predictors performed on a representative sample of the screening population.

Results

The best classification had an accuracy of 80.7%. CRS was the predominant predictor in the rules, creating a risk stratum containing 31.2% of the cancer cases and 1.9% of the controls. A CRS of 10 or above had an OR of 45.8 (43.4 - 48.3) and covered 11.3% of cancer cases, and no CRS value at 2 or above had an OR below 10, indicating a risk of developing cervical cancer at least 10 times the normal for all individuals in that range. The lowest risk defined by CRS was found in CRS -1 with an OR of 0.51 (0.49 - 0.54), half as likely to develop cervical cancer as the normal, and CRS -2 with an OR of 0.80 (0.75 - 0.85), covering 18.2% and 2.9% of cancer cases, respectively. In the study population, 61.2% of the controls and 43.9% of the cancer cases had only benign diagnoses in their screening histories. The distribution of CRS values peaked around 0 for both cases and controls (Figure 9).

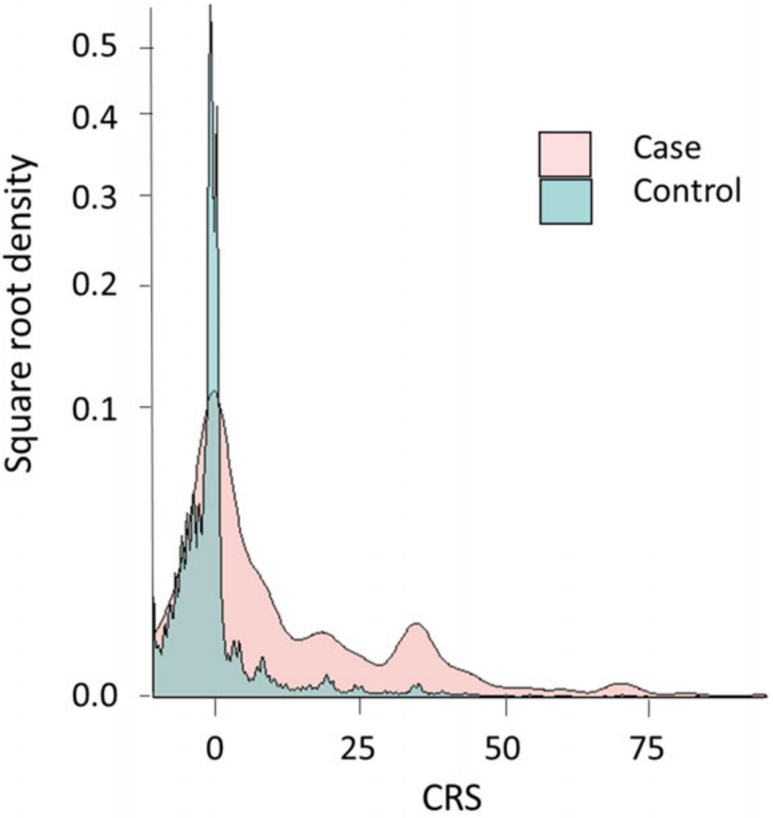


Figure 9 The distribution of the Cumulative Risk Score (CRS) by cancer case and control history. There is a distinct peak around 0, which is the starting point for new screening programme attendants as well as what individuals tend towards if their attendance is drastically lower than the recommended screening schedule.

There was no discernable difference in CRS between squamous cell carcinoma and adenocarcinoma. Intervention histories had the highest mean CRS at 2.96, cases were lower at 0.24, and controls the lowest at -1.33. Intervention histories were also far more likely to contain diagnoses warranting further testing or recall, with CIN1 and LSIL having incidence rates 4.5 and 2.3 times higher than the cancer cases.

The patterns discovered were consistent with the previous model developed based on Swedish screening data, suggesting that trends and behavioral markers are consistent in the two countries for cervical cancer development and avoidance. The pervasiveness of attendance markers in the classifier rules for both Norwegian and Swedish data show the same patterns as well, indicating that perfect attendance in the screening schedule is a signifier of higher cancer risk. The data-driven modelling for CRS calculations has increased the accuracy of the predictions considerably over the previous model that relied on expert assumptions, creating a greater separation between cases and controls. The new approach to screening recall based on this prediction model is likely to improve the use of resources in the screening programme and detect additional cancer developments at an earlier stage.

Paper IV

For definitions and descriptions of terms, see Some Biology Concepts.

Aim

Using our pipeline for identifying putative candidates of gene regulation, we wanted to identify AS-SNPs in the liver likely to affect the development of Type 2 Diabetes (T2D) and other metabolic diseases.

Methods

We acquired liver tissue and sequenced it using 10X Genomics tech at a mean depth coverage of 36x. We used the ALEA toolbox[48] to generate two *in silico* personal genomes of the liver sample. The two genomes were then processed with the pipeline from Paper II[17] using ChIP-seq data from three different transcription factors. The resulting AS-SNPs were tested for possible effects on transcription factor motifs using the funMotifs framework[60].

Results

We discovered 2,329 heterozygous AS-SNPs in putative regulatory elements using our established pipeline. Of these, 25 were associated with liver and metabolic related traits at 17 different genomic loci. Four of the AS-SNPs were found in loci associated with T2D on chromosomes 6 and 17. In both cases two AS-SNPs were found in a regulatory region and in LD with a GWAS SNP that is not in a regulatory region, suggesting a likely functional relationship that can explain why there is an SNP in an exon that is identified in GWAS as associated with T2D.

Overlapping the AS-SNPs with the motif map of the funMotifs framework resulted in 595 motifs altered by the AS-SNPs in the liver tissue. We identified 134 variants in 166 functional motifs, the majority being in transcription start site (TSS) regions. The most recurrent motifs altered were for factors EGR1, CTCF, KLF5, and ZNF263.

The systematic strategy presented in our previous work and demonstrated here with the addition of funMotifs has found several functional gene regulatory variants and possible target genes in human liver tissue. The AS-SNPs discovered offer a set of candidate regulatory variants supported by several layers of evidence ready for experimental validation for understanding the molecular mechanisms of many metabolic and liver diseases.

Paper V

Aim

The two aims of the paper were to create a practical approach to classification of large datasets and to promote the application of interpretable classification via Rough Set theory. Given the needs of accountability and explanation in medical applications, any classification used as the basis for a medical decision should be traceable throughout the decision process. The use of Rough Sets as the basis for classification provides a set of rules that can be analyzed and assessed individually for relevance and relation to other factors in the disease progression. Increasing the computation speed of this process is of considerable interest as it might take more than six months to construct a classifier with large datasets.

Methods

The ROSETTA C++ source code was updated with multi-threading capabilities using OpenMP. The update focused on computational speed and framework modularity, as it was important to retain the simplicity of implementing new algorithms and modules into the framework. This was managed through a separation of computational resources within the program that increased the memory requirements of execution.

The resulting program was evaluated for efficacy and speed using four datasets with different dimensions; a balanced dataset on histone modifications, an object-focused dataset on the likelihood of credit card payment defaults, a feature-focused dataset on Systemic Lupus Erythematosus (SLE), and a large synthetic balanced dataset with simplistic data. All tests were repeated with multiple thread numbers. Memory peaks and computational speeds were measured for all tests. The evaluation consisted of a ten-fold cross-validation (10CV) using one, two, three, four, five, six, and ten threads. The 10CV test had to be run ten times to reach completion, resulting in expected speed gains at two, three, four, five, and ten threads when compared to the previous recorded speed.

Results

The threading was successful in increasing the speed of the computational process. With ten threads, the time required to run the pipeline was reduced to 15.8% of the single-threaded time for the Histone Modifications dataset, to 29.6% of the single-threaded time for the SLE dataset, to 16.4% of the single-threaded time for the Credit Card Default dataset, and to 10.9% of the single-threaded time for the Synthetic dataset, with the theoretical optima being 10% of the single-threaded time. The reduction in time was observed at the expected thread numbers, with no significant reduction when increasing the thread number from five to six (Figure 10).

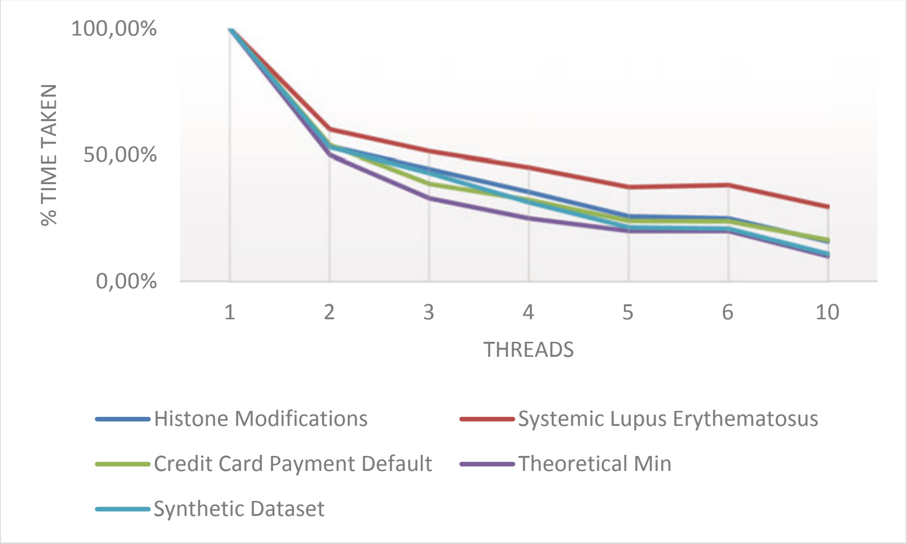


Figure 10 The time needed to complete a tenfold cross-validation (10CV) based on the number of threads. The reference is the time needed to run the test using only one thread (100%). The lowest line is the theoretical minimum time needed given the reference point.

Memory peaks increased by a factor of 6.9 for the Histone Modifications dataset, 6.1 for the credit card default dataset, 1.2 for the SLE dataset, and 4.0 for the Synthetic dataset as one thread was increased to ten. The increase was linear from one to five threads and from five to ten threads for two datasets, with the latter progression considerably lower than the first (Figure 11). The systemic erythematosus dataset led to almost no increase in memory peak usage while the Synthetic dataset had a consistent increase.

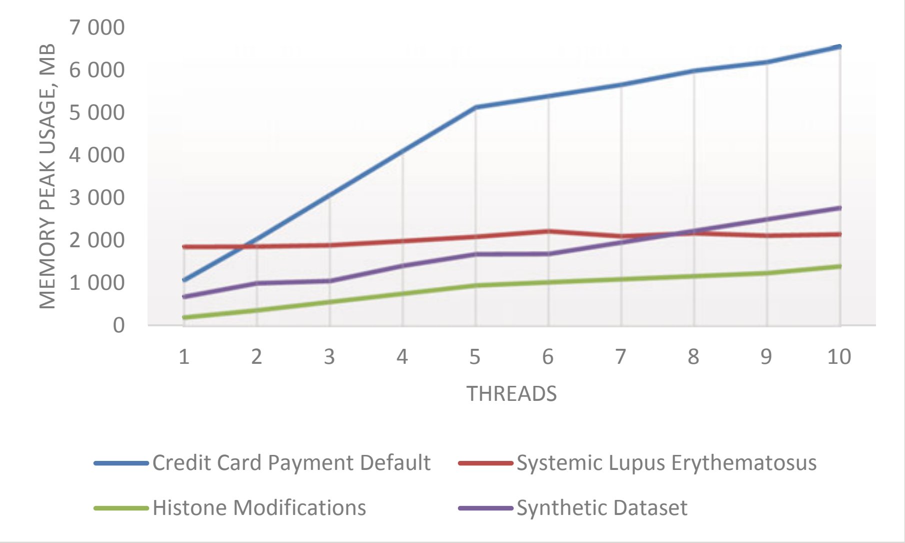


Figure 11 Memory peak usage in megabytes based on the number of threads when executing a ten-fold cross-validation (10CV) on different datasets.

Overall, the results show a significant gain in program speed, with lower than expected memory peak usage due to memory lane congestion. The Synthetic dataset approached the theoretical optimum with a linear memory increase, showing that the performance of the threaded classification was close to optimal when memory access was not a limiting factor. The end result is that ||-ROSETTA represents a useful tool in efficiently classifying modern quantities of data in a transparent fashion.

Conclusions

The creation of a screening model developed as a proof of concept on Swedish screening data and further refined on a Norwegian screening cohort was successful in stratifying individuals by risk. The final model had an accuracy of around 80% using primarily cytology tests, a type of test known for lacking predictive power[61]. The model identified 32.1% of cancer cases and 2% of control cases as high-risk histories. The identified control cases would likely benefit from an increased screening density. The data-driven scoring schema used was a functional approach to an unbiased screening model assessing purely on statistical merit, and its inclusion made the screening model internationally applicable by computing the score from local screening data instead of relying on established practices or experts.

The AS-SNP pipeline is not only suitable for discovering candidates for further analysis; it can be used to find likely biomarkers in screening-related predictions as well. The reduction of 17,293 AS-SNPs to 58 provided a sharp focus on the most relevant SNPs, a useful function given the vast amounts of genetic data produced even from a single sample. The AS-SNPs selected for further annotation were all associated on multiple levels with the traits from GWAS and eQTL SNPs they were linked to, identifying putative regulatory elements involved in the process as well.

The creation of ||-ROSETTA added efficiency needed for classification of the screening data, both Swedish and Norwegian. With more than 200 classifications to date on the screening data, the time per classification was reduced from around 3.5 hours per test to around 30 minutes per test.

The conclusions from these studies point to a considerable contribution towards predictive healthcare needs in the field of cervical cancer screening, ready to adopt new data from screening registries. Further, the models can filter through genetic data to find the markers that may indicate the likelihood of developing cancer, and extend to incorporate those into the predictive algorithms. These models are adjustable to any form of disease with a continuous development given that the relevant data is available.

Summary in Swedish

Användningen av maskininlärning (ML) inom vetenskap har ökat kraftigt det senaste decenniet och denna ökning har setts inte bara i de traditionella områdena som datavetenskap och matematik, utan även inom områden som biologi, kemi, och medicin. ML baseras på statistik och ger därmed bättre resultat ju mer data som finns tillgängligt. Det medicinska fältet har sedan länge samlat data från patienter via sjukhus, biobanker, och register, och mycket av denna data har aldrig tidigare analyserats med de metoder som ML erbjuder. Inom biologi har genetiken avancerat i takt med den datavetenskapliga utvecklingen, och idag finns enorma databaser med gendata tillgängliga för analys och vidareutveckling av metoder. I samma takt som genetiken har utvecklats har även kostnaderna för sekvensering minskat till den grad att genomdata nu används inom sjukvård och i mindre utsträckning inom personlig hälsovård. Inom sjukvården så används genomdata för att identifiera cancertyper, möjliga behandlingar, prognos, riskfaktorer, ärftliga sjukdomar, samt genterapi.

Denna avhandling adresserar tre olika aspekter av prediktiv sjukvård som kan förutspå risk för livmoderhalscancer och behandla patienter innan de utvecklar sjukdomen. Den första delen är en analys och klassifikation av diagnostiska data från nordiska screeningprogram i syfte att upptäcka de grupper som har en hög risk att utveckla livmoderhalscancer så att dessa kan få ett mer intensivt screeningschema och testas oftare för att bromsa eller förhindra en negativ utveckling. Den andra delen är en filtreringsmodell för genetiska varianter (SNPs) som har en stark koppling till reglerande proteiner som är involverade i olika sjukdomar. Dessa markörer kan användas antingen som riskmarkörer för klassifikation och prediktion eller som kandidater för att vidare undersöka de genetiska mekanismerna som ligger bakom många ärftliga sjukdomar. Den tredje delen är utvecklingen av ett program som heter ||-ROSETTA som snabbt kan klassificera stora mängder data och möjliggöra användningen av många olika algoritmer inom ML på ett förståeligt och transparent sätt.

||-ROSETTA är ett program som har lett till kraftigt ökad hastighet i beräkningarna för klassifikationen av screening data. Med över 200 klassifikationer gjorda så har tiden per klassifikation minskat från runt 3.5 timmar till ca 30 minuter. Denna ökning har möjliggjort en väsentligt större analys av data än vad som tidigare var möjligt.

Konceptmodellen för stratifiering av screeningbefolkningen på svenska data var lyckad och visade potentialen med att individualisera screeningprogrammet baserat på risk. Riskbedömningen inkluderade faktorer som inte fanns med i kliniska data såsom en individs oro för cancer eller en individs egen riskbedömning. Vidareutveckling och validering som gjordes med hjälp av data från det norska cancerregistret ledde till en markant förbättring av klassifikation och riskbedömningen tack vare utvecklingen av en data-driven modell för riskberäkningar. Denna modell gjorde även projektet oberoende av lokala experter, dvs läkare på lokala sjukhus, då cancerrisken från varje diagnos inte längre behöver specificeras utan kan beräknas utifrån de data som finns. Likheten i mönster mellan svenska och norska data indikerar att de kliniska och sociala faktorerna som påverkar risk för livmoderhalscancer är desamma i Norge och Sverige. De liknande resultaten visar att detta är en fungerande modell för att bygga ett individuellt screeningschema baserat på riskbedömningen av individen. Vidare betyder detta att om likheterna mellan Sverige och Norge beror på närheten så kan även länder utan ett etablerat screeningregister använda en riskbedömning från grannlandet om detta skulle ha ett screeningregister.

Filtrering av gendata i avsikt att identifiera mutationer som ger ökad risk för vissa sjukdomar har visat sig ha stor potential både utifrån allmäntillgängliga data och levervävnadsprover. De varianter som valts ut var alla associerade med de sjukdomar som pekats ut från GWAS och eQTL, två databaser varav den första listar genetiska varianter kopplade till sjukdomar och den andra listar varianter associerade till förändrad uttrycksnivå av proteiner. Användningen av s.k. ChIP teknologi visade inte bara aktiviteten hos inbindningsregionerna utan även mer specifikt den filtrerade aktiviteten i de regioner av DNA där inbindningen av proteiner var signifikant viktad och indikerade en funktionell förändring hos den genetiska varianten. Denna systematiska strategi för att hitta funktionella kandidater inom ärvda sjukdomar gjorde det möjligt att hitta varianter som påverkar sjukdomsprocessen, möjliga associationer mellan olika sjukdomar, den troliga effekten av en variants störning, och kandidatvarianter för att vidare utforska mekanismerna bakom dessa sjukdomar.

Tillsammans så skapar dessa resultat en funktionell modell att användas inom prediktiv hälsovård som effektivt kan förutspå riskerna för utvecklingen av sjukdom, i detta fall livmoderhalscancer, i ett tidigt stadium så att behandling och vård kan sättas in på ett sätt som minimerar både kortsiktiga och långsiktiga hälsorisker till en lägre kostnad. Framtida tillägg av ytterligare genetiska data kan bara förbättra resultaten.

Acknowledgements

The works herein would not be possible without the aid of my compatriots. I have been fortunate to find so many passionate minds and supportive hands in my life and in my studies. In matters grand to mundane, from cancer to coffee, I have been gifted with inspiring contacts to widen my horizons and challenge my preconceptions.

For the greatest inspiration in my life I have to thank my father and mother, Lars Baltzer and Inger Mattsby-Baltzer. They showed me from a young age that the intricacies of our world are infinitely exciting, and every small piece in it is seamlessly linked in a web from physics to biology, from philosophy to politics. I found this inspiration not in what they told me, but in their consistent actions of work, reflection, and pursuit of knowledge. For this I will be ever grateful.

I would like to offer my most sincere appreciation and gratitude to Jan Komorowski for planting me on the path of bioinformatics, a path I would never have discovered were it not for his enticing teaching abilities. I have learned much from him in the past five years, on matters from Machine Learning to sailing.

As a new student in a new job in a new field, Karin Sundström somehow managed to keep me afloat until I found my feet. For a Computer Scientist graduate most at home in the cold light of a screen at night, cancer, ethics, and biobanks, are not fields of expertise. But Karin somehow managed to make it so even though I had no experience of anything even associated to these. She also somehow managed to not yell at me even when I sent her 44 versions of my first poster for validation.

Of my colleagues there is much to say. Marcin Kruczyk encouraged me to continue with PhD studies. Conversations with Susanne Bornelöv flared what eventually became my interest in combinatorial mechanics. Husen Umer was inspiring in his determination, refusing to give up whether it came to ill-prepared skiing or slides that malfunctioned at the very last minute. Behrooz Torabi always got me psyched for morning meetings and tempting Zeeshan Khaliq with delicious treats during Ramadan was a devilish pleasure.

Klev Diamanti has been a perfect colleague with complementary skills and ever ready to discuss, develop, and debate, in the office and in the pub. Hopefully his child will inherit his C# skills. Mateusz Garbulowski and Karolina Smolinska-Garbulowska have been an excellent duo, both laughing dutifully at my hilarious jokes and ever ready to help. Mateusz has shouldered many a task because of my aversion to pro-active scheduling. Karolina deserves a special encomium for not only enduring my hubris but actively encouraging it. Conversations with Sara Younes have been a grand pleasure, each one turning from the mundane to the bizarre on the most pedantic of tangents. Fredrik Barrenäs has been a great colleague of sharp wit and debate, finding clever solutions to both statistical problems and long meetings. Marco Cavalli has been a grand co-author, with a tongue as sharp as his biology and ever on a deadline. Claes Wadelius has impressed every time with his unfathomable command of genetics, knowing something about every possible nucleotide sequence.

In Oslo I owe much to Jan Nygård who has helped me a great deal both with writing and understanding and his wife Mari Nygård for explaining to me so many practical details of screening. I would be remiss to not mention faster Bompa, who along with Arne and Martin Osvik managed to make me feel just as home in Oslo as in Uppsala.

In Stockholm I am grateful for the immense experience of Joakim Dillner, who showed me that you can finish a revision in less than a day if you know everything there is to know about your field.

References

1. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* 2016;375:1216–9.
2. Glossary of Terms. *Mach Learn* 1998;30:271–4.
3. Ahmad MA, Panicker NG, Rizvi TA, Mustafa F. Electrical detection and quantification of single and mixed DNA nucleotides in suspension. *Sci Rep* 2016;6:34016.
4. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature* 2015;526:68–74.
5. Liu L, Muralidhar S, Singh M, Sylvan C, Kalra IS, Quinn CT, Onyekwere OC, Pace BS. High-density SNP genotyping to define beta-globin locus haplotypes. *Blood Cells Mol Dis* 2009;42:16–24.
6. Eram SM, Azimifar B, Abolghasemi H, Foulady P, Lotfi V, Masrouri M, Hosseini M, Abdolhosseini A, Zeinali S. The IVS-II-1 (G → A) β0-Thalassemia Mutation in CIS with Hb A2-Troodos [δ116(G18)Arg → Cys (CGC → TGC)] Causes a Complex Prenatal Diagnosis in an Iranian Family. *Hemoglobin* 2005;29:289–92.
7. Norton HK, Phillips-Cremins JE. Crossed wires: 3D genome misfolding in human disease. *J Cell Biol* 2017;216:3441.
8. de Wit E, de Laat W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev* 2012;26:11–24.
9. Brambilla E, Gazdar A. Pathogenesis of lung cancer signalling pathways: roadmap for therapies. *Eur Respir J* 2009;33:1485–97.
10. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2013;42:D1001–6.
11. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447:661–78.
12. Jansen PR, Watanabe K, Stringer S, Skene N, Bryois J, Hammerschlag AR, de Leeuw CA, Benjamins J, Muñoz-Manchado AB, Nagel M, Savage JE, Tiemeier H, et al. Genome-wide Analysis of Insomnia (N=1,331,010) Identifies Novel Loci and Functional Pathways. *bioRxiv* 2018;214973.
13. Behjati S, Tarpey PS. What is next generation sequencing? *Arch Dis Child Educ Pract Ed* 2013;98:236–8.
14. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* 2007;316:1497.
15. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. *Mol Cell* 2015;58:586–97.

16. Rockman MV, Kruglyak L. Genetics of global gene expression. *Nat Rev Genet* 2006;7:862–72.
17. Cavalli M, Baltzer N, Umer HM, Grau J, Lemnian I, Pan G, Wallerman O, Spalinskas R, Sahlén P, Grosse I, Komorowski J, Wadelius C. Allele specific chromatin signals, 3D interactions, and motif predictions for immune and B cell related diseases. *Sci Rep* 2019;9:2695.
18. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. *Sci Adv* 2019;5:eaaw1668.
19. Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 2008;9:477–85.
20. Seifert M, Gohr A, Strickert M, Grosse I. Parsimonious Higher-Order Hidden Markov Models for Improved Array-CGH Analysis with Applications to Arabidopsis thaliana. *PLOS Comput Biol* 2012;8:e1002286.
21. Van Doorslaer K, Chen Z, Bernard H-U, Chan PKS, DeSalle R, Dillner J, Forslund O, Haga T, McBride AA, Villa LL, Burk RD, Consortium IR. ICTV Virus Taxonomy Profile: Papillomaviridae. *J Gen Virol* 2018;99:989–90.
22. de Villiers E-M, Fauquet C, Broker TR, Bernard H-U, zur Hausen H. Classification of papillomaviruses. *Virology* 2004;324:17–27.
23. Herbst LH, Lenz J, Van Doorslaer K, Chen Z, Stacy BA, Wellehan JFX, Manire CA, Burk RD. Genomic characterization of two novel reptilian papillomaviruses, Chelonia mydas papillomavirus 1 and Caretta caretta papillomavirus 1. *Virology* 2009;383:131–5.
24. Brianti P, De Flammineis E, Mercuri SR. Review of HPV-related diseases and cancers. *New Microbiol* 2017;40:80–5.
25. Kobayashi K, Hisamatsu K, Suzui N, Hara A, Tomita H, Miyazaki T. A Review of HPV-Related Head and Neck Cancer. *J Clin Med* 2018;7:241.
26. Stanley M. Immune responses to human papillomavirus. *Prev Cerv Cancer Hum Papillomavirus-Relat Dis Recent Adv Prophyl Vaccin* 2006;24:S16–22.
27. Muñoz N, Bosch FX, de Sanjosé S, Herrero R, Castellsagué X, Shah KV, Snijders PJF, Meijer CJLM. Epidemiologic Classification of Human Papillomavirus Types Associated with Cervical Cancer. *N Engl J Med* 2003;348:518–27.
28. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018;68:394–424.
29. Mirabello L, Yeager M, Yu K, Clifford GM, Xiao Y, Zhu B, Cullen M, Boland JF, Wentzensen N, Nelson CW, Raine-Bennett T, Chen Z, et al. HPV16 E7 Genetic Conservation Is Critical to Carcinogenesis. *Cell* 2017;170:1164-1174.e6.
30. Tokino T, Nakamura Y. The role of p53-target genes in human cancer. *Crit Rev Oncol Hematol* 2000;33:1–6.
31. DeFilippis RA, Goodwin EC, Wu L, DiMaio D. Endogenous human papillomavirus E6 and E7 proteins differentially regulate proliferation, senescence, and apoptosis in HeLa cervical carcinoma cells. *J Virol* 2003;77:1551–63.
32. Patel C, Brotherton JM, Pillsbury A, Jayasinghe S, Donovan B, Macartney K, Marshall H. The impact of 10 years of human papillomavirus (HPV) vaccination in Australia: what additional disease burden will a nonavalent vaccine prevent? *Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull* 2018;23:1700737.

33. Mühlberger N, Boskovic K, Krahn MD, Bremner KE, Oberaigner W, Klocker H, Horninger W, Sroczynski G, Siebert U. Benefits and harms of prostate cancer screening - predictions of the ONCOTYROL prostate cancer outcome and policy model. *BMC Public Health* 2017;17:596–596.
34. Shahyad S, Saadat SH, Hosseini-Zijoud S-M. The Clinical Efficacy of Prostate Cancer Screening in Worldwide and Iran: Narrative Review. *World J Oncol* 2018;9:5–12.
35. Autier P, Boniol M, Koechlin A, Pizot C, Boniol M. Effectiveness of and overdiagnosis from mammography screening in the Netherlands: population based study. *BMJ* 2017;359:j5224.
36. Peirson L, Fitzpatrick-Lewis D, Ciliska D, Warren R. Screening for cervical cancer: a systematic review and meta-analysis. *Syst Rev* 2013;2:35–35.
37. Cote RA. Architecture of SNOMED: Its Contribution to Medical Language Processing. *Proc Annu Symp Comput Appl Med Care* 1986;74–80.
38. International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). 2nd ed. Geneva: World Health Organization, 2004. 1200p
39. Szumilas M. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry J Acad Can Psychiatr Enfant Adolesc* 2010;19:227–9.
40. Cox DR, Hinkley DV. Theoretical statistics. Chapman and Hall/CRC, 1979.
41. Øhrn A, Komorowski J. Rosetta--a rough set toolkit for analysis of data. In: Proc. Third International Joint Conference on Information Sciences. Citeseer, 1997.
42. Anscombe FJ. Graphs in Statistical Analysis. *Am Stat* 1973;27:17–21.
43. Khaliq Z, Leijon M, Belák S, Komorowski J. A complete map of potential pathogenicity markers of avian influenza virus subtype H5 predicted from 11 expressed proteins. *BMC Microbiol* 2015;15:128.
44. Colussi D, Brandi G, Bazzoli F, Ricciardiello L. Molecular pathways involved in colorectal cancer: implications for disease behavior and prevention. *Int J Mol Sci* 2013;14:16365–85.
45. Andrae B, Kemetli L, Sparén P, Silfverdal L, Strander B, Ryd W, Dillner J, Törnberg S. Screening-Preventable Cervical Cancer Risks: Evidence From a Nationwide Audit in Sweden. *J Natl Cancer Inst* 2008;100:622–9.
46. Tung AKH. Rule-based Classification [Internet]. In: LIU L, ÖZSU MT, eds. Encyclopedia of Database Systems. Boston, MA: Springer US, 2009. 2459–62.Available from: https://doi.org/10.1007/978-0-387-39940-9_559
47. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2012;41:D991–5.
48. Younesy H, Möller T, Heravi-Moussavi A, Cheng JB, Costello JF, Lorincz MC, Karimi MM, Jones SJM. ALEA: a toolbox for allele-specific epigenomics analysis. *Bioinformatics* 2013;30:1172–4.
49. The ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57.
50. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* 2016;45:D896–901.

51. Lappalainen T, Sammeth M, Friedländer MR, ‘t Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 2013;501:506.

52. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 2012;22:1790–7.

53. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012;9:215.

54. Diamanti K, Umer HM, Kruczyk M, Dąbrowski MJ, Cavalli M, Wadelius C, Komorowski J. Maps of context-dependent putative regulatory regions and genomic signal interactions. *Nucleic Acids Res* 2016;44:9110–20.

55. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 2014;159:1665–80.

56. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, Kolpakov FA, Makeev VJ. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* 2017;46:D252–9.

57. Eggeling R, Grosse I, Grau J. InMoDe: tools for learning and visualizing intra-motif dependencies of DNA binding sites. *Bioinforma Oxf Engl* 2017;33:580–2.

58. Baltzer N, Sundström K, Nygård JF, Dillner J, Komorowski J. Risk stratification in cervical cancer screening by complete screening history: Applying bioinformatics to a general screening population. *Int J Cancer* 2017;141:200–9.

59. Schulte C, Tack G, Lagerkvist MZ. Modeling and programming with gecode. *Schulte Christ Tack Guido Lagerkvist Mikael* 2010;

60. Umer HM, Smolinska-Garbulowska K, Marzouka N, Khaliq Z, Wadelius C, Komorowski J. funMotifs: Tissue-specific transcription factor motifs. *bioRxiv* 2019;683722.

61. Dillner J, Rebolj M, Birembaut P, Petry K-U, Szarewski A, Munk C, de Sanjose S, Naucler P, Lloveras B, Kjaer S, Cuzick J, van Ballegooijen M, et al. Long term predictive values of cytology and human papillomavirus testing in cervical cancer screening: joint European cohort study. *BMJ* 2008;337:a1754.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1862

Editor: The Dean of the Faculty of Science and Technology

Distribution: publications.uu.se
urn:nbn:se:uu:diva-394293

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019