KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Messy data in life sciences

A discussion based on case studies

Winand Raf

Supervisor:
Prof. dr. ir. Jan Aerts
Prof. dr. ir. Yves Moreau, co-supervisor

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering Science

March 2016

Messy data in life sciences

A discussion based on case studies

Winand RAF

Examination committee:
Prof. dr. ir. Jean Berlamont, chair
Prof. dr. ir. Jan Aerts, supervisor
Prof. dr. ir. Yves Moreau, co-supervisor
Prof. dr. Pascal Borry
Prof. dr. ir. Rob Jelier
Prof. dr. Philippe Lemey
Prof. dr. ir. Joris Vermeesch

Dr. Kristien Hens
(University of Antwerp (BE))
Prof. dr. ir. Kathleen Marchal
(Ghent University (BE))

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering Science

March 2016

© 2016 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Winand Raf, Kasteelpark Arenberg 10 box 2446, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Preface

Het is nu ongeveer vijf jaar geleden dat ik in Lissabon mijn masterthesis aan het afwerken was en ik een vacature zag voor een PhD positie in de pas opgerichte groep van Jan Aerts. Ondertussen ben ik bij deze de laatste woorden aan het schrijven van mijn doctoraat. Voor ik reageerde op de vacature had iedereen me al gewaarschuwd dat een doctoraat nooit loopt zoals verwacht en dat je steeds ergens anders uitkomt dan waar je eerst dacht uit te komen. De voorbije jaren hebben inderdaad bewezen dat dit geen loze waarschuwing was en dat niet alles loopt zoals eerst verwacht. Aan de andere kant betekent dit ook dat ik de kans gekregen heb heel uiteenlopende studies te kunnen doen en dat ik enorm veel heb kunnen bijleren over verschillende domeinen. Het zijn jaren geweest met veel mooie en soms moeilijke momenten maar dit maakt ook deel uit van de reis die een doctoraat eigenlijk is.

Hieronder wil ik graag de mensen bedanken die me de voorbije jaren geholpen en bijgestaan hebben. Ik heb geprobeerd volledig te zijn maar ik wil bij voorbaat mijn verontschuldigingen aanbieden aan diegenen die ik vergeten zou zijn.

In de eerste plaats wil ik mijn promotor en co-promotor bedanken. Jan, bedankt om me de kans te geven om dit doctoraat te beginnen en af te ronden. Het is niet altijd makkelijk geweest maar na vele brainstorm sessies zijn we er wel uitgeraakt. Door in een groep van visualisatie mensen te zitten heb ik dingen kunnen leren die ik nooit gedacht had bij te leren toen ik begon. Yves, bedankt voor je hulp en opmerkingen op de momenten dat ik het echt nodig had.

Daarnaast wil ik de leden van mijn supervisory committee en examenjury, bestaande uit Jean Berlamont, Pascal Borry, Rob Jelier, Philippe Lemey, Joris Vermeesch, Kristien Hens en Kathleen Marchal, bedanken voor de tijd die ze genomen hebben om mijn thesis te beoordelen, voor hun opmerkingen en voor de samenwerking tijdens mijn doctoraat.

I also want to thank Annemie Vandamme, Ricardo Camacho, Ana Abecasis and Kristof Theys for giving me the opportunity to continue the work on

transmission of HIV-1 drug resistance that I started during my master's thesis. My gratitude also goes to Anne Rochtus and Benedetta Izzi for the great collaboration on the spina bifida paper. In general I would like to thank all co-authors that I worked with, your comments, advice, and support were greatly appreciated.

Ook wil ik een dankjewel richten aan de collega's in de toren voor de fijne werkomgeving en leuke lunches met hoogstaande conversaties. Bedankt Alejandro, Griet, Dusan, Jaak, Adam, Nico, Amin, Sarah, Georgios, Leo, Babis, Peter, Olivier, Gorana, Pooya, Xian, Jansi, Thomas, Daniel. In het bijzonder wil ik Toni bedanken voor al zijn hulp in de laatste fase van mijn doctoraat, om mij gezelschap te houden in de vroege uren op ESAT en om mij kennis te laten maken met het concept 'conservation of misery'. Ook een bedankt voor Marc en Arnaud om me te helpen met praktische tips ter voorbereiding van mijn verdediging. すべての私の可視化やその他の問題（私はGoogleが私にそれはあなたを失敗したと同じように失敗していない翻訳を願って）を私に同時に私と同じ旅を通過し、支援するため亮堺にも感謝

Natuurlijk wil ik ook de mensen bedanken die me door alle administratie hielpen, dus bij deze een dankjewel voor Elsy, Ida, John en Wim. Ook wil ik Maarten en Liesbeth bedanken voor de IT ondersteuning die regelmatig heel erg nodig was.

Ik zou ook nog graag mijn ouders willen bedanken om mij te steunen in al mijn keuzes en ook om mij de mogelijkheid gegeven te hebben om opnieuw te gaan studeren. Oma, bedankt om je bezorgdheid en betrokkenheid tijdens alle belangrijke momenten in mijn leven. Han en Nathalie, bedankt voor alles wat jullie gedaan hebben en om mij ook nog een huis te laten bouwen tijdens een doctoraat ;-). Bedankt ook aan de rest van mijn familie en vrienden. Ik vrees dat ik nooit echt duidelijk heb kunnen maken wat ik exact deed maar nu kunnen jullie dit manuscript lezen en alles zal duidelijk worden ;-).

Tot slot wil ik nog Stephanie en Casper bedanken. Casper om me te laten zien wat echt belangrijk is in het leven en om me te laten glimlachen zelfs wanneer het tegen zat. Stephanie, zonder jou had ik dit nooit gehaald, bedankt om mijn rots in de branding te zijn en me te steunen wanneer ik het het hardst nodig had.

Abstract

In the last decades we have witnessed an enormous increase in the amount of data being generated in every imaginable field. Where the bottleneck used to be the *creation* of the raw data it has now moved to the *analysis* of the data. Indeed, producing the raw data does not always take too much effort anymore, while extracting the relevant information contained in the data and drawing relevant conclusions can take an entire team of specialists in their own field. In this thesis we propose a conceptual framework that can be used to contemplate possible challenges that may arise during the analysis of data and more specifically biomedical data. The proposed framework consists of two dimensions: the amount of data in relevant dimensions and the messiness of the data.

While in general it is true that more data yields better models we discuss what 'more' data actually means and possible pitfalls of increasing the amount of data. For the second dimension in the framework, we consider data to be messy when it contains errors, violates statistical assumptions or is influenced by stochastic processes, non-linear interactions and feedback loops, environmental effects, human behavior, missing data, and temporal effects and when these factors can not be easily modeled, abstracted away, or are unknown. Studies with a large amount of data available and low messiness are more likely to yield accurate and reliable results while studies with a low amount of data and high messiness can lead to unexpected results and/or wrong predictions.

To illustrate the framework we discuss different case studies and where they are situated in the framework. These case studies include an analysis of transmission of HIV-1 drug resistance, the use of whole genome sequencing in the context of embryo selection, and epigenetic modifications associated with neural tube defects. In addition we also discuss the challenges faced in the analysis of personal health record data, and the development of a digital coach to help achieve sustainable weight loss when the data from these projects becomes available.

Beknopte samenvatting

In de voorbije jaren hebben we in zowat elke discipline de hoeveelheid gecreëerde data enorm zien toenemen. Waar tot voor kort de bottleneck lag in het *creëren* van de data verschuift deze nu naar de *analyse* ervan. We zien inderdaad dat het creëren van de data niet altijd zoveel moeite meer kost maar dat het extraheren van relevante informatie en het komen tot relevante conclusies een volledig team van specialisten kan vergen. In deze thesis stellen we een conceptueel raamwerk voor dat gebruikt kan worden om na te denken over mogelijke problemen die zich kunnen stellen bij het analyseren van de data en meer in het bijzonder biomedische data. Dit raamwerk bestaat uit twee dimensies: de hoeveelheid data in relevante dimensies en de 'messiness' van de data.

Hoewel we over het algemeen kunnen stellen dat meer data tot betere modellen leidt, bespreken we wat 'meer' data eigenlijk betekent en wat de mogelijke valkuilen zijn bij een toenemende hoeveelheid data. Voor de tweede dimensie in het raamwerk veronderstellen we data 'messy' te zijn wanneer ze fouten bevat, er niet wordt voldaan aan statistische assumpties of wanneer het beïnvloed wordt door stochastische processen, niet-lineare interactie en terugkoppelingen, omgevingseffecten, menselijk gedrag, ontbrekende data en temporele effecten en wanneer deze factoren niet kunnen gemodelleerd of geabstraheerd worden of wanneer ze onbekend zijn. Onderzoeken met veel beschikbare data en lage 'messiness', zullen eerder tot accurate en betrouwbare resultaten leiden terwijl studies met weinig data en hoge 'messiness' kunnen leiden tot onverwachte resultaten en/of verkeerde voorspellingen.

Ter illustratie bespreken we verschillende casussen en waar deze gesitueerd zijn in het raamwerk. Tot deze casussen behoren een analyse van de transmissie van geneesmiddelen resistentie in HIV-1, het gebruik van genoomsequenties in de context van embryo-selectie en de rol van epigenetische modificaties in neuralebuisdefecten. Daarnaast bespreken we ook de mogelijke uitdagingen in twee projecten aangaande de analyse van persoonlijke medische dossiers en de ontwikkeling van een digitale coach die kan bijdragen tot blijvend gewichtsverlies.

Acronyms

1kG 1000 Genomes project.

AIDS acquired immune deficiency syndrome.

ATP adenosine triphosphate.

cDNA complementary DNA.

CG Complete Genomics.

CI confidence interval.

CpG cytosine-guanine pair.

ddNTP dideoxynucleoside triphosphate analog.

DN drug-naive individuals.

DNA deoxyribonucleic acid.

dNTP deoxynucleoside triphosphate analog.

DRM drug resistance mutation.

FDA food and drug administration.

GWAS genome-wide association study.

HAART highly active antiretroviral therapy.

HGMD Human Gene Mutation Database.

HIV human immunodeficiency virus.

HOX homeobox.

IVF in vitro fertilization.

MAF minor allele frequency.

MAR missing at random.

MCAR missing completely at random.

MMC myelomeningocele.

MNAR missing not at random.

MO morpholino.

mRNA messenger RNA.

NGS next-generation sequencing.

NNRTI non-nucleoside reverse transcriptase inhibitor.

NRTI nucleoside/nucleotide reverse transcriptase inhibitor.

NTD neural tube defect.

OMIM online Mendelian Inheritance in Man.

PCA principal component analysis.

PCR polymerase chain reaction.

PGD preimplantation genetic diagnosis.

PGS preimplantation genetic screening.

PI protease inhibitor.

PP precautionary principle.

RNA ribonucleic acid.

RT reverse transcriptase.

SARS severe acute respiratory syndrome.

SDRM surveillance drug resistance mutation.

SNP single-nucleotide polymorphism.

TDR transmitted drug resistance.

TR patients failing treatment.

UTR untranslated region.

VL viral load.

WGS whole genome sequencing.

WISH whole mount in situ hybridization.

Contents

Abstract iii

Contents xi

List of Figures xv

List of Tables xix

1 Introduction 1

1.1 Pitfalls of data analysis 2

1.2 Big data 4

1.3 Ethical, legal and societal issues 8

1.3.1 Precaution 8

1.3.2 Privacy 10

1.3.3 Incidental findings 11

1.3.4 Predictions 12

1.4 Primers 12

1.4.1 HIV 13

1.4.1.1 HIV and AIDS 13

1.4.1.2 HIV Evolution 14

1.4.1.3 Antiretroviral drugs . . . 15
1.4.1.4 Drug resistance . . . 18
1.4.2 Next-generation Sequencing . . . 19
1.4.2.1 Medical genetics . . . 19
1.4.2.2 Secondary use . . . 24
1.4.3 Epigenetics . . . 25
1.4.3.1 Methylation . . . 26
1.4.3.2 Association studies . . . 26
1.4.3.3 Technology . . . 27
1.5 Chapter-by-chapter overview . . . 27

2 HIV-1 Transmitted Drug Resistance 29
2.1 Introduction . . . 29
2.2 Methods . . . 31
2.3 Results . . . 32
2.4 Discussion . . . 38

3 Embryo screening 47
3.1 Introduction . . . 48
3.2 Materials and Methods . . . 49
3.2.1 Disease selection . . . 49
3.2.2 Samples . . . 49
3.2.3 Transcripts . . . 50
3.2.4 Mutations predicted to be damaging . . . 50
3.2.5 Mutations described to be damaging in literature . . . 50
3.3 Results . . . 51
3.3.1 Autosomal dominant disorders: mutations predicted to be damaging . . . 51

3.3.2 Autosomal dominant disorders: mutations present in HGMD 54
3.3.3 Autosomal recessive disorders: mutations predicted to be damaging . . . 55
3.3.4 Autosomal recessive disorders: mutations present in HGMD 56
3.4 Discussion . . . 58
3.4.1 Analytical validity . . . 58
3.4.2 Clinical validity . . . 58
3.4.3 Clinical utility . . . 60
3.5 Conclusion and final remarks . . . 60
3.6 Supplementary data . . . 61
3.6.1 Prediction Algorithms . . . 61
3.6.2 Quality . . . 61

4 Epigenetics - Neural Tube Defects 83
4.1 Introduction . . . 83
4.2 Materials and methods . . . 85
4.3 Results . . . 88
4.4 Discussion . . . 99
4.5 Conclusion . . . 102
4.6 Supplementary data . . . 102

5 Messy data 111
5.1 Messiness and data size . . . 111
5.1.1 Data volume and dimensions . . . 111
5.1.1.1 Power analysis . . . 113
5.1.2 Messiness . . . 114
5.1.2.1 Errors . . . 115
5.1.2.2 Statistical assumptions . . . 116

5.1.2.3 Stochastic processes 118
5.1.2.4 Non-linear relationships 119
5.1.2.5 Environmental effects 121
5.1.2.6 Temporal effects 121
5.1.2.7 Missing data 122
5.1.2.8 Human behavior 123
5.2 Framework 125
5.2.1 HIV 128
5.2.2 Embryo selection 129
5.2.3 Neural tube defects 132
5.2.4 b-SLIM - MyHealthData 134
5.2.4.1 b-SLIM 134
5.2.4.2 MyHealthData 135
5.2.5 Visual analytics 136
5.2.6 Summary 139

6 General Conclusions **142**
6.1 Discussion 142
6.2 Future research 147
6.3 Conclusion 149

Bibliography **151**

List of publications **189**

List of Figures

1.1 Spurious correlation 4
1.2 Lie factor 5
1.3 HIV-1 genomic structure 14
1.4 HIV Lifecycle 15
1.5 Sequencing work flow 21
1.6 Emulsion PCR 22
1.7 Solid-phase amplification 22
1.8 GWAS Diagram 24

2.1 HIV-1 PI-NRTI SDRM Robust regression model 37
2.2 HIV-1 NNRTI SDRM Robust regression model 38
2.3 HIV-1 PI-NRTI SDRM Robust regression model (normalized) . 39
2.4 HIV-1 NNRTI SDRM Robust regression model (normalized) . . 40
2.5 HIV-1 PI-NRTI SDRM Robust regression model (DE) 41
2.6 HIV-1 NNRTI SDRM Robust regression model (DE) 42
2.7 Transmission ratio vs. median viral load 43

3.1 Histogram: predicted damaging mutations (dominant) 52
3.2 Histogram: predicted damaging mutations (MAF <1%, dominant) 53

3.3 Histogram: predicted damaging mutations (recessive) 56
3.4 Histogram: predicted damaging mutations (MAF <1%, recessive) 57

4.1 *HOXB7* methylation studies by Sequenom EpiTYPER in MMC patients. 90
4.2 *HOXB7* methylation studies by Sequenom EpiTYPER in pairs of unaffected siblings versus MMC patients. 98
4.3 Phenotype analysis of Hoxb7a-overexpression in zebrafish embryos. 100
4.S1 Genomic organization and expression patterns of *HOX* genes for humans and zebrafish. 106
4.S2 Genomic context of CpG methylation by HumanMethylation 450K BeadChip. 107
4.S3 *HOXB7* methylation is not significantly different between 70 Caucasian and 10 non-Caucasian MMC patients. 108
4.S4 *HOXB7* cg06493080 methylation versus expression in normal tissue extracted from brain and blood. 109
4.S5 Phenotype analysis of Hoxb7a depletion in zebrafish embryos. . 110

5.1 Model generation 112
5.2 Number of people infected with Ebola 118
5.3 Stochastic Resonance 120
5.4 Framework graph 126
5.5 Power analysis zone 127
5.6 HIV study in the framework 129
5.7 Embryo study in the framework 131
5.8 NTDs study in the framework 133
5.9 b-SLIM and MyHealthData in the framework 137
5.10 Visual analytics process 138
5.11 Overview of all projects in the framework 140

6.1 Robustness of models 146

List of Tables

2.1 Identified PI-SDRMs 33

2.2 Identified NRTI-SDRMs 34

2.3 Identified NNRTI-SDRMs 35

2.4 Outlier SDRMs 36

3.S1 List of dominant disorders 63

3.S2 List of recessive disorders 66

3.S3 List of Coriel IDs 71

3.S4 Number of damaging mutations (autosomal dominant) 71

3.S5 Number of damaging mutations (autosomal recessive) 72

3.S6 List of genes with mutations predicted to be damaging (dominant) 73

3.S7 Description of genes with the highest number of damaging mutations (dominant) 74

3.S8 List of genes with mutations predicted to be disease causing (dominant) 76

3.S9 List of genes with mutations predicted to be damaging (recessive) 77

3.S10 Description of genes with the highest number of damaging mutations (recessive) 78

3.S11 List of genes with mutations predicted to be damaging (compound heterozygous) 80

3.S12 List of genes with mutations predicted to be disease causing (recessive) 81

3.S13 List of genes with mutations predicted to be disease causing (compound heterozygous) 82

4.1 Background information of MMC patients 91

4.2 Methylation of the HOX genes 95

4.S1 Number of probes for the 4 HOX clusters included in the HumanMethylation 450K BeadChip platform. 103

4.S2 *HOXB7* methylation by Sequenom EpiTYPER analysis for single CpG units within the different cohorts. 104

4.S3 *HOXB7* methylation by Sequenom EpiTYPER for the 12 pairs of MMC patients and their unaffected siblings. 105

5.1 Overview of factors influencing messiness 141

6.1 HGMD Statistics 144

Chapter 1

Introduction

The last few decades have seen an enormous increase in the amount of data that is being generated and collected. In almost all sectors data from any imaginable source is collected and stored for further analysis. Think about going to the supermarket where all your purchases are stored in a central data store based on your customer loyalty card. On the road, traffic movements are monitored and stored as they do in the Netherlands with the National Data Warehouse (Viti et al., 2008). Also online, an enormous amount of data is being collected. For instance Google collects everything you search for, videos you watch, information on devices you use, etc. (Google Inc., 2015) and Facebook has been in the news lately for tracking everyone who visits sites with a 'Like'-button and has been sued by the Belgian privacy commission for these practices (Clapson, 2015).

In digital healthcare and biomedical research we have also seen a tremendous evolution. New techniques have been developed that allow us to generate more data than ever before. The first human genome was sequenced in the Human Genome Project and was completed in April 2003 with a total cost of $2.7 billion (National Human Genome Research Institute, 2010). In January 2014 Illumina announced a new sequencer making it possible to sequence a genome for less than $1,000 with the Illumina HiSeqX (Illumina Inc., 2015). Not only is more data generated, we can also generate it faster than ever before. While the first human genome was sequenced in about 13 years by a consortium of 20 institutions, with the Illumina HiSeqX 10, 18,000 human genomes can be sequenced per year or just over 340 per week. But next-generation sequencing (NGS) is not the only technology that has seen a rapid evolution. In epigenetics research the initial DNA methylation profiling by using gel electrophoresis has been replaced by array-based and sequence-based technologies, reducing

cost and allowing single-base pair resolutions and faster analysis (Laird, 2010). Another example of a technology that outputs vast amounts of data is mass spectrometry imaging. With this technology it is possible to create a complete spectrogram for each cell and for each slice of a sample, generating an output of several gigabytes for a typical experiment.

In addition to all this data being generated by corporations and researchers all over the world, also laymen have now started to obtain and record their own data. With the $1,000 genome, getting your own genome sequenced starts to get within reach for many people. While a whole genome sequence spans all 3 billion base pairs, it is already possible to get information for ∼600,000 base pairs for $99 from commercial companies like 23andMe (23andMe Inc, 2014). Other data is being recorded as well with a large increase in tracking technologies in recent years allowing users to monitor (some of) their own health parameters. Sensors like the fitbit and the announced angel sensor allow users to monitor heart rates, activity, body temperature and even blood oxygen level (Fitbit Inc., 2015; Seraphim Sense Ltd., 2015). Even when people are not actively monitoring their own health, electronic health records kept by general practitioners contain lots of data about every individual.

The biggest problem with many of these advances and collected data is that all that generated data also has to be analyzed in order to extract correct and relevant information. So while it is technically possible and not too hard anymore to produce these vasts amounts of data, sometimes called a tsunami of data, the challenge now lies in interpreting and analyzing the data. A rather well known statement from the field of genetics that relates this message is from Mardis (2008): “The $1,000 genome, the $100,000 analysis?”. Also Laird (2010) says that “The bottleneck in DNA methylation advances will increasingly shift from data production to data analysis”. In mass spectrometry imaging some of the more recent experiments produce more than a terabyte of complex and high-dimensional data that is impossible to analyze without the development of new computational methods (Verbeeck, 2014). Indeed, producing the raw data does not always take too much effort anymore, while extracting the relevant information contained in the data and drawing relevant conclusions can take an entire team of specialists in their own field.

1.1 Pitfalls of data analysis

Before big data and some possible challenges that are found in big data analysis will be discussed, it is important to know some general pitfalls that apply to data analysis in general. There are in general three classes of statistical pitfalls,

each containing several elements: sources of bias, errors in methodology and interpretation of results (Helberg, 1995).

Sources of bias can be introduced during sampling or by violating statistical assumptions. When predictions are made for the entire population it is imperative that the sample these predictions are based on is representative for that population. If one were for instance to conduct an online survey one would exclude all people without access to the internet or a lack of interest in filling out surveys. Conclusions from this kind of survey on a population level are therefore likely to be biased. As will be explained in Section 5.1.2.2 the violation of statistical assumptions can also have an influence on the model. Most models assume independence of the data points but as will be explained this is not always the case and can have large consequences.

Also the methodology can be a pitfall in a statistical analysis. A first element in the methodology is the power of the statistical test. The power of a statistical test indicates the probability of rejecting the null hypothesis when it really is false or in other words, the probability of detecting a real difference. To know how many samples are needed to detect an effect of a given size or how large the effect is that is detectable given a sample size, a power analysis can be conducted (UCLA: Statistical Consulting Group, 2015). Too little power means running the risk of overlooking a real effect that is there. With too many samples on the other hand, any difference will become statistically significant while the effect size might be too small to be of any practical use (Helberg, 1995). A second element is the problem of multiple comparisons because when a large number of statistical tests is performed on a data set, the chance of finding a spurious correlation increases. Finally, most statistical methods assume an error-free measurement and this is not always the case and might have an effect on the conclusion of the analysis.

The last pitfall lies in the interpretation of results. One of the concepts that is frequently misunderstood in statistics is the 'significance'. Something that is statistically significant might not be of any practical significance as mentioned in the previous paragraph. However, many people still consider a very low p-value to be equal to a larger effect. This is why Cohen (1994) proposes to report effect sizes so that it becomes easier to interpret the effect of what was discovered. Something else that is often misunderstood is the difference between precision and accuracy. A result can have a precision of many digits after the comma while accuracy says how reliable the result actually is. A last concept that has resulted in much confusion and misunderstandings is the concept of causality. When a statistical test reveals a correlation between two variables it does not mean that one causes the other, only that they are correlated. Only when the predictor variables are assigned by the experimenter can a causal relationship be identified (Helberg, 1995). Some nice examples of spurious correlations that

also show the danger of treating correlation as causation are given by Vigen (2015a), an example of which can be seen in Figure 1.1.

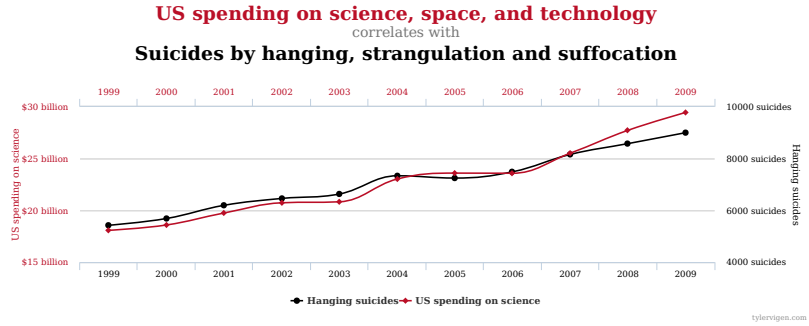


Figure 1.1: Example of a spurious correlation showing the correlation (r = 0.99789) between the US spending on science, space and technology and suicides by hanging, strangulation and suffocation. From Vigen (2015b)

The interpretation of results can also be affected by the visualizations used to display those results. A well known example of this is given by Tufte (1991) in his book 'The Visual Display of Quantitative Information' where he introduces the 'lie factor'. This factor is based on the principle that "The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the quantities represented." The lie factor then becomes the effect size shown in the graph divided by the real effect size in the data. In Figure 1.2 an example can be seen with a lie factor of 14.8 as the lines that represent the difference between 18 and 27.5 are respectively 0.6 inches and 5.3 inches long which is 14.8 times larger than the actual difference.

1.2 Big data

Now that some general pitfalls of data analysis have been introduced we can return to the explosive increase in amounts of data described in the first part of the introduction and identify some challenges specific to that field. The term 'big data' is often used to refer to these large amounts of data that are being created and although the term is used very often, there is no clear definition that is generally used. Most often it is defined to be characterized by a number of Vs with the amount of Vs depending on the source (Hashem et al., 2015; Demchenko et al., 2013; Tsai & Bell, 2015). Below the five most commonly used Vs to describe big data are explained. The first three are shared by all definitions while the last two are added in other definitions.

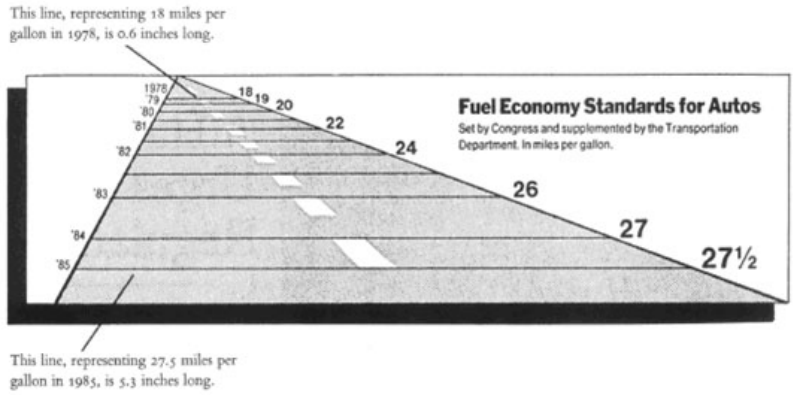


Figure 1.2: An example of the lie factor. As the lines that represent the difference between 18 and 27.5 are respectively 0.6 inches and 5.3 inches long this graph has a lie factor of 14.8. From Tufte (1991)

1. **Volume:**
   This is related to the vast amounts of data. There is however some discussion on how much data actually is needed to call it 'big' data. While some argue that several terabytes or petabytes of data have to be considered big data others say it is hard putting a number on it as what is considered big today might be considered normal tomorrow (Villanova University, 2015). In many cases it means that the data cannot be handled or processed in a straightforward way (Fisher et al., 2012).

2. **Velocity:**
   Velocity refers to the speed at which new data is generated.

3. **Variety:**
   Variety refers to the complexity and different types of data that are collected. The data being collected from many different sources amounts to an enormous variety of structured and unstructured data that is collected.

4. **Veracity:**
   Not all data that is being collected is of high quality. Veracity refers to the fact that data can of varying quality. Especially unstructured data can be of questionable quality.

5. **Value:**
   The value of big data lies in the ability to extract valuable information and conclusions from that data. For example, should company X buy company Y based on a social network analysis?

Although these characteristics describe big data, data sets that do not fit all five of them can still pose the computational challenges that are generally associated with big data analysis some of which will be explained below. Offline analysis of extremely large data sets might not show a high velocity but analyzing this data still poses significant challenges. On the other hand, some data sets might not have such a high volume while the complexity and quality of the data form the obstacles.

One of the big advantages of big data is that it contains much more information compared to classical experiments. Even relatively rare events can be detected as the number of samples is very high. On the other hand, big data comes with a number of computational and statistical challenges such as for example scalability, storage, noise accumulation and spurious correlations (Fan et al., 2014).

The first challenge lies in the fact that many algorithms and techniques developed for data analysis *do not scale well* as they are mostly designed with the assumption that all data will be loaded in memory on a single machine (Tsai et al., 2015). When the amount of data and/or dimensionality become too large these algorithms start to break down. In this case a dimensionality reduction by selecting only the most informative dimensions or by applying statistical procedures such as principal component analysis (PCA) can sometimes make the data analyzable again by standard techniques. Also, reducing the amount of data points by pre-processing the data can lower hardware and memory requirements for the analysis. An example of this is the khmer software that is used for metagenomic analysis and which trims, discards and bins genomic reads so that redundant and erroneous reads are removed from the original data set making it much smaller (McDonald & Brown, 2013).

Storing massive amounts of data can also become a very costly and complex bottleneck in big data analysis as more and more experiments are conducted leading to ever increasing amounts of data. The large hadron collider computing grid is composed of 132,994 physical CPUs, 553,611 logical CPUs, 300 petabytes of online storage and 230 petabytes of magnetic tape storage (Mearian, 2015). Also the data warehouse at Facebook stores an impressive amount of data with an available storage of up to 300 petabytes growing with a daily rate of 600 terabytes (Vagata & Wilfong, 2014). Another example of how much data has to be stored can be found in healthcare. In 2011 the total amount of data generated by healthcare organizations was estimated to be 150 exabytes (Hughes, 2011). It was also expected to increase with between 1.2 and 2.4 exabytes per year from then on. One way to cope with the enormous amounts of data being generated is by discarding the data that is not needed as soon as possible. The sensors in the large hadron collider for instance produce approximately 1 petabyte of data each second which is too much to handle for any computer system

available. Therefore an electronic pre-selection is performed only retaining 1 in 10,000 data points while subsequently only 1% of those remaining points are further selected for analysis, still amounting to more than 25 petabytes per year (CERN, 2015). This strategy to limit the amount of storage needed can also be applied in other fields and for other technologies. For example, in NGS where the sequencing cost keeps dropping, the infrastructure to keep all raw read information as well as any intermediate analysis results might become too complex and costly compared to resequencing if the original data would ever be needed again.

Noise accumulation in higher dimensions is a third challenge of big data analysis. When predictions are made from a large number of parameters and not all these parameters are informative, these non-informative parameters only increase the noise of the prediction (Fan et al., 2014). Therefore it becomes very important to select the right features to perform an analysis. Related to this noise accumulation is *error propagation* in higher dimensions (Taleb, 2015). Small measurement errors in some dimensions can accumulate in very large errors when the whole model is considered.

Another challenge that can be encountered when analyzing a very high dimensional data set is that some random unrelated variables might be found to be correlated during the analysis. These *spurious correlations* pose a major challenge as they can lead to wrong statistical inference or the wrong variables being retained during variable selection. This problem was also mentioned in the previous section but it becomes more important here as the chance of finding spurious correlations increases with the dimensionality of the data set.

An additional practical problem that might occur is the fact that there are *multiple standards* to store data. A lot of work can go into making data sets ready for analysis by combining them and converting them into a suitable format (Fisher et al., 2012). Results from a big data analysis can also be *less robust* than initially expected or can be gamed by users (e.g. Google bombs).

These and many more examples have lead to a number of critiques on the analysis of big data that try to warn people not to place all their faith into this science as a lot of challenges still have to be overcome and for some issues big data might not bring the answers we are looking for (Boyd & Crawford, 2012; Marcus & Davis, 2014; Harford, 2014; O'Neil, 2015)

1.3 Ethical, legal and societal issues

Up to now we only discussed the possible challenges that can arise when analyzing data but not *why* it should be analyzed or what the results of the analysis may lead to. In life sciences, researchers often analyze data and create algorithms with the goal of directly or indirectly improving the life and health of people. As noble as this may seem, it is important to keep in mind some ethical, legal and societal issues as not everyones' motives may be so innocent and technology could also potentially cause harm to people. In history, there are quite a few examples where technology that was first developed to do good and to help people was later used for more sinister purposes. This is the so-called *dual use* concept which is defined as "goods, software and technology that can be used for both civilian and military applications and/or can contribute to the proliferation of Weapons of Mass Destruction (WMD)" by the European Commission (2015). One of the most famous examples of dual use technology is the invention of dynamite by Alfred Noble who made handling of the otherwise extremely unstable nitroglycerin much safer thus leading to less deaths and making previously impossible constructions possible. The dark side of that invention was that in the late 19^{th} century, dynamite sparked a wave of terrorism by anarchists, killing many people (Jensen, 2004). Also more recently, when the 3D-printer became widely available, the technology gave access to customized implants and other medical devices but was also used to print parts for a gun (Greenberg, 2013).

In the next paragraphs a brief overview will be given of some issues that arise when using biomedical data for research and specifically with the data used in this thesis.

1.3.1 Precaution

While we are working with biomedical data we always have to ask ourselves why we are doing this and whether or not we are doing enough to protect the people participating in the study and the general public from adverse consequences. An example of research that was performed with the best intentions but that could lead to possible abuse can be found not so long ago when researchers studying the transmission of the H5N1 influenza strain from birds to humans developed a way to make the strain more easily transmittable between mammals. Before they published their paper, they were asked to remove all references to the methodology used to create this new strain as it was thought that this knowledge could be used to do harm (NIH, 2011). In the 1980's the German government introduced a philosophy in their law making called 'vorsorgeprinzip' which

was later translated as the precautionary principle (PP) (Jordan & O'Riordan, 2004). Although governments had already implemented legislation based on the same underlying elements, an exact definition of 'precaution' was not available. A much used definition of the PP can be found in the 1992 Rio Declaration on Environment and Development: "Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation." (Ellis et al., 2006). Another definition that covers not only environmental risks is given by Taleb et al. (2014) who state that "if an action or policy has a suspected risk of causing severe harm to the public domain (affecting general health or the environment globally), the action should not be taken in the absence of scientific near-certainty about its safety". As will be discussed later, these are the events that are located in the tails of a fat tail distribution and as they are not locally confined, they can cause catastrophic failure on a global scale. If we look back at the example of the H5N1 strain, we can wonder what would happen if this strain would escape the laboratory by accident and turn out to be highly lethal. Although the chance of the virus escaping might be very low, if the result would be a decimation of the human population the question should not be how likely it is that the virus escapes but whether or not the research should be done at all. First of all, debating the chance of the virus escaping and putting a number on that risk is extremely hard. Second, as long as the chance of the virus escaping is non-zero the question is not *if* the virus will escape but rather *when* it will escape thereby possibly killing billions of people. The PP has already been implemented in many jurisdictions worldwide covering legislation on a wide variety of topics including mobile phone regulations, fishery, food safety, etc. (Fisher, Elizabeth Charlotte Jones & von Schomberg, 2006).

Also on a smaller scale the underlying elements of the PP can play a role as we all have an obligation to the environment and the people around us. In some cases it might be clear what precautions have to be taken but as we will discuss in Chapter 5 things are not always clear and can be messier than expected. The question then becomes just how precautious we have to be to remain feasible. For example, there is a clear risk for congential defects in the unborn child when taking isotretinoin during a pregnancy and several guidelines exist to prevent exposure to this drug during pregnancy (Choi et al., 2013). Also for Toxoplasma infections several risk factors are identified (Baril et al., 1999) which can relatively easily be avoided. On the other hand, many more risk factors are constantly being discovered for a variety of conditions. For example air pollution is associated with a higher risk of developing autism (Raz et al., 2014) or exposure to dust mite allergens increases the risk of developing atopic dermatitis (Hagendorens et al., 2004). When all possible risk factors have to be considered, being precautious becomes near to impossible. So while in the

case of global ruin as described by Taleb et al. (2014) the PP certainly has to be applied, it is impossible for both governments and individuals to completely prevent exposure to risk factors.

In Chapter 3 a large part of the study deals with the ethical questions arising from the use of NGS as a technique for embryo selection. There we discuss what a selection of the 'best' embryo would mean in that context and what the problems associated with determining the 'best' embryo are. The discussion on so-called 'new eugenics' is very much alive with many scientists debating the necessity of defending genetic diversity versus the elimination of certain variants causing disability for instance (Sparrow, 2015; Mertes & Hens, 2015; Garland-Thomson, 2012). In light of the previous paragraphs one would have to wonder what would happen if we were able to proof that current technology would be able to perfectly predict human phenotypes. This is certainly not evident as in that case the technology could undoubtedly be used to improve people's lives but on the other hand could also be used to remove certain genetic variants from the population deemed to be unwanted by whoever decides on this. Applying the PP to this problem tells us that we should not start removing variants from the population or 'enhancing' the population as we cannot say with near-certainty that there will be no adverse consequences for the human race in the future. If, for instance, we were to find in the future that a certain variant now deemed damaging is actually crucial to the survival of the human race at some point, we would have already sealed its fate.

1.3.2 Privacy

Working with NGS data gives the advantage of instantly having access to the entire genome of an individual but comes with the disadvantage that this also means it contains confidential information about e.g. susceptibility to disease and carrier status without that information being necessarily related to an original clinical question. Although this information could be used to help people live longer and healthier lives by possibly detecting disease in an early and treatable stage, it can also be used by insurance companies for instance to refuse people with a certain genotype or even by employers who do not want to hire people with an increased risk of becoming sick. If this personal information would be made public it can also lead to wider discrimination where certain individuals or populations might be considered 'less' than others just because of their genotype. Therefore, keeping biomedical data secure and removing possible identifiers can prevent unauthorized use and consequent abuse of this information by third parties. As a side note is has to be mentioned however that removing identifiers in case of genomic sequences can be hard as the genomic

sequence itself is already a unique identifier of an individual. At this moment it remains rather difficult - although certainly not impossible (Gymrek et al., 2013) - for any person to link a genomic sequence with an individual so we should think about the possible consequences of genomes becoming publicly available and ensure that appropriate measures are taken to protect sequences being used in research.

1.3.3 Incidental findings

Even when the genomic information is secure, a researcher working with NGS data can stumble upon certain variants that may not be of interest for the study but that might be of interest to the individual being sequenced. These so-called incidental findings raise another question and this is whether or not to report these findings back. This has to be done taking into account whether the sample was collected for research purposes or for clinical purposes, where feedback of incidental findings in a research setting is much more ethically debated (Hallowell et al., 2015). Even without taking the setting into account, for some mutations reporting incidental findings might seem straightforward, especially when these mutations are clinically actionable and pose a severe risk to the individual, such as BRCA1/2 mutations for instance. On the other hand, some people might not want to know their genetic status for some diseases and indeed, people also have the right 'not to know' (Andorno, 2004). One way of avoiding this problem is by not looking at parts of the genome outside the parts needed for the current research project. In the other case, when mutations are found when looking at the entire genome, a system of 'binning' could be used that divides the possible findings into several categories ranging from 'clinically actionable' over 'low-medium high risk factors' to 'unknown implications' (Berg et al., 2011). Depending on the bin, the participant would either be contacted or not. This also means that an informed consent would have to be obtained from all participants and the incidental findings have to be returned to those participants in an appropriate way which can cost a great deal of time and effort (Mayer et al., 2011; Ormond et al., 2010).

These questions become even harder when children or fetuses are the ones being tested. They also have the right to privacy and the right not to know, but testing them will give the parents access to all information which they can decide to use or share at their own discretion. Surely one would not be opposed to revealing genetic information to the parents if it concerns a serious early onset disease that is treatable but how do we define a 'serious' disease and where do we draw the line on which diseases to report and which ones not? Even the definition of 'disease' can be open for discussion as for instance some would consider autism a disease while others would not. Also, some people

would not want a child born with Down Syndrome while others would welcome it regardless. Additional questions are raised when the information concerns an increased risk for a disease instead of a certainty. Would the parents need to be informed in that case? If so, would it be possible that this leads to increased stress for both parents and children as they might always assume the worst when possible symptoms of that disease surface even when caused by an unrelated benign infection? There are many more questions that are equally difficult that need to be answered by people holding very different views concerning these issues. For instance, while some argue that information should be disclosed only when it directly benefits the children themselves at an early age (Institute of Medicine, 1994), others advocate also testing for late onset diseases as early as possible in order to inform the children (Malpas, 2006). These views are very far apart as in the former case only very limited genetic information would be disclosed while in the latter there is full disclosure of any genetic risk even if it involves untreatable diseases. In any case, coming to a mutual agreement about these issues will be challenging and probably will never be reached.

1.3.4 Predictions

In case of the development of prediction algorithms there is a problem with false positives and false negatives. In case of false negatives, the algorithm might predict that someone would not benefit from a certain therapy while in reality they do; or that someone needs to change their eating habits to loose weight while in reality they are already losing weight. With false positive predictions the opposite scenario presents itself, i.e. predicting a beneficial effect of a therapy while it is not and saying the eating habits are OK while in reality they are not. In both cases the question has to be asked what the cost of each wrong prediction will be and weigh that cost against the possible benefit of a correct prediction. Giving a wrong prediction that causes stress and maybe unnecessary treatment can be as bad as giving a treatment that is not going to work while another one might.

1.4 Primers

As the different chapters in this dissertation cover a wide variety of topics, this section introduces these different topics with a short introduction. The information in these sections provides a basic overview of the different fields and is by no means intended to be exhaustive. It will however provide enough background information needed to understand the concepts that are used in the following chapters.

1.4.1 HIV

1.4.1.1 HIV and AIDS

Acquired immune deficiency syndrome (AIDS) is a disease that was first described in the beginning of the 1980's with the first article on five previously healthy homosexual patients suffering from Pneumocystis pneumonia being published on June 5, 1981 (Centers for Disease Control and Prevention, 1981b). In the following months several papers were published on patients with similar symptoms including multiple viral infections and a form a skin cancer and by 1982 the disease became known by the acronym AIDS (Gottlieb et al., 1981; Centers for Disease Control and Prevention, 1981a; Masur et al., 1981; du Bois et al., 1981; Marx, 1982). A year later the virus causing AIDS was isolated by Barré-Sinoussi et al. (1983) although there is some discussion whether they really identified the new virus or that is was discovered by Gallo *et. al* (Vahlne, 2009; Gallo et al., 1984; Sarngadharan et al., 1984). The name human immunodeficiency virus (HIV) for the newly discovered virus was adopted in 1986 by the International Committee on the Taxonomy of Viruses (Coffin et al., 1986). From the approximately 470 cases described by 1982, currently 35.3 million (32.2 million - 38.8 million) people are infected with HIV worldwide with 1.6 million (1.4 million - 1.9 million) dying from AIDS each year. It is only because of access to anti retro-viral treatment and improvements of this therapy that the number of deaths declined from 2.3 million (2.1 million - 2.6 million) in 2005 (UNAIDS, 2013).

HIV is an enveloped retrovirus of the lentiviral family containing an RNA of almost 10kb with 9 genes coding for 15 distinct proteins (Figure 1.3) (Los Alamos National Security LLC, 2014). A schematic overview of the life cycle of HIV can be seen in Figure 1.4. Infection of a host T cell starts with the binding of the viral particle to a CD4 receptor and a co-receptor, usually CCR5 or CXCR5 (Dalgleish et al., 1984; Feng et al., 1996; Deng et al., 1996). By binding to CD4 the viral envelope undergoes a conformational change ultimately leading to the fusion of the virus and the cell membrane (Malashkevich et al., 1998). After this fusion the viral core is released into the cell where the RNA genome is first uncoated before being transcribed to DNA by the viral reverse transcriptase (RT) (Goff, 2001). The reverse transcriptase complex will then be transported to the nucleus. During this journey a complete double stranded cDNA complexed with proteins will be created. This pre-integration complex will be transported into the nucleus where the double stranded cDNA will be integrated in the host DNA by the viral integrase (Murphy et al., 2008). The virus can then remain latent for a long time if the infected cell is not dividing. However, once a T cell becomes active, the viral genome is transcribed and

the spliced and unspliced mRNAs are exported to the cytoplasm helped by the viral protein Rev (Pollard & Malim, 1998). Once all the different viral proteins have been produced, new viral particles containing two RNA genomes are constructed and bud from the infected cell. However, these viral particles are not yet infective as the viral protease still has to cleave the viral polyproteins into functional subunits hereby yielding the mature and infectious particles (Laskey & Siliciano, 2014).

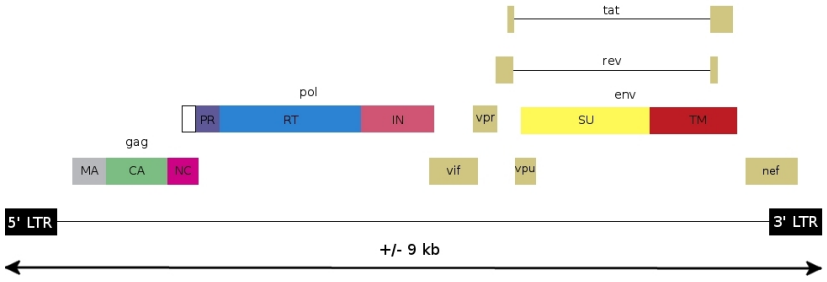


Figure 1.3: Genomic structure of HIV-1. The structural genes *gag*, *pol* and *env* code for poly-proteins that are later cleaved by the viral protease into their functional parts (coloured boxes). The accessory proteins *vif*, *vpr*, *tat*, *rev*, *vpu* and *nef* are also shown.

1.4.1.2 HIV Evolution

HIV is one of the fastest evolving organisms known today (Skar et al., 2011). There are several reasons why HIV reaches such a high rate of evolution. First, partly because of a lack of an error-correcting function, the viral RT makes between 3.0 x 10^{-4} and 3.4 x 10^{-5} errors per base and per generation in vivo (Mansky & Temin, 1995; Mansky & Bernard, 2000). In addition, HIV shows a generation time of 2.6 days and a release of 10.3 x 10^{9} virions per day in untreated individuals (Perelson et al., 1996). On top of this, RNA Pol II polymerase makes errors while transcribing the viral DNA to RNA in the host's cells. Finally, switching between the two available RNA copies of the viral genome by RT during reverse transcription results in a recombined genome and a very high rate of minimum 2.8 crossovers per genome per generation (Rambaut et al., 2004; Zhuang et al., 2002). This high mutation rate has important consequences for the development of resistance to antiretroviral drugs.

The high mutation rate in HIV also results in the fact that an HIV infection has to be seen as an infection by closely related but not identical viral strains. With each replication cycle, new errors are introduced in the viral RNA resulting in

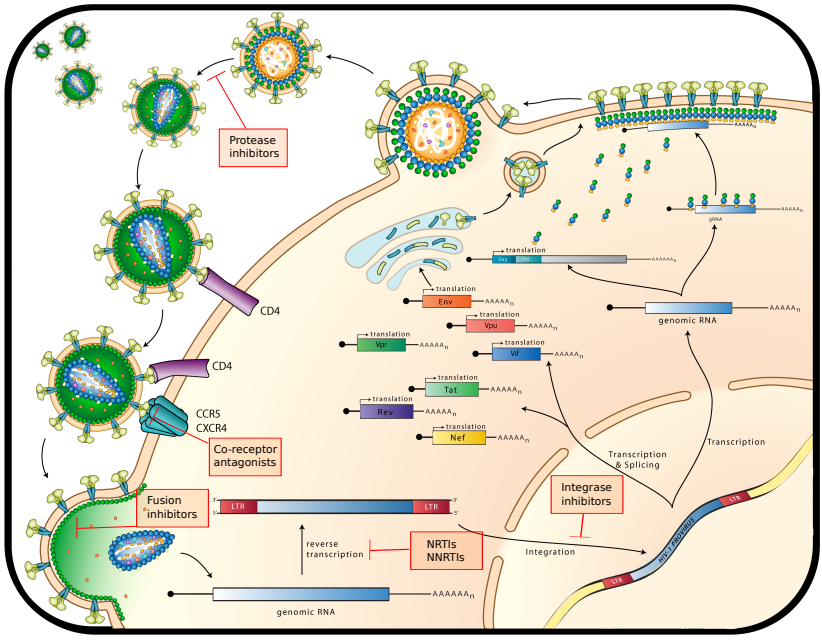


Figure 1.4: Schematic overview of the replication cycle of HIV and the processes on which the different antiretroviral drugs have an effect. Please refer to the text for a detailed explanation. Adapted from Ramdohr (2009)

variants that are genetically further apart from each other. The distribution of these variants is called the quasispecies (Holland et al., 1992). This in turn means that when a quasispecies is subjected to antiretroviral therapy, it can quickly shift to variants that exist in the quasispecies collection and that have a higher fitness under drug selective pressure.

1.4.1.3 Antiretroviral drugs

The life cycle of HIV provides several opportunities to halt the spread of the virus into new uninfected cells. Currently there are six classes of antiretroviral drugs that act on different parts of the life cycle as can be seen in Figure 1.4: (1) co-receptor antagonists, (2) fusion inhibitors, (3) nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs), (4) non-nucleoside reverse transcriptase inhibitors (NNRTIs), (5) integrase inhibitors, (6) protease inhibitors (PIs). The first drug found to be able to control HIV replication was zidovudine (AZT), a nucleoside reverse transcriptase inhibitor that was first synthezised as an

anti-cancer drug in the 1960's but failed selection (Broder, 2010). After that first drug was approved in 1987 more and more drugs became available and now doctors have access to around 30 FDA approved drugs in the battle against HIV (U.S. Department of Health and Human Services, 2015).

The newest drug classes have only been approved for the first time around 10 years ago. The first integrase inhibitor called raltegravir and the first co-receptor antagonist called maraviroc have only been licensed since 2007 and the first fusion inhibitor called enfuvirtide has been licensed since 2003. On the other hand, PIs (1995), NRTIs (1987), and NNRTIs (1996) have been around the longest (U.S. Department of Health and Human Services, 2015). As this also means that more data is available for these drug classes, we limited our analyses to these drug classes.

Because the first antiretroviral drug regimens only used one drug, this quickly lead to the development of drug resistance. Therefore treatment has evolved to a combination therapy using three to six different drugs from at least two different drug classes. This combination therapy is called highly active antiretroviral therapy (HAART). Current recommendations for HAART are to start with two NRTIs combined with a single NNRTI, PI boosted with ritonavir, or most recently an integrase inhibitor (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2014).

Protease inhibitors

Protease plays an essential role in the production of infectious viral particles in the life cycle of HIV. Without protease activity, the production of viral particles does not stop but the produced particles are not infective. The function of the viral protease is to cleave the Gag-Pol polyprotein into the functional viral proteins. For *gag* these are the structural proteins matrix (MA), capsid (CA), nucleocapsid (NC), p6 and two spacer proteins p2 and p1. For *pol* these are the enzymatic proteins protease (PR), reverse transcriptase (RT) and integrase (IN) (Wensing et al., 2010). Remarkable in the activity of the viral protease is that instead of recognizing a specific amino acid sequence, it recognizes an asymmetric secondary structure of the polyprotein (Prabu-Jeyabalan et al., 2002). Because the structure and substrates of the viral protease and its substrate have been extensively studied, it became possible to block the activity of the enzyme. Most PIs are competitive peptidomimetic inhibitors that resemble the natural substrate of the enzyme (Wensing et al., 2010).

Development of resistance against PIs is believed to be a stepwise process. In a first step a mutation in the substrate-binding cleft of the enzyme enlarges the cleft leading to a decreased binding of the inhibitor but also to the natural substrate which in turn results in a lower viral fitness. In a second step other

mutations can occur that (partially) restore the viral fitness to wild-type levels (Wensing et al., 2010). Two strategies have been employed to reduce the chance of resistant viruses emerging. A first one is to increase plasma levels of PI by also administering a cytochrome P450 3A4 enzyme inhibitor which reduces PI metabolization of the drug (Kempf et al., 1997). The second strategy is to develop new drugs with high potency against PI-resistant viruses and with a high genetic barrier (Dierynck et al., 2007).

Nucleoside/nucleotide reverse transcriptase inhibitors

NRTIs were the first drug class to be used in the battle against HIV starting in 1987 (Broder, 2010). They are analogous to the naturally occurring 2'-deoxynucleosides and nucleotides in the cell but all lack a 3'-hydroxyl group on their sugar moiety making them in fact chain terminators when incorporated during DNA replication. Because of this the virus is unable to produce a complete DNA copy of its genome to be inserted in the host genome. NRTIs are taken by the patient in their inactive form and are converted by host cell kinases and phosphotransferases into their active form, deoxynucleoside triphosphate analogs (dNTPs). It is in this form that they compete with the endogenous dNTPs for incorporation into the DNA strand created by the viral RT (Cihlar & Ray, 2010).

Resistance against NRTIs occurs through two different mechanisms. In the first mechanism the binding and rate of incorporation of the analogs is affected by mutations in the RT. For example the mutation M184V/I causes a selective reduction in the incorporation of the analogs by steric hindrance (Sarafianos et al., 2009). The second mechanism causes the analog chain terminator to be excised from the newly created viral DNA strand once it has been incorporated (Arion et al., 1998).

Non-nucleoside reverse transcriptase inhibitors

NNRTIs were the third drug class to be approved after NRTIs and PIs and this class is characterized by a large diversity of molecules which only interact with the RT of HIV-1. They inhibit the function of RT by binding in a hydrophobic pocket located near the catalytic site of the enzyme resulting in a conformational change of the enzyme and subsequent diminished catalytic activity (Sluis-Cremer et al., 2004).

Because of the low genetic barrier - one mutation is enough to confer resistance - of many NNRTIs, they need to be given together with at least two non-NNRTI drugs (de Béthune, 2010). Newer NNRTIs such as etravirine have been specifically developed to raise the genetic barrier to resistance mutations (Ludovici et al., 2001c,b,a). For some drugs and resistance mutations in this drug

class the molecular mechanisms for resistance have been elucidated. For instance, first generation NNRTIs rely on interactions with Y181 and Y188 resulting in a significantly reduced binding affinity when these residues are substituted by non-aromatic amino acids. The structural role of other mutations such as K103N still has to be elucidated for their effect on the resistance to first and second generation NNRTIs (Ren & Stammers, 2008; de Béthune, 2010).

1.4.1.4 Drug resistance

The viral load is the most important measurement to determine the successfulness of therapy (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2014). Patients strictly adhering to their therapy and who have an undetectable viral load for some time have a 99% chance that a subsequent test also has undetectable viral loads (Combescure et al., 2009). It is even allowed for infected (heterosexual) individuals to have unprotected sex when they: 1) adhere to the antiretroviral therapy, 2) have undetectable viral loads for at least six months and 3) have no additional sexually transmitted diseases (Vernazza et al., 2008). However, because of the previously mentioned high mutation rate in HIV, there can be a quick selection for drug resistance mutations in patients that do not adhere strictly to the prescribed therapy. Once the virus develops a mutation that reduces the effects of the drugs, it will quickly start to replicate again resulting a higher viral loads in the patient's blood (Wainberg & Friedland, 1998). This higher viral load in turn can lead to (increased) transmission of this resistant virus to other infected and non-infected individuals (Quinn et al., 2000).

The first report describing transmitted drug resistance (TDR) was published by Erice et al. (1993) where they described a patient who had never received treatment to be infected with a zidovudine resistant viral strain. About five years later Hecht et al. (1998) described transmission of a strain resistant to several PIs and RT-inhibitors from a patient not adhering his therapy to a drug naive patient. Quickly, TDR had been reported to occur through several infection routes: homosexual and heterosexual contacts, intravenous drug use, mother-child transmission and exposure to infected blood (de Ronde et al., 1996; Veenstra et al., 1995; Conlon et al., 1994; Colgrove et al., 1998). Currently, the prevalence of antiretroviral drug resistance in drug-naive individuals (DN) varies across geographic regions but can be found in approximately 10% of DN (Vercauteren et al., 2009; Wensing et al., 2005; Frentz et al., 2012; Bennett et al., 2008; Booth & Geretti, 2007; Stadeli & Richman, 2013; Rhee et al., 2015).

1.4.2 Next-generation Sequencing

1.4.2.1 Medical genetics

The field of medical genetics has seen an amazing evolution since the first visualization of chromosomes was published over 130 years ago by Walther Flemming in his book *Zellsubstanz, Kern und Zelltheilung* (Flemming, 1882). Twenty years later Boveri and Sutton would combine the discovery of chromosomes and the laws of inheritance described by Mendel by proposing the "chromosomal theory of heredity" (Sutton, 1903; Harper, 2008). It would take another 50 years before the correct number of human chromosomes was published (Tjio & Levan, 1956). But as the techniques for analyzing and counting human chromosomes became available, tremendous progress was made in the identification of chromosomal abnormalities. One of the first chromosomal abnormalities to be associated with a disease was the extra chromosome 21 in patients with Down syndrome (Lejeune et al., 1959). In the same time period also the cause of Turner syndrome and Klinefelter syndrome were discovered to be an abnormal number of sex chromosomes (Ford et al., 1959; Jacobs & Strong, 1959). In the decades following these discoveries several technical improvements allowed the field of medical genetics to evolve from a low resolution view showing just the chromosomes to a very high resolution view where we can identify single base pair alterations.

Sanger sequencing

It was over 20 years after the discovery of the double helix structure by Watson & Crick in 1953 that Sanger & Coulson developed the technology called Sanger sequencing that would become the gold standard for DNA sequencing in the future (Watson & Crick, 1953; Sanger & Coulson, 1975). After this first publication in 1975, they already published an improved version of the method two years later in which they started using chain terminating nucleotides (Sanger et al., 1977). In 1986 a variant of Sanger sequencing was developed in the lab of Leroy Hood to allow for automated sequencing by using dye terminators instead of the ones used in the original method (Smith et al., 1986). This improved method would form the basis for the sequencing of the first human genome and still remains to be a gold-standard method to date (Voelkerding et al., 2009; Ladouceur et al., 2012). The techniques available in the 80's and 90's made it possible to identify the variants in genes responsible for certain diseases which was impossible with the lower resolution cytogenetics that existed before. In this time period, the genes responsible for e.g. Duchenne muscular dystrophy (Koenig et al., 1987), cystic fibrosis (Riordan et al., 1989; Rommens et al., 1989), and Huntington's disease (The Huntington's Disease Collaborative Research

Group, 1993) were discovered. Because of the high cost and time needed for sequencing a single genome with Sanger sequencing, efforts were made to reduce the cost and time needed to sequence a complete genome. It was believed that achieving this would make it possible for whole genome sequencing to be routinely used in a clinical diagnosis setting, making it possible to move from a gene-centric view to a full genomic view (Hert et al., 2008).

Next-generation sequencing

Several successors to Sanger sequencing were developed by companies such as Roche, Illumina, and Applied Biosystems who have all introduced their implementation of NGS techniques, respectively 454, Solexa and SOLiD. While they all use different techniques, the basic work flow is similar between all of them and can be seen in Figure 1.5. For the amplification step there are two common methods: emulsion PCR and solid-phase amplification (Metzker, 2010). A brief overview of these methods can be seen in Figure 1.6 and Figure 1.7. Both methods have in common that they use a polymerase chain reaction (PCR) for amplification. A drawback of using a PCR reaction for amplification is the fact that it can introduce a bias based on the GC-content of the sequence as PCR favors regions with GC-neutral content so AT-rich and GC-rich regions can be underrepresented (van Dijk et al., 2014). Another drawback is that a PCR reaction can create errors which are subsequently amplified and can be perceived as sequence variants while they are in fact PCR artifacts (Eckert & Kunkel, 1991; Metzker, 2010).

As the sequencing technique differs widely between the different technologies they are all prone to different errors that can occur during sequencing (Shendure & Ji, 2008; Metzker, 2010). For instance, 454 sequencing makes use of pyrosequencing where a polymerase reaction is used and light is emitted following the incorporation of a nucleotide by the use of ATP sulphurylase and luciferase. As each nucleotide is added to the system in a repeated cycle, recording the sequence of light pulses yields the DNA sequence of the fragment. However, it is possible that multiple nucleotides are incorporated in the same cycle if they form homopolymers, i.e. stretches of the same nucleotide. As it becomes more difficult to distinguish the number of identical nucleotides that are added in longer homopolymer stretches, more errors start to occur (Hodkinson & Grice, 2015). The technique used by Illumina on the other hand makes use of reversible terminators with a fluorescent dye for each nucleotide that are added to the system in a repeated cycle, allowing a single nucleotide to be added in each cycle. Before the dye and the terminator are removed and a new cycle starts, a picture of the slide is taken in order to identify the nucleotide that was added during that cycle. This technique is more prone to substitution errors because it is possible that one of the amplified fragments does not elongate

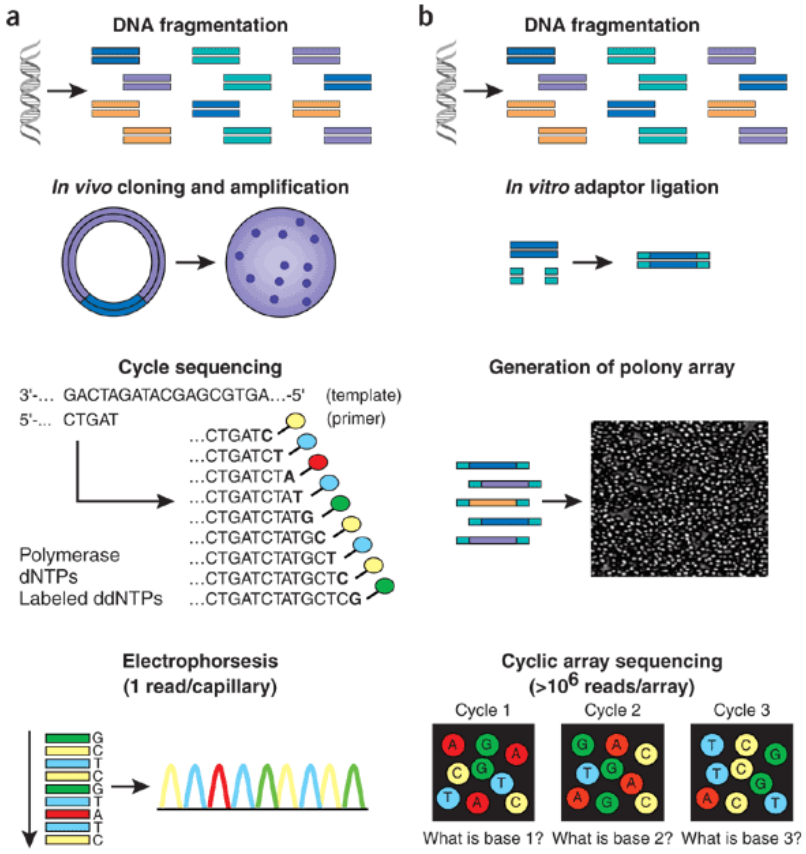


Figure 1.5: Overview of **(a)** Sanger sequencing and **(b)** next-generation sequencing work flow. In the first step of the Sanger sequencing work flow DNA is fragmented before being cloned and amplified. Subsequently each resulting colony is then sequenced by first using dye labeled ddNTPs to create copies of the template with different lengths which are separated by electrophoresis. As the fragments are of different lengths, the order in which the different dyes pass a detector yields the sequence. In a next-generation sequencing work flow the DNA is also fragmented but instead of cloning, adaptors are ligated to the resulting fragments. These constructs are then immobilized and amplified to form millions of distinct colonies or 'polonies'. Depending on the technology used, different kinds of fluorescent labels are added in a cyclic process and an image based analysis of the incorporation these labels yields the sequence. From Shendure & Ji (2008)

properly during a cycle, and therefore moves out of phase with the rest of the fragments within that cluster. When enough fragments are out of phase, it becomes extremely difficult to correctly identify the nucleotide (Dohm et al., 2008; Schirmer et al., 2015).

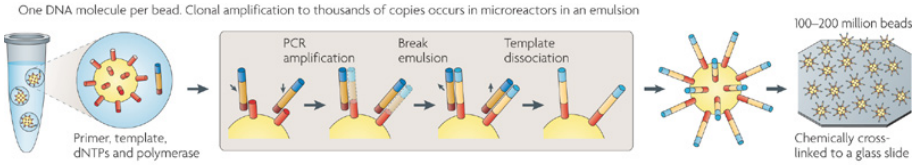


Figure 1.6: In emulsion PCR an oil-aqueous emulsion is created where in each droplet a bead, primer, template, dNTPs and polymerase is present. In this microreactor the template is amplified to create beads that are covered in identical sequences. These beads can then be immobilized for sequencing. Adapted with permission from Macmillan Publishers Ltd from Metzker (2010)

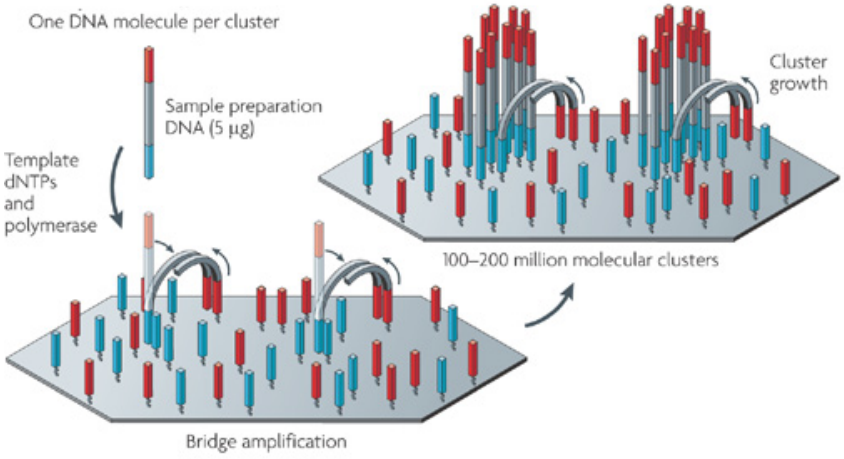


Figure 1.7: In solid-phase amplification a template is first immobilized on a solid support followed by bridge amplification. The result is a large number of small clusters of identical templates. Adapted with permission from Macmillan Publishers Ltd from Metzker (2010)

The output of a sequencing run consists of a large number of sequences called 'reads', coming from each of the fragmented templates that were created in the first step of the process. As these short reads offer no real information on a genomic level, they are mapped to a reference genome in order to be able to identify the complete genomic sequence and possible mutations. One drawback of these short reads being mapped to a reference is that it is hard to map them to repetitive regions that are longer than the length of the reads as each of those reads could be coming from anywhere in the repetitive sequence. Paired end sequencing tries to partially resolve this by sequencing two ends of a region which are separated by a known distance (Sims et al., 2014). Another

possible problem lies in the fact that some of the reads will contain sequencing errors. Therefore it is not enough to have just one read covering a genomic position as it is unknow whether a possible mutation is real or a sequencing error. This problem can be overcome by increasing the number of reads covering a certain position, i.e. increasing the read depth. By sampling the same position a number of times, sequencing errors can be identified as it is not expected to observe the same error to occur multiple times. The read depth is one of the parameters that is taken into account by variant callers, such as GATK (McKenna et al., 2010) and SAMtools (Li et al., 2009), when they assign a probability score to a mutation.

This is only a very brief overview of some of the most popular methods for NGS and some possible errors that may arise when using these methods. As discussing all possible techniques at length would fall outside the scope of this primer, a complete overview of all techniques with their benefits and drawbacks can be found in papers by Metzker (2010) and Shendure & Ji (2008).

Genome-wide association studies

Although Sanger sequencing is still used in clinical diagnostics, the benefits of NGS with the low cost and fast throughput offers an unprecedented opportunity to look beyond a single gene and beyond the original clinical question. This also means that instead of doing linkage analysis on a single gene to help identify the genetic cause for a disease, we can now look at hundreds of thousands of single-nucleotide polymorphisms (SNPs) in order to find associations with more complex and polygenic diseases. These studies are called genome-wide association studies (GWAS) and they have offered insights in the mechanisms that can lead to a number of more common and complex diseases. This way, a huge number of SNPs were identified to be associated with diseases like diabetes, high blood pressure, Crohn's disease, and many more. The NHGRI-EBI GWAS catalog is a publicly available and manually curated database containing 1,778 publications and 12,123 SNPs associated with complex diseases (Welter et al., 2014). In Figure 1.8 an overview is given of where SNPs with a SNP-trait association p-value of $< 5 \times 10^{-8}$ can be found in the genome. But this expanded view also comes with its own problems as most of these SNPs are not causative for a certain phenotype but are part of a more complex non-linear system where several SNPs influence each other and are influenced by environmental factors (Marjoram et al., 2014). So instead of finding a single (or a few) causative mutation(s) one is now confronted with a number of SNPs, each of which increases or decreases the odds of getting a particular disease or has an effect on the measurement of metabolic markers.

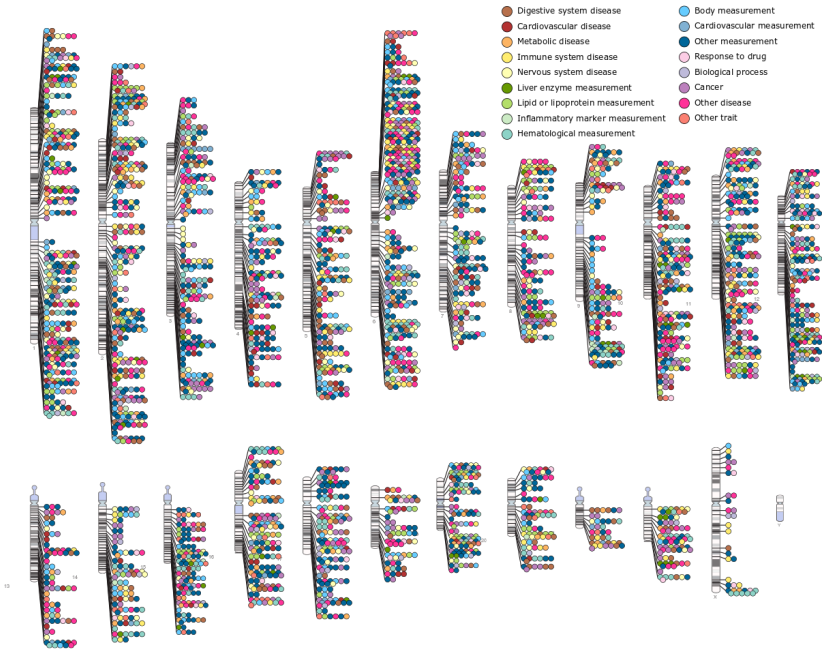


Figure 1.8: Visualization of all SNP-trait associations with a p-value $< 5 \times 10^{-8}$ mapped to the SNPs corresponding cytogenetic band and colored according to disease or measurement. From Welter et al. (2014)

1.4.2.2 Secondary use

Now that NGS is becoming affordable, even for the general public, we are confronted with a large amount of secondary information that is not related to an initial clinical question but that could still prove to be very useful for both patients and the healthcare system. First, if an individual knows to be carrier for a mutation causing a serious but treatable disease it would be possible to start treatment in time and hence improve prognosis. A well known example of such a mutation and disease could be predisposition to breast and ovarian cancer due to BRCA1 and BRCA2 mutations. Second, certain mutations can affect the efficacy or side effects of certain drugs which would help caregivers to adjust the medicine or the dosage depending on the genotype of the patient. This could for instance be useful for patients having to receive Abacavir treatment following an HIV infection where the presence of HLA-B*5701 strongly predicts hypersensitivity to this drug (Martin et al., 2004). Third, couples with a child wish could benefit from knowledge about their carrier status for known recessive

mutations.

1.4.3 Epigenetics

The term epigenetics was first used in the first half of the 20^{th} century by Waddington (1942) who described an epigenotype as a whole complex of developmental processes between genotype and phenotype, and connecting them to each other. In the following decades, the definition of epigenetics has evolved and in 1994 two definitions were proposed (Holliday, 2006): "1) The study of changes in gene expression which occur in organisms with differentiated cells, and the mitotic inheritance of given patterns of gene expression, and 2) Nuclear inheritance which is not based on changes in DNA sequence". Today, these definitions have evolved even further with minor variations still existing between them. For instance, Langevin & Kelsey (2013) describe it as "any mitotically stable modification that influences gene expression without altering the underlying DNA sequence. Epigenetic modifications include DNA methylation, histone modifications and the now large field encompassing the many effects of non-coding RNAs". Rozek et al. (2014) give a slightly different definition: "the study of DNA methylation and patterns of histone modifications (histone marks) which lead to changes in gene expression not accompanied by alterations in DNA sequence". So although there are some minor differences between these definitions concerning the mechanisms, they agree on the point that epigenetics studies changes in gene expression not related to changes in the DNA sequence.

While environmental influences do not alter the DNA sequence except for certain mutagenic compounds, epigenetic changes can occur due to normal chemical communication between cells, for example when cells are differentiating into a specific tissue (Holliday, 2006). Even more remarkable is the fact that epigenetic changes can also occur because of social and behavioral influences. For instance, working in shifts has been shown to lead to a significant increase of *TNF-α* promotor methylation (Rozek et al., 2014). In contrast with changes in the DNA sequence that are not so easily reversed, epigenetic changes are reversed more often, for example when they are reset in germ cells and the zygote (Lees-Murdock & Walsh, 2008).

As seen in the different definitions of epigenetics, different mechanisms underlie changes in gene expression. In the following subsections a brief overview of the importance of DNA methylation is given as this is the focus of the study performed in this chapter. There are many more modifications like histone modification, chromatine remodeling, and small and non-coding RNA but discussing these would take us too far. A complete overview and description

of all epigenetic modifications can be found in the book “Epigenetics” by Tost (2008).

1.4.3.1 Methylation

The most widely studied epigenetic modification to date is DNA methylation which was the first one to be described on a molecular basis. This methylation consists of an added methyl group to a cytosine leading to the formation of a 5-methylcytosine. This happens most often in a cytosine-guanine pair (CpG) (Holliday & Pugh, 1975; Langevin & Kelsey, 2013). In total about 75% of CpG nucleotides are methylated in human DNA (Tost, 2010). Regions that are relatively rich in these CpG pairs are called CpG islands and are found in the 5’ promotor region and first exons of many transcribed genes. Methylation of CpG pairs in these promotor regions has been associated with inhibition of transcription of the gene and an estimated 8-12% of gene promotors are methylated under normal circumstances (Meaburn & Schulz, 2012). This does not mean however that when a region is hypomethylated, the gene will be actively transcribed but just that it is possible for the gene to be transcribed when needed (Tost, 2010).

Methylation plays an important role in several processes such as the inactivation of an X chromosome in female cells, genomic imprinting, genomic stability and transcription (Tost, 2010). The methylation profile of cells is also associated with susceptibility and development of diseases, especially cancer. It has been shown that aberrant DNA methylation plays an important role in carcinogenesis and that approximately 5-10% of CpG islands in the promotor regions of hundreds of genes are hypermethylated in cancer cells while they are not in normal cells (Brennan & Flanagan, 2008; Baylin & Jones, 2011). But also in more complex diseases and conditions such as schizophrenia, autism, lupus, and metabolic syndrome there are signs that epigenetic modifications play a role in the development of the disease (Van Vliet et al., 2007). In autism for instance folic acid intake and methylation are suspected to have an effect on the risk of developing autism spectrum disorders (Neggers, 2014)

1.4.3.2 Association studies

An important complication of searching for epigenetic modifications associated with disease is that in contrast to the genome which essentially stays the same in an individual, the epigenomic profile can differ between cells within a person. For instance, in monozygotic twins the difference in epigenetic profiles increases over time as these twin are separated (Fraga et al., 2005). Also, compared to

the errors made when copying the DNA sequence during mitosis, more errors occur in the transmission of methylation patterns with an error rate of 1 in 10^3 instead of 1 in 10^6 for DNA. This also means that this higher error rate may be responsible for the discordance between twins and not only the environment that is no longer shared between these twins (Petronis, 2010). As epigenetic modifications occur during aging under the influence of different environmental factors and errors in transmission, it is important to take this into account when analyzing epigenetic data. Another factor complicating association studies is the fact that the epigenetic profile differs between cell types which is not the case with the genotype (Reinius et al., 2012; Baron et al., 2006). In this case, differences that are found between samples could just be differences between the composition of cell types in the samples and not necessarily a difference associated with a disease of interest. This may prove to be a complication particularly when using blood samples to perform association studies (Langevin & Kelsey, 2013).

1.4.3.3 Technology

Many technologies exist for the analysis of DNA methylation (Laird, 2010). One of the most commonly used methods involves bisulphite conversion. By treating DNA with sodium bisulphite, unmethylated cytosines are deaminated more quickly than their methylated counterparts. When cytosines are deaminated, they are converted to uracils which in turn are converted to thymidines during amplification. Therefore, because of this conversion a methylation signal is converted into a genetic mutation allowing it to be captured by sequencing or array methods. An example of a technology that uses bisulphite converted DNA is the Illumina HumanMethylation450 BeadChip. This widely used array based method allows the interrogation of around 480.000 methylation sites per sample (Wang et al., 2015).

1.5 Chapter-by-chapter overview

The remainder of this dissertation is divided into the following chapters.

In Chapter 2 we investigate the relation between the prevalence of HIV-1 surveillance drug resistance mutations in treated and drug naive individuals in order to test whether the prevalence in treated patients can contribute to assessing transmission of drug resistance, transmissibility and transmission source of surveillance drug resistance mutations.

Chapter 3 describes an assessment of the analytical and clinical validity and clinical utility of *in vitro* screening of embryos by whole genome sequencing. This was done by looking for mutations in genes associated with severe Mendelian disorders (i.e. characterized by extreme dysmorphology and early-onset) that are predicted to be damaging and/or disease causing in individuals considered healthy at the time of screening. As this concerns the screening of embryos, part of the paper addresses possible ethical issues that can arise when applying these techniques outside a research setting.

In Chapter 4 we try to identify epigenetic modifications that increase the risk of neural tube defects during pregnancy by investigating methylation levels in *HOX* genes.

Chapter 5 introduces a conceptual framework that can be used to get an indication of how likely it will be to extract accurate and relevant information from data sets without being confronted with unexpected results.

Chapter 6 concludes this dissertation with a general discussion, overview of future research possibilities and a general conclusion.

Chapter 2

HIV-1 Transmitted Drug Resistance

The remainder of this chapter has been published in abbreviated form as:

Winand R, Theys K, Eusébio M, Aerts J, Camacho RJ, Gomes P, Suchard MA, Vandamme A-M, Abecasis A on behalf of the Portuguese HIV-1 Resistance Study Group. Assessing transmissibility of HIV-1 drug resistance mutations from treated to and among drug naïve individuals *AIDS* 2015; **29**(15):2045-2052.

Author's contribution to this project: database implementation, data analysis, figures, data interpretation, writing

Reprinted with permission of Wolters Kluwer Health

2.1 Introduction

Studies in Europe and US show that antiretroviral drug resistance can be found in approximately 10% of drug-naive individuals (DN) infected with HIV-1. In Europe, the prevalence of transmitted drug resistance (TDR) was 8.4% in patients newly diagnosed between 2002 and 2005 (Vercauteren et al., 2009; Wensing et al., 2005), and 7.7% between 2003 and 2009 (Frentz et al., 2012). In Sub-Saharan Africa it is up to 5.7% while in the US numbers ranging between 8% and 24.5% are reported (Frentz et al., 2012; Bennett et al., 2008; Booth & Geretti, 2007; Stadeli & Richman, 2013). Most of these studies count the

number of transmitted drug resistance mutations (DRMs) in newly diagnosed DN using a list of surveillance drug resistance mutations (SDRMs) that excludes known polymorphisms (Bennett et al., 2009).

These aggregated numbers offer no further insights into transmissibility and transmission source of these SDRMs. Differences in transmissibility of wild-type and SDRM-containing viruses have been reported. For instance, one study showed that strains with SDRMs are found in DN less often than expected from the population of patients failing treatment (TR) (Leigh Brown et al., 2003). Others deduced from mathematical models that the relative fitness of the strain and the treatment-induced reduction in viral load have an impact on transmission of SDRMs (Blower et al., 2001). de Mendoza et al. (2004) showed some SDRMs to be transmitted more efficiently. SDRMs in DN are mostly singletons, which are rare in TR. It is currently assumed this is the result of reversion of some SDRMs to wild type after transmission, since such mutations have a fitness cost for the virus in absence of drugs. In such cases, minor variants still carrying all transmitted SDRMs are to be expected. When after reversal, the major variant with fewer SDRMs is transmitted further among DN, minor variants with more SDRMs are less likely to exist and the frequency of singletons among DN will increase further. The distinction between acquisition of TDR from TR or from other DN is important for first line treatment strategies. Current guidelines assume that DN with TDR are likely to harbor undetected SDRMs as minor variants and therefore suggest avoiding low genetic barrier regimens, even if mutations against such regimens are not observed (Vandamme et al., 2011; Panel on Antiretroviral Guidelines for Adults and Adolescents, 2014). Evidence of onward transmission between DN could therefore influence treatment decisions. To investigate the mode of transmission of TDR, directly from TR vs onwards between DN, most studies rely on transmission chain investigations (Yerly et al., 2009; Audelin et al., 2011; Drescher et al., 2014), which are complicated and for which the methodology is still under development.

To guide strategies to avoid TDR and improve treatment strategies for DN with TDR, a deeper understanding of transmissibility and reversion dynamics of transmitted SDRMs, and the extent of acquisition of TDR from TR versus from other DN is needed. Here we describe an innovative and simple approach of correlating prevalence of SDRMs in DN and TR, while using outlier detection to measure transmissibility and source of SDRMs. High or low transmissibility is defined as higher or lower prevalence in DN compared to what would be expected from the regression model. We suggest that high transmissibility is associated with persistence of SDRMs (slower reversion dynamics) and higher rates of onward transmission among DN, and low transmissibility is associated with faster reversion dynamics and acquisition of TDR from TR. This approach does no longer require a dense sample of patients to study transmission chains

of TDR since general epidemiological data for the regression analysis can be extracted from a sparser sample.

2.2 Methods

The protocol was in accordance with the Declaration of Helsinki and approved by Ethical Committees of Centro Hospitalar de Lisboa Ocidental (108/CES-2014) and KU Leuven, Faculty of Medicine (NH019/2015-06-01). We analyzed a clinical database containing anonymized patients' clinical and HIV-1 genetic sequence data sampled in Portugal (PT) between April 2001 and March 2013 in the context of antiretroviral resistance testing, either before starting treatment (DN) or at treatment failure (TR). In order to ensure that both populations were part of the same epidemic, the analysis was limited to subtype B, as identified by Rega subtyping tool v2.0 using protease (PR) and reverse transcriptase (RT)sequences. Of a total of 3606 subtype B sequences, obtained from 3354 patients, 1685 sequences were from DN (counting only the first DN sequence per patient) and 1921 from TR (counting only the last viral sequence per patient). Median age and gender proportion was similar in DN (43 years IQR = 15 years; 75.7% males) and TR (48 years IQR = 13 years; 74.1% males). Reproducibility was tested on a previously published German cohort (DE) (Schmidt et al., 2014).

A mutation was considered a TDR mutation when it was included in the SDRM list (Bennett et al., 2009). When reporting prevalence of SDRMs, DN were patients naïve for all drug classes while TR (e.g. PI-TR) were patients with experience with the drug class of interest. Prevalence in TR is presented per drug class or, for comparison between drug classes, normalized for the complete treated population (Tables 2.1 to 2.3). We used robust linear regression to estimate a regression model correlating prevalence of SDRMs in TR and DN for each drug class, allowing for outlier detection.

For each SDRM, we evaluated two measures of transmissibility: a) its transmission ratio (Tables 2.1 to 2.3): prevalence in DN divided by prevalence in TR experienced with the respective drug class; b) whether or not it was an outlier with respect to the regression models (Figures 2.1 to 2.4 for PT and Figure S2 for DE). Mutations positioned above or below the regression line have, respectively, higher or lower prevalence in DN than would be expected from the prevalence in TR, and are therefore interpreted as having higher or lower transmissibility. To evaluate the effect of viral load (VL) as a confounding factor, we correlated VL of DN, TR or DN+TR with specific SDRMs and the transmission ratio for those same SDRMs.

Data analyses and manipulation were performed using in-house Python scripts (van Rossum & Drake, 2001). For statistical analyses, SAS, R and the Rpy2 interface for Python were used (SAS Institute Inc., 2014; R Core Team, 2013; Gautier and Rpy2 contributors, 2014). To identify outliers in the regression model, we performed a robust linear regression assuming the prevalence in TR to be fixed and independent. Robust linear regression and outlier detection were performed using ROBUSTREG, as implemented in SAS.

2.3 Results

In this dataset of subtype B infected patients, the overall prevalence of TDR in DN was 10.0% [95% confidence interval (CI), 8.6–11.5], while 67.7% [95% CI, 65.5–69.8] of TR carried SDRMs. Most patients with TDR showed single class resistance resulting in a prevalence of 7.4% [95% CI, 6.2–8.8], the majority (5.4%) being singletons. Double class resistance was found in 2.2% [95% CI, 1.6–3.0] and triple class resistance in 0.4% [95% CI, 0.1–0.8] of DN. In TR failing any drug class, 18.6% [95% CI, 16.9–20.4] of individuals showed single class resistance, a minority (1.9%) being singletons; double class resistance was found in 38.4% [95% CI, 36.2–40.6] and triple class in 10.7% [95% CI, 9.3–12.1]. The SDRMs in DN with the highest prevalence were for PI L90M with 1.4%, for NRTI T215Y/F+Rev (Rev = C,D,E,I,S,V) with 2.6% and for NNRTI K103N/S with 2.3%. In TR experienced with the respective drug class we find that the SDRMs with the highest prevalence were for PI L90M with 19.7%, for NRTI M184I/V with 44.0%, for NNRTI K103N/S with 32.3%. Tables 2.1 to 2.3 list all SDRMs.

We then looked at the transmission ratio of each SDRM separately (Tables 2.1 to 2.3). The highest transmission ratios were observed for PI N83D (0.165), N88D/S (0.134), and D30N (0.130); for NRTI T215Rev (0.727) (while T215Y/F transmission ratio was much lower, 0.029) and V75A/M/S/T (0.120); for NNRTI K101E/P (0.092) and G190A/E/S (0.091).

By plotting the prevalence of mutations found in DN versus in TR experienced with the respective drug class, a significant linear correlation indicated that the prevalence of SDRMs in DN can be predicted from its prevalence in TR belonging to the same epidemic. In Figures 2.1 and 2.2 (drug class specific) and Figures 2.3 and 2.4 (normalized) the robust regression lines are drawn together with a 95% CI. The R-squared values range between 0.52 (for PIs) and 0.68 (for NNRTIs). After normalizing for treatment, the regression slope was higher for NNRTIs (0.116), followed by NRTIs (0.084) and PIs (0.073) (Figures 2.3 and 2.4). A number of SDRMs are classified as outliers indicating that their

Table 2.1: Overview of the PI-SDRMs identified in the Portuguese cohort, at each position with prevalence and transmission ratios. Bold entries indicate an SDRM located above or below the 95% CI in the robust linear regression (Figure 2.1). Variants at a single position are pooled (e.g. 88D/S). SDRM – Surveillance Drug Resistance Mutations, DN – Drug Naïve patients, TR – Patients failing treatment, PI – Protease Inhibitors, PI-TR – Patients failing a PI-containing treatment.

SDRM	Prevalence (%) in DN (N=1685) [95% CI]	Prevalence (%) in PI-TR (N=1352) [95% CI]	Normalized prevalence (%) in TR (N=1921) [95% CI]	Transmission ratio
83D	0.1 [0.0–0.3] (N=1)	0.4 [0.1–0.9] (N=5)	0.3 [0.1–0.6] (N=5)	0.2
88D/S	**0.7 [0.4–1.2] (N=12)**	**5.3 [4.2–6.7] (N=72)**	**3.7 [2.9–4.7] (N=72)**	**0.1**
30N	**0.7 [0.4–1.2] (N=12)**	**5.5 [4.3–6.8] (N=74)**	**3.9 [3.0–4.8] (N=74)**	**0.1**
32I	**0.3 [0.1–0.7] (N=5)**	**3.0 [2.2–4.1] (N=41)**	**2.1 [1.5–2.9] (N=41)**	**0.1**
24I	0.2 [0.1–0.6] (N=4)	2.7 [1.9–3.8] (N=37)	1.9 [1.4–2.6] (N=37)	0.1
85V	0.2 [0.0–0.5] (N=3)	2.5 [1.7–3.5] (N=34)	1.7 [1.2–2.5] (N=34)	0.1
90M	**1.4 [0.9–2.0] (N=23)**	**19.7 [17.6–21.9] (N=266)**	**13.8 [12.3–15.5] (N=266)**	**0.1**
46I/L	0.9 [0.5–1.5] (N=16)	17.5 [15.5–19.7] (N=237)	12.3 [10.9–13.9] (N=237)	0.1
53L/Y	0.1 [0.0–0.4] (N=2)	2.4 [1.6–3.3] (N=32)	1.7 [1.1–2.3] (N=32)	0.1
82A/C/F/L/M/S/T	0.8 [0.4–1.3] (N=13)	15.5 [13.6–17.5] (N=209)	10.8 [9.5–12.4] (N=209)	0.1
54A/L/M/S/T/V	**0.5 [0.2–1.0] (N=9)**	**16.0 [14.1–18.0] (N=216)**	**11.2 [9.9–12.7] (N=216)**	**0.0**
48M/V	0.1 [0.0–0.3] (N=1)	2.0 [1.3–2.9] (N=27)	1.4 [0.9–2.0] (N=27)	0.0
84V	**0.1 [0.0–0.4] (N=2)**	**6.2 [5.0–7.6] (N=84)**	**4.4 [3.5–5.4] (N=84)**	**0.0**
73A/C/S/T	**0.1 [0.0–0.3] (N=1)**	**7.0 [5.7–8.5] (N=95)**	**4.9 [4.0–6.0] (N=95)**	**0.0**
23I	0.0 [0.0–0.2] (N=0)	0.7 [0.3–1.3] (N=9)	0.5 [0.2–0.9] (N=9)	0.0
47A/V	0.0 [0.0–0.2] (N=0)	2.2 [1.5–3.2] (N=30)	1.6 [1.1–2.2] (N=30)	0.0
50L/V	0.0 [0.0–0.2] (N=0)	1.2 [0.7–1.9] (N=16)	0.8 [0.5–1.3] (N=16)	0.0
76V	0.0 [0.0–0.2] (N=0)	1.5 [0.9–2.3] (N=20)	1.0 [0.6–1.6] (N=20)	0.0
Any PI-SDRM	3.5 [2.7–4.5] (N=59)	38.3 [35.7–41.0] (N=518)	27.0 [25.0–29.0] (N=518)	

Table 2.2: Overview of the NRTI-SDRMs identified in the Portuguese cohort, at each position with prevalence and transmission ratios. Bold entries indicate an SDRM located above or below the 95% CI in the robust linear regression (Figure 2.1). Variants at a single position are pooled (e.g. 184I/V). SDRM – Surveillance Drug Resistance Mutations, DN – Drug Naïve patients, TR – Patients failing treatment, NRTI – Nucleoside/nucleotide Reverse Transcriptase Inhibitors, NRTI-TR – Patients failing a NRTI-containing treatment.

SDRM	Prevalence (%) in DN (N=1685) [95% CI]	Prevalence (%) in NRTI-TR (N=1920) [95% CI]	Normalized prevalence (%) in TR (N=1921) [95% CI]	Transmission ratio
215rev	**1.8 [1.2–2.5] (N=30)**	**2.4 [1.8–3.2] (N=47)**	**2.4 [1.8–3.2] (N=47)**	**0.7**
75A/M/S/T	0.2 [0.1–0.6] (N=4)	2.0 [1.4–2.7] (N=38)	2.0 [1.4–2.7] (N=38)	0.1
219E/N/Q/R	**1.2 [0.8–1.9] (N=21)**	**13.8 [12.2–15.4] (N=264)**	**13.7 [12.2–15.4] (N=264)**	**0.1**
41L	**2.1 [1.5–2.9] (N=35)**	**23.3 [21.4–25.2] (N=447)**	**23.3 [21.4–25.2] (N=447)**	**0.1**
69D	0.2 [0.1–0.6] (N=4)	2.7 [2.0–3.5] (N=52)	2.7 [2.0–3.5] (N=52)	0.1
215Y/F+rev	2.6 [1.9–3.5] (N=44)	31.3 [29.2–33.4] (N=600)	31.2 [29.2–33.4] (N=600)	0.1
67E/G/N	1.4 [0.9–2.1] (N=24)	20.7 [18.9–22.6] (N=398)	20.7 [18.9–22.6] (N=398)	0.1
210W	**0.8 [0.4–1.3] (N=13)**	**15.6 [14.0–17.3] (N=300)**	**15.6 [14.0–17.3] (N=300)**	**0.0**
151M	0.1 [0.0–0.3] (N=1)	1.7 [1.1–2.3] (N=32)	1.7 [1.1–2.3] (N=32)	0.0
215Y/F	**0.8 [0.5–1.4] (N=14)**	**28.8 [26.8–30.9] (N=553)**	**28.8 [26.8–30.9] (N=553)**	**0.0**
184I/V	**1.2 [0.8–1.9] (N=21)**	**44.0 [41.7–46.2] (N=844)**	**43.9 [41.7–46.2] (N=844)**	**0.0**
70E/R	**0.2 [0.1–0.6] (N=4)**	**12.4 [11.0–14.0] (N=238)**	**12.4 [10.9–13.9] (N=238)**	**0.0**
74I/V	**0.1 [0.0–0.3] (N=1)**	**7.4 [6.3–8.7] (N=142)**	**7.4 [6.3–8.7] (N=142)**	**0.0**
65R	0.0 [0.0–0.2] (N=0)	2.7 [2.0–3.5] (N=52)	2.7 [2.0–3.5] (N=52)	0.0
77L	0.0 [0.0–0.2] (N=0)	0.9 [0.5–1.4] (N=17)	0.9 [0.5–1.4] (N=17)	0.0
115F	0.0 [0.0–0.2] (N=0)	1.6 [1.1–2.3] (N=31)	1.6 [1.1–2.3] (N=31)	0.0
116Y	0.0 [0.0–0.2] (N=0)	1.4 [0.9–2.0] (N=26)	1.4 [0.9–2.0] (N=26)	0.0
Any NRTI-SDRM	5.0 [4.0–6.2] (N=85)	61.1 [58.9–63.3] (N=1174)	61.1 [58.9–63.3] (N=1174)	

Table 2.3: Overview of the NNRTI-SDRMs identified in the Portuguese cohort, at each position with prevalence and transmission ratios. Bold entries indicate an SDRM located above or below the 95% CI in the robust linear regression (Figure 2.1). Variants at a single position are pooled (e.g. 103N/S). SDRM – Surveillance Drug Resistance Mutations, DN – Drug Naïve patients, TR – Patients failing treatment, NNRTI – Non-Nucleoside Reverse Transcriptase Inhibitors, NNRTI-TR – Patients failing a NNRTI-containing treatment.

SDRM	Prevalence (%) in DN (N=1685) [95% CI]	Prevalence (%) in NNRTI-TR (N=1237) [95% CI]	Normalized prevalence (%) in TR (N=1921) [95% CI]	Transmission ratio
101E/P	**0.6 [0.3–1.1] (N=10)**	**6.5 [5.2–8.0] (N=80)**	**4.2 [3.3–5.2] (N=80)**	**0.1**
190A/E/S	**1.4 [0.9–2.1] (N=24)**	**15.6 [13.6–17.7] (N=193)**	**10.0 [8.7–11.5] (N=193)**	**0.1**
103N/S	2.3 [1.6–3.1] (N=38)	32.3 [29.7–34.9] (N=399)	20.8 [19.0–22.7] (N=399)	0.1
225H	0.2 [0.0–0.5] (N=3)	3.2 [2.3–4.3] (N=39)	2.0 [1.4–2.8] (N=39)	0.1
181C/I/V	**0.6 [0.3–1.1] (N=10)**	**12.4 [10.7–14.4] (N=154)**	**8.0 [6.8–9.3] (N=154)**	**0.0**
188C/H/L	0.2 [0.0–0.5] (N=3)	5.2 [4.0–6.6] (N=64)	3.3 [2.6–4.2] (N=64)	0.0
100I	0.0 [0.0–0.2] (N=0)	4.1 [3.1–5.4] (N=51)	2.7 [1.9–3.5] (N=51)	0.0
106A/M	0.0 [0.0–0.2] (N=0)	2.7 [1.8–3.7] (N=33)	1.7 [1.2–2.4] (N=33)	0.0
230L	0.0 [0.0–0.2] (N=0)	1.1 [0.6–1.8] (N=13)	0.7 [0.4–1.2] (N=13)	0.0
Any NNRTI-SDRM	4.3 [3.4–5.4] (N=73)	57.8 [55.0–60.6] (N=715)	37.2 [35.1–39.4] (N=715)	

Table 2.4: Overview of the SDRMs identified as outliers in the robust linear regression model of the Portuguese cohort. The Mahalanobis distance and standardized robust residual are given. Cutoff level for outliers is 3.00. PI – Protease Inhibitors, NRTI – Nucleoside/nucleotide Reverse Transcriptase Inhibitors, NNRTI – Non-Nucleoside Reverse Transcriptase Inhibitors

Drug class	Position	Mahalanobis distance	Standardized robust residual
PI	30	0.11	3.74
PI	88	0.13	3.8
PI	90	2.12	3.07
NRTI	184	2.46	9.44
NNRTI	190	0.65	3.24

transmissibility is either higher or lower compared to the other mutations in the same drug class. For PIs D30N, N88D/S, and L90M have a higher transmissibility; for NRTIs M184I/V has a lower transmissibility; for NNRTIs G190A/E/S has a higher transmissibility (Table 2.4).

In the normalized regression model for DE (Schmidt et al., 2014) (Figures 2.5 and 2.6), the slope was also higher for NNRTIs (0.08) followed by NRTIs (0.07) Outliers and positioning of mutations were consistent with PT for NRTIs (Figure 2.5 Bottom) and for NNRTIs (Figure 2.6). For NRTIs, K219N and T215rev were consistently above while K70E, M184I/V, L210W and T215Y/F were below the CI (see Figures 2.1 and 2.5 for outlier classification). Mutations M41L (above) and L74I/V (below) were outside the CI for PT but not for DE, and T69N (above) and K65R (below) outside the CI for DE but not for PT. For NNRTI, Y181C/I/V was below the CI and G190A/E/S was above the CI in both cohorts. K101P was above the CI for PT and not for DE and V106I below the CI for DE but not PT. There were no SDRMs significantly on opposite sides of the regression line when comparing cohorts. For PIs (Figures 2.1 and 2.5), preliminary analysis showed important differences between cohorts that we could not analyze further, since access to DE data was limited (only DRMs with a prevalence >0.3% were reported in (Schmidt et al., 2014)). It has to be noted that in the German cohort all subtypes were considered together, while we only analyzed the subtype B epidemic. Additionally, other SDRMs may have spread in transmission clusters among DN.

To test the confounding effect of viral load (VL), we plotted the transmission ratio of specific SDRMs against the median VL of individuals carrying such SDRMs in samples taken at most 30 days before or after the resistance test. The correlation was not significant, neither when plotting the median VL in TR carrying the mutation (Figure 2.7), nor when plotting the median VL in

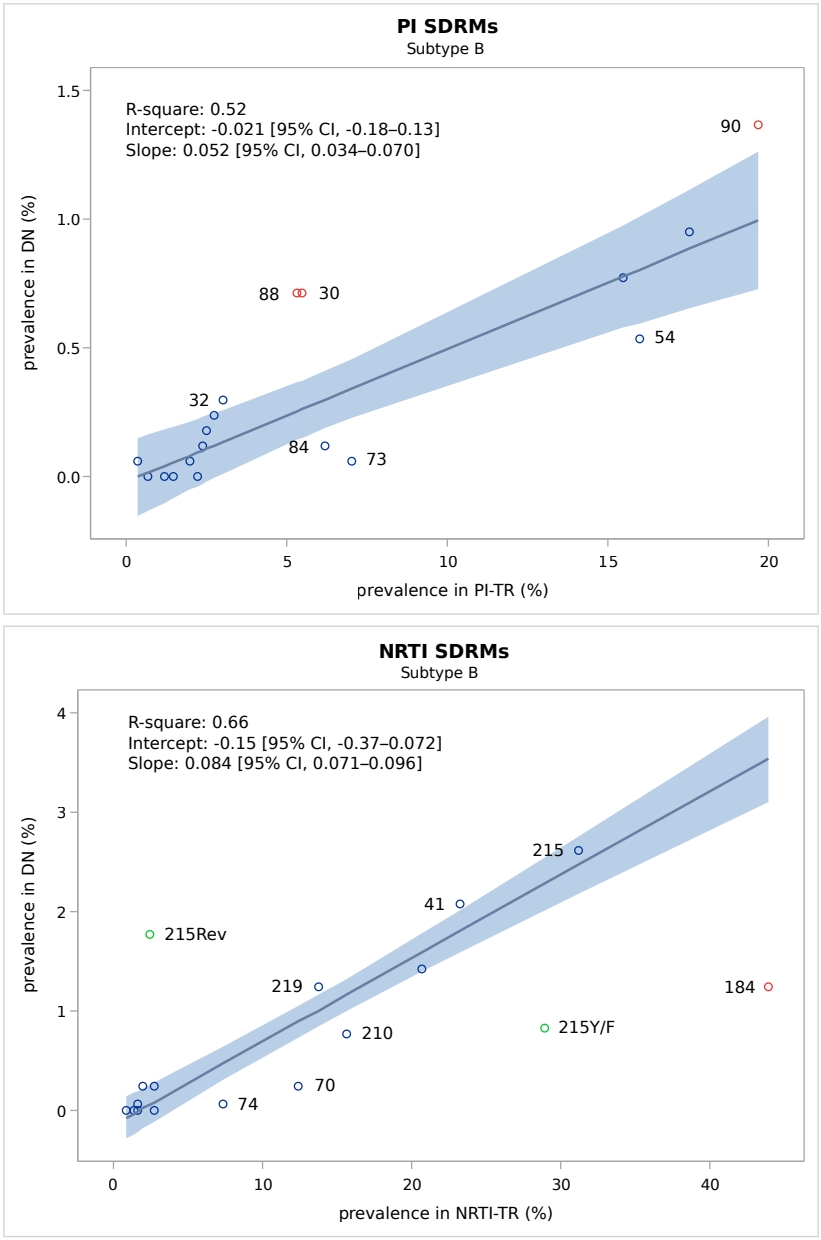


Figure 2.1: Robust regression model relating prevalence of SDRMs (codon position) in TR per drug class versus in prevalence of SDRMs in DN. The linear regression line is shown together with the 95% CI, slope, intercept, and R-Square. SDRMs above or below the 95% CI are labeled with codon position, robust outliers (see Table 2.4) are shown as red circles. Top: Prevalence of each PI SDRM in PI-TR (i.e. [# PI SDRM in PI-TR]/# PI-TR, see Table 2.1) versus prevalence of each PI SDRM in DN i.e. [# PI SDRM in DN]/# DN; Bottom: Prevalence of each NRTI SDRM in NRTI-TR versus prevalence of each NRTI SDRM in DN. T215Y/F and T215Rev are shown in green but were taken together (T215) for the linear regression.

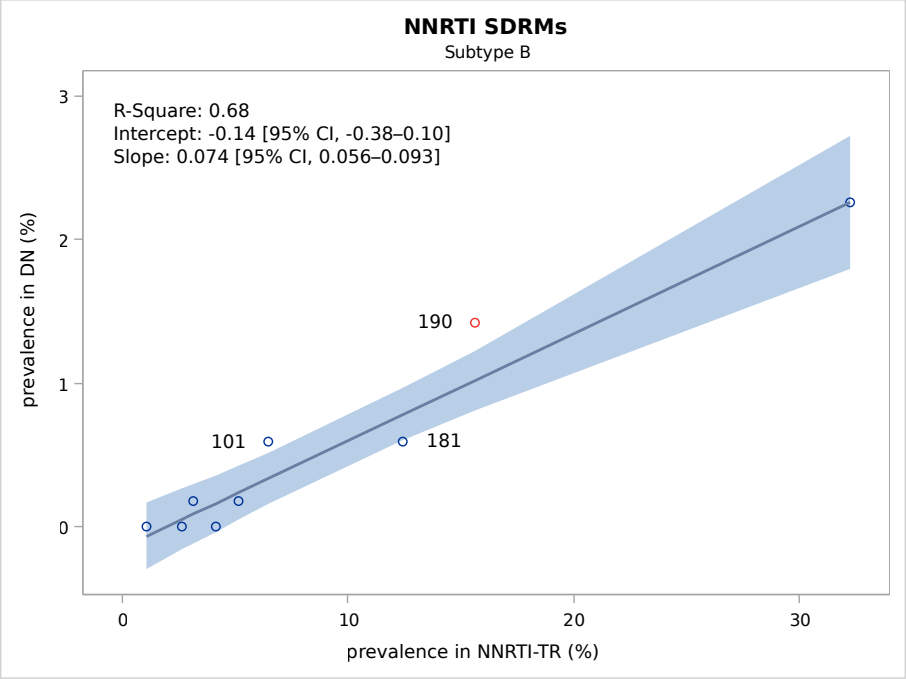


Figure 2.2: Robust regression model relating prevalence of SDRMs (codon position) in TR per drug class versus in prevalence of SDRMs in DN. The linear regression line is shown together with the 95% CI, slope, intercept, and R-Square. SDRMs above or below the 95% CI are labeled with codon position, robust outliers (see Table 2.4) are shown as red circles. Prevalence of each NNRTI SDRM in NNRTI-TR versus prevalence of each NNRTI SDRM in DN.

DN carrying the mutation or both DN and TR together (data not shown). In addition we performed a Wilcoxon rank sum test on the VL of patients failing therapy for different drug classes and found no significant difference between the VL.

2.4 Discussion

We propose to capture differences in transmissibility of DRMs by comparing the prevalence of each surveillance DRM (SDRM) in DN with its prevalence in TR. Such an analysis could convey whether an SDRM is found more often than could be expected from its prevalence in TR, thereby suggesting it may

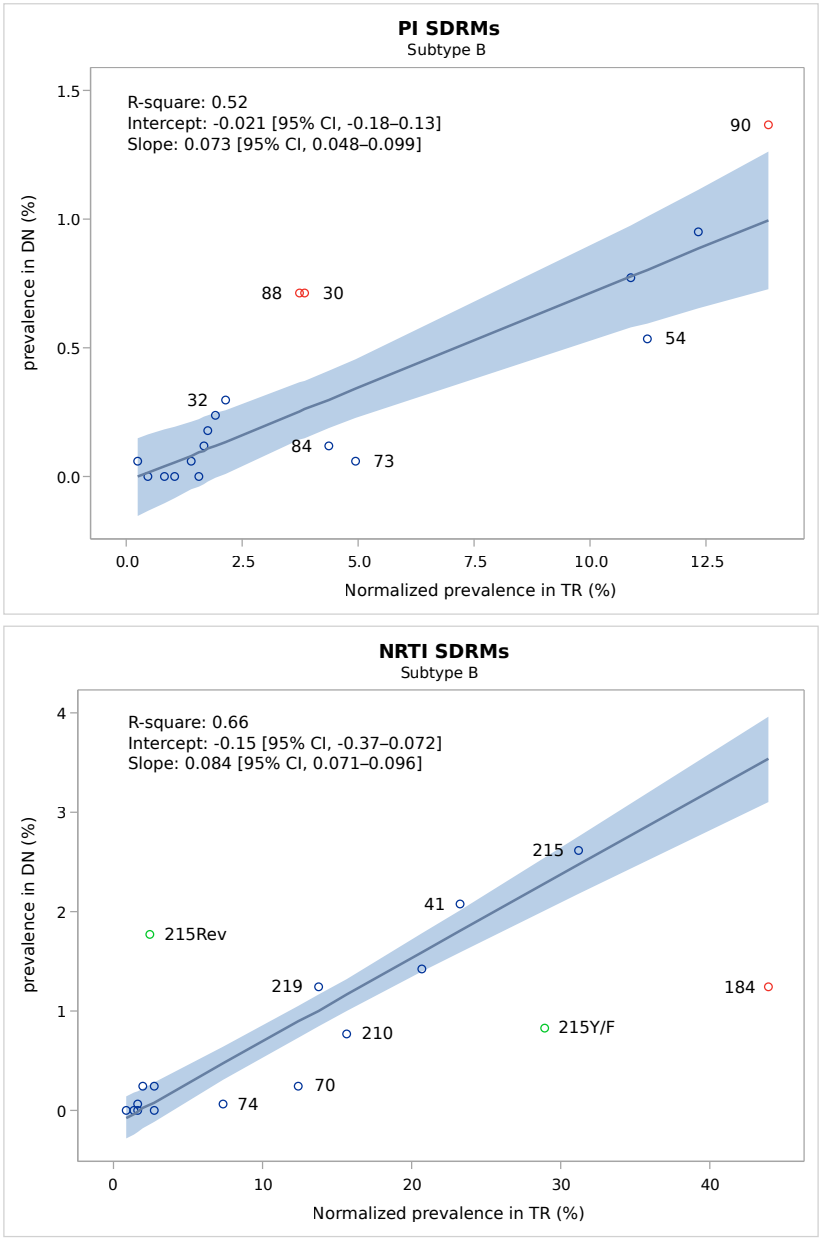


Figure 2.3: Robust regression model relating normalized prevalence of SDRMs (codon position) in TR per drug class versus in prevalence of SDRMs in DN. The linear regression line is shown together with the 95% CI, slope, intercept, and R-Square. SDRMs above or below the 95% CI are labeled with codon position, robust outliers (see Table 2.4) are shown as red circles. Top: Normalized prevalence of each PI SDRM in PI-TR (i.e. [# PI SDRM in PI-TR]/# PI-TR, see Table 2.1) versus prevalence of each PI SDRM in DN i.e. [# PI SDRM in DN]/# DN; Bottom: Normalized prevalence of each NRTI SDRM in NRTI-TR versus prevalence of each NRTI SDRM in DN. T215Y/F and T215Rev are shown in green but were taken together (T215) for the linear regression.

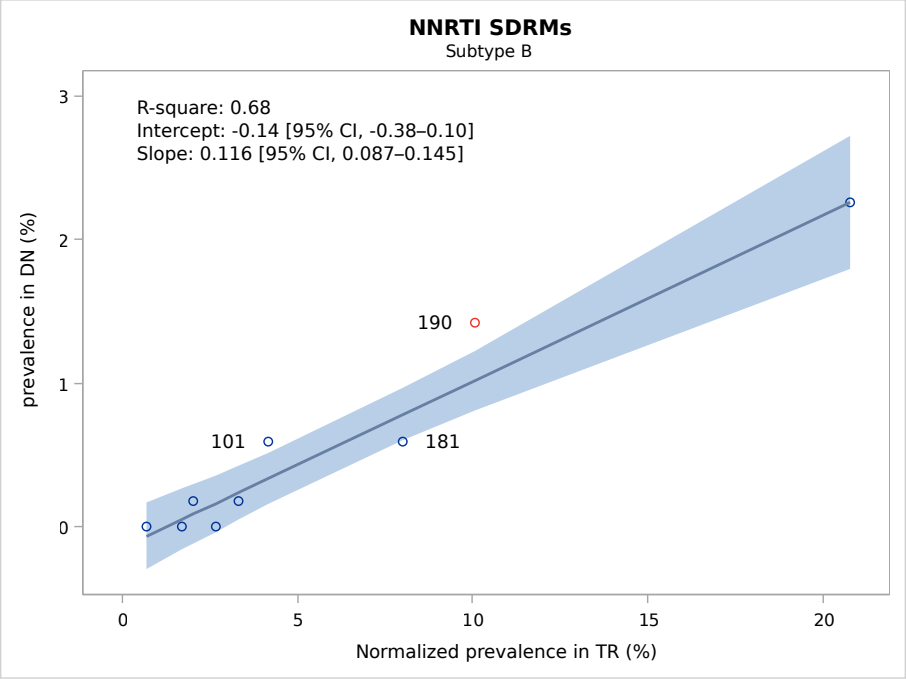


Figure 2.4: Robust regression model relating normalized prevalence of SDRMs (codon position) in TR per drug class versus in prevalence of SDRMs in DN. The linear regression line is shown together with the 95% CI, slope, intercept, and R-Square. SDRMs above or below the 95% CI are labeled with codon position, robust outliers (see Table 2.4) are shown as red circles. Normalized prevalence of each NNRTI SDRM in NNRTI-TR versus prevalence of each NNRTI SDRM in DN.

have increased its prevalence in DN by onward transmission among DN. This simple and innovative way to analyze TDR is however dependent on access to the treated population of the same epidemic, as prevalence of SDRMs in TR becomes an important parameter to analyze. As far as we know, such an approach has never been suggested.

We tested this in Portuguese subtype B infected patients, representing the bulk of the native infections (thus DN and TR belong to the same epidemic) (Palma et al., 2007). Patients were counted among TR (e.g. NNRTI-TR) only when having experience with the drug class of interest, thereby avoiding counting e.g. NNRTI SDRMs in a NNRTI naïve patient albeit experienced with other drug classes that had acquired this NNRTI SDRM through historical TDR. This approach requires knowledge of the treatment regimen. If such information is

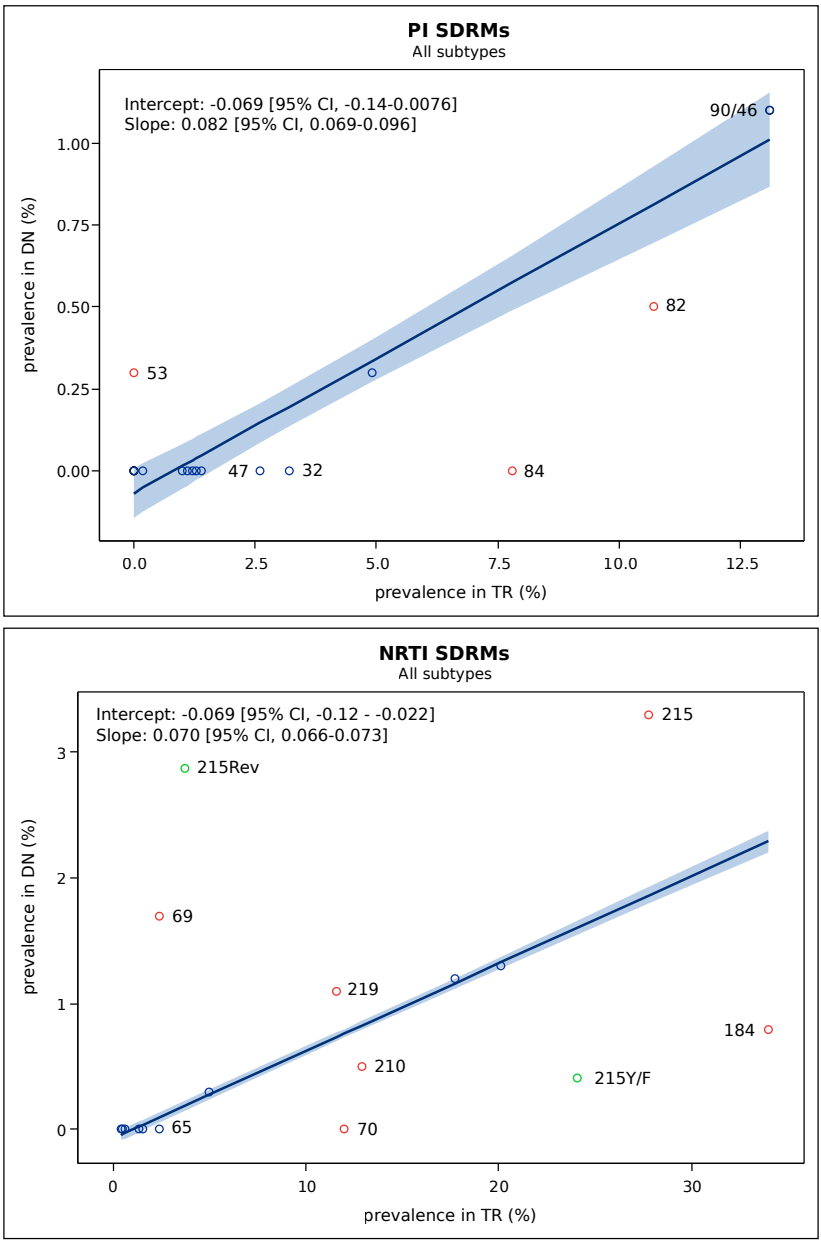


Figure 2.5: Robust regression model for the German cohort data, relating normalized prevalence of SDRMs (codon position) in TR per drug class versus in prevalence of SDRMs in DN. The linear regression line is shown together with the 95% CI, slope, intercept, and R-Square. SDRMs above or below the 95% CI are labeled with codon position, robust outliers (see Table 2.4) are shown as red circles. Top: Normalized prevalence of each PI SDRM in PI-TR (i.e. [# PI SDRM in PI-TR]/# PI-TR, see Table 2.1) versus prevalence of each PI SDRM in DN i.e. [# PI SDRM in DN]/# DN; Bottom: Normalized prevalence of each NRTI SDRM in NRTI-TR versus prevalence of each NRTI SDRM in DN. T215Y/F and T215Rev are shown in green but were taken together (T215) for the linear regression.

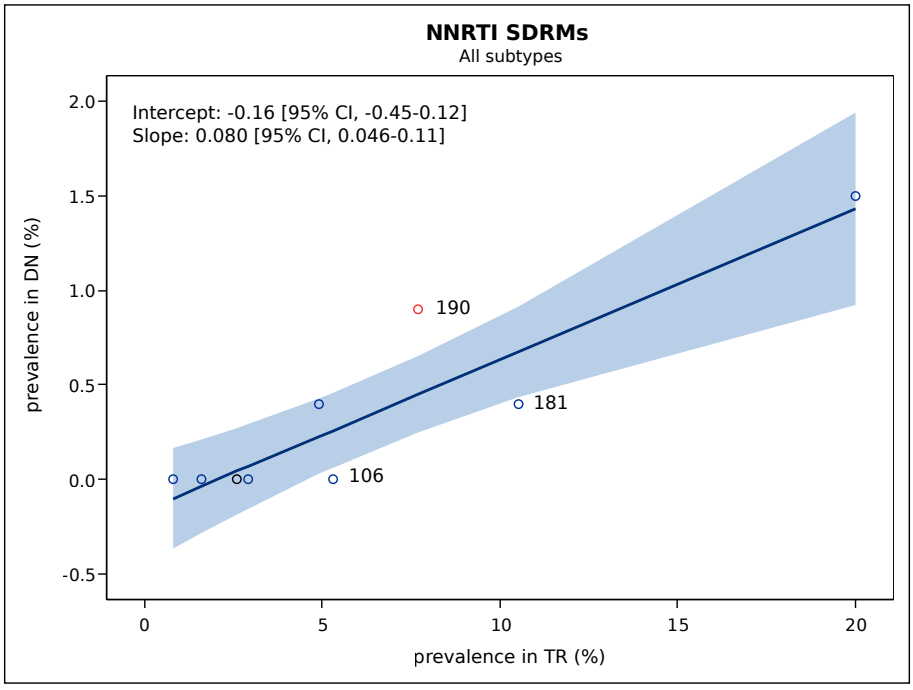


Figure 2.6: Robust regression model for the German cohort data, relating normalized prevalence of SDRMs (codon position) in TR per drug class versus in prevalence of SDRMs in DN. The linear regression line is shown together with the 95% CI, slope, intercept, and R-Square. SDRMs above or below the 95% CI are labeled with codon position, robust outliers (see Table 2.4) are shown as red circles. Normalized prevalence of each NNRTI SDRM in NNRTI-TR versus prevalence of each NNRTI SDRM in DN.

not available counting among all TR still gives reliable results, however in our hands this was less sensitive (results not shown).

SDRM transmission ratios (prevalence in DN/prevalence in TR) were high for SDRMs with known high transmissibility (e.g. 0.727 for T215Rev, Tables 2.1 to 2.3), however the precision was low for SDRMs with very low prevalence in TR. For instance, the N83D mutation has a high transmission ratio (0.165) but was only found in 1 DN and 5 TR, consequently the CI is very large. Thus, transmission ratios can only give a reliable indication of transmissibility for mutations with high prevalence in TR.

We then built a robust linear model and detected outliers. We found that the prevalence of SDRMs in DN is linearly correlated with its prevalence in TR (Figures 2.1 and 2.2), indicating that the prevalence in TR can be used for

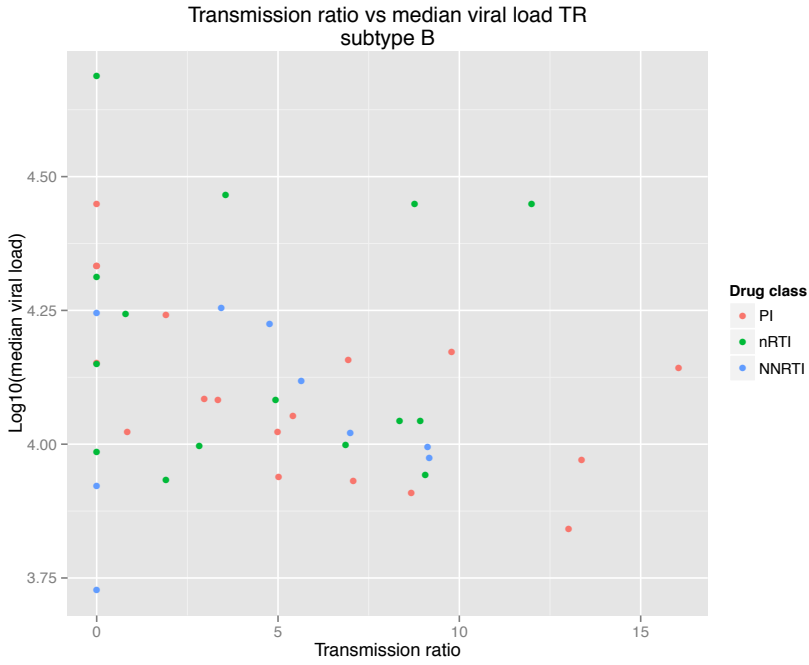


Figure 2.7: Transmission ratio of each SDRM versus the log10 of the median viral load found in TR for samples taken at most 30 days before or after the resistance test, calculated from the Portuguese cohort data. SDRM – Surveillance Drug Resistance Mutations, TR – Patients failing treatment, PI – Protease Inhibitors, NRTI – Nucleoside Reverse Transcriptase Inhibitors, NNRTI – Non-Nucleoside Reverse Transcriptase Inhibitors

surveillance purposes and predicts levels of TDR. Interpretation of the slope is difficult. While faster reversion dynamics will result in a reduced slope, and increased onward transmission among DN will result in a steeper slope, the slope will also vary according to the treatment strategies encountered in the cohort, such that slopes from different drug classes cannot be compared with each other or across cohorts. We therefore built an additional robust linear model normalizing the prevalence in TR to the total population failing treatment (Figures 2.3 and 2.4). If SDRMs would not affect fitness in absence of drugs, the normalized regression slope should be similar between drug classes. The differences can be therefore interpreted as differences in reversion dynamics among SDRMs affecting different drug classes, which will consequently impact the rates of onward transmission of such mutations. Remarkably, the NNRTI normalized slope (0.116) is higher than the PI (0.073) or NRTI (0.084)

normalized slope, consistent with the higher persistence of NNRTI mutations (de Béthune, 2010). Consequently, NNRTI SDRMs are more likely to be part of DN transmission chains.

Within a drug class, some SDRMs were found to be outliers below or above the regression line. Outlier detection also yielded significant results for some mutations with low prevalence in TR. It could be argued that the higher the viral load of a patient, the more often the virus will be transmitted, and this may confound our interpretation of transmissibility of specific SDRMs (Quinn et al., 2000). We tested this hypothesis and this was not the case, suggesting it was rather related to the intrinsic transmission efficiency and/or reversion dynamics of variants carrying those mutations.

If the prevalence in DN of an SDRM is below the regression line, this SDRM is transmitted less often and/or reverses more quickly after transmission due to a higher fitness cost in absence of the drug. In both cases it means that the transmission fitness of that mutation is lower than for other SDRMs of the same drug class. If the prevalence in DN is above the regression line, then the mutation is more frequently transmitted, has a lower rate of reversion after transmission to DN due to a lower fitness cost, and/or has a longer persistence in transmission chains of DN. Therefore outlier detection can inform about transmission fitness. Such an interpretation is consistent with Yang et al. (2013), who stated that it is likely that SDRMs with a high fitness cost are transmitted mainly from TR and SDRMs with a low fitness cost may more likely be transmitted by DN. However some differences may also be due to actual differences in transmission frequency as was suggested by Chateau et al. (2013) who observed a possible reduced vaginal transmission efficiency for K65R. Both effects together can be expressed as transmissibility.

Our results are consistent with other knowledge in the field. We show the NRTI SDRM M184I/V to be an outlier with lower transmissibility. This is consistent with Castro et al. (2013) who showed the same mutation to be less persistent after transmission to DN. In the same study G190A was found to have a higher persistence and this mutation is here identified as an outlier with higher transmissibility. Castro et al. did not present results on all PI SDRMs, while our approach suggests the outliers D30N, N88D/S, and L90M to have high transmissibility. Although L90M shows a decreased replicative fitness, the polymorphism L63P has been shown to increase the fitness of the virus carrying L90M up to wild-type levels (Suñé et al., 2004). In our dataset we found that 82.6% of DN with L90M also carry L63P explaining the higher transmissibility of outlier L90M despite the replicative cost of the mutation in absence of therapy.

In addition to these robust outliers we find a number of additional SDRMs

above or below the 95% confidence interval of the regression line. For instance, the NRTI SDRMs M41L, T215Rev and K219E/N/Q/R are above and K70E/R, L74I/V, and T215Y/F below the CI of the regression line. For the NNRTI SDRMs, K101E/P is above and Y181C/I/V below the CI of the regression line. The PI SDRMs I84V and I54A/L/M/S/T/V are below the CI of the regression line with Castro et al. showing a lower persistence for I84V and I54V. While these findings are consistent with Castro et al.'s results, statistical significance needs to be confirmed using models built from larger datasets. Our findings are also confirmed with the German cohort, where mostly the same mutations fall above or below the regression line.

We pooled T215Y/F and T215 revertants as they can all be traced back to development of T215Y/F under NRTI selective pressure in TR. While the combined mutations can be found within the CI of the regression line, T215Rev and T215Y/F mutations are located above and below the regression line respectively. These findings support our model, as T215Y/F has a fitness cost in absence of drug and will revert rather quickly after transmission into one of these other non-wild-type SDRMs (together called T215Rev) and will show less onward transmission between DN. Any T215Y/F in a DN is therefore a testimony of a relatively recent acquisition from TR. Conversely, T215Rev located well above the regression line, is confirming that such outliers are more persistent and an indicator of more onward transmission between DN. The higher persistence and consequent higher rates of onward transmission of T215 revertants has recently been described by Dolling et al. (2012).

As far as we know, this is the first time the correlation of SDRM prevalence in TR and DN is used to make claims on transmissibility and acquisition of TDR from TR or from DN. While other studies have investigated transmission ratios, they only considered TR as source of SDRMs (Leigh Brown et al., 2003; de Mendoza et al., 2004). Our approach enables the identification of SDRMs for which DN are a likely additional source. Future efforts should include the development of a more sophisticated statistical model to address the uncertainty associated with less frequent SDRMs.

We anticipate that our population of DN is mainly chronically infected, even though we know that we have a significant proportion of acutely infected DN (Palma et al., 2007). Both for mutations with high or low persistence, the time since infection will affect its observed prevalence in DN: the longer the time, the higher the chance that such mutation will have reverted and vice-versa. A dataset with more chronically infected patients will therefore result in a more sensitive outlier detection.

Our results support that for 1st line treatment strategies, prevalence of SDRMs in local TR should be considered, as such SDRMs will generally be more

prevalent in local DN. We deem this more realistic and more practical than only counting SDRMs in DN as is currently done because: 1) the prevalence of SDRM in TR is higher than in DN and therefore the CI is smaller; 2) many DN have not been diagnosed and have therefore not been sampled; 3) while most TR are tested for drug resistance immediately at treatment failure, in many countries most DN are either not tested or only just before starting treatment when eventual reversion already took place.

In conclusion, correlating prevalence of SDRMs in DN with the prevalence in TR gives a refined view of the relationship between transmission dynamics, persistence, viral fitness, and source of TDR. In addition robust outlier detection could inform the extent of transmission of particular SDRMs among DN, which could help public health and strategic measures to avoid increase of TDR.

Chapter 3

Embryo screening

In this chapter we investigate whether the current next-generation sequencing (NGS) techniques and knowledge about the genome can be used to accurately predict phenotypes. We chose the setting of embryo selection for our analysis as the use of whole genome sequences outside clinical diagnostics leads to many (ethical) questions that need to be resolved before these techniques can be adopted outside the clinic. Although this is the setting we chose, the questions that are raised go beyond the setting of embryo selection as they are also relevant for carrier screening or donor selection. While in these cases there is not yet an embryo whose rights have to be protected, the other issues that are raised in this chapter are equally important.

This remainder of this chapter has been previously published as:

Winand R*, Hens K*, Dondorp W, de Wert G, Moreau Y, Vermeesch JR, Liebaers I & Aerts J. In vitro screening of embryos by whole-genome sequencing: now, in the future or never? *Hum Reprod* 2014; **29**:842–51.
* *Equally contributing first authors*

Author's contribution to this project: study design, disease selection, database implementation, data analysis, figures, data interpretation, writing

Reprinted with permission from Oxford University Press

3.1 Introduction

Genetic testing of preimplantation embryos is a generally accepted approach in the context of preimplantation genetic diagnosis (PGD): patients with a known risk to transmit a specific genetic condition, or with a known chromosomal rearrangement, can opt for PGD to select embryos without the relevant disease-causing mutation. A second application of genetic testing of embryos is preimplantation genetic screening (PGS) for aneuploidy. Although not yet sufficiently proven by randomized controlled trials, this is offered by some centers to subfertile patients undergoing in vitro fertilization (IVF) as a treatment of infertility or as part of PGD, with the aim of improving their chances of a successful pregnancy. Whole genome sequencing (WGS) and analysis, which determines and analyzes the entire DNA sequence of an individual in one procedure, has been performed in single cells and single blastomeres (Navin et al., 2011; Xu et al., 2012; Voet et al., 2013). WGS (followed by a targeted analysis) could be a generic approach for PGD, avoiding time-consuming and labor-intensive customized PGD workups. WGS might also be an elegant alternative method for PGS, since full chromosomal aneuploidies and expected segmental imbalances associated with chromosome rearrangements can be easily identified (Harper & Sengupta, 2012). As WGS costs are now approaching the costs for array-based single-cell PGD or PGS, WGS may become a useful auxiliary technique for embryo testing as already performed in such context (Baslan et al., 2012; Hens et al., 2013b; Simpson et al., 2013).

In addition to these applications, WGS could theoretically be used to extend considerably the scope of embryo testing. This would entail a widening of the aims of the procedure. In addition to helping people either to have a child without a specific disorder (as in PGD), or to attempt to increase the chances of a successful IVF pregnancy (as in PGS for aneuploidy), the aim of WGS in embryo testing would be to ensure that children born after IVF or IVF/PGD are free from major disorders. Unlike classical PGD, embryo testing with these aims would be a form of medical screening, as it is a form of indiscriminate genetic testing without clinical data. One of the accepted criteria for responsible screening is that there should be a suitable test (Wilson et al., 1968). This means that both the analytical and clinical validity of the test must have been demonstrated. The analytical validity of a genetic test is its ability to determine accurately the genotype of interest. Clinical validity is the accuracy with which the test can then predict a phenotype. If the test performs poorly in these regards, this will adversely affect the clinical utility of the screening. This last concept refers to the balance of aim-related advantages and unavoidable disadvantages (drawbacks and costs), and is as such directly linked with the ethical acceptability of screening programs (Sanderson et al., 2005; Dondorp

et al., 2010). In this paper, we assess the analytical and clinical validity of WGS-based testing as a necessary (though not a sufficient) condition for the clinical utility and ethical acceptability of extended or comprehensive embryo screening using this technology.

3.2 Materials and Methods

We ascertained whether state-of-the-art technology and know-how could adequately distinguish benign polymorphisms from disease-causing mutations and help in deciding which embryo to transfer. We investigated how many apparently healthy adults carry mutations predicted to be damaging or annotated as disease causing—under the rationale that if such mutations were sufficient to cause the severe early-onset phenotype (and therefore predictive in a screening setting), they should be absent from the exome of apparently healthy adults.

3.2.1 Disease selection

We obtained the online Mendelian Inheritance in Man (OMIM) (McKusick-Nathans Institute of Genetic Medicine, 2012) list of diseases with a known molecular basis, consisting of >3000 diseases, from the OMIM website. This list was cross-referenced with the Human Phenotype Ontology database (Robinson & Mundlos, 2010) to provide an overview of diseases and their associated phenotypes. The resulting set contained 2172 diseases from which we selected diseases characterized by dysmorphology and early-onset symptoms. By selecting early-onset disorders we ensure that individuals affected by one of these disorders would already show symptoms at the time of sequencing. In addition, only diseases that had the inheritance annotated in OMIM as 'autosomal dominant', 'autosomal dominant type; high penetrance' and 'autosomal recessive' were selected. As a result 132 autosomal dominant and 215 autosomal recessive diseases were retained. The complete list of these diseases can be found in Supplementary data, Tables 3.S1 and 3.S2.

3.2.2 Samples

For our analysis, we used both private and publicly available samples from adult individuals who were considered healthy at the time of sequencing (i.e. who did not exhibit clear signs of a congenital disorder at the time of sampling). We had access to 42 exome sequences from in-house research samples. These are

trio samples where the parents are considered healthy, as they are symptom free. The sequences of affected children were also available but were not used in the initial analysis. In addition to our high-quality exome sequences, we downloaded two freely available sets from the 1000 Genomes project (1kG) (The 1000 Genomes Project Consortium, 2010) and Complete Genomics (CG) (Drmanac et al., 2010). The 1kG data came from the phase 1 integrated release version 3 from 30 April 2012, containing genotype calls for >1000 individuals. The data from CG were from 69 individuals, the sample names of which can be found in Supplementary data, Table 3.S3. All these sequences are from people who do not express symptoms of the selected disorders and therefore considered 'healthy'. Because some of the samples of these publicly available sets are from trios, we removed the samples from related individuals from the data sets leaving 1004 samples from the 1kG data and 50 from the CG data.

3.2.3 Transcripts

To localize the mutations in the genes and the different transcripts of those genes, we used the Ruby Ensembl API (Strozzi & Aerts, 2011) to connect to the Ensembl core database version 70. In addition, we also looked at transcripts present in the Consensus CDS Project (CCDS) database Hs37.3 (Pruitt et al., 2009).

3.2.4 Mutations predicted to be damaging

When analyzing genomes in the search of causative mutations, prediction algorithms are often used to predict the effect of a mutation on the protein. We used an in-house database called Annotate-it (Sifrim et al., 2012) that holds detailed information on the selected genes, to retrieve prediction scores from SIFT (Ng, 2003) and PolyPhen (Adzhubei et al., 2010) for the identified mutations. More information about these algorithms can be found in the Supplementary data. Because PolyPhen and SIFT only give scores to missense mutations, we considered nonsense and splice site mutations also to be damaging in the analysis. Indels were not included in the analysis.

3.2.5 Mutations described to be damaging in literature

In addition to the prediction algorithms, we also looked for mutations that are described as disease causing in the literature. For information on these mutations, we used the Human Gene Mutation Database (HGMD) containing

>70 000 disease-causing mutations (Stenson et al., 2003). For our analysis, we only selected those mutations with associations not considered tenuous by the curators of the HGMD and in the disease-causing category 'DM'.

3.3 Results

The results from this analysis show that many healthy individuals have mutations predicted to be damaging or annotated as disease causing in HGMD in genes associated with severe developmental disorders. For the 1kG samples all mutations are included, i.e. mutations that were found in both the low coverage and exome sequences. For the in-house data sets we only retained mutations with a read depth of at least 30× and phred score of 30. Relaxing these constraints leads to a higher number of mutations, which are described in Supplementary data, Tables 3.S4 and 3.S5.

3.3.1 Autosomal dominant disorders: mutations predicted to be damaging

When looking for mutations predicted to be damaging, we found that depending on the data set, 98–100% of healthy individuals had damaging mutations in genes associated with the selected autosomal dominant disorders. In the 1kG data set, we found a median of 8 mutations (min. 3, max. 14), in the Complete Genomics data set a median of 9 mutations (min. 4, max. 14) and in the in-house data set a median of 2 mutations (min. 0, max. 5) per individual. In only one sample from the in-house data set no damaging mutations were found. The distribution of the number of mutations can be seen in Figure 3.1. Because variants that are frequently found in the population are unlikely to cause these severe Mendelian disorders, we only retained the variants with a minor allele frequency (MAF) in the 1000 genomes of <1%. Applying this filter leads to a large decrease in identified variants but still 40–94% of individuals were found to carry damaging mutations. In the 1kG data set we found a median of 0 (min. 0, max. 5) mutations, in the Complete Genomics data set a median of 2 mutations (min. 0, max. 6) and in the in-house data set a median of 0 mutation (min. 0, max. 4) per individuals. The distribution of the number of mutations can be seen in Figure 3.2.

As explained by MacArthur et al. (2012) faulty gene annotation is a likely cause for these genes containing a large number of deleterious mutations. An example of this can be found in EDARADD (Ectodysplasin-A receptor-associated death domain), which is the gene with the highest number of unfiltered damaging

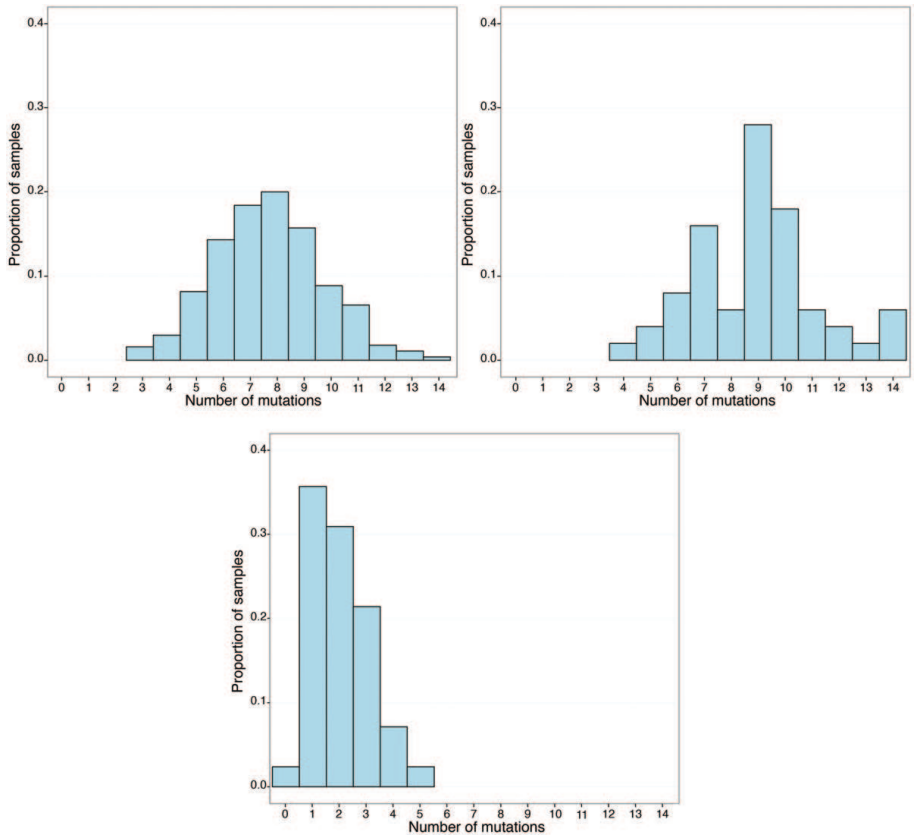


Figure 3.1: Histograms showing the proportion of samples with a certain amount of mutations that are predicted to be damaging. Samples are from the 1000 genomes, Complete Genomics and in-house (clockwise starting top left).

mutations in all data sets. Mutations in EDARADD cause the autosomal dominant form of ectodermal dysplasia, characterized by sparse hair, missing or abnormal teeth and the inability to sweat (Cluzeau et al., 2011). For this gene we found mutations in 83–98% of the samples. These mutations however are not annotated as disease causing in HGMD. In the 1kG data the most occurring mutation in EDARADD is at chromosome 1, position 236557771 G>A (dbSNP id: rs966365). So even though it is predicted to be damaging by both SIFT and PolyPhen, it is frequently found in that population (91%) suggesting an annotation error in the reference genome. Another observation we made was that most of the genes with the most common mutations have multiple transcripts and most have at least one transcript that is not affected by these

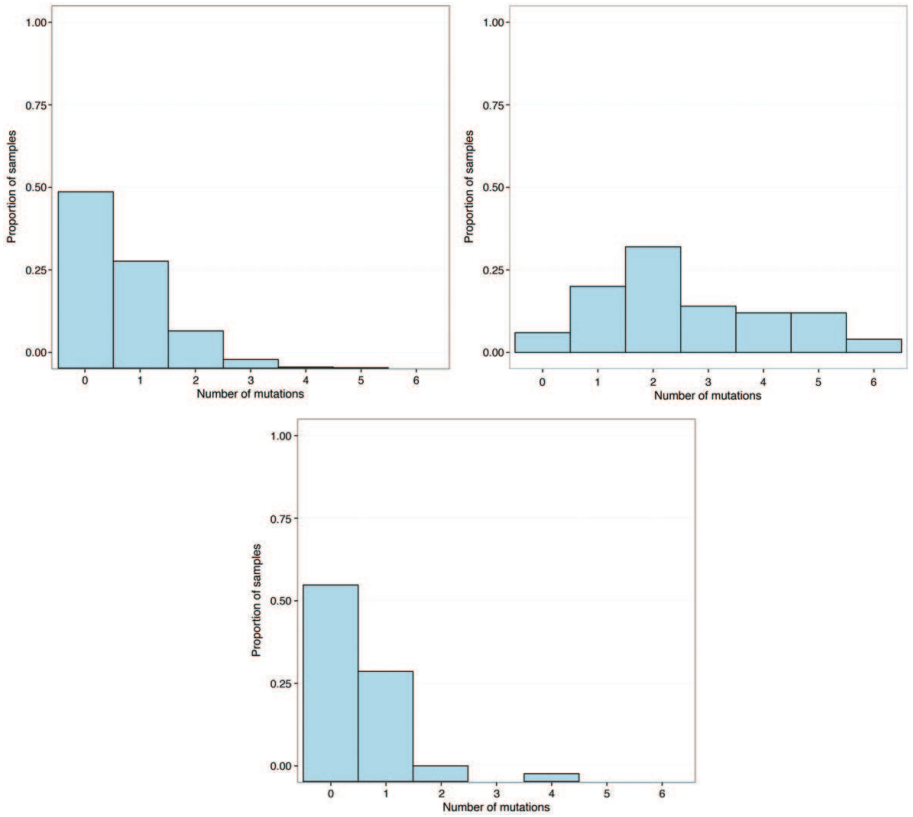


Figure 3.2: Histograms showing the proportion of samples with a certain amount of mutations that are predicted to be damaging and are found with a minor allele frequency of <1% in the 1000 genomes data. Samples are from the 1000 genomes, Complete Genomics and in-house (clockwise starting top left).

mutations.

After filtering out the variants with a MAF in the 1000 genomes of <1%, we find differences in the percentage of samples showing mutations between the data sets. For instance, the most occurring variant in the in-house data set was found in NOTCH2, which is associated with Hajdu–Cheney syndrome, a disease characterized by coarse face, short neck, hirsutism, joint laxity and bone dysplasias (Ramos et al., 1998). In this case, mutations were found in 14% of the in-house samples, in 1% of the 1kG samples and 6% of the CG samples. The fact that a predicted damaging variant is frequently found in the local population but not in the 1000 genomes data set and dbSNP can indicate that

the variant is a benign polymorphism in this population. An example of this is a variant that we identified at chromosome 3, position 98300354 (A>C) in CPOX, which is associated with hereditary coproporphyria. This mutation is found in almost 10% of the in-house samples but does not occur in the other data sets and is not found in dbSNP. In total we identified 323, 120 and 21 distinct mutations with a MAF of <1% in the 1kG data in 69, 59 and 18 genes in, respectively, the 1kG, CG and in-house data sets. An overview of the genes with the largest number of predicted damaging mutations and some phenotypic information about the disorder is shown in Supplementary data, Tables 3.S6 and 3.S7.

3.3.2 Autosomal dominant disorders: mutations present in HGMD

Taking only the mutations into account that are annotated as disease causing in HGMD for the severe disorders from our list, we identified mutations in 20% of the 1kG samples, 22% of the Complete Genomics samples and 12% of the in-house samples. The number of affected samples with curated disease-causing variants is thus much lower compared with those identified by the prediction programs. When looking at the in-house samples, we found that the mutations were considered to be damaging by either PolyPhen or SIFT but not by both. A list of the diseases of which causative mutations were identified can be found in Supplementary data, Table 3.S8.

In the in-house data, we identified a total of six heterozygous mutations in five individuals. One of these mutations (CM023740) had a MAF >1% in the 1000 genomes indicating a possibly suspicious annotation or a variant with reduced penetrance. At the time of the initial analysis, this mutation was annotated as causing extrahepatic biliary atresia, a feature of Alagille syndrome, but it was removed from HGMD at a later time. For Greig cephalopolysyndactyly syndrome two distinct mutations were found in GLI3, i.e. P707S (CM970684) (Wild et al., 1997) and I808M (CM990707) (Kalff-Suske et al., 1999). In a functional analysis, both mutations were found to cause misregulation of GLI3-localization by Krauss et al. (2009). The mutation causing Rubinstein-Taybi syndrome — A981T (CM021081) in CREBBP — was identified by Coupry et al. (2002) in a set of 60 patients. Because these samples are part of trios, we also had access to the samples of the children. Out of a total of five children, four children were heterozygous for the same mutation as their parent(s) but also did not express the disease. The fact that these variants are present in apparently healthy individuals may hint towards (1) sequencing errors, (2) false-positive entries in HGMD or (3) incomplete penetrance of certain variants that would make them of low predictive value in a PGS context.

3.3.3 Autosomal recessive disorders: mutations predicted to be damaging

For autosomal recessive disorders, there are two categories of affected individuals. Either they are homozygous for a mutation or they are compound heterozygous. We found almost the same number of samples that were homozygous for damaging mutations in genes associated with autosomal recessive disorders as we found mutations for the autosomal dominant disorders. Approximately 98–100% of samples were homozygous for at least one mutation predicted to be damaging. The number of damaging mutations per individual was lower however. In the 1kG data set we found a median of 4 mutations (min. 0, max. 9), in the Complete Genomics data set a median of 4 mutations (min. 1, max. 7) and also in the in-house data set a median of 4 mutations (min. 0, max. 7) per individual. The distribution of the mutation counts can be seen in Figure 3.3. Limiting the variants to those with a MAF of <1% in the 1000 genomes data produced a large decrease especially in the 1kG data set. Less than 2% of the samples are homozygous for a damaging mutation in the 1kG data set and 56–69% in the CG and in-house data set. The median in the 1kG data sets is 0 mutations (min. 0, max. 1), in the Complete Genomics data set 1 mutation (min. 0, max. 2) and in the in-house data set 1 mutation (min. 0, max. 1). The distribution of the mutation counts can be seen in Figure 3.4. In total, we identified 17, 4 and 1 distinct homozygous mutations with MAF in 1kG of <1% in 16, 4 and 1 genes in, respectively, the 1kG, CG and in-house data set. As for the autosomal dominant disorders, Supplementary data, Tables 3.S9 and 3.S10 show an overview of the genes with the largest number of mutations and the corresponding diseases.

In addition to homozygosity, we also investigated the case of compound heterozygosity. Because no haplotype information was available for the in-house data set, we identified the number of samples that have two different mutations predicted to be damaging in each gene. In this case, the difference between the data sets is also large. While for some diseases a number of individuals are identified in one data set, there are no individuals with damaging mutations in another and vice versa. In this case, the gene NEB, which is associated with nemaline myopathy, showed the highest number of individuals that had at least two distinct mutations with 2% of samples in the 1kG data set, 2% in the CG data set but no samples in the in-house data set affected. An overview of the number of affected samples and genes can be found in Supplementary data, Table 3.S11.

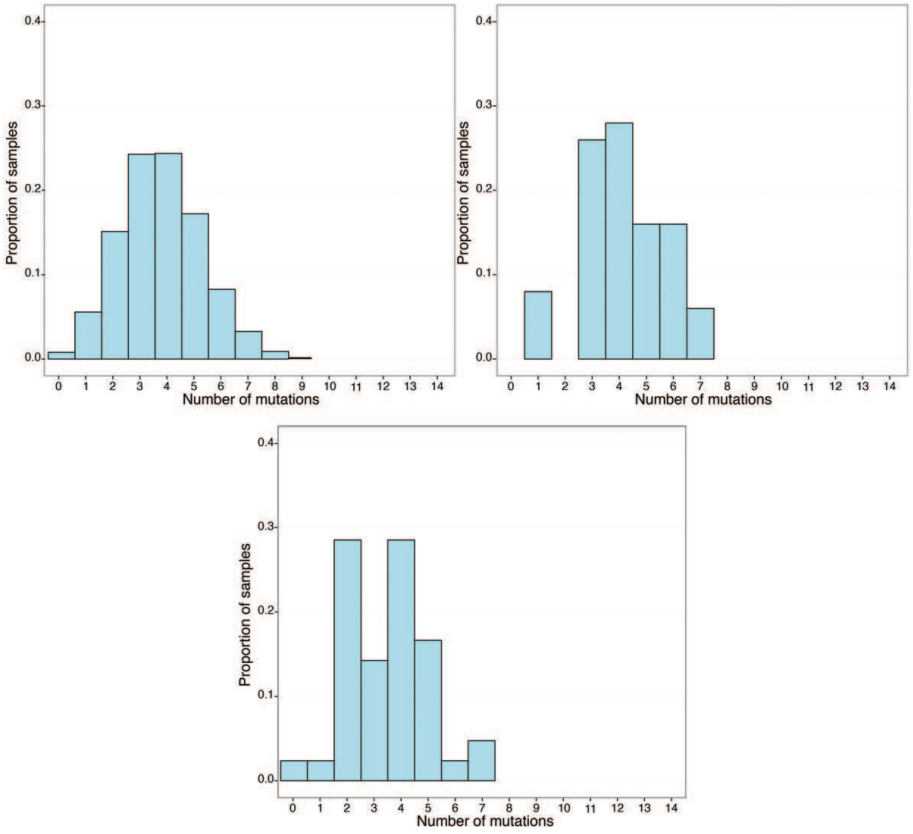


Figure 3.3: Histograms showing the proportion of samples with a certain amount of mutations that are predicted to be damaging. Samples are from the 1000 genomes, Complete Genomics and in-house (clockwise starting top left). Only homozygous mutations are counted.

3.3.4 Autosomal recessive disorders: mutations present in HGMD

For HGMD mutations, we found that the percentage of samples that are homozygous and therefore are expected to express the disease is much lower than with autosomal dominant diseases. In the in-house data set, no homozygous HGMD mutations were found. An overview of all affected genes can be found in Supplementary data, Table 3.S12. The most frequently identified mutation was found in 16 individuals in the 1kG data (1.8%) on chromosome 17, position 7915912 C>T (dbSNP id: rs34598902). This mutation was thought to be

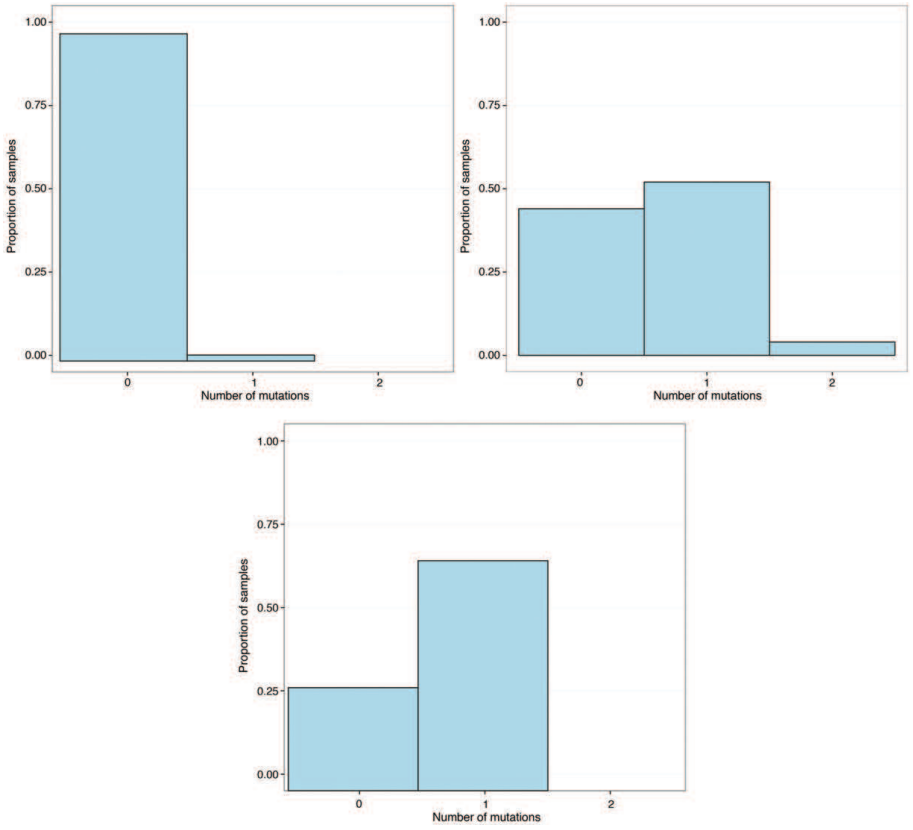


Figure 3.4: Histograms showing the proportion of samples with a certain amount of mutations that are predicted to be damaging and are found with a minor allele frequency of <1% in the 1000 genomes data. Samples are from the 1000 genomes, Complete Genomics and in-house (clockwise starting top left). Only homozygous mutations are counted.

causative for Leber Congenital Amaurosis (Zernant et al., 2005), a disease characterized by congenital blindness that affects 10–20% of all blind children (INSERM, 1997). However, it was later found in the normal population (Ito, 2004) with a MAF in the 1000 genomes of 8.26%. While the mutation was annotated as disease causing in HGMD at the time of our analysis, it has been reclassified at a later time to the category 'DM?' indicating a tenuous association. Another mutation in the same gene at position 7912879 C>T (dbSNP id: rs28743021) was found in two individuals in the 1kG data who were homozygous for this mutation. This mutation is annotated as disease causing

in HGMD based on a study by Koenekoop et al. (2002).

For compound heterozygous mutations, again counted as two distinct mutations, no positive samples were identified in the in-house data, only one in the CG data set with a MAF >1% and a few in the 1kG data set. An overview of the number of affected samples is shown in Supplementary data, Table 3.S13.

3.4 Discussion

3.4.1 Analytical validity

The analytical validity of WGS-based embryo testing is constituted by the quality of single-cell sequencing and the accuracy of interpretation. At present, single-cell genome-wide sequencing is not as good as sequencing based on multiple cells. On the assumption that this limitation will be overcome, we focus here on the accuracy of the interpretation of sequencing data.

We found that with increasing quality thresholds the number of individuals that carry mutations predicted to be damaging decreases in the samples from our in-house data set. Except for a common mutation found in the CG and in-house data set, we found fewer individuals with mutations predicted to be damaging in genes associated with autosomal recessive disorders. This may be because homozygous mutations occur less frequently than heterozygous mutations or it may be linked to the clinical validity of the test: autosomal dominant disorders are more likely to show reduced penetrance or variable expressivity and therefore symptoms of the disorder might not be present in these individuals although the mutations are.

3.4.2 Clinical validity

More than 40% of genomes of healthy individuals in our study had a genetic mutation thought to be causative for severe autosomal dominant congenital disorders. Moreover, some healthy individuals were homozygous or compound heterozygous for mutations associated with autosomal recessive disorders. These results are in line with the results of Xue et al. (2012) and MacArthur & Tyler-Smith (2010) who studied loss-of-function and disease-causing variants in healthy individuals. Our analysis relies on current prediction programs and databases. It is obvious from this analysis that current programs predicting protein damage based on exonic sequence information alone show a number of false positives and will need to improve to enable proper future phenotype prediction (Sifrim et al.,

2013). In our results we see that filtering the identified mutations on the MAF in the 1000 genomes significantly reduces the number of identified mutations per individual. In the CG and in-house data set however, a larger percentage of individuals still show mutations. This suggests that these mutations might be local polymorphisms that are not identified in the 1000 genomes project. An in-depth knowledge of the variants found in the local population can therefore lead to improved filters. Finally, the HGMD database is frequently updated and thus the results of each analysis might be different with each version of the database.

A specific challenge for embryo screening is that predictions have to be made in the absence of phenotypic information. This is different from the use of WGS-based testing in the post-natal context where this technology is being introduced as a means of finding a diagnosis for existing patients with a phenotype that could not be clarified using more traditional diagnostic approaches. However, in embryo selection, genetic information is all that is available, except for classical PGD where the phenotype of the parents can be taken into account. We considered to what extent adding WGS-based preconception testing of the prospective parents might help fill in this lacuna by providing additional context information. This approach may perhaps be helpful where findings related to dominant disorders are concerned, as information about genotype and phenotype of the parents will contribute to making better predictions about the health of children resulting from embryos with the relevant mutations found in WGS-based screening. However, preconception testing of the prospective parents seems less useful for the interpretation of findings related to recessive disorders, given our observation that healthy persons may be compound heterozygous or even homozygous for mutations in this category. Advancing knowledge in genomics may enhance the clinical validity of WGS-based embryo testing, as more will be known about the influence of modifier and protective genes and possible epigenetic influences that may explain our findings (von Kanel et al., 2013). Also, other projects, such as the 'Deciphering Developmental Disorders' (Firth & Wright, 2011) study, may contribute to identifying the genetic cause of disorders, making knowledge about the genome more complete. Although it can be expected that with increasing up-to-date knowledge about the relation between the genome and the phenotype it will become possible in the future to make better predictions about the health of children resulting from the transfer of an embryo with a specific genotype, we conclude that at present the clinical validity of WGS-based embryo screening is limited.

3.4.3 Clinical utility

As the analytical and clinical validity are still insufficient, a necessary condition for the introduction of extended or comprehensive embryo screening ('suitable test') is not met. Clearly, this would adversely affect the clinical utility of the screening, as it would lead to discarding embryos that may well develop into healthy children. Some may still defend the rapid introduction of WGS-based screening, arguing that mutations will be found that are thought to be causative at least in some cases, and that non-transfer of embryos with such mutations may still be warranted. However, as the number of available good-quality embryos in an IVF cycle is limited, there may not be much scope for choosing a mutation-free embryo, and the option of a new cycle just to avoid a 'suspected' embryo may well be disproportional. Moreover, with the current state of the art, prospective parents would be faced with choices based on unreliable predictions about the health of the children they could have as a result of transferring this or that embryo. Overestimation of the predictive value of adverse findings of WGS-based embryo screening may lead to the couple remaining childless or to undermining of their confidence in the health of any children they may still decide to have. In addition to this, they may also be confronted with equally unreliable information suggesting that they themselves are carriers of a potentially severe disease.

3.5 Conclusion and final remarks

At present, the drawbacks of WGS-based embryo screening appear to outweigh the possible benefits for prospective parents, making the introduction of such screening in clinical practice unwarranted and at best premature. It may be that further scientific developments will lead to improving the predictive accuracy of WGS-based embryo screening. Although that would take away the drawback of decision-making based on unreliable information, it does not automatically follow that WGS-based screening would then be unproblematic.

As screening aimed at simultaneously excluding a more than a limited number of genetic risk factors would very soon run into the problem of leaving no embryo for transfer (Hens et al., 2013a), a possible approach would be to always select the embryo(s) with the best health profile (while maintaining a threshold of at least not transferring high-risk embryos). However, the clinical utility of that approach would depend on whether meaningful choices between embryos with various health profiles can indeed be made. To say the least it is not clear whether that is the case. And even if it were, the amount of relevant data and the fact that (where genetic susceptibilities are concerned) even reliable

information would be about risks rather than certainties, the feasibility of well-considered decision-making would not be obvious. Moreover, whose decisions should this be? As is it can be argued that both prospective parents (whose child it will be) and professionals (given their active involvement in the creation of the child) have a say in this matter, WGS-based screening may have the potential of leading to conflicts between those stakeholders. Last but not least, a difficult problem is that choosing the embryo with the best profile will inevitably mean that children are born for whom some health prospects might already be known. The ethical question here is whether the future child should be allowed to decide for herself what she wants to know about her genome (Hens et al., 2013b). We conclude that even if current limitations in terms of analytical and clinical validity can be overcome, the notion of WGS-based embryo screening still raises some difficult questions. It is presently unclear whether these can be satisfactorily answered. A possible alternative approach with its own advantages and disadvantages (de Wert, 2009) would consist of preconception screening of the prospective parents followed by targeted PGD.

3.6 Supplementary data

3.6.1 Prediction Algorithms

When analyzing genomes in the search of causative mutations, prediction algorithms are often used to predict the effect of a mutation on the protein. We applied some of the most widely used prediction algorithms, i.e. PolyPhen (Adzhubei et al., 2010) and SIFT (Ng, 2003) to predict whether a mutation would be damaging. The accuracy of the SIFT and PolyPhen algorithms is reported to be as high as 91.7% (Zou et al., 2011) but also as low as 60% depending on the setting in which these algorithms were used (Di et al., 2009). The accuracy of different prediction algorithms is also dependent on the conservation score of the position. While up to 92% accuracy is obtained for laboratory-induced mutations that severely disrupt protein function, only 26–46% of neutral mutations are correctly classified as such at highly conserved positions (Gray et al., 2012).

3.6.2 Quality

The graphs from the in-house data set in the paper show the distribution of mutations with at least 30× coverage and a minimum phred score of 30. When relaxing these constraints the number of identified mutations and/or

affected samples increases as can be seen in Supplementary data, Table 3.S4 for autosomal dominant disorders and in Supplementary data, Table 3.S5 for autosomal recessive disorders.

Table 3.S1: List of Online Mendelian Inheritance in Man (OMIM) IDs and corresponding disorders with autosomal dominant inheritance used in the analysis.

OMIM ID	Description	OMIM ID	Description
200610	Achondrogenesis, type II	214800	Charge syndrome
100800	Achondroplasia	118400	Cherubism
101800	Acrodysostosis with hormone resistance	119600	Cleidocranial dysplasia
102370	Acromicric dysplasia	262700	Combined pituitary hormone deficiency 4
100300	Adams-oliver syndrome	255310	Congenital myopathy with fiber-type disproportion
103285	Adult syndrome	122470	Cornelia de Lange syndrome 1
118450	Alagille syndrome1	218040	Costello syndrome
170390	Andersen cardiodysrhythmic periodic paralysis	123000	Craniometaphyseal dysplasia
106260	Ankyloblepharon-ectodermal defects-cleft lip/palate	123100	Craniosynostosis, type 1
101200	Apert syndrome	123500	Crouzon syndrome
108721	Atelosteogenesis type III	108120	Distal arthrogryposis type Ia
180500	Axenfeld-rieger syndrome, type I	193700	Distal arthrogryposis type IIa
153480	Bannayan-Riley-Ruvalcaba syndrome	158300	Distal arthrogryposis type IX
123790	Beare-Stevenson cutis gyrata syndrome	121050	Distal arthrogryposis type VII
130650	Beckwith-Wiedemann syndrome	607323	Duane-Radial ray syndrome
110100	Blepharophimosis, ptosis, and epicanthus inversus	604292	Ectrodactyly, ectodermal dysplasia, and cleft lip/palate syndrome
605039	Bohring-Opitz syndrome		
112310	Boomerang dysplasia	160120	Episodic ataxia, type I
113500	Brachyolmia type III	164280	Feingold syndrome 1
113620	Branchiooculofacial syndrome	614524	Fibrochondrogenesis 2
613071	Bronchiectasis with or without elevated sweat chloride 3	135100	Fibrodysplasia ossificans progressiva
114290	Campomelic dysplasia	175700	Greig cephalopolysyndactyly syndrome
131300	Camurati-Engelmann disease	102500	Hajdu-Cheney syndrome
115150	Cardiofaciocutaneous syndrome	121300	Hereditary coproporphyria
169100	Char syndrome	129500	Hidrotic ectodermal dysplasia
142946	Holoprosencephaly 4	186500	Multiple synostoses syndrome 1

Continued on next page

Table 3.S1 – continued from previous page

OMIM ID	Description	OMIM ID	Description
609637	Holoprosencephaly 5	139210	Myhre syndrome
176670	Hutchinson-Gilford progeria syndrome	161200	Nail-patella syndrome
145900	Hypertrophic neuropathy of Dejerine-Sottas	601321	Neurofibromatosis-Noonan syndrome
146000	Hypochondroplasia	163950	Noonan syndrome 1
129490	Hypohidrotic ectodermal dysplasia	164200	Oculodentodigital dysplasia
147750	Ivic syndrome	165800	Osteochondritis dissecans, short stature, and early-onset
123150	Jackson-Weiss syndrome		osteoarthritis
147920	Kabuki syndrome 1	166200	Osteogenesis imperfecta, type I
148050	KBG syndrome	166210	Osteogenesis imperfecta, type II
610253	Kleefstra syndrome	259420	Osteogenesis imperfecta, type III
118100	Klippel-Feil syndrome 1	166220	Osteogenesis imperfecta, type IV
156550	Kniest dysplasia	166250	Osteoglophonic dysplasia
150250	Larsen syndrome	146510	Pallister-Hall syndrome
151100	Leopard syndrome 1	168400	Parastremmatic dwarfism
127300	Leri-Weill dyschondrosteosis	168500	Parietal foramina
609192	Loeys-Dietz syndrome, type 1a	168550	Parietal foramina with cleidocranial dysplasia
610168	Loeys-Dietz syndrome, type 1b	118800	Paroxysmal nonkinesigenic dyskinesia 1
153400	Lymphedema-distichiasis syndrome	169400	Pelger-Huet anomaly
109150	Machado-Joseph disease	101600	Pfeiffer syndrome
154700	Marfan syndrome	151210	Platyspondylic lethal skeletal dysplasia, Torrance type
154780	Marshall syndrome	119500	Popliteal pterygium syndrome
156400	Metaphyseal chondrodysplasia, jansen type	177170	Pseudoachondroplasia
156500	Metaphyseal chondrodysplasia, schmid type	129400	Rapp-Hodgkin syndrome
156530	Metatropic dysplasia	270450	Resistance to insulin-like growth factor I
607932	Microphthalmia, syndromic 6	180700	Robinow syndrome
235730	Mowat-Wilson syndrome	180849	Rubinstein-Taybi syndrome 1
602849	Muenke syndrome	101400	Saethre-Chotzen syndrome
269150	Schinzel-Giedion midface retraction syndrome	184840	Stickler syndrome, type III

Continued on next page

Table 3.S1 – continued from previous page

OMIM ID	Description	OMIM ID	Description
269921	Sialuria	184900	Stiff skin syndrome
147891	Small patella syndrome	206900	Syndromic microphthalmia 3
182290	Smith-Magenis syndrome	187600	Thanatophoric dysplasia, type I
147250	Solitary median maxillary central incisor	187601	Thanatophoric dysplasia, type II
603546	Spondyloepimetaphyseal dysplasia with joint laxity type 2	107480	Townes-Brocks syndrome
602111	Spondyloepimetaphyseal dysplasia, Missouri type	154500	Treacher Collins syndrome 1
184250	Spondyloepimetaphyseal dysplasia, Strudwick type	190350	Trichorhinophalangeal syndrome, type I
183900	Spondyloepiphyseal dysplasia congenita	190351	Trichorhinophalangeal syndrome, type III
184095	Spondyloepiphyseal dysplasia, Maroteaux type	119300	Van der Woude syndrome 1
184252	Spondylometaphyseal dysplasia, Kozlowski type	148820	Waardenburg syndrome type III
271700	Spondyloperipheral dysplasia	277590	Weaver syndrome 1
108300	Stickler syndrome type I	608328	Weill-marchesani syndrome 2
604841	Stickler syndrome, type II	277610	Weissenbacher-Zweymuller syndrome

Table 3.S2: List of Online Mendelian Inheritance in Man (OMIM) IDs and corresponding disorders with autosomal recessive inheritance used in the analysis.

OMIM ID	Description	OMIM ID	Description
246450	3-@hydroxy-3-methylglutaryl-CoA lyase deficiency		polymicrogyria
257920	3MC syndrome 1	209900	Bardet-Biedl syndrome
265050	3MC syndrome 2	211530	Brown-Vialetto-van Laere syndrome
100100	Absence of abdominal muscles with urinary tract abnormality	211750	C syndrome
276820	Absence of ulna and fibula with severe limb deficiency	271900	Canavan disease
231550	Achalasia-addisonianism-alacrima syndrome	608836	Carnitine palmitoyltransferase II deficiency, lethal neonatal
200500	Acheiropody	201000	Carpenter syndrome
200600	Achondrogenesis, type Ia	224050	Cerebellar ataxia, mental retardation, and dysequilibrium
200990	Acrocallosal syndrome		syndrome
607778	Acrocapitofemoral dysplasia	214150	Cerebrooculofacioskeletal syndrome 1
602875	Acromesomelic dysplasia, Maroteaux type	214500	Chediak-Higashi syndrome
202370	Adrenoleukodystrophy	215045	Chondrodysplasia, Blomstrand type
218000	Agenesis of the corpus callosum with peripheral neuropathy	200700	Chondrodysplasia, Grebe type
608688	Aicar transformylase/imp cyclohydrolase deficiency	225060	Cleft lip/palate-ectodermal dysplasia syndrome
225750	Aicardi-Goutieres syndrome 1	216360	Coach syndrome
203800	Alstrom syndrome	216400	Cockayne syndrome, type a
205100	Amyotrophic lateral sclerosis 2, juvenile	133540	Cockayne syndrome, type b
241200	Antenatal Bartter syndrome, type 2	216550	Cohen syndrome
201750	Antley-Bixler syndrome with genital anomalies and	262600	Combined pituitary hormone deficiency 2
	disordered steroidogenesis	212065	Congenital disorder of glycosylation, type Ia
207410	Antley-Bixler syndrome without genital anomalies or	601110	Congenital disorder of glycosylation, type Id
	disordered steroidogenesis	608799	Congenital disorder of glycosylation, type Ie
208400	Aspartylglucosaminuria	212066	Congenital disorder of glycosylation, type IIa
256050	Atelosteogenesis, type II	608540	Congenital disorder of glycosylation, type Ik
251290	Band-like calcification with simplified gyration and	610768	Congenital disorder of glycosylation, type Im
263700	Congenital erythropoietic porphyria	226600	Epidermolysis bullosa dystrophica

Continued on next page

Table 3.S2 – continued from previous page

OMIM ID	Description	OMIM ID	Description
612540	Congenital myopathy Compton-North	226730	Epidermolysis bullosa junctionalis with pyloric atresia
275100	Congential nongoitrous hypothyroidism 4	226900	Epiphyseal dysplasia, multiple, 4
260660	Cousin syndrome	226980	Epiphyseal dysplasia, multiple, with early-onset
218330	Cranioectodermal dysplasia 1		diabetes mellitus
607812	Craniolenticulosutural dysplasia	211500	Fazio-Londe disease
601378	Crisponi syndrome	208150	Fetal akinesia deformation sequence
219700	Cystic fibrosis	228520	Fibrochondrogenesis
261515	D-bifunctional protein deficiency	228930	Fibular aplasia or hypoplasia, femoral bowing and
600060	Deafness 2		poly-, syn-, and oligodactyly
251450	Desbuquois dysplasia	219000	Fraser syndrome
602398	Desmosterolosis	256540	Galactosialidosis
608022	Diaphanospondylodysostosis	230800	Gaucher disease, type I
251850	Diarrhea 2, with microvillus atrophy	230900	Gaucher disease, type II
222600	Diastrophic dysplasia	231005	Gaucher disease, type IIIc
604320	Distal spinal muscular atrophy 1	231050	Geleophysic dysplasia 1
246200	Donohue syndrome	231670	Glutaric acidemia I
223800	Dyggve-Melchior-Clausen disease	232200	Glycogen storage disease Ia
224230	Dyskeratosis congenita 1	232240	Glycogen storage disease Ic
224410	Dyssegmental dysplasia, Silverman-Handmaker type	232500	Glycogen storage disease IV
609304	Early infantile epileptic encephalopathy 3	230500	GM1-gangliosidosis, type I
225320	Ehlers-Danlos syndrome, cardiac valvular form	230600	GM1-gangliosidosis, type II
225400	Ehlers-Danlos syndrome, type VI	230650	GM1-gangliosidosis, type III
225410	Ehlers-Danlos syndrome, type VII	609460	Goldberg-Shprintzen megacolon syndrome
600002	Eiken skeletal dysplasia	235510	Hennekam lymphangiectasia-lymphedema syndrome
225500	Ellis-van Creveld syndrome	241530	Hereditary hypophosphatemic rickets with hypercalciuria
612651	Endocrine-cerebroosteodysplasia	608233	Hermansky-Pudlak syndrome 2
603034	Endplate acetylcholinesterase deficiency	236490	Hyalinosis, infantile systemic
236680	Hydrolethalus syndrome 1	608728	Matrilin-3 related spondyloepimetaphyseal dysplasia

Continued on next page

Table 3.S2 – continued from previous page

OMIM ID	Description	OMIM ID	Description
215140	Hydrops-ectopic calcification-moth-eaten skeletal dysplasia	249000	Meckel syndrome, type 1
243700	Hyper-IgE recurrent infection syndrome	224690	Meier-Gorlin syndrome 1
239300	Hyperphosphatasia with mental retardation	613803	Meier-Gorlin syndrome 3
241410	Hypoparathyroidism-retardation-dysmorphism syndrome	613804	Meier-Gorlin syndrome 4
242900	Immunoosseous dysplasia, Schimke type	613805	Meier-Gorlin syndrome 5
241500	Infantile hypophosphatasia	277380	Methylmalonic aciduria and homocystinuria, cblF type
269920	Infantile sialic acid storage disorder	610377	Mevalonic aciduria
262400	Isolated growth hormone deficiency, type Ia	210720	Microcephalic osteodysplastic primordial dwarfism, type II
243800	Johanson-Blizzard syndrome	206920	Microphthalmia with limb anomalies
213300	Joubert syndrome	255320	Minicore myopathy with external ophthalmoplegia
608091	Joubert syndrome 2	257300	Mosaic variegated aneuploidy syndrome 1
608629	Joubert syndrome 3	252500	Mucolipidosis II alpha/beta
226700	Junctional epidermolysis bullosa, Herlitz type	252900	Mucopolysaccharidosis type IIIa
262650	Kowarski syndrome	252920	Mucopolysaccharidosis type IIIb
611722	Krabbe disease, atypical, due to saposin a deficiency	252930	Mucopolysaccharidosis type IIIc
249700	Langer mesomelic dysplasia	252940	Mucopolysaccharidosis type IIId
262500	Laron syndrome	253000	Mucopolysaccharidosis type IVa
204000	Leber congenital amaurosis 1	253010	Mucopolysaccharidosis type IVb
204100	Leber congenital amaurosis 2	253200	Mucopolysaccharidosis type VI
253310	Lethal congenital contracture syndrome 1	253220	Mucopolysaccharidosis type VII
260600	Leukodystrophy, hypomyelinating, 3	253250	Mulibrey nanism
278000	Lysosomal acid lipase deficiency	245600	Multiple joint dislocations, short stature, craniofacial
248370	Mandibuloacral dysplasia with type a lipodystrophy		dysmorphism
608612	Mandibuloacral dysplasia with type b lipodystrophy	265000	Multiple pterygium syndrome, Escobar variant
248500	Mannosidosis, alpha b, lysosomal	253290	Multiple pterygium syndrome, lethal type
248800	Marinesco-Sjogren syndrome	236670	Muscular dystrophy-dystroglycanopathy
212720	Martsolf syndrome	256030	Nemaline myopathy 2
256500	Netherton syndrome	222765	Rhizomelic chondrodysplasia punctata, type 2

Continued on next page

Table 3.S2 – continued from previous page

OMIM ID	Description	OMIM ID	Description
610127	Neuronal ceroid lipofuscinosis 10	600121	Rhizomelic chondrodysplasia punctata, type 3
204500	Neuronal ceroid lipofuscinosis 2	268300	Roberts syndrome
204200	Neuronal ceroid lipofuscinosis 3	268310	Robinow syndrome
256731	Neuronal ceroid lipofuscinosis, 5	269000	SC phocomelia syndrome
257200	Niemann-Pick disease, type a	269250	Schneckenbecken dysplasia
257220	Niemann-Pick disease, type c1	255800	Schwartz-Jampel syndrome, type 1
251260	Nijmegen breakage syndrome	269500	Sclerosteosis 1
258315	Omodysplasia 1	210600	Seckel syndrome 1
610682	Osteogenesis imperfecta, type VII	612541	Severe congenital neutropenia 4
610915	Osteogenesis imperfecta, type VIII	263520	Short rib-polydactyly syndrome, type II
259700	Osteopetrosis 1	260400	Shwachman-Diamond syndrome
259730	Osteopetrosis 3	270400	Smith-Lemli-Opitz syndrome
259720	Osteopetrosis 5	612936	Spastic quadriplegic cerebral palsy 3
215150	Otospondylomegaepiphyseal dysplasia	272460	Spondylocarpotarsal synostosis syndrome
239000	Paget disease, juvenile	612350	Spondylocheirodysplasia, Ehlers-Danlos syndrome-like
608013	Perinatal lethal Gaucher disease	277300	Spondylocostal dysostosis 1
261540	Peters-Plus syndrome	143095	Spondyloepiphyseal dysplasia with congenital joint
262190	Pineal hyperplasia, insulin-resistant diabetes mellitus		dislocations
263200	Polycystic kidney disease	271665	Spondylometaepiphyseal dysplasia, short limb-hand type
225753	Pontocerebellar hypoplasia, type IV	601559	Stuve-Wiedemann syndrome
263650	Popliteal pterygium syndrome, lethal type	272300	Sulfocysteinuria
263750	Postaxial acrofacial dysostosis	272800	Tay-Sachs disease
259900	Primary hyperoxaluria, type I	273395	Tetraamelia
251200	Primary microcephaly 1	273750	Three M syndrome 1
259775	Raine syndrome	259600	Torg-Winchester syndrome
275210	Restrictive dermopathy, lethal	609015	Trifunctional protein deficiency
215100	Rhizomelic chondrodysplasia punctata, type 1	264700	Vitamin D hydroxylation-deficient rickets, type 1a
600081	Vitamin D hydroxylation-deficient rickets, type 1b	277600	Weill-Marchesani syndrome 1

Continued on next page

Table 3.S2 – continued from previous page

OMIM ID	Description	OMIM ID	Description
277440	Vitamin D-dependent rickets, type 2a	214100	Zellweger syndrome
600118	Warburg micro syndrome 1		

Table 3.S3: List of Coriel IDs used from the Complete Genomics dataset.

NA19240-L2-200-37ASM	NA12888-200-37-ASM	NA19700-200-37-ASM
NA19649-200-37-ASM	NA12890-200-37-ASM	NA19701-200-37-ASM
NA18505-200-37-ASM	NA12891-200-37-ASM	NA19703-200-37-ASM
NA19025-200-37-ASM	NA12893-200-37-ASM	NA19704-200-37-ASM
NA19238-L2-200-37-ASM	NA18501-200-37-ASM	NA19735-200-37-ASM
NA20846-200-37-ASM	NA18502-200-37-ASM	NA20502-200-37-ASM
NA19670-200-37-ASM	NA18504-200-37-ASM	NA20509-200-37-ASM
NA19834-200-37-ASM	NA18508-200-37-ASM	NA20510-200-37-ASM
NA18947-200-37-ASM	NA18517-200-37-ASM	NA20511-200-37-ASM
HG00731-200-37-ASM	NA18526-200-37-ASM	NA20845-200-37-ASM
HG00732-200-37-ASM	NA18537-200-37-ASM	NA20847-200-37-ASM
HG00733-200-37-ASM	NA18555-200-37-ASM	NA20850-200-37-ASM
NA06985-200-37-ASM	NA18558-200-37-ASM	NA21732-200-37-ASM
NA06994-200-37-ASM	NA18940-200-37-ASM	NA21733-200-37-ASM
NA07357-200-37-ASM	NA18942-200-37-ASM	NA21737-200-37-ASM
NA10851-200-37-ASM	NA18956-200-37-ASM	NA21767-200-37-ASM
NA12004-200-37-ASM	NA19017-200-37-ASM	NA12880-200-37-ASM
NA12877-200-37-ASM	NA19020-200-37-ASM	NA12881-200-37-ASM
NA12879-200-37-ASM	NA19026-200-37-ASM	NA12883-200-37-ASM
NA12882-200-37-ASM	NA19129-200-37-ASM	NA12886-L2-200-37-ASM
NA12884-200-37-ASM	NA19239-L2-200-37-ASM	NA12889-L2-200-37-ASM
NA12885-L2-200-37-ASM	NA19648-200-37-ASM	NA12878-200-37-ASM
NA12887-L2-200-37-ASM	NA19669-200-37-ASM	NA12892-L2-200-37-ASM

Table 3.S4: Number of damaging mutations found in in-house samples in genes associated with autosomal dominant disorders. The minimum (Min), maximum (Max), and median number of mutations is given for the different threshold of read depth and Phred score. The number between brackets indicate the number found without filtering on a minor allele frequency of <1% in the 1000 genomes.

Phred score/read depth	**10**	**20**	**30**
Number of Individuals	22 (42)	19 (42)	17 (41)
Min	0 (1)	0 (1)	0 (0)
Max	4 (8)	4 (7)	4 (5)
Median	1 (4)	0 (3)	0 (2)

Table 3.S5: Number of damaging, homozygous mutations found in in-house samples in genes associated with autosomal recessive disorders. The minimum (Min), maximum (Max), and median number of mutations is given for the different thresholds of read depth and Phred score. The number between brackets indicate the number found without filtering on a minor allele frequency of <1% in the 1000 genomes.

Phred/read depth	**10**	**20**	**30**
Individuals	34 (42)	33 (42)	29 (41)
Min	0 (1)	0 (1)	0 (0)
Max	1 (8)	1 (8)	1 (7)
Median	1 (4)	1 (4)	1 (4)

Table 3.S6: List of genes with mutations predicted to be damaging, the corresponding diseases and the number and proportion of samples in which these mutations are found. The number between brackets indicate the number found without filtering on a minor allele frequency of <1% in the 1000 genomes.

Disease	**Gene**	**1kG**	**% 1kG**	**CG**	**% CG**	**IH**	**% IH**
Hajdu-Cheney syndrome	NOTCH2	14 (675)	1.39% (67.23%)	3 (7)	6.00% (14.00%)	6 (6)	14.29% (14.29%)
Coproporphyria, hereditary	CPOX	3 (26)	0.30% (2.59%)	1 (1)	2.00% (2.00%)	4 (4)	9.52% (9.52%)
Hypertrophic Neuropathy of Dejerine-Sotas	EGR2/MPZ/PMP22/PRX	10 (10)	1.00% (1.00%)	3 (3)	6.00% (6.00%)	2 (2)	4.76% (4.76%)
Arthrogryposis, distal, type 2a	MYH3	21 (21)	2.09% (2.09%)	2 (2)	4.00% (4.00%)	2 (2)	4.76% (4.76%)
Smith-Magenis syndrome	RAI1	39 (729)	3.88% (72.61%)	11 (37)	22.00% (74.00%)	0 (4)	0.00% (9.52%)
Ectodermal dysplasia	EDARADD EDAR	23 (988)	2.29% (98.41%)	1 (49)	2.00% (98.00%)	1 (35)	2.38% (83.33%)
KBG Syndrome	ANKRD11	23 (157)	2.29% (15.64%)	8 (13)	16.00% (26.00%)	0 (0)	0.00% (0.00%)
Osteochondritis dissecans, short stature and early-onset osteoarthritis	ACAN	70 (70)	6.97% (6.97%)	6 (6)	12.00% (12.00%)	0 (0)	0.00% (0.00%)
Trichorhinophalangeal syndrome, type I and II	TRPS1	17 (17)	1.69% (1.69%)	2 (2)	4.00% (4.00%)	1 (1)	2.38% (2.38%)
Marshall syndrome, Stickler syndrome, type II	COL11A1	15 (15)	1.49% (1.49%)	1 (1)	2.00% (2.00%)	1 (1)	2.38% (2.38%)

Table 3.S7: Overview of genes with the highest number of damaging mutations, the corresponding diseases and a short description of that disease. Compiled from information from Online Mendelian Inheritance in Man (McKusick-Nathans Institute of Genetic Medicine, 2012), Orphanet (INSERM, 1997) and U.S. National Library of Medicine. (National Institutes of Health, 1993)

Gene	Disease	Description
NOTCH2	Hajdu-Cheney Syndrome	Skeletal disorder characterized by short stature, coarse and dysmorphic facies, bowing of the long bones, and vertebral anomalies. Facial features include hypertelorism, bushy eyebrows, micrognathia, small mouth with dental anomalies, low-set ears, and short neck. There is progressive focal bone destruction, including acroosteolysis and generalized osteoporosis. Additional and variable features include hearing loss, renal cysts, and cardiovascular anomalies.
CPOX	Hereditary coproporphyria	Acute hepatic porphyria (see this term) characterized by the occurrence of neuro-visceral attacks and, more rarely, by the presence of cutaneous lesions. The prevalence in Europe is estimated at about 1/1,000,000. The disease manifests after puberty and preferentially affects women.
EGR2 MPZ PMP22 PRX	Hypertrophic Neuropathy of Dejerine-Sottas	Dejerine-Sottas neuropathy is a demyelinating peripheral neuropathy with onset in infancy. Affected individuals have delayed motor development due to severe distal motor and sensory impairment, resulting in difficulties in gait. Some patients have generalized hypotonia in infancy. Other features may include pes cavus, scoliosis, and sensory ataxia.
MYH3	Arthrogryposis, Distal, Type 2A	A rare congenital disorder characterised by dysmorphic status combining bone anomalies and joint contractures with typical facies. Affects less than 1 in 1,000,000.
RAI1	Smith-Magenis Syndrome	Variable intellectual deficit, sleep disturbance, craniofacial and skeletal anomalies, psychiatric disorders, and speech and motor delay. Affects 1 in 15,000 people.
EDARADD EDAR	Hypohidrotic Ectodermal Dysplasia	Hypotrichosis, hypodontia, and anhidrosis. Affects 1 in 17,000 people

Continued on next page

Table 3.S7 – continued from previous page

Gene	Disease	Description
ANKRD11	KBG Syndrome	KBG syndrome is a rare condition characterised by a typical facial dysmorphism, macrodontia of the upper central incisors, skeletal (mainly costovertebral) anomalies and developmental delay. To date it has been reported in 45 individuals.
ACAN	Osteochondritis dissecans, short stature and early-onset osteoarthritis	Familial osteochondritis dissecans is a rare genetic skeletal disorder characterized clinically by abnormal chondro-skeletal development, disproportionate short stature and skeletal deformation mainly affecting the knees, hips, ankles and elbows with onset generally in late childhood or adolescence.
TRPS1	Trichorhinophalangeal syndrome, type I	Trichorhinophalangeal syndrome type 1 is a malformation syndrome characterized by short stature, sparse hair, a bulbous nasal tip and cone-shaped epiphyses, as well as severe generalized shortening of all phalanges, metacarpals and metatarsal bones. Affects less than 1 in 1,000,000
	Trichorhinophalangeal syndrome, type II	Trichorhinophalangeal syndrome type 2 is characterized by the association of intellectual deficit and numerous other anomalies including redundant skin, multiple cartilaginous exostoses, characteristic facies and cone-shaped phalangeal epiphyses.
COL11A1	Marshall syndrome	Marshall syndrome is a multisystem disease characterized by ocular abnormalities, sensorineural hearing loss, craniofacial anomalies and anhidrotic ectodermal dysplasia. Affects less than 1 in 1,000,000
	Stickler syndrome, type III	Stickler syndrome is a clinically variable and genetically heterogeneous disorder characterized by ocular, auditory, skeletal, and orofacial abnormalities. Most forms of Stickler syndrome are characterized by the eye findings of high myopia, vitreoretinal degeneration, retinal detachment, and cataracts. Additional findings may include midline clefting (cleft palate or bifid uvula), Pierre Robin sequence, flat midface, sensorineural or conductive hearing loss, mild spondyloepiphyseal dysplasia, and early-onset osteoarthritis.

Table 3.S8: List of genes with mutations predicted to be disease causing (DM) in the Human Gene Mutation Database (HGMD), the corresponding diseases and the number and proportion of samples in which these mutations are found. The number between brackets indicate the number found without filtering on a minor allele frequency of <1% in the 1000 genomes.

Disease	**Gene**	**1kG**	**% 1kG**	**CG**	**% CG**	**IH**	**% IH**
Alagille syndrome (extrahepatic biliary atresia)	JAG1	6 (85)	0.60% (8.47%)	0 (4)	0.00% (8.00%)	0 (1)	0.00% (2.38%)
Greig cephalopolysyndactyly syndrome	GLI3	2 (2)	0.20% (0.20%)	0 (0)	0.00% (0.00%)	3 (3)	7.14% (7.14%)
Kallmann syndrome 3	PROKR2	16 (16)	1.59% (1.59%)	3 (3)	6.00% (6.00%)	0 (0)	0.00% (0.00%)
Marfan syndrome	FBN1	8 (8)	0.80% (0.80%)	0 (0)	0.00% (0.00%)	0 (0)	0.00% (0.00%)
Rubinstein-Taybi syndrome 1	CREBBP	18 (18)	1.79% (1.79%)	0 (0)	0.00% (0.00%)	2 (2)	4.76% (4.76%)

Table 3.S9: List of genes with homozygous mutations predicted to be damaging, the corresponding diseases and the number and proportion of samples in which these mutations are found. The number between brackets indicate the number found without filtering on a minor allele frequency of <1% in the 1000 genomes.

Disease	**Gene**	**1kG**	**% 1kG**	**CG**	**% CG**	**IH**	**% IH**
Joubert syndrome 2	TMEM216	0 (0)	0.00% (0.00%)	27 (27)	54.00% (54.00%)	29 (29)	69.05% (69.05%)
Alstrom syndrome	ALMS1	2 (4)	0.20% (0.40%)	1 (1)	2.00% (2.00%)	0 (0)	0.00% (0.00%)
Epidermolysis bullosa, junctional, Herlitz type	LAMA3 LAMB3 LAMC2	1 (12)	0.10% (1.20%)	1 (1)	2.00% (2.00%)	0 (0)	0.00% (0.00%)
Mannosidosis, Alpha B	MAN2B1	0 (1)	0.00% (0.10%)	1 (1)	2.00% (2.00%)	0 (0)	0.00% (0.00%)
Band-like calcification with simplified gyration	OCLN	2 (2)	0.20% (0.20%)	0 (0)	0.00% (0.00%)	0 (0)	0.00% (0.00%)
Ellis-Van Creveld syndrome	EVC/EVC2	2 (11)	0.20% (1.10%)	0 (0)	0.00% (0.00%)	0 (0)	0.00% (0.00%)

Table 3.S10: Overview of genes associated with autosomal recessive disorders with the highest number of damaging mutations, the corresponding diseases and a short description of that disease. Compiled from information from Online Mendelian Inheritance in Man (McKusick-Nathans Institute of Genetic Medicine, 2012), Orphanet (INSERM, 1997) and U.S. National Library of Medicine. (National Institutes of Health, 1993)

Gene	Disease	Description
NTMEM216	Joubert Syndrome 2	Joubert syndrome is a genetically heterogeneous autosomal recessive disorder characterized by a specific hindbrain malformation, which is referred to as the 'molar tooth sign' (MTS) on brain MRI, hypotonia, developmental delay, oculomotor apraxia, and breathing abnormalities.
ALMS1	Alstrom syndrome	Alström syndrome is a multisystemic disorder characterized by cone-rod dystrophy, hearing loss, obesity, insulin resistance and hyperinsulinemia, type 2 diabetes mellitus, dilated cardiomyopathy (DCM), and progressive hepatic and renal dysfunction. Approximately 450 cases have been identified worldwide.
LAMA3 LAMB3 LAMC2	Epidermolysis bullosa, junctional, Herlitz type	Junctional epidermolysis bullosa, Herlitz-type is a severe subtype of junctional epidermolysis bullosa (JEB, see this term) characterized by blisters and extensive erosions, localized to the skin and mucous membranes. Prevalence is 1-9 in 1,000,000.
MAN2B1	Alpha-mannosidosis B	Alpha-mannosidosis is an inherited lysosomal storage disorder characterized by immune deficiency, facial and skeletal abnormalities, hearing impairment, and intellectual deficit. It occurs in approximately 1 in 500,000 live births.
OCLN	Band-like calcification with simplified gyration and polymicrogyria	Congenital intrauterine infection-like syndrome is characterised by the presence of microcephaly and intracranial calcifications at birth accompanied by neurological delay, seizures and a clinical course similar to that seen in patients after intrauterine infection with Toxoplasma gondii, Rubella, Cytomegalovirus, Herpes simplex. Prevalence is less than 1 in 1,000,000.

Continued on next page

Table 3.S10 – continued from previous page

Gene	Disease	Description
EVC/EVC2	Ellis-Van Creveld syndrome	Ellis-van Creveld syndrome (EVC) is a chondral and ectodermal dysplasia characterized by short ribs, polydactyly, growth retardation, and ectodermal and heart defects. It is a rare disease with approximately 150 cases reported worldwide.

Table 3.S11: List of genes that have at least two distinct mutations that are predicted to be damaging. The corresponding diseases and the number and proportion of samples in which these mutations are found is also given. The number between brackets indicate the number found without filtering on a minor allele frequency of <1% in the 1000 genomes.

Disease	**Gene**	**1kG**	**% 1kG**	**CG**	**% CG**	**IH**	**% IH**
Acrocallosal syndrome	KIF7	0 (20)	1.99% (1.99%)	0 (2)	0.00% (2.00%)	2 (2)	4.76% (4.76%)
Dyssegmental dysplasia, Silverman-Handmaker type	HSPG2	2 (10)	0.20% (1.00%)	0 (0)	0.00% (0.00%)	1 (3)	2.38% (7.14%)
Hyper-IgE recurrent infection syndrome	DOCK8	0 (2)	0.00% (0.20%)	0 (1)	0.00% (2.00%)	1 (2)	2.38% (4.76%)
Schwartz-Jampel syndrome, type I	HSPG2	2 (10)	0.20% (1.00%)	0 (0)	0.00% (0.00%)	1 (3)	2.38% (7.14%)
Nemaline myopathy 2	NEB	23 (68)	2.29% (6.77%)	1 (14)	2.00% (28.00%)	0 (0)	0.00% (0.00%)
Fraser syndrome	FRAS1/ FREM2	10 (16)	1.00% (1.59%)	0 (0)	0.00% (0.00%)	0 (0)	0.00% (0.00%)
Cystic fibrosis	CFTR	8 (8)	0.80% (0.80%)	1 (1)	2.00% (2.00%)	0 (0)	0.00% (0.00%)
Pontocerebellar hypoplasia, type IV	TSEN54	0 (156)	0.00% (15.54%)	0 (3)	0.00% (6.00%)	0 (2)	0.00% (4.76%)

Table 3.S12: List of genes with mutations predicted to be disease causing (DM) in the Human Gene Mutation Database (HGMD), the corresponding diseases and the number and proportion of samples in which these mutations are found. The number between brackets indicate the number found without filtering on a minor allele frequency of <1% in the 1000 genomes.

Disease	**Gene**	**1kG**	**% 1kG**	**CG**	**% CG**	**IH**	**% IH**
Polycystic kidney disease	PKHD1	1	0.10%	0	0.00%	0	0.00%
		(2)	(0.20%)	(0)	(0.00%)	(0)	(0.00%)
Leber congenital amaurosis I	GUCY2D	0	0.00%	0	0.00%	0	0.00%
		(18)	(1.79%)	(0)	(0.00%)	(0)	(0.00%)
Ellis-van Creveld syndrome	EVC	0	0.00%	0	0.00%	0	0.00%
		(10)	(1.00%)	(1)	(2.00%)	(0)	(0.00%)
Trifunctional protein deficiency	HADHA	0	0.00%	0	0.00%	0	0.00%
		(5)	(0.50%)	(1)	(2.00%)	(0)	(0.00%)
Pituitary hormone deficiency, combined, 2	PROP1	0	0.00%	0	0.00%	0	0.00%
		(4)	(0.40%)	(0)	(0.00%)	(0)	(0.00%)
Mucopolysaccharidosis type IIIc	HGSNAT	0	0.00%	0	0.00%	0	0.00%
		(3)	(0.30%)	(0)	(0.00%)	(0)	(0.00%)
Aspartylglucosaminuria	AGA	0	0.00%	0	0.00%	0	0.00%
		(1)	(0.10%)	(1)	(2.00%)	(0)	(0.00%)
Mucopolysaccharidosis type VI	ARSB	0	0.00%	0	0.00%	0	0.00%
		(1)	(0.10%)	(0)	(0.00%)	(0)	(0.00%)
Cystic fibrosis	CFTR	1	0.10%	0	0.00%	0	0.00%
		(1)	(0.10%)	(0)	(0.00%)	(0)	(0.00%)
Carnitine palmitoyltransferase II deficiency	CPT2	1	0.10%	0	0.00%	0	0.00%
		(1)	(0.10%)	(0)	(0.00%)	(0)	(0.00%)
Mucopolysaccharidosis type IVa	GALNS	1	0.10%	0	0.00%	0	0.00%
		(1)	(0.10%)	(0)	(0.00%)	(0)	(0.00%)

Table 3.S13: List of genes that have at least two distinct mutations that are predicted to be disease causing (DM) in the Human Gene Mutation Database (HGMD). The corresponding diseases and the number and proportion of samples in which these mutations are found is also given. The number between brackets indicate the number found without filtering on a minor allele frequency of <% in the 1000 genomes

Disease	**Gene**	**1kG**	**% 1kG**	**CG**	**% CG**	**IH**	**% IH**
Polycystic kidney disease	PKHD1	4 (14)	0.40% (1.39%)	0 (0)	0.00% (0.00%)	0 (0)	0.00% (0.00%)
Cystic fibrosis	CFTR	4 (4)	0.40% (0.40%)	0 (0)	0.00% (0.00%)	0 (0)	0.00% (0.00%)
Leber congenital amaurosis 1	GUCY2D	0 (3)	0.00% (0.30%)	0 (1)	0.00% (2.00%)	0 (0)	0.00% (0.00%)
Aspartylglucosaminuria	AGA	2 (2)	0.20% (0.20%)	0 (0)	0.00% (0.00%)	0 (0)	0.00% (0.00%)
Congenital disorder of glycosylation, type Ia	PMM2	1 (1)	0.10% (0.10%)	0 (0)	0.00% (0.00%)	0 (0)	0.00% (0.00%)
Mucopolysaccharidosis type IIIc	HGSNAT	0 (1)	0.00 (0.10%)	0 (0)	0.00% (0.00%)	0 (0)	0.00% (0.00%)

Chapter 4

Epigenetics - Neural Tube Defects

This remainder of this chapter has been published as:

Rochtus A, Izzi B, Vangeel E, Louwette S, Wittevrongel C, Lambrechts D, Moreau Y, Winand R, Verpoorten C, Jansen K, Van Geet C, Freson K. DNA methylation analysis of Homeobox genes implicates *HOXB7* hypomethylation as risk factor for neural tube defects. *Epigenetics* 2015; **10**(1): 92-101

Author's contribution to this project: data analysis, script writing

Reprinted with permission from Taylor & Francis Group

4.1 Introduction

Neural tube defects (NTDs), affecting 0.5-2 per 1000 pregnancies, arise as a failure of the neural tube to close in the cranial (anencephaly) or the caudal (myelomeningocele) region (Wallingford et al., 2013; Copp et al., 2013; Au et al., 2010). The nature and severity of NTDs is determined by the stage and axial level at which closure fails. Cranial NTDs are mostly not compatible with life while caudal NTDs give rise to lifelong disabilities. Although more than 250 genes are known to cause NTDs in mice (Harris & Juriloff, 2007, 2010) and many candidate genes have been studied in patient cohorts, the molecular basis underlying NTDs still remains largely unknown. Folic acid

reduced the incidence of NTDs by 50-75% (Blom et al., 2006). However, in most NTD-affected pregnancies maternal folic acid levels are within the normal range (Kirke et al., 1993) and despite optimal supplementation a significant proportion of NTDs are even unresponsive to folic acid (Blom et al., 2006; Cavalli & Copp, 2002). This would suggest the existence of folic acid resistance in mothers at risk for NTD-affected pregnancies but this hypothesis is not supported by genetic and/or environmental risk factors. Folic acid is central to the one-carbon metabolism that produces pyrimidines and purines for DNA synthesis and for the generation of S-adenosyl-methionine that is a methyl donor for DNA, RNA and protein methylation. The only well-characterized genetic risk factor for NTDs is the *677C>T* variant in the 5,10-methylene tetrahydrofolate reductase gene (*MTHFR*), causing thermolability of the enzyme and predicted to divert the available methyl groups from the DNA methylation towards the DNA synthesis pathway (Blom et al., 2006). Interestingly, the *MTHFR 677C>T* variant is associated with global DNA hypomethylation in both controls and NTDs (Blom et al., 2006; Chen et al., 2010), and this seems to be more pronounced under low folic acid conditions (Friso et al., 2002). Most DNA methylation studies in patients with NTDs were performed independent of the presence of the *MTHFR 677C>T* variant. Findings of global DNA and *LINE-1* hypomethylation were found in fetal neural tissue DNA from NTD patients, suggesting that disruption of genomic stability may lead to abnormal neural tube closure (Wang et al., 2010; Chang et al., 2011).

The methylation hypothesis suggests that folic acid prevents NTDs by enhancing cellular methylation reactions. It is known that a tight regulation of genome-wide erasure of epigenetic footprints with resetting of the methylation signature is critical for normal embryogenesis and therefore, it is believed that DNA methylation changes and genomic instability may disturb neural tube folding (Greene et al., 2011). Immediately post fertilization, rapid de-methylation takes place, followed by re-methylation in the blastocyst and early embryo. It is expected that changes in cytosine methylation are not randomly distributed in the genome but are preferentially located at loci that are more sensitive to these processes. Methylome analysis during early embryonic differentiation showed changes in the methylation patterns for developmental regulatory genes, such as Homeobox (*HOX*) genes (Laurent et al., 2010). The *HOX* gene clusters comprise a family of genes assembled in four clusters (*HOX A, B, C,* and *D* located on chromosomes 7, 17, 12, and 2, respectively). *HOX* genes encode highly conserved transcription factors expressed in the brain and spinal cord that play a central role in establishing the anterior-posterior body axis during embryogenesis (Figure 4.S1) (Volcik et al., 2002; Carpenter, 2002). Their expression is tightly regulated in a spatiotemporal and collinear manner, partly by chromatin structure and epigenetic modifications (Barber & Rastegar, 2010; Chambeyron & Bickmore, 2004; Boudadi et al., 2013). Though genetic studies

could not show an association between variants in *HOX* genes and NTDs (Volcik et al., 2002), DNA methylation studies for the *HOX* cluster genes have not yet been performed.

We hypothesize that children born from mothers with folic acid resistance and a disturbed methylation cycle can present with an abnormal DNA methylation profile for *HOX* genes, resulting in increased risk for abnormal embryonic development and NTDs. Therefore, the aim of this study was to investigate DNA methylation of *HOX* genes as mediator of NTD risk using data extracted from a genome-wide DNA methylation analysis study performed for 10 patients with lumbosacral MMC and 6 healthy unrelated controls. A validation study was performed to quantify locus-specific methylation differences in larger cohorts. The functional characterization of the candidate *HOX* gene was finally analyzed using a zebrafish model of neural tube formation.

4.2 Materials and methods

Ethics statement

Written informed consent to collect blood samples for (epi)genetic studies was obtained from all participants and/or their legal representatives. This study was approved by the Medical Ethics Committee of the University of Leuven (study ML9193).

Description of MMC patients, related healthy siblings and unrelated healthy controls

A total of 85 MMC patients and 12 healthy related siblings enrolled in this study are followed at the pediatric neurology department of the University Hospital Leuven (all <18 years). Detailed clinical and general characteristics for all these subjects are reported in Table 4.1. As sensory and motor functions at and below the level of the spinal cord defect are impaired; paralysis, bowel and bladder dysfunction is present in most of the patients. Folic acid supplementation was recommended, but red blood cell folate was not measured during pregnancy. Table 4.1 also indicates which MMC patients were included in the 450K array and/or the Sequenom validation study. In addition, we have recruited 30 age- and gender-matched non-related healthy control subjects with no family history of NTDs (15 males and 15 females).

Genome-wide DNA methylation analysis using the Illumina 450K BeadChip array

Genome-wide DNA methylation analysis was assessed using Illumina Infinium

HumanMethylation 450K BeadChip (Illumina, Inc., California, USA) that provides a genome-wide coverage of CpG sites (99% of RefSeq genes, covering the promoter region, 5'UTR, first exon, gene body and 3'UTR; Figure 4.S1A) (Bibikova et al., 2011). Bisulfite conversion of leukocyte DNA (1 μg) was performed using the EZ DNA methylation kit (Zymo Research, Irvine CA, USA). Control nested PCR reactions were done on both unconverted and converted DNA to verify DNA conversion. Arrays were processed according to the manufacturer's protocol. Samples were randomly distributed to control for batch effects. Before analyzing the data, possible sources of technical bias were excluded. Probes were excluded from further analysis if >95% of samples had a detection value >0.01 (Dedeurwaerder et al., 2014). The software GenomeStudio (Illumina) was used to convert on-chip fluorescent methylation values into numerical values (β-value). Methylation, described as a β-value, is a continuous variable ranging between 0 (no methylation) and 1 (full methylation) for each CpG site. From this genome-wide analysis, we extracted the methylation levels for the different CpGs that cover all regions within the *HOX* clusters (for overview see Table 4.S1). We discarded the following probes (608 in total): i) probes with absent signals in one or more of the DNA samples analyzed, ii) non-CpG probes, iii) probes containing SNPs and iv) leukocyte-specific probes (Dedeurwaerder et al., 2014). The signal processing was conducted using the Illumina Methylation Analyzer (IMA) package implicated in the open source statistical environment R (Wang et al., 2012; R Core Team, 2013) Two filters were applied to identify differentially methylated CpGs between MMC patients and controls: i) absolute β-value difference >0.05 and ii) p-value <0.05 as calculated with the Wilcoxon rank-sum test.

Methylation of CpGs within the *HOXB7* gene body using the Sequenom EpiTYPER

Leukocyte DNA (1 μg) was subjected to bisulfite treatment using the MethylDetector^{TM} bisulfite modification kit (Active Motif, Carlsbad CA, USA) as we described (Izzi et al., 2010, 2012). The Sequenom MassARRAY (Sequenom, San Diego, CA, USA) was used for quantitative DNA methylation analysis of the CpG island within the *HOXB7* gene body using conditions described (Izzi et al., 2010). Long cycling incubation was applied to further optimize the conversion reaction (Izzi et al., 2014). Primers were designed using the Sequenom EpiDesigner BETA software (www.epidesigner.com), taking into account amplicon coverage, number of CpGs, fragment size and number of nucleotide repeats in the primer sequence. The primers were: 5'-aggaagagagGTGTTGGGATTATAGGTTTGAGTTT-3' and 5'-cagtaatacgactcactatagggagaaggctACTAAACTTCTCTTCCTCTCCCTTTC-3'. This 395 bp long amplicon covers 26 CpGs but the EpiTYPER analysis only detected 10 separate analytical units that comprise single, duplicate or triplicate

CpGs as shown in Figure 4.1A. The Illumina probe cg07547765 is similar as CpG6. The 2 other Illumina probes cg11041817 and cg22622477 are located within the studied CpG Sequenom amplicon but were not detected by the EpiTYPER. PCR steps were performed in triplicate for each DNA sample and a standard deviation between replicates was mostly <10%. When triplicate measurements had a SD >10% or when only one of the triplicates was available, data for that sample were excluded. The mean of 3 values was used for further analyses. The EpiTYPER analysis method reports CpG methylation values as percentage. Statistical analyses to quantify DNA methylation differences were performed using the Prism 6 software (GraphPad Software Inc., San Diego, CA, USA). A two-tailed T-test was used to assess differences in mean DNA methylation levels between cohorts for the overall *HOXB7* amplicon considered as methylation average and for each CpG unit within this amplicon separately.

MTHFR *677C>T* genotyping

Leukocyte DNA from MMC patients, related healthy siblings and unrelated healthy controls was screened for the presence of the *MTHFR* 677C>T variant by PCR and restriction digestion as described (Frosst et al., 1995).

Functional in silico analysis of *HOXB7* methylation versus expression

A correlation between *HOXB7* CpG promoter methylation and gene expression was studied by data mining using the open source database MENT (Methylation and Expression database of Normal and Tumor tissues) (Baek et al., 2013). The database only included Illumina 27K BeadChip CpG probes (cg09357097 and cg06493080 as shown in Figure 4.1A) that are located in the *HOXB7* promoter and not in the gene body.

Hoxb7a overexpression and depletion in zebrafish

Wild-type AB zebrafish strains were maintained according to standard protocols (Westerfield, 1995). Embryos were produced by natural mating and collected and fixed at different stages based on standard morphological criteria (Kimmel et al., 1995). To produce Hoxb7a mRNA, the full coding Hoxb7a transcript (NM_001115091.2) was PCR amplified and cloned in the pGEM T Easy vector (Promega, Madison, WI, USA). Forward and reverse primers were 5'-ATGAGTTCATTGTATTATGCGA-3' and 5'-GTAGTTTATACATCTATATTAA-3'. Next, capped and polyadenylated Hoxb7a mRNAs were synthesized using mMESSAGE mMACHINE® High Yield Capped RNA Transcription Kit and Poly(A) Tailing Kit (both from Ambion, Austin, TX, USA) according to the manufacturer's protocol. The synthesized mRNA was diluted in phenol red to different concentrations as indicated in the figure legends. Morpholino (MO) injection was performed with a splice Hoxb7a-MO (5'-AGCACCTGTGAAAAGCGCAGAATGA-3'). This

MO was designed against Chr17: 46,685,144-46,685,550. Off-target effects were assessed by injecting with a standard control MO against beta-globin (5'-CCT CTT ACC TCA GTT ACA ATT TAT A 3'). MOs were designed by Gene Tools, LLC (Philomath, OR, USA). All injected embryos were life-screened at 24, 48 and 72 hours post-fertilization (hpf) using a Zeiss Lumar V12 (Carl Zeiss Microscopy, Thornwood, NY, USA) and images were captured with a Leica DFC310 FX digital color camera (Leica Microsystems, Wetzlar, Germany). Overexpression and depletion experiments were performed in duplicate. Ethical approval was obtained for these studies.

Pax2a whole mount in situ hybridization

Whole mount in situ hybridization (WISH) with a probe for the paired box gene 2a (pax2a) was performed 24h after injection of MOs or Hoxb7a mRNA. Pax2a cDNA obtained from Dr. W. Driever (University of Freiburg, Germany) was cloned in the pGEM-3zf+ for the synthesis of a digoxygenin (DIG) labeled antisense RNA probe as described (Krauss et al., 1991). The Pax2a probe was subsequently used to analyze the influence of Hoxb7a overexpression and inhibition on spinal cord and notochord formation using standard morphological criteria (Kimmel et al., 1995). WISH experiments were performed in duplicate. Embryos were screened using a Zeiss Lumar V12 (Carl Zeiss Microscopy, Thornwood, NY, USA) and images were captured with a Leica DFC310 FX digital color camera (Leica Microsystems, Wetzlar, Germany).

4.3 Results

DNA methylation analysis of the different *HOX* genes in MMC case-control study

Methylation values for all CpGs located within the 4 *HOX* clusters were extracted from data obtained from a 450K array-based genome-wide methylation analysis, using leukocyte DNA from 10 MMC cases and 6 unrelated age- and gender-matched healthy controls. Detailed clinical characteristics of these MMC patients are reported in Table 4.1. Pie charts were made to show the equal distribution of the filtered CpG probes (n=967) based on i) location within the 4 different *HOX* clusters, ii) location with respect to gene transcripts and iii) location with respect to the CpG Island (Figure 4.S2). From the 967 filtered CpGs profiled on the 450K array (Table 4.S1), only 26 CpGs were found to be differentially methylated between MMC patients and controls (Table 4.2 and Figure 4.S2). Interestingly, 25 of these 26 CpGs were hypomethylated for the MMC patients and 20 of the 26 CpGs were located within the *HOXA* or *HOXB* clusters. Only for the *HOXB7* gene, 3 different CpG probes (cg11041817, cg22622477 and

cg07547765) were significantly hypomethylated in MMC patients (β-differences of -0.29; -0.16 and -0.26, respectively with all p-values of 0.007). These 3 probes are located within a single CpG island at chr17:46,685,244-46,685,449 within the *HOXB7* gene body (Figure 4.1A).

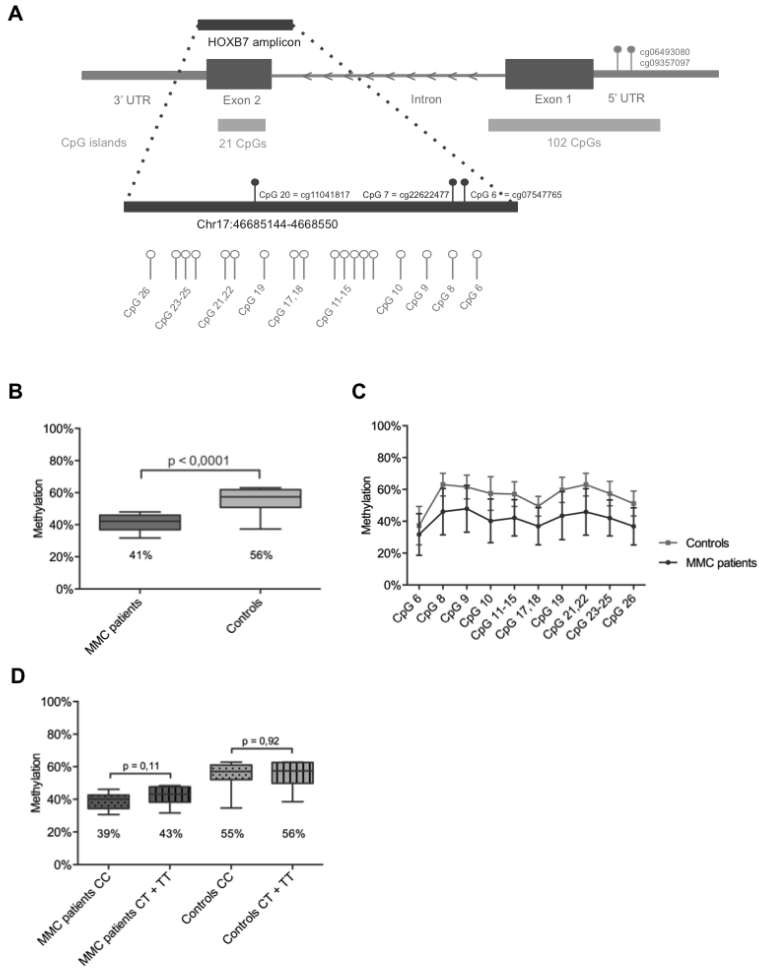


Figure 4.1: *HOXB7* methylation studies by Sequenom EpiTYPER in MMC patients. **A**: Localization of the studied amplicon (Chr17:46,685,144-46,685,550) within *HOXB7* Exon 2. The amplicon covers 26 single CpGs and our assay provides data 10 analytical CpG units. Nucleotide positions accord to the NCBI build 37/hg19. The CpG units studied by 450K Array (cg11041817, cg22622477 and cg07547765) and the *in silico* analysis (cg06493080, cg09357097) are also indicated. **B**: Boxplot representing the methylation pattern of MMC patients and controls with box = 25th and 75th percentiles; bars = min and max values. The mean methylation level of each group is shown below the plot. **C**: Methylation pattern for each CpG unit within the amplicon. Wilcoxon Rank-Sum test was performed. **D**: Boxplot representing the methylation pattern of MMC patients and controls divided according to *MTHFR 677C>T* genotype with box = 25^{th} and 75^{th} percentiles; bars = min and max values. The mean methylation level of each group is shown below the plot.

Table 4.1: Background information of MMC patients included in the HumanMethylation 450K BeadChip and Sequenom EpiTYPER.

MMC patient	Type MMC	Hy/VP	ACM	Scoliosis	ADL	UI	Ethnicity	Maternal age (years) at birth of MMC patient	MMC patient		Sibling		MMC Patient versus sibling age (months)
									MTHFR 677C>T	Gender	MTHFR 677C>T	Gender	
1*+	S	yes	1	yes	2	yes	Belgian	36	CC	F			
2*+	LS	yes	2	yes	3	yes	Belgian	29	CT	F			
3*+	LS	yes	2	yes	3	yes	Moroccan	32	CT	M			
4*+	LS	yes	1	yes	3	yes	Indonesian	26	CC	F			
5*+	LS	yes	2	no	2	yes	Belgian		CC	M	CC	F	-18
6*+	LS	yes	2	yes	3	yes	Belgian	27	CC	F			
7*+	LS	yes	1	yes	3	yes	Belgian		CC	M			
8*+	S	yes	2	no	1	yes	Belgian	27	CC	F			
9*	S	no	0	no	1	yes	Belgian		TT	M			
10*	LS	yes	2	no	2	yes	Belgian	25	CC	M			
11+	LS	yes	2	yes	3	yes	Belgian	27	CC	M			
12+	LS	yes	2	yes	3	yes	Turkish	28	CC	F			
13+	S	yes	0	no	1	yes	Belgian	33	CT	M			
14+	LS	yes	2	yes	3	yes	Belgian		CC	F			
15+	LS	yes	2	yes	3	yes	Turkish	20	CT	F			
16+	S	yes	2	yes	3	yes	Belgian		TT	M			
17+	LS	yes	2	yes	3	yes	Belgian	25	CT	F			
18+	S	yes	2	yes	2	yes	Belgian	28	CC	M			
19+	LS	yes	2	no	1	yes	Belgian	36	CT	M			
20+	LS	yes	2	yes	1	yes	Belgian		CT	M			
21+	LS	yes	2	yes	3	yes	Belgian	36	CT	F			
22+	TL & LS	yes	2	yes	3	yes	Belgian	23	CC	M			

Continued on next page

Table 4.1 – continued from previous page

MMC patient	Type MMC	Hy/VP	ACM	Scoliosis	ADL	UI	Ethnicity	Maternal age (years) at birth of MMC patient	MMC patient		Sibling		MMC Patient versus sibling age (months)
									MTHFR 677C>T	Gender	MTHFR 677C>T	Gender	
23+	S	no	0	yes	1	yes	Belgian	30	CT	M			
24+	LS	yes	2	yes	2	yes	Belgian	25	CT	F			
25+	LS	yes	2	no	2	yes	Belgian	34	CC	F			
26+	LS	no	0	no	1	yes	Belgian		CC	F			
27+	LS	yes	2	yes	3	yes	Belgian		CT	M	CC	F	-20
28+	LS	yes	2	yes	3	yes	Moroccan		CT	F			
29+	LS	yes	2	yes	3	yes	Belgian	31	CC	M	CC	M	-19
30+	LS	yes	0	no	1	yes	Mongolian		CC	F			
31+	LS	no	0	no	1	no	Belgian	20	TT	M	CC	M	-20
32+	LS	yes	2	no	1	yes	Belgian	22	CC	M			
33+	LS	yes	2	yes	2	yes	Belgian	28	CT	M			
34+	LS	yes	2	yes	3	yes	Belgian		CT	F			
35+	S	yes	2	no	1	yes	Belgian	33	CT	M			
36+	LS	yes	2	no	3	yes	Belgian	27	CC	M			
37+	LS	yes	NA	yes	2	yes	Belgian	26	CT	F			
38+	LS	yes	2	yes	2	yes	Belgian		CT	F			
39+	LS	yes	2	yes	3	yes	Belgian	28	CT	F			
40+	LS	yes	2	no	1	yes	Belgian	32	CC	M			
41+	LS	no	NA	no	3	yes	Belgian	36	CC	M	CT	M	-26
42+	L	yes	2	yes	3	yes	Ukrainian	20	CC	F			
43+	LS	yes	2	no	1	yes	Belgian	25	CC	F	TT	M	20
44+	LS	yes	2	yes	3	yes	Belgian		CC	M			
45+	LS	yes	2	yes	3	yes	Belgian	29	CT	F			
46+	LS	yes	2	yes	3	yes	Belgian		CC	F			

Continued on next page

Table 4.1 – continued from previous page

MMC patient	Type MMC	Hy/VP	ACM	Scoliosis	ADL	UI	Ethnicity	Maternal age (years) at birth of MMC patient	MMC patient		Sibling		MMC Patient versus sibling age (months)
									MTHFR 677C>T	Gender	MTHFR 677C>T	Gender	
47+	S	yes	2	no	1	yes	Moroccan	31	CT	F			
48+	LS	yes	2	yes	2	yes	Belgian	27	CT	F			
49+	S	yes	2	yes	3	yes	Belgian	34	CC	F			
50+	S	no	NA	no	1	yes	Belgian	36	CT	F	CC	F	17
51+	LS	yes	2	no	3	yes	Belgian	30	CC	F			
52+	LS	yes	2	no	1	yes	Moroccan	24	CC	F			
53+	S	yes	2	no	1	yes	Belgian	28	CC	F			
54+	S	no	NA	no	1	yes	Belgian		CT	F			
55+	LS	yes	2	no	2	yes	Bosnian	31	CT	F	CC	M	53
56+	S	no	0	no	1	yes	Belgian	30	CC	F			
57+	LS	yes	2	no	1	yes	Belgian	24	CC	F			
58+	S	no	2	no		yes	Moroccan	26	CC	F			
59+	LS	yes	2	yes	3	yes	Belgian	27	CC	F			
60+	LS	yes	2	no	1	yes	Turkish	35	CT	F	CC	F	-101
61+	LS	yes	2	yes	3	yes	Belgian		CT	F			
62+	LS	yes	2	no	1	yes	Turkish	29	CT	F			
63+	LS	yes	2	yes	3	yes	Belgian	35	CT	F			
64+	L & CP	yes	2	no	2	yes	Belgian	40	CT	F	CT	F	-27
65+	LS	yes	2	yes	2	yes	Belgian		CT	F			
66+	LS	yes	2	yes	2	yes	Belgian	24	CT	F			
67+	LS	yes	2	yes	2	yes	Belgian		CT	F			
68+	S	yes	2	no	1	no	Belgian		CT	M			
69+	LS	yes	2	yes	3	yes	Belgian	24	CT	M			
70+	LS	yes	2	yes	3	yes	Belgian	29	CC	M			

Continued on next page

Table 4.1 – continued from previous page

MMC patient	Type MMC	Hy/VP	ACM	Scoliosis	ADL	UI	Ethnicity	Maternal age (years) at birth of MMC patient	MMC patient		Sibling		MMC Patient versus sibling age (months)
									MTHFR 677C>T	Gender	MTHFR 677C>T	Gender	
71+	LS	yes	2	yes	3	yes	Belgian	33	CC	F			
72+	LS	yes	2	yes	2	yes	Belgian	29	CC	M			
73+	LS	yes	2	yes	3	yes	Belgian	29	CT	F	CT	F	-24
74+	LS	yes	2	yes	2	yes	Belgian		CT	M			
75+	LS	yes	2	no	1	yes	Belgian		CC	F			
76+	LS	yes	2	yes	2	yes	Belgian	31	CT	F			
77+	LS	yes	2	yes	3	yes	Belgian	26	CC	M			
78+	LS	yes	2	yes	3	yes	Belgian	30	CT	M			
79+	S	no	0	no	1	yes	Belgian		CT	M			
80+	LS	yes	2	no	1	yes	Belgian	30	CC	F			
81+	LS	yes	NA	no	2	yes	Turkish		CT	M			
82+	TL	yes	2	yes	3	yes	Belgian		TT	F			
83+	TL	yes	2	yes	1	yes	Belgian	24	TT	F			
84+	TL	yes	1	yes	3	yes	Belgian	37	CC	F			
85+	TL	yes	2	no	1	no	Belgian		TT	F			

*Inclusion in HumanMethylation 450K BeadChip (10 MMC patients) and +Sequenom EpiTYPER (83 MMC patients); MMC: myelomeningocele; M: male; F: female; S: Sacral; LS: Lumbosacral; TL: Thoracolumbal; CP: Cheilopalatoschisis; Hy/VP: presence of hydrocephaly and ventriculoperitoneal drain; ACM: Arnold Chiari Malformation: 1 = type 1, 2 = type 2; NA: Not Available; ADL: Activities of daily life: 1 = ambulatory, 2 = household ambulatory with wheelchair for longer distances, 3 = wheelchair dependent; UI: urinary incontinence; *MTHFR 677C>T* genotype (*CC/CT/TT*) in 85 MMC patients: 47% *CC* - 53% *CT/TT* versus 40% *CC* - 60% *CT/TT* in 30 healthy unrelated controls (P-value = ns).

Table 4.2: Methylation of the HOX genes using the HumanMethylation 450K BeadChip and analysis.

				P-value	β-diff	Ctrls (n = 6)		MMC (n = 10)	
Cluster	**Gene**	**Chr**	**Illumina ID**	**<0.05**	**>0.05**	**Mean**	**SD**	**Mean**	**SD**
A: Chr.7p14	HOXA2	7	cg06055873	0.016	-0.06	0.32	0.03	0.26	0.04
	HOXA2	7	cg19432993	0.031	-0.12	0.69	0.05	0.57	0.12
	HOXA2	7	cg06166490	0.016	-0.12	0.70	0.06	0.58	0.11
	HOXA2	7	cg04027736	0.022	-0.11	0.61	0.06	0.50	0.11
	HOXA2	7	cg00445443	0.042	-0.10	0.41	0.08	0.31	0.11
		7	cg15037137	0.007	-0.05	0.85	0.02	0.80	0.07
	HOXA4	7	cg25952581	0.042	-0.09	0.45	0.06	0.37	0.11
	HOXA4	7	cg17591595	0.042	-0.08	0.73	0.05	0.65	0.09
	HOXA11	7	cg24709033	0.011	-0.06	0.27	0.03	0.21	0.04
B: Chr.17q21	HOXB5	17	cg01405107	0.042	0.09	0.49	0.05	0.58	0.08
	HOXB6	17	cg09983216	0.016	-0.13	0.55	0.08	0.42	0.10
	HOXB7	17	cg11041817	0.007	-0.29	0.70	0.08	0.41	0.20
	HOXB7	17	cg22622477	0.007	-0.16	0.36	0.06	0.20	0.09
	HOXB7	17	cg07547765	0.007	-0.26	0.71	0.10	0.44	0.18
		17	cg19051015	0.022	-0.09	0.72	0.05	0.63	0.07
	HOXB9	17	cg15117739	0.007	-0.10	0.68	0.04	0.57	0.07
	HOXB9	17	cg12057127	0.042	-0.06	0.72	0.01	0.67	0.07
		17	cg20454400	0.031	-0.06	0.37	0.04	0.31	0.05
		17	cg16654603	0.011	-0.09	0.67	0.02	0.58	0.08
		17	cg02052915	0.016	-0.05	0.36	0.03	0.31	0.04
C: Chr.12q13		12	cg08299265	0.016	-0.06	0.39	0.04	0.33	0.03
		12	cg26643142	0.031	-0.06	0.35	0.03	0.30	0.05
	HOXC4	12	cg18473521	0.011	-0.12	0.43	0.08	0.31	0.07
D: Chr.2q31	HOXD9	2	cg04730882	0.005	-0.07	0.35	0.03	0.28	0.05

Continued on next page

Table 4.2 – continued from previous page

Cluster	Gene	Chr	Illumina ID	P-value <0.05	β-diff >0.05	Ctrls (n = 6) Mean	Ctrls (n = 6) SD	MMC (n = 10) Mean	MMC (n = 10) SD
		2	cg07783843	0.011	-0.06	0.24	0.02	0.18	0.04
		2	cg05525812	0.007	-0.07	0.25	0.02	0.18	0.05

Nucleotide positions accord to NCBI build 37/hg19. Selection is performed along both β-value > 0.05 difference and p-value < 0.05; calculated with Wilcoxon Rank-Sum test. The 3 probes in bold are located within the same CpG island at Chr17:46,685,244-46,685,449 within the *HOXB7* gene body. This region was selected for the validation study using Sequenom EpiTYPER. β-diff: β-difference; Chr: chromosome; Ctrls: controls; MMC: myelomeningocele.

HOXB7 methylation analysis in MMC case-control study

A validation study using larger cohorts (83 MMC patients described in Table 4.1) was performed with the Sequenom EpiTYPER technology to quantify methylation of the above selected CpG island located in the *HOXB7* gene body. A Sequenom amplicon was developed that covers 26 CpGs (Figure 4.1A) including the 3 significant CpGs detected in the 450K array. Within this amplicon, the EpiTYPER detected 10 analytical CpG units for which CpG6 is similar to cg07547765. DNA methylation values for the amplicon for 83 MMC patients and 30 unrelated healthy controls were normally distributed (Shapiro-Wilk test, P>0.05). A significant *HOXB7* hypomethylation (p-value < 0.0001) was detected for MMC patients versus controls with mean methylation values of 41% (95% CI: 38-45%) versus 56% (95% CI: 50-61%), respectively (Figure 4.1B). The mean level of methylation for each CpG unit within the amplicon was also significantly different between MMC patients and controls (Figure 4.1C and Table 4.S2). To exclude an effect of changes in methylation due to differences in ethnicity, the *HOXB7* methylation pattern between 70 Belgian MMC patients was compared to 10 Non-Caucasian MMC patients without significant differences (Figure 4.S3).

As findings of global DNA hypomethylation and LINE-1 hypomethylation suggest that disruption of genomic stability may disrupt neural tube closure and the *MTHFR 677C>T* variant is associated with global DNA hypomethylation, we determined the *MTHFR 677C>T* variant for MMC patients and healthy controls (Table 1). Interestingly, an intrinsic defect in the folic acid pathway related to *MTHFR* activity seems not to be involved as no association was found between *MTHFR 677 CC* versus *CT+TT* carriers and *HOXB7* methylation (Figure 1D).

HOXB7 methylation analysis in unaffected siblings of MMC patients

For 12 out of 83 MMC patients, DNA was also collected from their healthy siblings (Table 4.1). Remarkably, the mean methylation level of the *HOXB7* amplicon was not different between MMC patients and their unaffected siblings with values of 37% (95% CI: 33-40%) versus 40% (95% CI: 37-42%) (Figure 4.2A). Multiple T-testing for each CpG within the *HOXB7* amplicon also showed no significant differences between patients and healthy siblings (Figure 4.2B and Table 4.S2).

HOXB7 methylation versus expression

Since leukocyte RNA was not collected for our cohorts, we used the MENT database to estimate a correlation between *HOXB7* methylation and gene expression. The 2 CpGs (cg09357097 and cg06493080) located in the *HOXB7* promoter (Figure 4.1A) showed no correlation with gene expression in normal

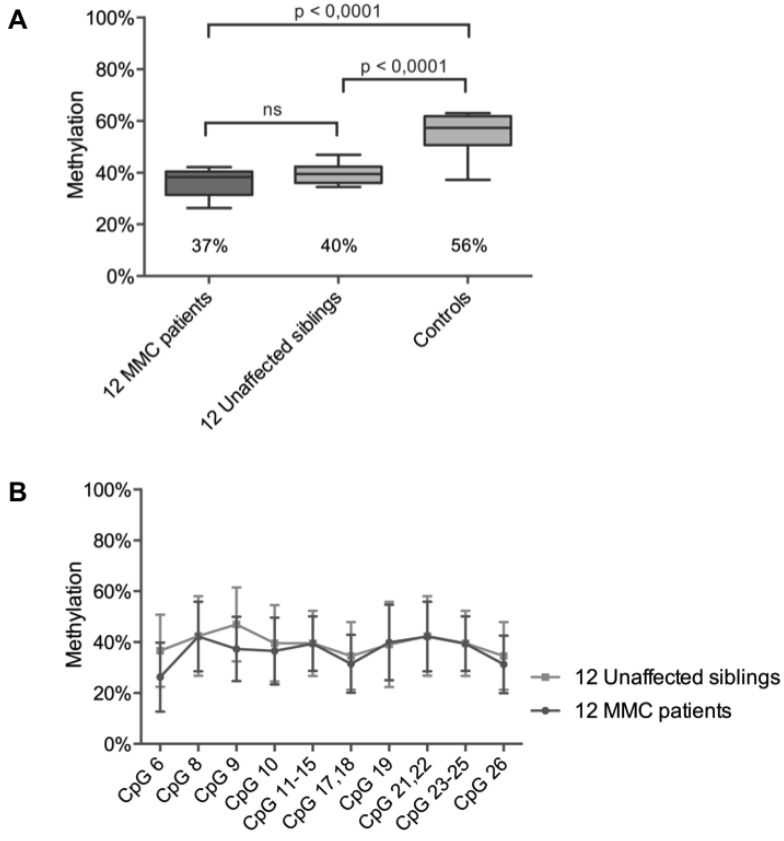


Figure 4.2: **A**: Boxplot representing the methylation pattern of affected siblings and unaffected siblings with box = 25^{th} and 75^{th} percentiles; bars = min and max values. The mean methylation level of each group is shown below the plot. **B**: Methylation pattern for each CpG unit within the amplicon. Wilcoxon Rank-Sum test was performed.

brain and blood tissues. However, there is evidence that lower *HOXB7* methylation values in brain tightly regulate higher and stable gene expession levels compared to the higher methylation levels in blood that are associated with a variable gene expression (Figure 4.S4). Interestingly, cg06493080 showed a strong negative correlation with gene expression in different cancer tissues, especially for brain (correlation -0.15, P-value 0.008).

Hoxb7a overexpression and depletion in zebrafish

Functional genetics was performed in zebrafish to study alterations in Hoxb7a expression during embryogenesis and neural tube formation. The regulation of the *HOX* clusters is highly conserved between humans and zebrafish (Figure 4.S1). Hoxb7a has an anterior expression limit adjacent to the somite 3-4 boundary at the 20 somite stage (Prince et al., 1998). We analyzed embryos with Hoxb7a depletion and overexpression using microinjection of a splice morpholino (MO) and synthetic Hoxb7a mRNA, respectively. MO-induced Hoxb7a depletion resulted in hypopigmentation and developmental delay with dysmorphy in 83-94% of the embryos at 24 hours post fertilization (hpf) (Figure 4.S5). However, pax2a staining to visualize neural tube formation at 24 hpf was not different between Hoxb7a- or control-MO injected embryos, even for the severely affected Hoxb7a depleted embryos (Figure 4.S5). Embryos injected with different concentrations of Hoxb7a mRNA also present with severe to mild malformations in about 48-71% of the embryos at 24 hpf (Figure 4.3B). These embryos had shorter anterior/posterior axes as well as crooked or bent tails (Figure 4.3A). Interestingly, pax2a staining after overexpression of Hoxb7a for different concentrations showed a neural tube was that was absent or completely disorganized (Figure 4.3C).

4.4 Discussion

As *HOX* genes play key roles in neural tube closure and many studies have shown that folic acid prevents NTDs by stimulating cellular methylation reactions, we extracted methylation data for the different *HOX* genes from a genome-wide DNA methylation analysis performed for 10 MMC patients and 6 unrelated healthy controls. Interestingly, 25 of the 26 CpGs were hypomethylated for the MMC patients with the *HOXB7* gene body as most significant locus. Interestingly, *HOXB7* hypomethylation was not only confirmed in a larger MMC cohort but was also detected in 12 healthy siblings each related to a MMC patient. These results are suggestive for a maternal effect that contributes to *HOXB7* hypomethylation. Additional healthy siblings must be recruited but gender, age, *MTHFR 677C>T* genotype or MMC being the firstborn or not,

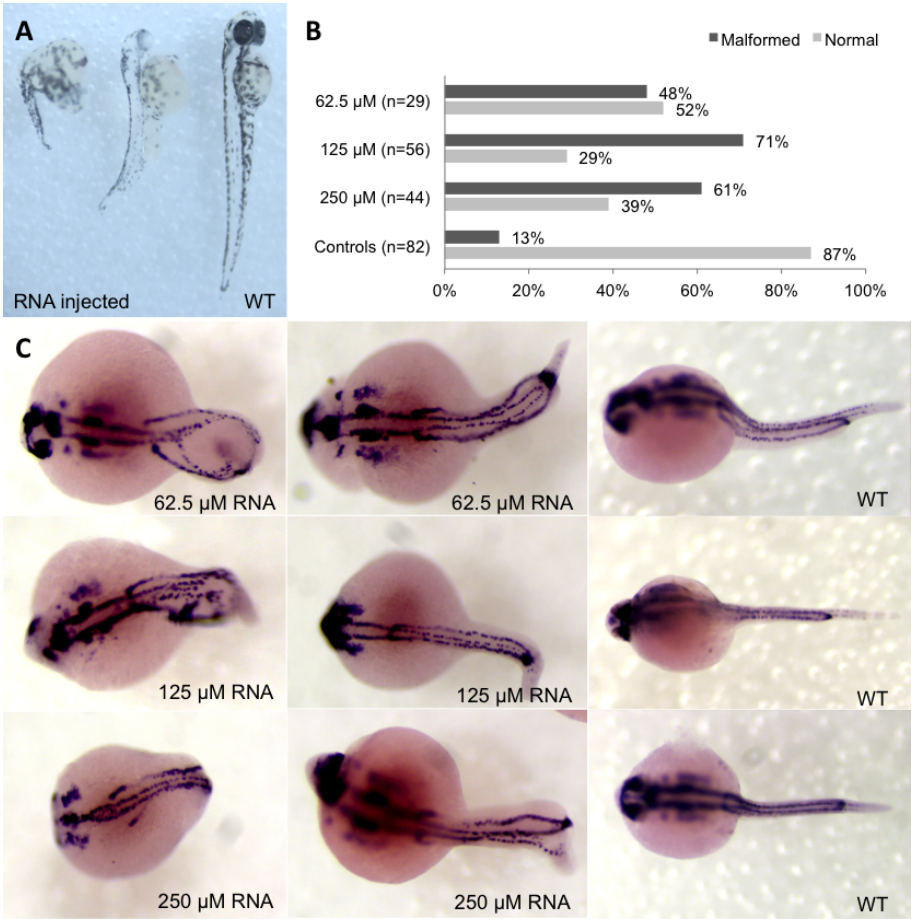


Figure 4.3: Phenotype analysis at 72 hpf of Hoxb7a mRNA injected zebrafish resulted in significant hypopigmentation and malformation in 66% of the injected zebrafish. These embryos had shorter anterior/posterior axes as well as crooked or bent tails. B: Phenotype analysis after pax2a at 24 hpf resulted in about 63% embryos with a mild or severe affected phenotype after Hoxb7a overexpression compared to 13% in injected controls. C: Pax2a staining after microinjection of different concentrations of mRNA. From left to right severe, mild affected and wild type (WT) embryos at 24 hpf. WT zebrafish show expression in the hindbrain, hindbrain-midbrain boundary, neural tube, mesoderm, optic stalk, otic vesicle and pronephric duct. Microinjection 62.5 μM mRNA, 125 μM mRNA and 250 μM mRNA resulted in respectively 48%, 71% and 61% malformed zebrafish. There was no correlation between mRNA dosage and severity of malformation.

seem no predictive risk factors for NTDs based on data for these sibling pairs (Table 4.1). *HOXB7* hypomethylation by itself is likely not to be causative for NTDs but rather be part of a complex combination of environmental and (epi)genetic risk factors. We found no association between *MTHFR CC* versus *CT+TT* carriers and *HOXB7* methylation suggesting that an intrinsic defect in the folic acid pathway related to *MTHFR* activity is not involved. Though we had no measurements of maternal folic acid levels or uptake, it is known that folic acid levels in most affected pregnancies are within the normal range (Kirke et al., 1993) and despite optimal supplementation a significant proportion of NTDs are unresponsive to folic acid (Blom et al., 2006; Cavalli & Copp, 2002). We therefore hypothesize that these mothers have folic acid resistance leading to a disturbed methylation cycle with alterations in DNA methylation and an increased risk for abnormal embryonic development. Additional studies must be undertaken to study the association between maternal folic acid intake and *HOXB7* methylation in DNA from the mother and her offspring.

HOX genes encode for evolutionary highly conserved transcription factors expressed in the central nervous system (Figure 4.S1) . They are tightly regulated in a spatiotemporal and collinear manner (Barber & Rastegar, 2010; Chambeyron & Bickmore, 2004; Boudadi et al., 2013), patterning the embryo along the rostro-caudal axis. The *HOXA* and *HOXB* clusters have a closer phylogenetic relationship and hence share more functionality than with either the *HOXC* or *HOXD* cluster (Prince et al., 1998). Their cooperative functioning is necessary for the generation of the cranial neural crest and craniofacial diversity (Philippidou & Dasen, 2013; Soshnikova et al., 2013; Vieux-Rochas et al., 2013). The spinal cord is a caudal structure, but the neural cells from which it derives initially express rostral, forebrain-like characteristics. The caudal character emerges soon after neural induction, through different extrinsic signals (Jessell, 2000; Doniach, 1995). According to our study, differential *HOX* gene methylation in MMC patients occurs in both anterior and posterior *HOX* genes. Moreover, failure in establishing correct *HOX* gene methylation in the HOXA and HOXB clusters may result in disturbances in neural cell identity that ultimately leads to neural malformations. *HOX* gene clusters are evolutionary highly conserved between human and zebrafish and a neural tube formation zebrafish model was previously used to study *VANGL1* (Reynolds et al., 2010). *HOX7* has two paralog members in humans and only one in zebrafish (Figure 4.S1) but the zebrafish Hoxb7a gene shares 60% homology with the human *HOXB7* sequence. As *HOXB7* hypomethytlation is suggestive for *HOXB7* overexpression, Hoxb7a overexpression experiments were performed in zebrafish. Overexpression of Hoxb7a in zebrafish resulted in developmental abnormalities and pax2a staining showed abnormal neural tube formation in about 60% of the embryos.

In the present study, we were of course not able to use patients DNA samples

from brain or spinal cord tissue. Concordant DNA methylation profiles in brain and blood samples from the same individuals suggest that blood might hold promise as surrogate for brain tissue to detect DNA methylation (Masliah et al., 2013; Tylee et al., 2013; Ikegame et al., 2013). Genome-wide methylation arrays revealed similar methylation patterns for the *HOX* genes in breast cancers and white blood cells, which suggests that methylation is more likely to be a normal developmental and tissue-specific process that does not directly relate to the malignant mechanism (Avraham et al., 2010). Interestingly, functional in silico analysis using the MENT database showed no correlation with gene expression in normal brain and blood tissues for the methylation of two *HOXB7* promoter CpGs but there is evidence that lower *HOXB7* methylation values in brain tighly regulate higher and stable gene expession levels compared to the higher methylation levels in blood that are associated with a variable *HOXB7* gene expression. These data would suggest that *HOXB7* hypomethylation is associated with higher gene expression. A limitation of our study was the lack of *HOXB7* gene expression studies using leukocyte RNA from MMC cases and unrelated healthy controls but RNA samples were not collected. Additional studies are needed to correlate the methylation levels of the *HOXB7* gene body CpGs with *HOXB7* gene or protein expression values. Furthermore, it would be interesting to compare our findings in leukocytes with neural tissue.

4.5 Conclusion

This is the first study that uses genome-wide DNA methylation data for the locus-specific analysis of the different *HOX* genes in patients with NTDs. We found evidence that *HOXB7* hypomethylation is a potential risk factor for MMC but also that the underlying methylation defect is present in both affected and non-affected offspring. This could confirm the hypothesis that children born from mothers with folic acid resistance and a disturbed methylation cycle, can present with alterations in DNA methylation with high risk for abnormal embryonic development. Investigating the complex etiology of NTDs requires consideration of more DNA methylation studies; therefore genome-wide DNA methylation analysis without focusing on candidate pathways could reveal more epigenomic changes associated with NTDs. The challenge ahead is to determine which DNA regions are more sensitive to methylation changes during embryogenesis and lead to NTDs.

4.6 Supplementary data

Cluster	Gene	UCSC RefSeq	Gene region including shores/shelves	Total	Filtered
A	HOX A1	Chr7:27132614-27135625	Chr7:27131625-27240499	481	298
	HOX A13	Chr7:27236499-27239725			
B	HOX B1	Chr17:46606807-46608272	Chr17:46602807-46810111	409	236
	HOX B13	Chr17:46802127-46806111			
C	HOX C4	Chr12:54410636-54449814	Chr12:54328576-54453814	360	232
	HOX C13	Chr12:54332576-54340328			
D	HOX D1	Chr2:177053307-177055635	Chr2:176953532-177059635	325	201
	HOX D13	Chr2:176957532-176960666			
				1575	967

Table 4.S1: Number of probes for the 4 HOX clusters included in the HumanMethylation 450K BeadChip platform. Nucleotide positions accord to the NCBI build 37/hg19. *HOX* genes cover 1575 CpGs, 967/1575 CpGs (61%) remained after preprocessing and filtering of the samples.

A	MMC versus controls						
	Controls			Patients			P-value
	Mean (%)	SD (%)	N	Mean (%)	SD (%)	N	
CpG 6	37	12	20	32	13	71	0.002
CpG 8	63	7	29	46	15	83	****
CpG 9	62	7	28	48	15	76	****
CpG 10	57	11	26	40	14	76	****
CpG 11-15	57	8	29	42	11	83	****
CpG 17,18	49	6	25	37	12	81	****
CpG 19	60	8	26	43	15	81	****
CpG 21,22	63	7	29	46	15	82	****
CpG 23-25	57	8	28	42	11	83	****
CpG 26	51	8	27	37	12	81	****

B	Matched pairs						
	Unaffected siblings			MMC			P-value
	Mean (%)	SD (%)	N	Mean (%)	SD (%)	N	
CpG 6	37	14	11	26	14	12	0.069
CpG 8	42	16	12	42	14	12	0.968
CpG 9	47	14	11	37	13	10	0.104
CpG 10	40	15	10	37	13	12	0.605
CpG 11-15	40	13	12	39	11	12	0.986
CpG 17,18	35	13	12	31	11	11	0.583
CpG 19	39	17	12	40	15	12	0.886
CpG 21,22	42	16	12	42	14	12	0.968
CpG 23-25	40	13	12	39	11	12	0.986
CpG 26	35	13	12	31	11	11	0.558

C	Unaffected siblings versus controls						
	Unaffected siblings			Controls			P value
	Mean (%)	SD (%)	N	Mean (%)	SD (%)	N	
CpG 6	37	14	11	37	12	20	0.865
CpG 8	42	16	12	63	7	29	****
CpG 9	47	14	11	62	7	28	***
CpG 10	40	15	10	57	11	26	****
CpG 11-15	40	13	12	57	8	29	****
CpG 17,18	35	13	12	49	6	25	****
CpG 19	39	17	12	60	8	26	****
CpG 21,22	42	16	12	63	7	29	****
CpG 23-25	40	13	12	57	8	28	****
CpG 26	35	13	12	51	8	27	****

Table 4.S2: *HOXB7* methylation by Sequenom EpiTYPER analysis for single CpG units within the different cohorts. Methylation levels were compared between A: MMC patients and controls, B: MMC patients and their unaffected siblings, C: unaffected siblings and controls. MMC=myelomeningocele, N=number, SD=standard deviation. P-value: ***<0.001, ****<0.0001.

Pair	1		2		3		4		5		6		7		8		9		10		11		12	
	MMC	Sibling	MMC	Sibling	MMC	Sibling	MMC	Sibling	MMC	Sibling	MMC	Sibling	MMC	Sibling	MMC	Sibling	MMC	Sibling	MMC	Sibling	MMC	Sibling	MMC	Sibling
MTHFR genotype	CT	CC	CC	CC	TT	CC	CC	CT	CC	TT	CC	CC	CT	CC	CT	CC	CT	CC	CT	CC	CT	CT	CT	CT
Ethnicity	Belgian		Belgian		Belgian		Belgian		Belgian		Belgian		Belgian		Ukrainian		Turkish		Turkish		Belgian		Belgian	
Age difference	-20		-19		-20		-26		20		-18		17		53		-101		83		-27		-24	
CpG 8	61	54	36	38	45	52	25	24	55	50	45	47	48	54	53	15	11	62	40	20	40	35	47	56
CpG 9		58	40	25	42	57	15	35	53	57	41	53	47	55			19	66	37	25	30	34	49	52
CpG 10	50	37	37	30	41	48	19	26	49	47	37	43	39		51		7	67	30	14	31	32	47	52
CpG 11-15	54	50	38	33	46	47	24	27	50	41	37	46	48	49	41	16	16	57	37	21	39	38	43	50
CpG 17,18	48	44	26	31	37	45	18	23	43	37	35	36	34	41		14	8	53	30	12	28	29	40	50
CpG19	57	46	37	34	44	49	17	22	53	43	38	51	53	47	50	9	7	65	37	16	39	34	47	53
CpG21,22	61	54	36	38	45	52	25	24	55	50	45	47	48	54	53	15	11	62	40	20	40	35	47	56
CpG23-25	54	50	38	33	46	47	24	27	50	41	37	46	48	49	41	16	16	57	37	21	39	38	43	50
CpG26	48	44	26	31		45	18	23	43	37	35	36	34	41	35	14	8	53	30	12	28	29	40	50

Table 4.S3: *HOXB7* methylation by Sequenom EpiTYPER for the 12 pairs of MMC patients and their unaffected siblings.MMC: myelomeningocele; MTHFR 677C>T (CC/CT/TT); Age difference (in months) between patients and their unaffected siblings.

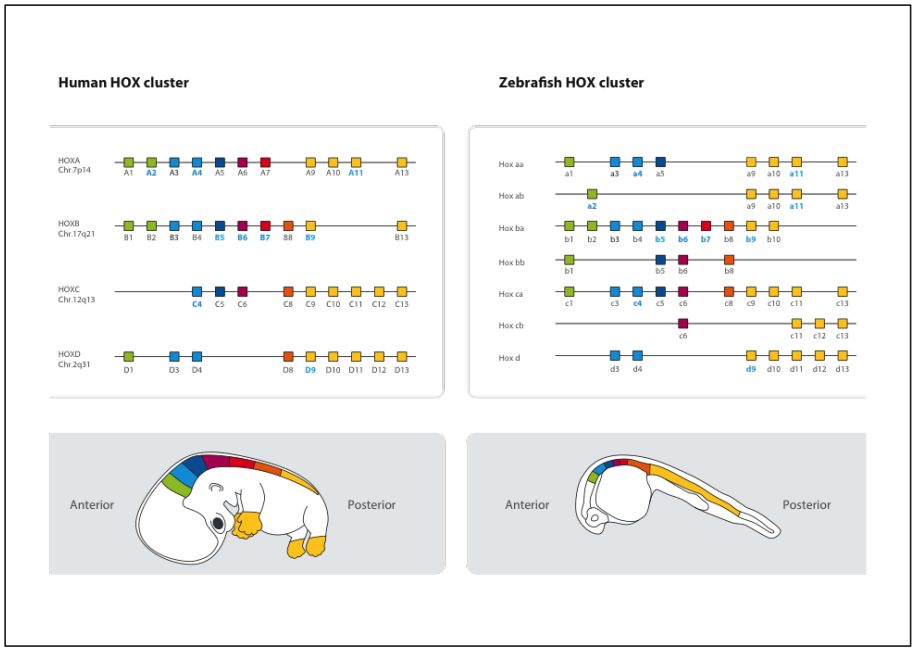


Figure 4.S1: Genomic organization and expression patterns of *HOX* genes for humans and zebrafish. *HOX* genes are evolutionary highly conserved, it appears that an ancestral *HOX* gene cluster was duplicated twice within the 2R genome duplication events, and was important to allow the emergence of vertebrates. Mammals have 39 *HOX* genes distributed in 4 clusters; they pattern the embryo along the rostro-caudal axis. Based on sequence similarities and position within the cluster, genes can be classified in 13 paralog groups (*HOX1-13*). They become activated during early embryogenesis in a time sequence that reflects their physical order within the clusters, referred to as 'temporo-spatial collinearity'. Genes at the 3' end of the clusters are activated first, whereas genes located at more 5' genomic positions are activated subsequently. The 3' paralog groups *HOX1-5* have restricted expression borders in general in the hindbrain, while *HOX4-11* genes are detected in the spinal cord. The overall number of *HOX* genes in zebrafish (48) is close to that in mammals (39), despite the existence of seven clusters. This is due to a massive elimination of *HOX* genes after the additional duplication. Genes indicated in blue refer to the *HOX* genes that are differentially methylated after analysis of the 450K data with the IMA R-package. [Figure adapted with permission from Veraksa et al. (2000)]

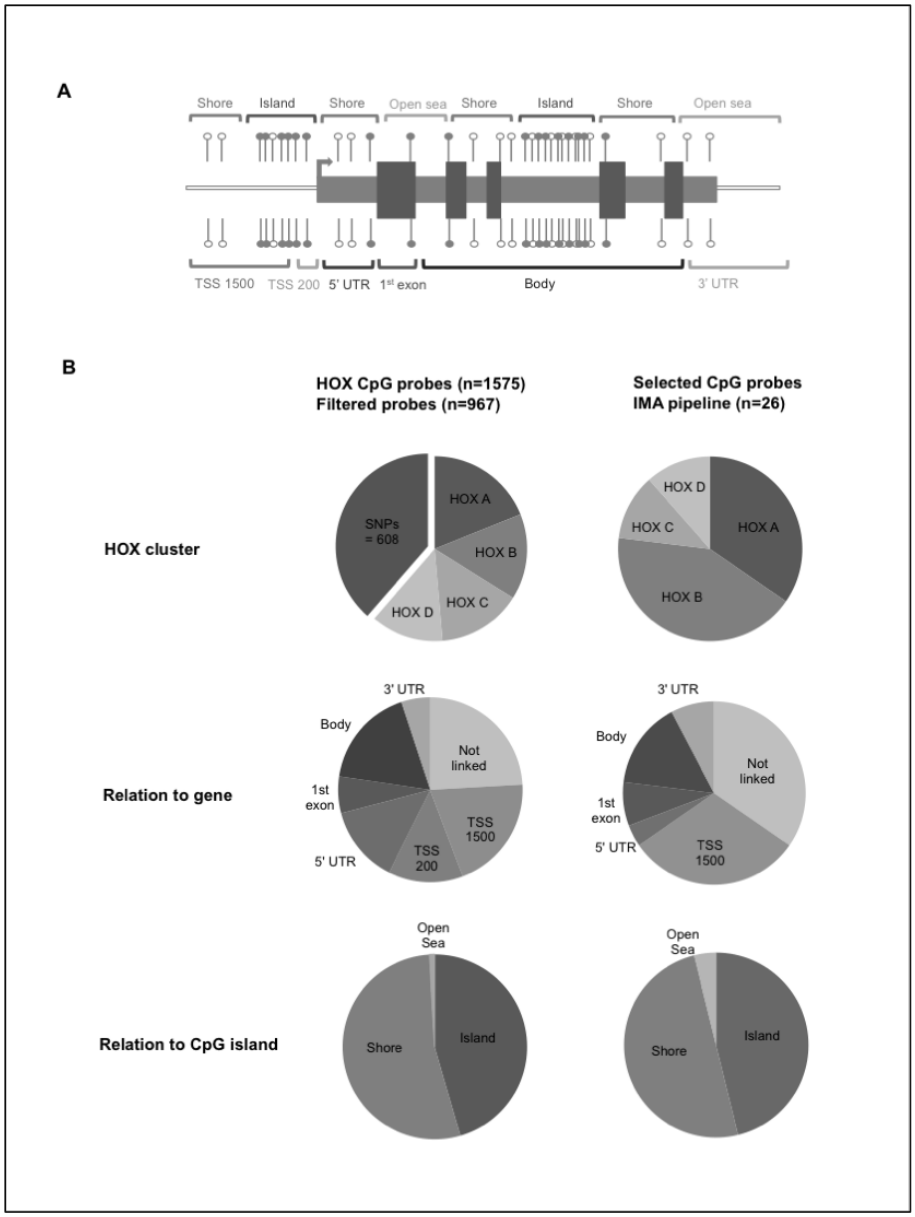


Figure 4.S2: Genomic context of CpG methylation by HumanMethylation 450K BeadChip. **A**. Schematic diagram of CpGs depicts their genomic context relative to the nearest CpG island (top) or gene (bottom). Island context is defined as being within an "island" (dark grey), in a 4kb "shore" (grey) flanking the island, or in an "open sea" (light grey) more than 4kb from an island. Gene context is defined relative to the nearest open reading frame: within 1500 (TSS1500) or 200 bp (TSS200) of a transcription start site, in the 5' UTR, the 1st exon of a transcript, in the body of gene or the 3' UTR. **B**. Pie charts show the distribution of the CpGs examined based upon their genomic context for all probes retained from the array (left column) and for all the differentially methylated probes (right column). Distribution within the *HOX* clusters and gene and island context after analysis with the IMA R package is presented.

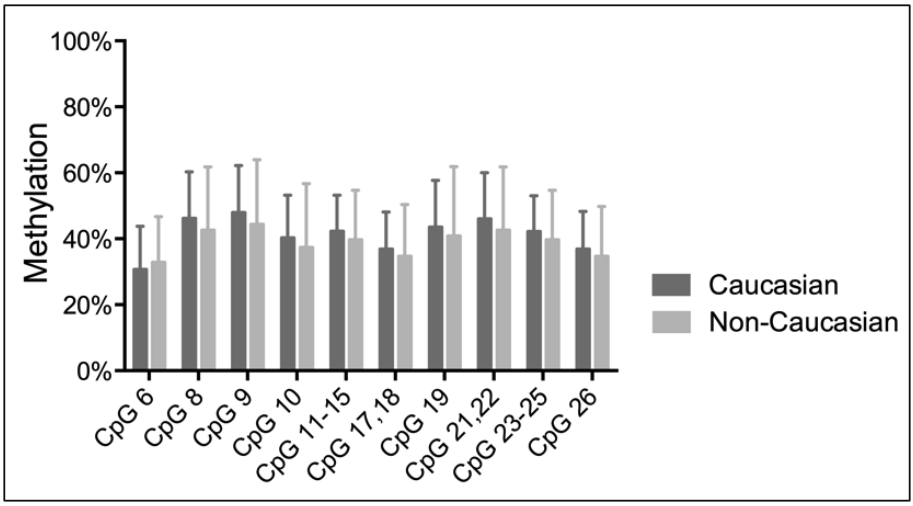


Figure 4.S3: Methylation pattern of *HOXB7* by Sequenom EpiTYPER in 70 Belgian MMC patients and 10 non-Caucasian (5 Moroccan and 5 Turkish) patients.

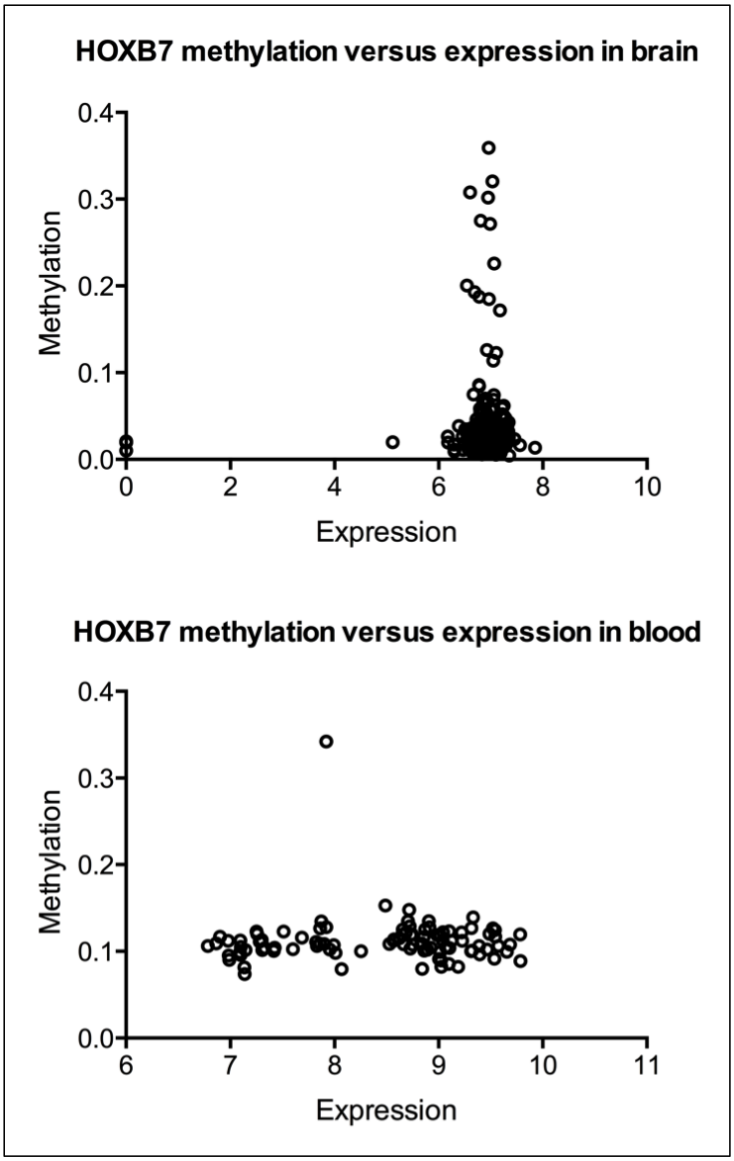


Figure 4.S4: *HOXB7* cg06493080 methylation versus expression in A: brain and B: blood. A low degree of methylation and high expression is observed for brain tissue. In blood *HOXB7* has a higher methylation compared to brain tissue and the expression pattern is highly variable for similar methylation levels. Data are extracted from the MENT database (Methylation and Expression database of Normal and Tumor tissues (http://mgrc.kribb.re.kr:8080/MENT/). Cg06493080 is located in the promotor region of *HOXB7* on position Chr17:46688310. Datasets GSE15745 (brain) and GSE26212 (blood) contain respectively 492 and 96 samples.

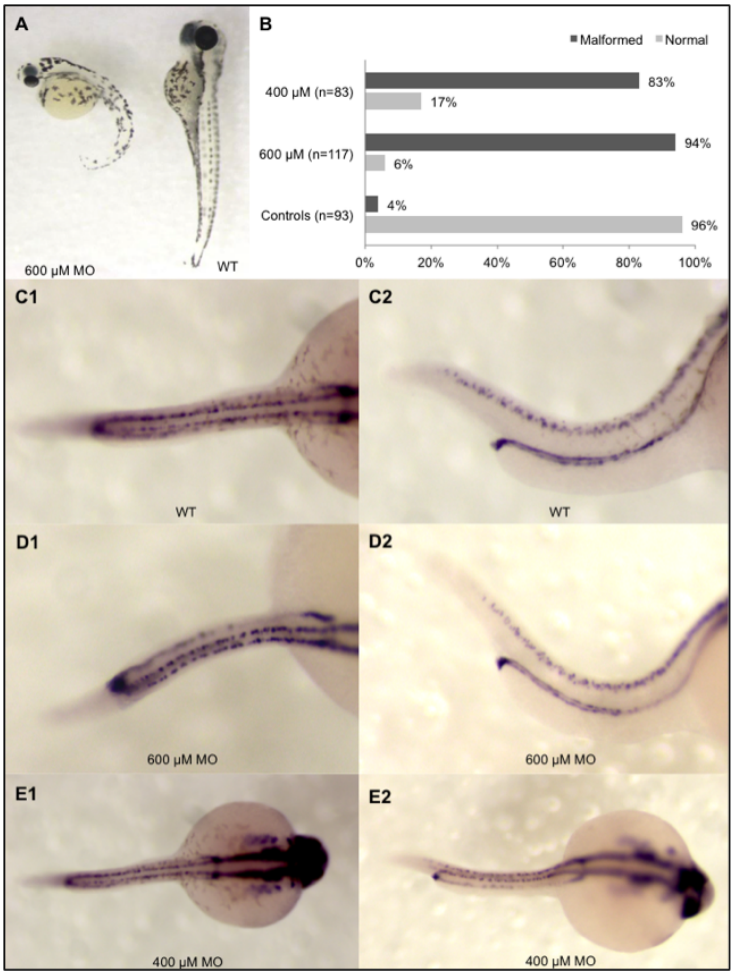


Figure 4.S5: Phenotype analysis of Hoxb7a depletion in zebrafish embryos. **A**: Phenotype analysis at 72hpf of 600 μM Hoxb7a MO injected zebrafish resulted in a developmental delay with hypopigmentation, shortening and bending of the tail. **B**: Phenotype analysis after pax2a staining at 24 hpf resulted in about 88% embryos with a developmental delay. **C1-2**: Pax2a WISH at 24hpf in control AB zebrafish; **D1-2**: Pax2a staining in 24hpf zebrafish after microinjection of 600 μM Hoxb7a MO is normal; **E1-2**: Pax2a staining in 24hpf zebrafish after microinjection of 400 μM Hoxb7a MO is normal.

Chapter 5

Messy data

From the previous chapters it is clear that unexpected results can surface during an analysis even when one tries to prevent such unexpected results from occurring. In this chapter an overview will be given on *why* some analyses can lead to unexpected results and *how* one can try to prevent this by using a conceptual framework that will be proposed and which can be used to get an indication of how likely it is that one will be able to extract meaningful conclusions from a given data set. In this framework, two dimensions are important: messiness and the amount of data, i.e. the volume of data combined with the number of dimensions. In Section 5.1 these dimensions will be explained in more detail before the framework is introduced in Section 5.2.

5.1 Messiness and data size

5.1.1 Data volume and dimensions

More data tends to lead to better models even with simple algorithms (Rajaraman, 2008; Garrett, 2013). There are two ways to interpret 'more' data: either the number of available data points is larger or the number of (independent) variables is larger. Adding more data points will lead to a point where adding more data points will no longer have a significant influence on the model. Once the relevant trends or patterns in the data have been identified, adding more points might refine the model a bit further but new insights will not be gained anymore. An example of this can be seen in Figure 5.1 where we first start with only two data points and the best model being a straight

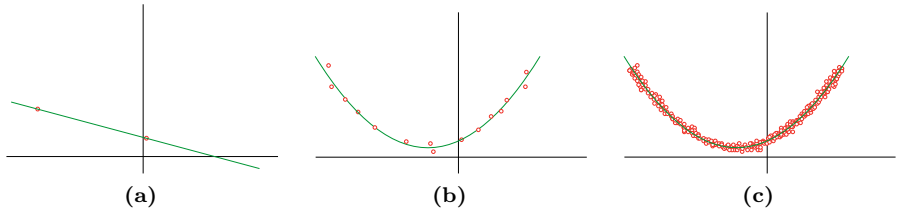


Figure 5.1: Generating a model from a number of data points. (a) Only having two points, the simplest model would be a straight line through those points. (b) Sampling more points reveals a quadratic relationship between the points. (c) Adding many more points no longer improves the model but just reinforces it.

line between those points (Figure 5.1a). Adding more points as in Figure 5.1b reveals a quadratic relationship between the variables. After a certain number of points, the parabolic relationship will only be reinforced by the new points but no further insights are gained as can be seen in Figure 5.1c.

When a model does not lead to satisfying results, adding more dimensions can give the algorithms the information needed to create a better model. For instance, the prediction algorithms PolyPhen (Adzhubei et al., 2010) and SIFT (Ng, 2003) have a hard time accurately predicting whether a mutation is going to be damaging with respect to a certain disease. By adding other dimensions to the data such as the phenotype and a range of other biological information about the gene, Sifrim et al. (2013) were able to construct an algorithm that has far less false positives compared to the techniques not using this additional information. Another example can be found in epidemiology where models that do not include travel patterns from infected patients lead to unexpected behavior of the epidemic when looking at the model (Knipl et al., 2015). One very important aspect however is that adding dimensions only improves the results when these dimensions are relevant to the problem being analyzed. As a very simple example we could say that when predicting weight loss, adding stock market data on all shares would most likely not improve the accuracy of the algorithm.

Although more data and dimensions might seem to always be better, it is necessary to keep in mind "the curse of dimensionality", a term introduced by Bellman (1957). This 'curse' generally means that when the number of dimensions increases, the amount of data needed to create and support models often increases exponentially with it. Often, because of this exponential increase, methods may start to break down because they do not scale well as the amount of data and time needed for analysis will also increase exponentially. In addition, adding variables can also lead to a decrease in power of a statistical test (Pirinen

et al., 2012). Finally, Taleb (2015) warns about error propagation in higher dimensions where adding more variables can lead to a dramatic increase in error. This increase in error comes from small errors in the measurements of the different variables. Especially in high dimensional problems such as economics or ecology these individual small errors can lead to much larger errors in the complete model. In light of the proposed framework we see that the messier the data becomes, the larger the error in the resulting model can be and the more difficult it will be to make accurate predictions. Possible solutions to these problems consist of either removing some dimensions from the data, reducing the number of dimensions by using techniques such as principal component analysis (PCA) or by using other algorithms that support a higher number of dimensions (Hotelling, 1933; Tibshirani, 1996; Bouveyron & Brunet-Saumard, 2014; Hall et al., 2005; Murtagh, 2009).

5.1.1.1 Power analysis

One way to ascertain whether there is enough data available to draw a conclusion is by performing a power analysis. As said before in Section 1.1 of the introduction, the power of a statistical test is the probability of rejecting the null hypothesis when it really is false. In other words, the higher the power of a statistical test, the more likely a difference can be detected when there really is one. A power analysis will either provide the minimal sample size needed to detect an effect of a given size or the smallest detectable effect given a certain sample size. More data will allow smaller effects to be detected but getting more data might not always be possible or economically affordable (UCLA: Statistical Consulting Group, 2015). For instance, when performing a genome-wide association study (GWAS) where the susceptibility associated alleles occur in less than 10% of the population and the odds ratio is less than 1.3, more than 10,000 cases and the same amount of controls would be needed (Wang et al., 2005).

There are several software packages available that can calculate the power of a statistical test. Statistical packages such as SAS (SAS Institute Inc., 2014), PASS (NCSS, 2015) and SPSS SamplePower (IBM, 2015) include functions to calculate power. In order to calculate the power a number of parameters have to be provided depending on the type of test that will be performed. For instance, in order to test the difference between two means with a t-test, the mean and the standard deviation for each population have to be known. For a chi-square test the proportion of positives in each condition has to be known. This also means that some statistical summaries of the populations have to be known before the power of the analysis can be calculated. While this is not always easy there are several ways to make an educated guess. First reviewing

literature may provide possible effect sizes and other summaries. Second, a small pilot study can be performed and the summaries from that smaller study can be used to enter in the power analysis keeping in mind that with a smaller study, the variance and error in the numbers may be larger. Finally, Cohen's recommendations can be used which depend on whether the expected effect size is small, medium or large (UCLA: Statistical Consulting Group, 2015).

In some cases and especially when doing exploratory research, it will be very hard to make an educated guess of these numbers and therefore extremely difficult to calculate the correct power. Once interaction effects come into play it can become even harder as many software packages do not provide the option to calculate power of a model including an interaction effect (UCLA: Statistical Consulting Group, 2015). Another problem is that the underlying distribution of the data has to be known beforehand to do a proper power analysis. For instance, with relatively symmetric distributions, the t-test gives more power compared to the nonparametric Wilcoxon Rank-Sum test but for non-symmetric distributions it is the other way around (Bridge & Sawilowsky, 1999). On the other hand, even with tests that give comparable results for unknown distributions such as the Baumgartner-Weiß-Schindler test, an estimate of the effect size still has to be provided (Neuhauser & Senske, 2004). A field where we are confronted with these problems is epigenetics. For epigenetic studies there is a lack of data and theory concerning epigenetic modifications to predict the distributions of certain modifications and the effect sizes that are relevant, although much work is being done in order to overcome these problems (Rakyan et al., 2011; Tsai & Bell, 2015; Li et al., 2015). This fact is also relevant for the study described in Chapter 4 and its location in the proposed framework (Section 5.2). In any case, while for many of these problems a correct power analysis may be obtainable in theory by performing a thorough analysis of domain data by statisticians, for many researchers this may prove to be too difficult.

5.1.2 Messiness

While the importance of the volume of data and the number of dimensions are relatively easy to understand and explain, messiness is rather hard to describe. Many studies mention the use of messy data but without giving an explanation or definition of what is really understood by messy data (Hauser, 2012; Burr & Cowan, 1992; Gould et al., 2006; Blommaert, 2013; Trenberth, 2010). Other studies do mention a definition or description of what they understand to be messy data but these definitions vary widely. For instance Urquhart & Weeks (1978) say that "Cross-classified data with unequal replication and perhaps empty cells often is called "messy data"" while Chi (1997) says "Messy data refer to such things as verbal explanations, observations, and videotapings,

as well as gestures". What most definitions do have in common however is the fact that they consider data to be messy when statistical assumptions are violated, observations are missing, sampling biases are introduced, data are from asymmetric distributions, and unexpected extreme results or errors are present in the data (Baguley, 2012; Milliken & Johnson, 1993; Longford, 2001; Reed & Stark, 1996; Hafley & Lewis, 1963; Legendre, 1993; Johnson & Slatkin, 2006; Harvey et al., 1999; Burke, 2001). As there is no clear definition available we will provide a definition of messiness that will be used throughout this dissertation. We define data to have high messiness when the data contains errors, violates statistical assumptions or is influenced by stochastic processes, non-linear interactions and feedback loops, environmental effects, human behavior, missing data, and temporal effects and when these factors can not be easily modeled, abstracted away, or are unknown.

Below we will briefly describe each of these factors.

5.1.2.1 Errors

When an analysis is performed on a data set the results of that analysis largely depend on the data that is used as input. If that data contains errors, it is very likely that the conclusions drawn from that data will be incorrect or as is said 'garbage in, garbage out'. These errors in the data can come from a variety of sources and they can either be completely random or systematic. Random errors are errors that can be revealed by repeating the measurements while systematic errors cannot be revealed this way (Taylor, 1997). To clarify the difference between these types of error a few examples and sources of errors are given in the next paragraph.

An important source of errors are *human errors*. Humans are prone to make mistakes on many different levels when they are collecting and generating data. A source of random errors could be the counting of birds in a flock with the naked eye. It is likely that one time the number of birds is overestimated and another time underestimated. By taking a large number of measurements these errors will cancel each other out and the mean would be correct although the variance might be larger. If the goal would be to identify the different species of birds in a forest, a random misclassification would not have a huge impact on the total number per species. On the other hand when a researcher systematically mis-classifies a certain species the final results would be biased. Also by using instruments in the wrong way when collecting measurements random or systematic errors can be introduced. Reading values of an analog scale for instance will introduce random errors as sometimes the value will be overestimated and other times underestimated. Improper use of an instrument

on the other hand will lead to systematic errors as the measurement will always be under- or overestimated. Another source of human error can be introduced during data entry when incorrect data is entered. Again this can either be random (e.g. sometimes pressing a random wrong button) or systematic (e.g. pressing the wrong button consistently for a certain class of observations).

Not only humans can make errors, also the *equipment* that is used for collecting data can introduce errors. A digital scale used to measure the weight of a certain compound may not be so accurate when measuring least significant digit which can lead to a random error. On the other hand, when that same machine is poorly calibrated, malfunctions, or shows a measuring error, systematic errors will be introduced in the data.

A last source of errors that is not necessarily caused by humans or machines and could be considered an extreme case of systematic error, is when the *wrong variables* are measured. In that case the measurements itself may be totally correct but as the analysis starts with the wrong input, the conclusions can never be correct. An example of this can be found in epigenetics where the epigenetic profile differs between different cell types. Performing measurements in the wrong cell type can give entirely different results compared with results gained from measuring in the correct cell type.

Only when the mentioned errors are consistent and/or predictable will it be possible to correct them. In any other case the messiness of the system will increase.

5.1.2.2 Statistical assumptions

Most statistical models assume that all data points are independent, leading to so called 'thin tailed' distributions. These are distributions where a single data point does not differ extremely from the other data points. However, for many systems, not all data points will be independent and a single data point can be extremely different from the other data points. These distributions are called 'fat tailed'. Examples that show the difference between these two classes of distributions are human weight and income. Human weight is thin tailed as it is impossible to find a person that would weigh twenty times the weight of an average adult. Income on the other hand is fat tailed as it is not unimaginable that in a random sample, the income of one person can be twenty times larger than the average income (Taleb et al., 2014).

An extreme event that occurs in fat tailed distributions was coined the term 'black swan' by Taleb (2007). The definition he gives is: "The black Swan is a) a highly unexpected event for a given observer, b) carries large consequences,

and c) is subjected to ex-post rationalization." (Taleb, 2015). Essential in this definition is that it is unexpected for a given observer. While a terrorist attack is completely unexpected by many people, for the terrorist and other accomplices it is not. As these events are extreme and carry large consequences, it is unlikely that they can be observed in past data but that does not mean that such an event will not occur. Observing a black swan can prove that the statement 'all swans are white' is false but not observing one does not mean that they do not exist. Finally, a black swan event can be explained or rationalized afterwards by hindsight. One of these black swan events is the financial crisis of 2008 where we saw a cascading set of worldwide events almost bringing down the entire system. In hindsight we were able to see why this was possible but at the moment it came as a complete surprise to many people.

Especially when models are used for risk prediction the difference between thin and fat tailed distributions becomes important. An example of this is given by Bar-Yam (2014) in a paper on the Ebola epidemic. Traditional epidemiological models provide an average number of uninfected people an infectious patient might infect (R_0). If R_0 is larger than one more and more people will become infected and when R_0 is smaller than one, the infection will eventually die out. As these epidemiological models use thin tailed distributions they assume that the chance of finding an individual infecting many more individuals than average is extremely small. In a fat tailed distribution this would not be the case and the chance of finding a 'superspreader' - as already documented in the SARS epidemic (Stein, 2011) - is much higher. As the epidemiological models of Ebola are based on findings in rural Africa with many small communities, the thin tailed distribution can be applied. When transferring that same model to huge cities that are interconnected by airplanes, it is easy to see that the model may break down. Indeed, a single infectious individual on an airport can infect any number of other passengers who in turn can infect an even higher number of people in other metropolitan areas. For instance, a security guard at the airport unknown to be infected with Ebola can come into close contact with several hundreds of passengers which is many times the average number from the African studies. The effects of this can be seen in Figure 5.2 where the number of infected people is followed over time but with small variations of the mean R_0 simulating a sometimes higher or sometimes lower rate of infection compared to the average. It is clear that even with small variations, there are some extreme scenarios where the number of infected people differs from the average by an order of magnitude. So, in our interconnected society the assumption of a thin tailed distribution no longer holds and the traditional models can give a false sense of security as they predict extreme events to be extremely unlikely while they are not. These fat tailed distributions can be found in many fields of both natural and social science with over 80 examples identified (Andriani & McKelvey, 2007).

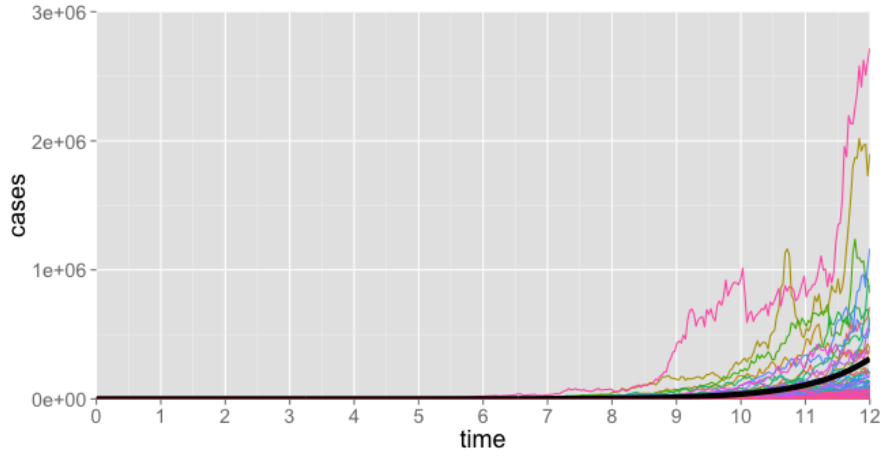


Figure 5.2: Number of infected individuals over time. The black graph shows the number of infected people using the average R_0 while the colored graphs show the distribution with small variations of R_0 at each time step. Figure from T. Verbeiren, with permission

Overall, we can say that when statistical assumptions are violated in a model, the messiness is higher as it can lead to unexpected results.

5.1.2.3 Stochastic processes

Also an important factor in determining the messiness of a system is the existence of stochastic processes, i.e. processes driven by random events. In many biological processes, these effects are also called 'noise' in the data. In non-linear biological dynamical systems this noise can completely obfuscate the underlying deterministic behavior of the system (Goel & Richter-Dyn, 2013). A well known example of a stochastic process is the inactivation of the X chromosome in human female embryos (Monkhorst et al., 2008). Another example is found in monozygotic twins where paradoxical differences can be explained by stochastic epigenetic changes (Wong et al., 2005). In quantitative genetic analysis where the prediction of a mean phenotype from the genotype is the ultimate goal, stochastic noise can make an accurate prediction almost

impossible (Jimenez-Gomez et al., 2011). But stochastic processes do not only play a role in biology, they are also found in traffic modeling (Chowdhury & Schadschneider, 1999; Kerner, 2013), weather prediction (Semenov & Barrow, 1997; Wilks & Wilby, 1999; Srikanthan & McMahon, 2001), chemistry and physics (Gillespie, 1976, 2007; Cao et al., 2004; Van Kampen, 1992; Freund & Pöschel, 2000; Cottis, 2001), and even studies on housing choices (Chung & Haurin, 2002).

It is important to note though that sometimes stochastic noise can make weak signals detectable. This effect, called stochastic resonance, has been found to improve the detection of signals in many fields such as medicine, physics, chemistry and biology (Hänggi, 2002; Benzi et al., 1982; Morse & Evans, 1996; Priplata et al., 2003; Guderian et al., 1996; Yilmaz et al., 2013; Chiou-Tan et al., 1997). A nice visual representation of stochastic resonance is given by Simonotto et al. (1997) who show that a signal that is generally below the detection threshold can be detected by the addition of noise. In Figure 5.3 the top part shows a signal that is below the threshold but the added noise makes the signal jump that threshold on some occasions. The bottom part of Figure 5.3 shows a picture where a pixel is drawn white when the signal is below the threshold and black when it is above the threshold. By adding three different levels of noise, some of the pixels jump above the threshold and are therefore drawn black instead of white. It is clear that a certain amount of noise (middle picture) gives the best results for interpretation. So although the intrinsic messiness of the system stays the same, the perceived messiness can be lower as the signal becomes detectable.

5.1.2.4 Non-linear relationships

Another factor that can increase the messiness of a system is the existence of non-linear relationships in the system such as feedback mechanisms. Analyzing a system while being unaware of certain feedback mechanisms or other forms of interaction, may lead to totally unexpected results or conclusions. For example, methamphetamine dependence causes some parts of the brain to increase in volume while other parts decrease in volume in HIV infected patients (Jernigan et al., 2005). In at least one part of the brain the combination of these processes can yield apparently normal volumes in HIV infected patients who are methamphetamine dependent.

A different example can be found when identifying the genetic components of diseases. It is possible to look at a single gene or protein as the cause for a certain disease, which essentially means looking at a linear system. However, in reality many processes are non-linear and involve a number of gene and/or

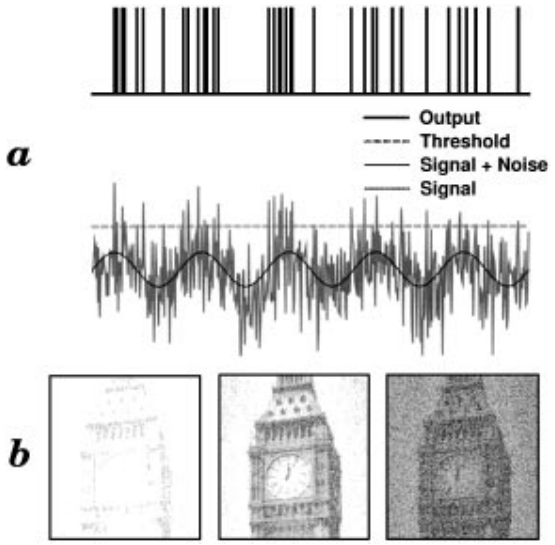


Figure 5.3: (a) A signal is shown below a certain detection threshold. By adding noise to that signal, the resulting signal sometimes jumps over the detection threshold yielding the output above. (b) All pixels of an image are placed below a certain threshold and painted white. By adding three different levels of noise from low to high (left to right) some pixels jump above the threshold and are painted black allowing the human eye to identify what was in the original picture. From Simonotto et al. (1997)

protein interactions. This has been a problem for many GWAS where it is assumed that much of the observed variation is additive and that the causal mutations have measurable effects on the phenotype (Marjoram et al., 2014). This also leads to the overestimation of the total heritability. For example, 71 risk loci have been identified associated with Crohn's disease in a meta-analysis (Franke et al., 2010). If we were to consider these loci to be strictly additive, they would explain 21.5% of the heritability. Using a model that allows for interaction between these loci, can increase the explained heritability to 58% or even more (Zuk et al., 2012). Even though for many genes and proteins interaction data is available from several databases (Jensen et al., 2009; Stark, 2006), many interactions are still unknown or poorly understood.

In addition to the examples described here, the importance of modeling interaction effects has also been addressed in several other studies where they investigated ecological footprint inequality (Duro & Teixidó-Figueras, 2013), learning characteristics and instructional conditions (Hwu & Sun, 2012), neonatal physical development (Tang et al., 2012), oscillations in blood pressure (Ringwood & Malpas, 2001), and family ownership of international businesses

(Sciascia et al., 2012).

5.1.2.5 Environmental effects

A third factor that can have a great influence on the messiness of a system is the environment. Several studies into the relationship between genotype and phenotype have found that environmental effects can have a large effect on the phenotype. For instance, new traits can appear when certain molecular chaperons aiding in protein folding are depleted due to environmental stress (Lindquist, 2009). Another example can be found in the study of the effects of two sets of alleles on the sporulation of yeast where researchers found - as they expected - these different alleles influence the efficiency of sporulation but also found that the explained variability between different environments varies (Gerke et al., 2010). Many studies have been performed on yeast but in humans environmental factors have also been found to play a role. For example, estrogen use has an effect on cognitive decline dependent on APOE-ϵ4 genotype, a genetic risk factor for Alzheimer's disease (Yaffe et al., 2000). In medicine there are many examples where the environment has a large influence such as in asthma (Björkstén, 1999; Gilliland et al., 2001), inflammatory bowel disease (Danese et al., 2004), cardiovascular disease (Navas-Acien et al., 2007) and cancer (Czene et al., 2002; Sasco, 2001). These examples all focused on biological effects but environmental effects can also play an important role in social sciences where we find environmental effects such as socioeconomic status (SES), peer pressure and enriched rearing (Hoff, 2003; Magolda & Astin, 1993; Rowe et al., 1999).

From these examples it is clear that environmental factors can have a large effect on a model. Ideally these effects can either be abstracted away or added to the model. However, when they are unknown, impossible to abstract away or difficult to describe, the messiness of the system can increase dramatically. A possible solution to counter the effect of the environmental effects is to conduct the experiments in as similar conditions as possible although this might not always be feasible, for instance when studying different populations.

5.1.2.6 Temporal effects

In a deterministic system the outcome of a certain process can exactly be predicted based on its past and in a Markov chain, the next state of a system only depends on the current state. But sometimes the complete past of a system is unknown or the next state can be influenced by a previous state or a combination of previous states. There are several examples where events in the

past have an influence on the state of a system later on. For diabetes type II for example, there is a greatly increased susceptibility to develop the disease when there is inadequate nutrition in the fetal and early post-natal stages (Hales & Barker, 2013). Also children born prematurely show a reduction in insulin resistance, a marker for diabetes type II (Hofman et al., 2004). Another study found that a low birth weight followed by accelerated weight gain during childhood is associated with coronary heart disease, diabetes type II and high blood pressure (Barker et al., 2002). But also in cancer, temporal effects can play an important role. One of the most known examples is that of asbestos exposure where there can be more than 40 years between exposure and the subsequent consequences (Cugell & Kamp, 2004). These temporal effects can increase the messiness of the system as they may not be known or their effects poorly understood at the time of analysis.

We have to remain cautious however not to confuse temporal effects or other historical data with other confounding factors. For example, in a well known paper by Angrist & Krueger (1991) they describe that the season of birth has a strong influence on school results and on earnings later in life. However, Buckles & Hungerman (2013) found that nearly half of this correlation can be explained by socioeconomic factors instead of seasonal influences on health and nutrition or by compulsory schooling laws.

5.1.2.7 Missing data

Many statistical tests assume that there is no missing data but this assumption may not hold in reality. There are three types of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Schafer & Graham, 2002). To explain these types we will start from a theoretical example where it is investigated whether people with a lower income visit the doctor more often compared to people with higher incomes. When the data is MCAR means that the missing values are not dependent on any other values in the data. In our example this would for instance mean that during some visits to the doctor the computer of the doctor malfunctioned and the visit was not recorded. This factor is completely independent of the number of visits so the obtained data from other patients can be considered a random subset of the complete data. In case of MAR there is a relation between another measured variable and the missing values. Going back to our example this would mean that there is an effect of age on the number of visits to the doctor. In this case the value of the number of visits can be predicted or estimated from the age of the patient. Finally, when the data is MNAR it means that the variable can not be predicted by any other measured variables but solely relates to the independent variable. Again, returning to our example, people

with low income might not have enough money to go to the doctor resulting in the visit not being recorded. Especially missing data that is MNAR poses problems as the resulting parameter estimates for the missing values will be biased (Graham, 2009; Rubin et al., 2007).

One way to deal with missing data is to remove all data points that have missing values. While this is surely one of the easiest ways, removing samples can reduce power and in case of small sample sizes there might not be enough data left when removing the incomplete samples. There is a myriad of other methods that can handle missing data such as listwise deletion, pairwise deletion, and maximum likelihood methods (Myers, 2011). Especially in the case of missing data being MNAR, the messiness of the system is higher. For MNAR data assumptions have to be made about the underlying model of the missing data and this may prove to be very difficult if not impossible in case of exploratory research. Choosing the wrong model can lead to a lower performance of the different methods (Graham, 2009).

5.1.2.8 Human behavior

A last factor that can lead to a huge increase in the messiness of a system is when the system being modeled includes human behavior. An increase in messiness is not always necessary because not all human behavior is completely random but can be driven by societal influences, instincts, and reflexes. Because some basic principles of human behavior are already known it is sometimes possible to model the behavior especially as more and more data is being collected that can be mined to detect individual and group behavior (Kennedy, 2012; Lazer et al., 2009). Recommender systems are a well known example of a technology that is aimed to predict what would likely be of use for an individual based on the behavior of other consumers (Ricci et al., 2011).

On the other hand, in emergency situations humans behave differently than would be expected leading to possibly wrong conclusions when trying to predict behavior in these circumstances. For instance, one would assume that during a fire people would try to find the nearest exit and move away from the source of the fire. In a fire in Manchester in 1979 however, nine out of ten deaths occurred in the canteen where eye witnesses said that a number of people were hesitant to leave before finishing their meals or paying for what they had consumed (Johnson, 2005). Simulations of real life evacuations in airplanes show that giving the participants a financial reward for exiting the airplane as fast as possible, drastically changes their behavior and can lead to blockages that are not observed without this motivation (Muir et al., 1996; McLean et al., 1996). Because of these unexpected behaviors, many studies have tried to identify

why people reacted a certain way in face of eminent danger and attempted to include this possible 'irrational' behavior in prediction models (Cornwell, 2003; Proulx, 2001; Schadschneider et al., 2011; Miyoshi et al., 2010). Behavior does not always have to be unexpected or contrary to expectations in order to increase messiness. Also in epidemiology attempts have been made to model the influence of human behavior on the dynamics of an epidemic and although these studies may offer insights into the spread of a disease, the exact behavior of large groups of people can remain unpredictable (Funk et al., 2010).

Also in social sciences human behavior can lead to unexpected results for example because of strong heuristic processes in the brain competing with analytical processes (Evans, 2006). For instance, even when participants of a study are specifically told to resort to analytical reasoning, heuristic processes can take over and lead the participants to give the wrong - heuristic - answer. This was found to be the case in a study where students were asked to interpret box plots. Even when the students are told that each area of the box represents the same number of observations, i.e. 25% of the total number of observations, they gave the heuristic answer that a bigger area of the box meant that there were more observations in that part (Lem et al., 2013). Another example from the social sciences is when people are aware of certain dangers or other relevant information but behave in a way that would not be expected from someone who is aware of this information. For example, the relation between knowledge about AIDS and practicing safe sex is very weak (Helweg-Larsen & Collins, 1997; Ananth & Koopman, 2003). In the same manner, knowing about the dangers of alcohol consumption has no effect on drinking behavior (amount and frequency) and knowledge about the environment had no effect on energy conservation (Ajzen et al., 2011). A final example can be found in learning and e-learning environments where teachers and developers are confronted with the difficulty to predict the students' behavior and perception of the developed program (Martens et al., 2007; Broekkamp et al., 2002; Vermetten et al., 2002).

In all examples above, researchers have tried to model or predict human behavior - with varying and sometimes unexpected results - but not modeling behavior can potentially lead to a model completely breaking down. For example, one of the many factors that ultimately led to the financial crisis is the lack of including the behavior of people who are active in the financial world and only relying on past numerical data to model the financial markets (Rajan et al., 2015).

5.2 Framework

Now that we understand what messiness is and what it means to have 'more data', it is possible to design a framework that gives an indication of the likelihood that an experiment with a certain messiness and amount of data will give consistent and accurate output. When talking about a framework in this case, we are not talking about an elaborate structure as is meant in for example software development but about a basic structure underlying the concept we want to discuss. The framework is purely descriptive and qualitative as it is meant to raise awareness and to encourage thinking about the data and the possible problems that may arise before performing the actual analysis.

In the graph shown in Figure 5.4 the messiness of the system is represented on the X-axis and the amount of data in relevant dimensions is on the Y-axis. It is important to note that both the amount of data and the number of relevant dimensions is important. While in one case more data points might be needed as can be seen in Figure 5.1, in another case more dimensions might be needed as in the example of Sifrim et al. (2013), and yet in another case both might be needed, always keeping in mind the 'curse of dimensionality' and the fact that only relevant dimensions add real information to the data. This is why in this framework, the Y-axis represents both volume and number of relevant dimensions. Scoring high on this axis means that there is more than enough data to work with and the number of dimensions is adequate to make predictions or draw conclusions from the data.

The color of the bands in the framework represent the difficulty one would expect when analyzing the data with these characteristics. To explain the differences, four zones (A-B-C-D) are identified in Figure 5.4. Zone A would be the ideal location for the system being investigated. In this corner the messiness of the system is low and enough data - with respect to both volume and relevant dimensions - is available. This also means that it is more likely that meaningful and unambiguous information can be extracted without too many problems or unexpected or paradoxical results. The next zones B and C are located in the same color so they are equally likely to yield meaningful results. The difference between them lies in the messiness and amount of data. In zone B, the system has high messiness but there is enough data - both volume and relevant dimensions - available. Zone C on the other hand, has much lower messiness so there is less data needed to get meaningful results. In the lower right corner, we find zone D which is the least favorable location as it implies that there is not enough data available, either due to a too low volume or too low number of dimensions, and the messiness of the system is high. This makes it hard if not impossible to gain any meaningful insights from the data.

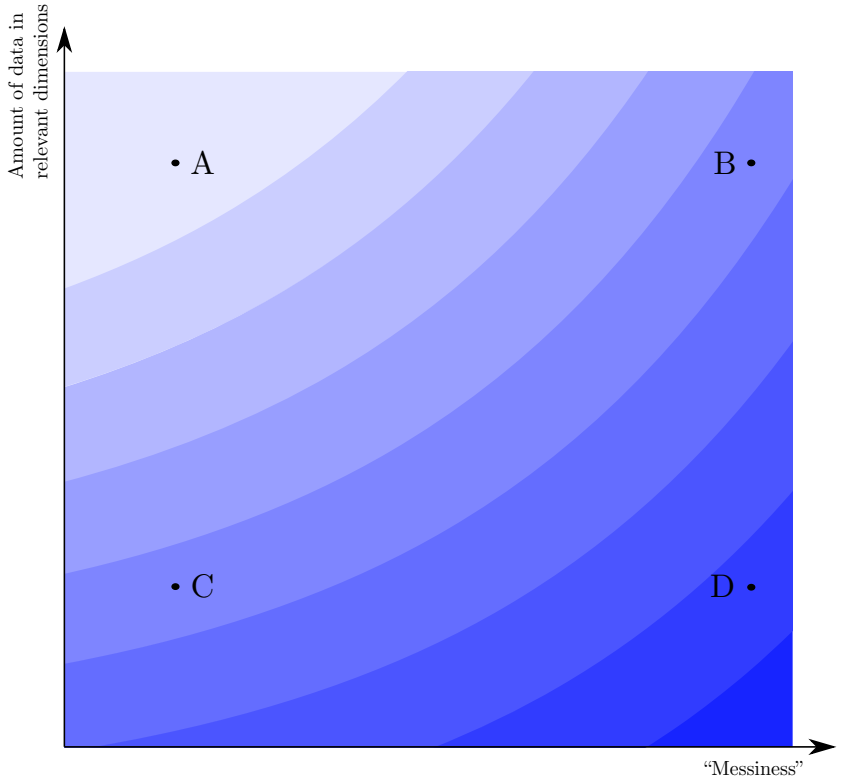


Figure 5.4: Graph of the framework with messiness of the system represented on the X-axis and volume of data and/or number of dimensions on the Y-axis. The color of the bands indicate how difficult extracting useful information from the data is expected to be. The darker the color, the harder this is expected to be. Four different zones of the graph are identified: A-B-C-D.

Referring back to the power analysis discussed in Section 5.1.1 we can say that a power analysis will only be feasible in the leftmost part of the graph. This is the region where the messiness of the data is low and it is therefore expected that there is a clear underlying distribution with very few to no interaction effects. This area is indicated in Figure 5.5. Once we start moving to the right of this region, the messiness increases and the statistical assumptions for many tests start to be violated and the predictions therefore also less reliable. In addition, the number of samples needed to detect a difference increases when moving further to the right. An example of a generalized linear model with simulated data for instance that would require 84 samples to find an effect between two

treatments without interaction would require an additional 192 samples if there were an interaction effect between the treatments (Lin, 2012). Moving even further to the right brings us in an area where the messiness is just too high to perform a proper power analysis and we are left with a qualitative indication that more data is needed.

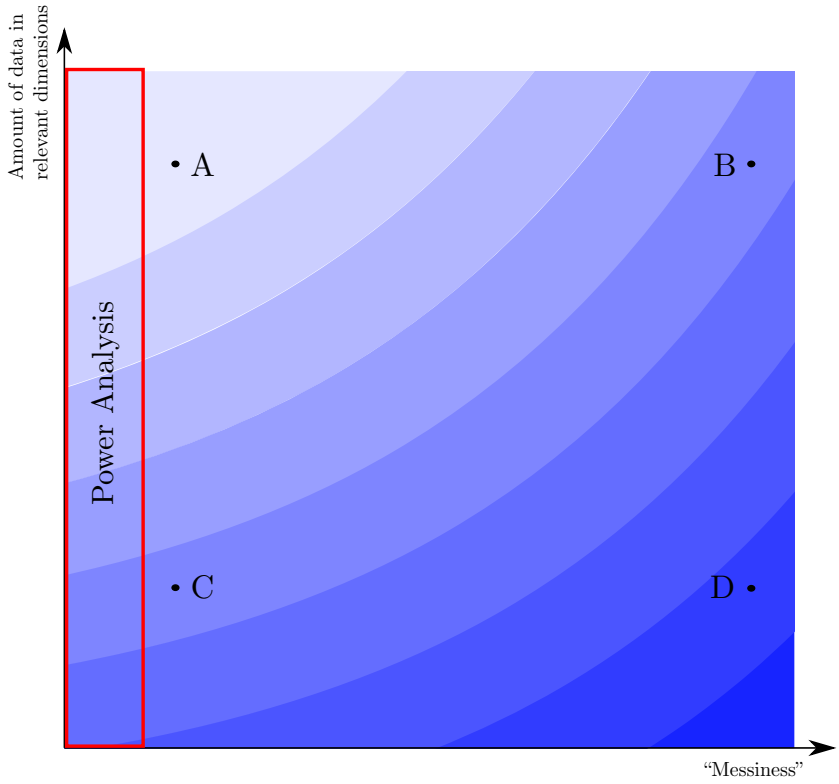


Figure 5.5: Zone where a power analysis is possible. Moving outside this area means that statistical assumptions are violated and the data too messy to perform a proper power analysis.

The following subsections will give an overview of the work conducted during the course of this doctoral project and will explain where it is situated in the graph. This overview serves as a reflection on the different case studies that were performed and the challenges that were encountered in each of the studies. It is important to emphasize beforehand however, that even though a study might be located in the lightest bands of the graph, it is by no means guaranteed

that no unexpected behavior can occur as can be seen in the different chapters concerning these studies.

5.2.1 HIV

Transmitted drug resistance (TDR) is a large problem in the treatment of patients infected with HIV-1, therefore a list of mutations was created in order to assist global efforts to chart the spread and effects of TDR (Bennett et al., 2009). The mutations that are added to this list are therefore called surveillance drug resistance mutations (SDRMs) and are only added to the list when they fulfill four criteria: (1) the mutation should be present on three or more out of five expert lists, (2) it should be non-polymorphic, (3) it has to apply to the most common HIV-1 subtypes, and (4) it should frequently arise from drug selective pressure. These mutations are all considered to be causing resistance in HIV-1.

Because the list of SDRMs only lists mutations that conform to strict criteria, because the mode of action of most antiretroviral drugs is well understood, and because this list is the result of thorough research and use in clinical diagnosis for years, TDR in this setting can be considered a system with low messiness. Therefore, patients harboring variants with one or more of the SDRMs will show resistance to a certain drug class and it is possible to use this information without having to worry too much about noise or other unexpected effects or interactions. Including the more polymorphic minor mutations would certainly increase the messiness as these mutations do not have a direct influence on the phenotype but can interact with the SDRMs to increase viral fitness (Johnson et al., 2013). The dataset that was available to us for our analysis contained samples of 3,354 patients infected with HIV-1 subtype B. As we are only looking at a limited number of mutations in two genes, the volume of data compared to the number of dimensions is also high. All of these observations combined, place the HIV study in zone A in the upper left corner of our graph (Figure 5.6).

Although we limited the messiness of the system by only looking at surveillance drug resistance mutations as these are causative for drug resistance, we still found one small interaction effect. The mutation L90M in the viral protease leads to a reduced viral fitness while our analysis suggested the opposite. This difference can be explained by the polymorphism L63P which is known to restore replication to wild type levels (Suñé et al., 2004). So even though we tried to exclude all factors that can increase the messiness of the system it is almost impossible in biomedical data to have no effects that increase messiness. If we would have expanded our research to all mutations in the virus, the data

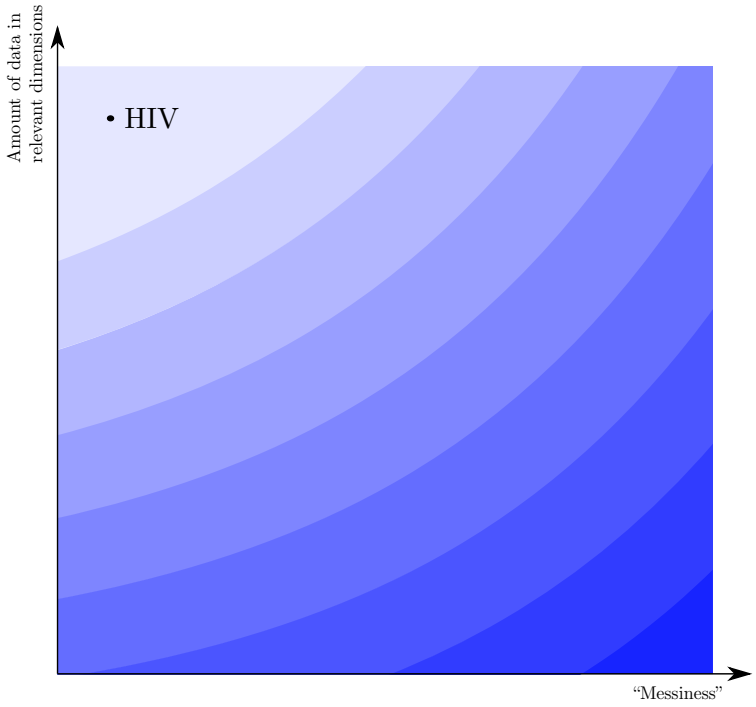


Figure 5.6: Location of the HIV study in the framework

for that analysis would have become much messier as there are many complex interactions and stochastic evolutionary processes at work (Deforche et al., 2007).

5.2.2 Embryo selection

In the last several decades we have seen a huge increase in our knowledge about the genome and genomic disorders. Now that next-generation sequencing (NGS) is becoming more and more affordable, it gives us the advantage of having information about the entire genome compared to the classic tests that only give information on a single or a few genes. With this progress comes the opportunity to analyze the resulting genomic data in order to improve healthcare. But before the technology can be adopted to be used on a larger scale as a screening tool, we have to make sure that the extracted information is correct and improves outcomes. According to the concept discussed here, this

would mean that there should be enough data and that the data should not be too messy as drawing conclusions from such data might prove difficult.

Huge efforts have already been made to identify the genetic basis for many diseases. For some diseases, such as Duchenne muscular dystrophy (Koenig et al., 1987), cystic fibrosis (Riordan et al., 1989; Rommens et al., 1989), and Huntington's disease (The Huntington's Disease Collaborative Research Group, 1993), the causal mutations were identified and the mechanisms behind the disease elucidated. These examples are however all from monogenic Mendelian diseases and in more recent years a large number of studies have focused on identifying mutations that are associated with more complex diseases. This is not straightforward as more complex and polygenic disorders show non-linear interactions between (causative) genes, phenotypes are influenced by environmental effects, and mutations are often found to be associated with the disease but may not cause the disease (Welter et al., 2014).

One of the goals of the study in Chapter 3 was to assess the clinical validity (i.e. does the test predict the phenotype of interest?) of NGS in a screening setting. Therefore, we opted to start from a system with a rather low messiness as this would be expected to yield better results. In order to achieve this low messiness, we only selected genes associated with monogenic disorders with extreme dysmorphologies and used a manually curated database with disease causing mutations. Indeed, as we only take monogenic disorders we expect there to be few or no non-linear interactions and because we only consider diseases with a known molecular basis, the environmental effects should be limited. Compared to the HIV study, this data is a bit messier as some of these factors will still play a (expected limited) role. The datasets that we had available were from over 1,000 healthy (or at least not suffering from the selected diseases) individuals, so we still expect this to be a large enough sample compared to the number of diseases and genes under investigation although smaller compared to the HIV study as the number of dimensions is higher and the number of samples lower. This places this study in the upper left part of the graph but a bit lower on the Y-axis. It is also a bit further to the right on the X-axis compared to the HIV study because of the higher messiness (Figure 5.7).

We discovered that although we assumed the messiness of data to be rather low, the results indicated that we were partially wrong about this assumption. The exact reason for the increase in messiness is unknown but is probably a combination of one or more possible causes. First, there might have been technical errors in the sequencing of the samples and the subsequent variant calling. Second, non-linear relations between the different gene products and mutations can obfuscate findings as many diseases are polygenic in nature instead of monogenic and modifier genes can have small effects on the expression of other genes. This can lead to reduced penetrance and variable expressivity which can

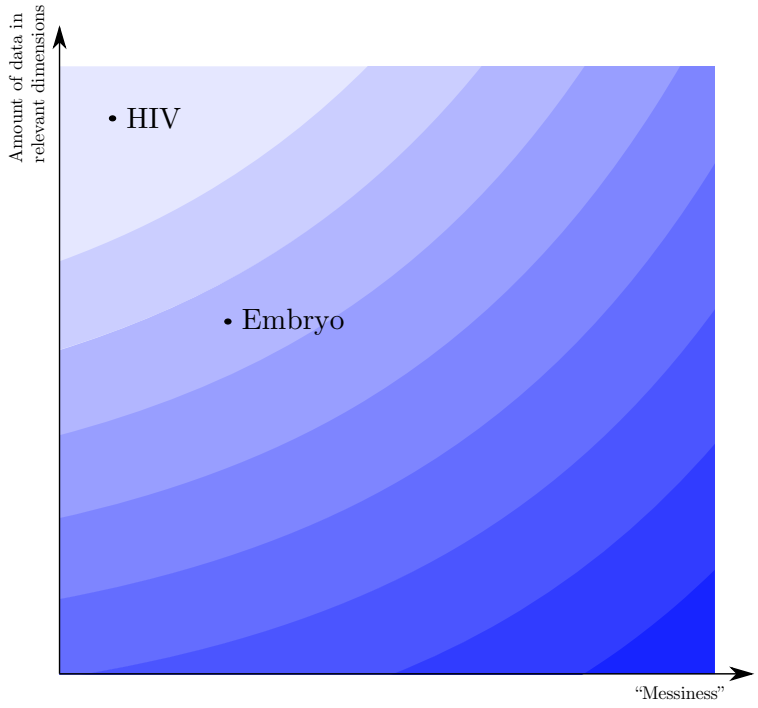


Figure 5.7: Location of the embryo study in the proposed framework

not be easily explained. Third, because of a possible ascertainment bias some of the mutations might be thought to be causative while in fact they are not. Fourth, some of the identified mutations appear to be local polymorphisms which are not found in the 1000 genomes (The 1000 Genomes Project Consortium, 2010) or Complete Genomics data (Drmanac et al., 2010). Finally, environmental factors and epigenetic modifications can also play an important role in the development and severity of diseases. Considering everything we discovered, the study on embryo selection should be located more to the right and somewhat lower in the graph because of the higher messiness and the fact that there is apparently a lack of data from local populations (see Figure 5.11 at the end of this chapter).

While the messiness of the system in this analysis proved to be too high to be of use in a screening setting, it is possible to achieve a lower messiness by imposing different constraints. This can be done for instance by limiting the analysis to diseases included in current preimplantation genetic diagnosis

(PGD)/preimplantation genetic screening (PGS) tests as they have already proven to be reliable. Another way to combat the messiness would be by developing new methods and models to describe the system. As more and more sequences of both healthy and sick people become available, our knowledge about the genome and local polymorphisms will increase and the ascertainment bias can be reduced. Also, advances in sequencing techniques and variant callers can reduce the number of wrong single-nucleotide polymorphism (SNP) calls leading to a better analytical validity. By doing research on the different pathways leading to a disease it might be possible to identify some of the non-linear relations between the different gene products. Also, epigenetic modifications can be assumed to play a role in the development and susceptibility to disease and combining knowledge from this field with NGS data can increase the predictive value

5.2.3 Neural tube defects

Although for many diseases genetic factors can be identified, for many more diseases few to no genetic factors were found that can explain the occurrence of these diseases. One of the possible mechanisms that can explain certain diseases is epigenetic modification. Epigenetic modifications have been shown to be associated with several diseases, most importantly cancer (Brennan & Flanagan, 2008; Baylin & Jones, 2011; Van Vliet et al., 2007). The epigenetic modification that was the subject of the analysis described in Chapter 4 is DNA methylation. Because of the effects that were described in Section 1.4.3, the messiness of epigenetic modifications can be rather high. There are multiple effects in play, each of which leads to an increase in messiness. For example, there is the influence of the environment and stochastic events on epigenetic changes which can lead to differences even between monozygotic twins (Fraga et al., 2005; Petronis, 2010). There are even behavioral influences that can play a role (Rozek et al., 2014). Another added complexity is the fact that epigenetic profiles differ between different cell types in contrast with the DNA sequence that is (practically) the same in all cells (Reinius et al., 2012; Baron et al., 2006). Therefore, this study is expected to have rather high messiness and is located more to the right in the proposed framework.

As there were only samples from 10 patients with lumbosacral MMC and 6 healthy controls available, a study including all ∼480.000 sites interrogated by the array would not yield any statistically significant results after multiple testing correction unless the difference of interest between these two groups would be enormous. This would place the study very low on the Y-axis of the proposed framework as there is clearly not enough data compared to the number of dimensions. In order to get this study higher on the Y-axis, we limited

our analysis to the region of the *HOX* genes and this dramatically reduces the number of dimensions. Although the number of dimensions is much lower in this case, it is still rather high compared to the available samples so it remains on the low end on the Y-axis (Figure 5.8). As mentioned before, performing a power analysis for this experiment is extremely hard as the underlying distribution and the expected effect sizes are unknown.

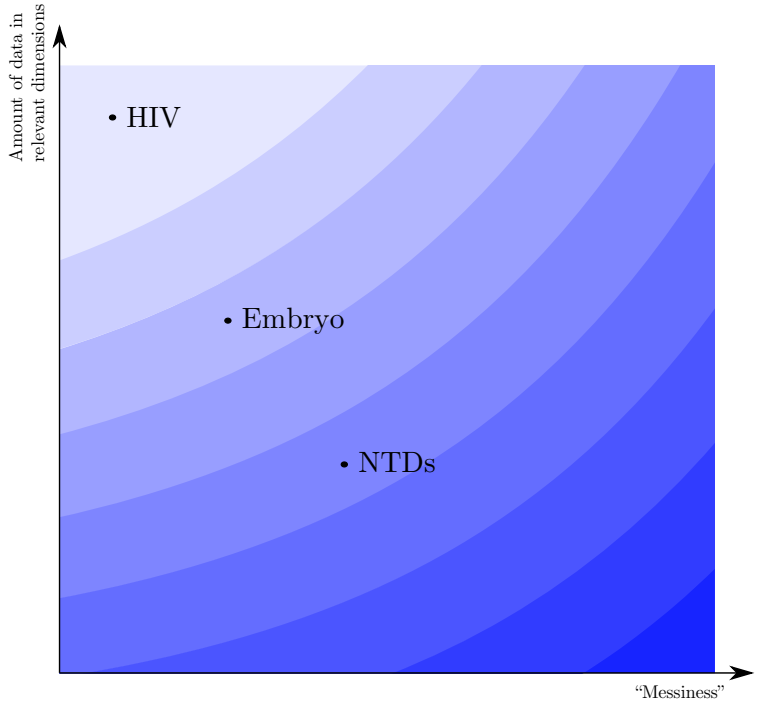


Figure 5.8: Location of the neural tube defects (NTDs) study in the proposed framework

Although we limited the analysis to 967 sites, it still remained too high to get any statistically significant results after correction for multiple testing as the effect size was too small. Because this analysis was only used to guide further analysis, this did not pose a serious problem as the results only needed to give an indication that an effect might be present. In order to detect epigenome-wide differences with smaller effect sizes, at least 800 case and control samples would be needed, while for large effects 65 cases and 56 controls have been shown to be enough to detect the difference (Rakyan et al., 2011). Because of the limited amount of data, the NTD study is located lower on the Y-axis than

first assumed (see Figure 5.11 at the end of this chapter). By validating the results of the initial analysis in a larger cohort of 83 cases and 30 controls and performing a functional genetics study, the limitation of having not enough data in the initial phase was overcome and insights into the mechanisms behind NTDs were still gained.

5.2.4 b-SLIM - MyHealthData

In all previous examples there were models available to describe the data or there was a clear question that needed to be answered. But what if there is no model or what if the patterns that exist in the data are completely unknown? As was the case in the study on neural tube defects, doing exploratory research means not knowing what to expect or what the underlying distributions will be. In this subsection two ongoing studies will be described that are located on the lower right corner of the graph and where the research is completely exploratory. As these studies are ongoing and we are still awaiting the full data sets, not enough results are available and therefore these studies are not covered by a separate chapter.

5.2.4.1 b-SLIM

Obesity and obesity-related diseases are on the rise all over the world with more than 1.9 billion adults being overweight and 600 million obese (WHO, 2015). One of the goals of the b-SLIM project is creating a digital coach to help people achieve sustainable weight loss. Part of this coach are algorithms to identify participant profiles and to identify triggers of behavioral shift leading to weight gain or weight loss. By investigating the different profiles we can learn more about underlying structure in the data and guide the development of the prediction algorithms.

The data collected for this project will consist of a wide variety of features. Physical activity of participants will be tracked by an activity tracker called FitBug (Fitbug Ltd., 2015), supermarket Delhaize will provide a list of all purchases done with the Delhaize Plus card linked to each participant, and online usage of the digital coach and answers to questionnaires in the coach are recorded for each individual. This data set has a very high messiness as it is influenced by many factors. First and foremost there is the effect of human behavior that needs to be modeled in order to identify triggers of behavioral shift. In addition, there will be noise in the data as for example the purchases done at Delhaize might not include all purchases done by that participant and will most likely also include items meant for other family members. The data

from Delhaize will also not be granular enough as products are aggregated into categories not tailored to this analysis. For example refrigerated pizza dough can be in the same category as low fat yogurt and this can obfuscate the interpretation of results. Another effect leading to higher messiness can be the existence of unknown medical conditions of the participant influencing the weight gain or weight loss. From these examples it is clear that this data is most probably very messy and therefore located on the far right of the X-axis (Figure 5.9).

As the data set is high dimensional, it is necessary to have enough samples to perform the clustering and to train and evaluate the prediction models. In this case there will only be data available on 100 participants approximately half of which will be using the digital coach and this is certainly low compared to the number of dimensions and possible profiles. In order to reduce the dimensionality, several features will be aggregated into one score, hereby lowering the number of samples needed to perform the analysis. Considering these elements, the b-SLIM study is located very low on the Y-axis (Figure 5.9) where we are unsure whether or not useful and robust prediction and clustering algorithms can be created.

5.2.4.2 MyHealthData

The MyHealthData project investigates the potential added value of allowing patients to access their own healthcare data in a contextualized way together with medically validated information. One of the work packages in this project deals with the analysis of healthcare data and the medical validation of the results. In this work package we want to develop a visual analytics tool to identify the underlying structures in the data and to help guide to subsequent development of data analysis tools.

The data available for this project is again rather messy. Although most information is structured, we are dealing with information that for the most part was entered manually by physicians, meaning that it is prone to contain mistakes, missing values, and unstructured data. Another factor lies in the fact that guidelines for physicians concerning certain care trajectories or treatments are written in natural language and therefore not directly interpretable by algorithms. The data will also be influenced by environmental factors as for instance some practices may be located in densely populated industrial areas while other are located in more rural areas. Differences in air pollution for example could cause very different prevalences of certain diseases in those practices. On the other hand, knowing this can help guide other analyses where the correlation between location and other variables is investigated. Also human

behavior can add messiness to the data. As said before, the general practitioner has to enter the data and they may not always enter it in a structured way making it nearly impossible to use that data in the analysis. In addition, behavior of the patients can add to the messiness, e.g. patients with a higher education level are more likely to visit a specialist directly without consulting their general practitioner first (WIV-ISP, 2013) which in turn means that information about their condition cannot be found in our data. Therefore, because some data might be MNAR, possible biases are introduced in the underlying distributions.

As we are dealing with all variables associated with a personal health record, the dimensionality of the entire dataset is very large. Therefore we limit ourselves to identifying patterns in smaller subsets of variables. We are for instance interested in certain co-morbidities and in that case we do not need the drug use information. Likewise, diagnoses can be excluded when looking at the geographical distribution of prescriptions for a certain drug. This makes the actual dimensionality that has to be worked with much smaller compared to the dimensionality of the entire data set. In the initial phase of the project we will be given access to data from three practices in Flanders with a total of $\sim$25.000 patients. While this is more than enough data for local analysis, it is not enough to extract accurate information on the level of the Belgian population. Combining all these facts puts the MyHealthData project currently on the lower end of the Y-axis but above the b-SLIM project as can be seen in Figure 5.9.

5.2.5 Visual analytics

In some cases the underlying distributions and interactions in a system are completely unknown and it is hard to formulate a hypothesis to test as not enough is known about the system. This also means that it can be hard to extract relevant information from the data in an automated manner. So instead of trying to prove certain hypotheses we are often in a position where we do not even know what all possible hypotheses might be. In this case, finding the question(s) has become as important or even more so than answering them. Even if some hypotheses are already known beforehand, it is likely that in these data sets many unknown or unexpected patterns exist that would warrant further investigation. This shift to data exploration and data-intensive science is also referred to as the 'fourth paradigm' following the empirical, theoretical and computational paradigms. Where initially science described observations (empirical) it went on to use models and generalizations (theoretical) followed by simulations (computational). With the huge increase in data being generated

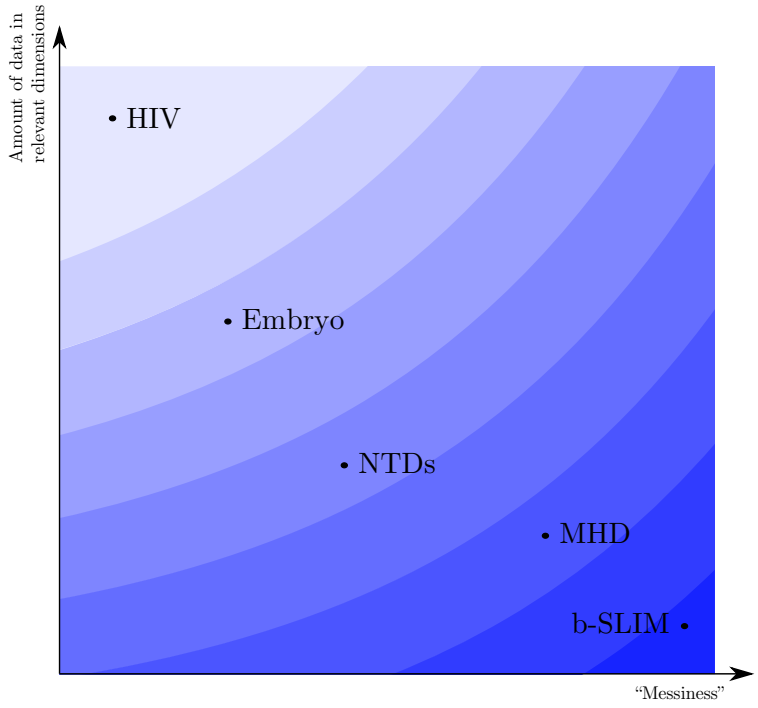


Figure 5.9: Location of the b-SLIM and MyHealthData projects in the proposed framework.

a new paradigm was needed as the previous ones cannot always be used in combination with very large data sets (Hey et al., 2009; Bell et al., 2009).

In this kind of environment, visual analytics can help to identify complex patterns in the data in order to gain a better understanding of the data. A definition given by Keim et al. (2010) says that "visual analytics combines automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets". The visual analytics process can be seen in Figure 5.10. The process always starts with the transformation and preprocessing of the data so that it can be used for the analysis. In the next step, either an automated data mining algorithm is used when a possible hypothesis is already known or a visual representation of the data is made when the goal is hypothesis generation. In the latter case, when a possible interesting hypothesis has been identified, an automated analysis can be performed in order to test that hypothesis. In

both cases, following the automated analysis, the resulting model is refined in an iterative loop where the parameters of the model are adjusted between each iteration. Also in this stage visualization plays an important role as it can help define the parameter space for the automated algorithms (Franken, 2009). The most important aspect of visual analytics is the fact that a human is put back into the loop. The visualization itself provides cognitive support for that human by making use of the advantages of human perception and by helping overcome the limitations of cognitive abilities (Tory & Möller, 2004).

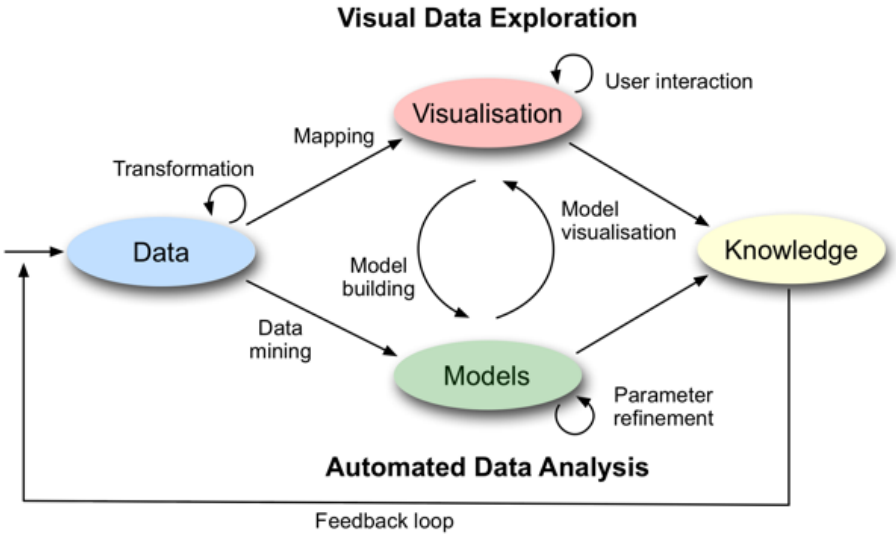


Figure 5.10: The visual analytics process combines automated data analysis and visual data exploration through human interaction in order to gain knowledge about the data. From Keim et al. (2010)

In both the b-SLIM and MyHealthData projects a visual analytics approach is used to formulate hypotheses and to guide the analysis of the data. As in general most visualizations will be used by more than one person a user-centered design process is usually followed (Rogers et al., 2011). By collaborating with domain experts and end users in these projects, it is possible to gain a deeper understanding of the field and to identify both the challenges and opportunities that are present in the data.

In the b-SLIM project, one of the biggest challenges is the low number of participants compared to the high number of dimensions of the available data. Therefore the number of dimensions will have to be reduced by removing non-informative dimensions and by aggregating similar dimensions. Visual analytics can play an important role in this process as visualizations of

different aggregations and sets of variables can be made in order to provide information on the underlying patterns in the data. By iterating through the different possible sets, the most effective set of variables for the prediction and clustering algorithms can be identified. So instead of going through all possible combinations with automated algorithms we can limit the parameter space by exploring it through visualizations. As the results of this exercise are not yet available, the b-SLIM study remains in the lower right corner of the graph. If the visual analytics provides us with more insight into the data and a way to limit the number of dimensions, it is possible that it will move up in the graph as the amount of data in relevant dimensions will become high enough.

For the MyHealthData project the biggest challenge lies in the fact that we do not know all hypotheses we want to test beforehand. In other words, there will be unknown and/or unexpected patterns or relations in the data that still need to be identified. By applying the visual analytics process to the data, it is possible to identify these unknown and unexpected patterns in the data and use that information to formulate hypotheses and to guide the further analysis by automated machine learning algorithms. Working closely together with domain experts, such as general practitioners and the people who developed the database, while developing the different iterations of visualizations allows all parties to gain a deeper understanding of the data and to discover patterns in the data that were previously unknown. Initially, data from only three practices will be made available which is not enough to extract information on a population level. It is however enough to design the first visualizations that can be used when the more data becomes available. In the near future much more data will become available and the amount of data in relevant dimensions will increase significantly.

5.2.6 Summary

As we saw in the previous subsections, studies are not always located in the framework where we expect them to be located even when we actively try to reduce the messiness or increase the amount of available data. Various unexpected factors can increase the messiness or reveal that not enough data is available to perform the analysis. In Figure 5.11 the positions of the different studies are adjusted according to what we learned during the respective analyses.

In Table 5.1 an overview is given of all the sources of messiness that were described together with a few examples that one could look for in order to ascertain whether this factor might pose a problem or not. This is by no means an exhaustive list but it gives a birds eye view of messiness together with some examples. It is also important to note that the system will only be messier when

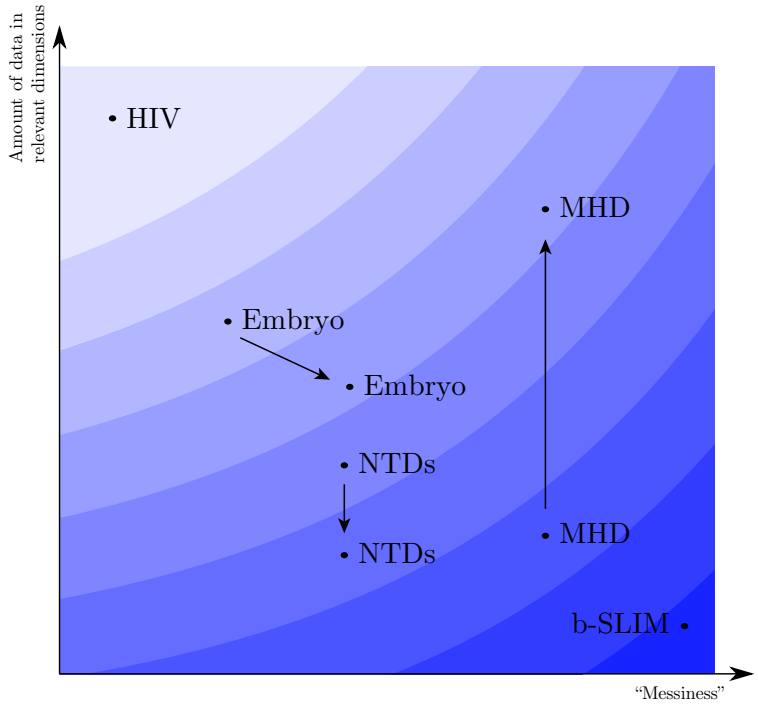


Figure 5.11: Location of all projects in the proposed framework. NTD - Neural Tube Defects; MHD - MyHealthData

these factors are unknown at the time of analysis. Knowing about these factors before the analysis provides an opportunity to model them or abstract them away leading to a reduction in messiness. It is also important to understand the *difference between messiness and complexity*. The messiness is only high when the data is influenced by factors that cannot be modeled or abstracted away or when they are unknown. When these factors are known and can be modeled, the messiness is reduced but the system can become much more complex as more variables have to be modeled which might not always prove to be straightforward.

Table 5.1: An overview of all the sources of messiness described in the manuscript. For each source a few examples are given.

Source	Example
Errors	Bad calibration of equipment
	Malfunctioning equipment
	Human error in measurement
	Human error on data input
	Measuring wrong variable
Statistical assumptions	Thin tailed vs fat tailed
	Independence of data points
	Underlying distribution
Stochastic process	Epigenetics
	X-chromosome inactivation
Non-linear relationships	Feedback loops
	Polygenic vs monogenic
	Interaction effects
Environmental effects	Different environments
	Socioeconomic status
	Peer pressure
	Epigenetics
Temporal effects	Measurements of long periods of time
Missing data	Missing not at random
	Opt-int vs opt-out
Human behavior	Heuristic vs analytical
	Societal influences
	Instincts
	Reflexes
	Circumstances (emergency vs normal)

Chapter 6

General Conclusions

6.1 Discussion

In the last decades we have seen a huge increase in the amount of biomedical data that is generated. This also means that the bottleneck starts to move from the creation of the raw data itself to the interpretation and analysis of that data. In this thesis we proposed a framework that allows one to contemplate how likely it will be to extract accurate and relevant information from a given data set. As was discussed in Chapter 5 we sometimes over- or underestimate the amount of data available and/or the messiness of the system because of several factors influencing the messiness and volume of data. The benefit of using this framework lies in the fact that it *encourages thinking about the available data and possible problems that can arise* before doing the actual analysis, thereby potentially avoiding certain problems.

Identifying these problems does not mean that they cannot be overcome because knowing them is the first step to solving them. The challenges described in the different studies for example are not static. Continued research into the different domains increases the amount of available data and decreases the messiness as the mechanisms underlying these systems become better understood. The study performed in Chapter 3 for instance was published two years ago and since then many things have already changed. As already mentioned, the price per sequenced genome has dropped considerably in the last decade and more genomes than ever are being sequenced. Last year it was estimated that about 100,000 genomes and exomes had been sequenced by 2,555 sequencers in 920 centers across the world (Ardeshirdavani et al., 2014). In June of 2015 there were

already 7,389 machines in 1,027 centers worldwide (Hadfield & Loman, 2015). While we had access to the 1,000 Genomes study many projects are underway to gather many times that number of sequences. In the UK 100,000 sequences will be acquired from 70,000 people (Genomics England, 2015), Saudi Arabia also plans to sequence around 100,000 Saudi genomes (Saudi Human Genome Program, 2013), and the US even wants to sequence 1,000,000 people (National Institutes of Health, 2015). As more and more genomic sequences become available, better models can be created to identify disease causing variants and it will be possible to identify local polymorphisms and to exclude them as possibly damaging. It has to be noted though that the 100,000 Genomes Project in the UK for example will only be sequencing people with a rare disorder or cancer so it will not deliver any data on healthy individuals. Projects like the Decoding Developmental Disorders Project are also looking into the genetic cause of yet undiagnosed developmental disorders. By studying 1,133 children 12 new genes associated with these disorders were found (Fitzgerald et al., 2014). Also HGMD is constantly being updated with many more variants associated with different diseases being discovered all the time (see Table 6.1). Even during the analysis in Chapter 3 some variants that were first classified as disease causing were reclassified. By having more genomes available of both healthy and diseased individuals more variants can be identified as disease causing and many of the false positive ones can be removed. In addition to the increased number of genomes available more and more epigenetic data is being collected which can also offer more insight into the underlying mechanisms thereby possibly explaining some of the false positives we observed.

Not only is the amount of data continuously increasing, the techniques used for sequencing also evolve rapidly. New third generation sequencers such as the ones from Pacific Biosciences or Oxford Nanopore can now sequence much longer reads although the error rate is still quite high (Koren & Phillippy, 2015). As the technologies used for sequencing keep maturing and being improved, the accuracy of the sequences will also keep increasing. Also the algorithms used for analysis and prediction are under continuous development. While we saw that PolyPhen and SIFT have a tendency to classify variants at highly conserved positions as damaging, new algorithms such as eXtasy (Sifrim et al., 2013) decrease the amount of false positives by adding more dimensions to the classification. Application suites that help guide the search for causative variants such as Annotate-it (Sifrim et al., 2012) or Exomiser (Smedley et al., 2015) can help researchers in their quest to identify causative variants.

Considering these advances it is clear that performing the same analysis as we did before would yield different results now and in the future. It would be expected that a number of the false positives that we identified would no longer be considered to be positives. In addition we would also include other

Table 6.1: HGMD Statistics: Number of entries for each category in HGMD version 2012 vs 2015

	Number of Entries	
Mutation Type	**2012**	**2015**
Micro Lesions:		
Missense / Nonsense	68,773	97,626
Splicing	11,525	15,990
Regulatory	2,355	3,297
Small Deletions	19,267	26,113
Small Insertions	7,923	10,927
Small Indels	1,847	2,497
Gross Lesions:		
Repeat Variations	391	482
Gross Insertions / Duplication	1,893	3,182
Complex Rearrangements	1,210	1,668
Gross Deletions	8,472	13,217
Total	**123,656**	**174,999**

prediction algorithms in the analysis such as MutationTaster (Schwarz et al., 2010), CAROL (Lopes et al., 2012), and the aforementioned eXtasy which have a lower false positive rate compared to SIFT and PolyPhen. It is however also likely that some of the new mutations that were entered into HGMD after our first analysis would also be false positives as spurious correlations or ascertainment bias might still lead to incorrect inductions of causality.

Also the study on transmission of HIV-1 drug resistance in Chapter 2 still faces some remaining challenges. For the analysis of the data a standard robust linear regression was used while the distribution of the occurrence of mutations is binomial. As the proper link function was not used, the statistical assumptions are violated and this might influence the results by for instance over- or underestimating the confidence intervals. Attempts to use the proper link function in both SAS and R where unsuccessful as no working modules for this problem were found. Using a binomial link function would make sure that the confidence interval does not go negative as negative prevalences cannot be observed in reality. Also forcing the regression line through zero which would be logical in this setting did not work with current tools. The fact that there are many points with a prevalence close to or equal to 0 in drug naive patients also has an influence on the model as they are common and thus have the greatest influence on the regression line. In addition, the prevalences in treated individuals were assumed to be fixed and independent but there is also an uncertainty associated with these prevalences as they also come from a sample of

the population. Considering all these challenges, a more sophisticated statistical model - that addresses both the existence of outliers and the uncertainty in both the dependent and the independent variables - would have to be developed to analyze this data. This was also one of the conclusions of Chapter 2 but fell beyond the scope of that study. On the other hand we do see that the results are in line with what could be expected considering the influence of these mutations on the viral fitness and we also see that when replicating the model with other data the results are comparable. This suggests that although the statistical assumptions might be violated for this model, the underlying concept can still be valid and additional (statistical) research has to be performed to verify these results.

The previous examples illustrate that as time passes more data will become available, more insights can be gained into a system, and better models can be built. This means that as more is known about the system, the messiness can decrease as it possible that some of the factors leading to increased messiness can be modeled where this was impossible before. With less messy and more data the study would move towards the upper left corner of the proposed framework meaning that the resulting models will be more robust. This will happen at different rates for different fields and studies. For the study in Chapter 3 for instance it will probably take longer to get to a robust model to explain the link between genotype and phenotype whereas custom statistical model for the study on transmission of HIV-1 drug resistance can be enough to create a robust model for the mutations that were selected. This difference is visualized in Figure 6.1 where some studies will approach the best possible model rather quickly while others might never get there. The latter can be caused by factors that increase messiness but can likely never be modeled. Completely random effects for instance or unpredictable human behavior will always lead to higher messiness and cause the results to be less robust. Other complexities such as non-linear interactions or feedback loops are more likely to be identified and modeled, leading to a decrease in messiness. But unless we can analyze the hypothetical 'real' data (i.e. all possible data, not just measured data) a model may never be able to always predict with complete accuracy.

This brings us to the question of *usefulness*. Especially the studies described in Chapter 3 and Chapter 4 raise the question what the usefulness of these technologies would be when applied as a screening tool or as a part of personalized medicine. Over a decade ago when the first human genome was completed it was believed that this would revolutionize medicine leading to personalized medicine. Since then we have learned that although many questions can be answered by looking at the genome, many more questions cannot be answered as the underlying mechanisms of some conditions are very complex and influenced many factors that were described as increasing messiness in

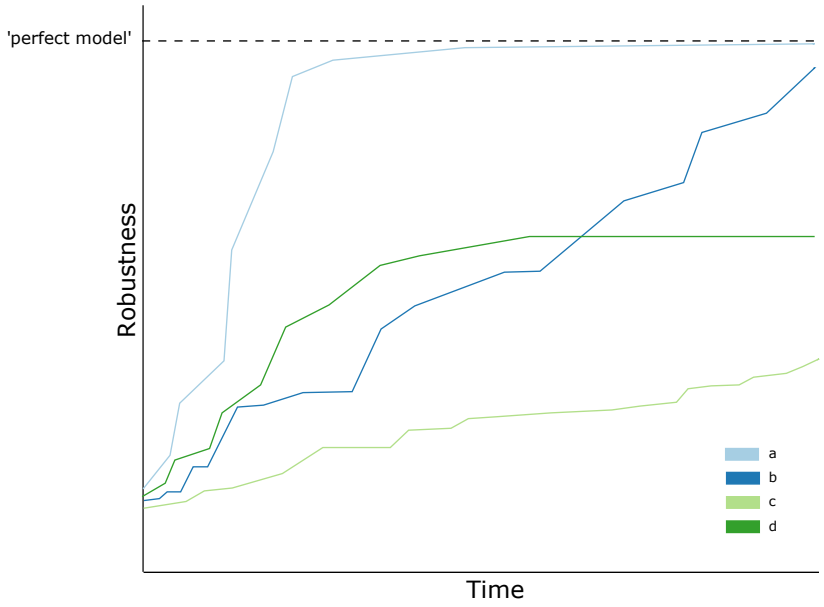


Figure 6.1: This conceptual graph shows how models can become more robust over time as more data becomes available and the messiness decreases. The robustness for a hypothetical 'perfect model', meaning that the predictions from these models are always completely accurate, is displayed. Some systems may quickly become very robust when more data is available or the messiness decreases (a) while others may take longer (b,c) or may reach a maximum robustness because not enough data is available or they are influenced by factors that increase messiness and that cannot be modeled (d).

this dissertation. As of this moment many of these mechanisms are not yet clear and giving people advise based on current knowledge might not be very useful or even do more harm than good. Take for instance the link between autism spectrum disorders and the intake of folic acid (Neggers, 2014). At this moment there is an indication that folic acid intake can have an effect on the development of autism but the mechanisms behind it remain unclear. The question is whether we will gain enough knowledge to fully elucidate the mechanisms behind autism or whether the system is just too messy or complex to unravel based on current knowledge and technologies at our disposal. Making general recommendations on a population level is still possible - e.g. folic acid intake decreases the chance of neural tube defects - and for many disorders we do understand the underlying mechanisms and are able to make predictions on an individual level but we are still a long way from accurately predicting many other disorders on an individual level. The mechanisms behind many disorders

that are as of now still unknown will certainly be elucidated and therefore become very useful while many other disorders and conditions might not give up their secrets so easily and remain very difficult to predict or give advise on.

Overall, from the different case studies it is clear that the proposed framework in Chapter 5 can be useful for anticipating the possible challenges that can arise during a study. By contemplating the messiness and availability of data in relevant dimensions beforehand it is possible to guide the analysis in a way that relevant conclusions can still be drawn. If this is not possible the challenges that need to be overcome are identified thus giving a clear goal for future research. In case of extremely messy and complex data visual analytics can be used to detect previously unknown patterns in the data that would be extremely hard to identify by regular automated analysis. Still, visual analytics would only indicate patterns to investigate further as they show a high "degree-of-interest" but it would not prove that those patterns are real. After the analyses have been performed, the framework can also help identify the cause of unexpected problems that were encountered during the analyses.

6.2 Future research

In general we can say that research opportunities remain to increase the amount of data in relevant dimensions and to create better methods and algorithms to interpret that data. By creating more and more accurate data, we keep increasing the amount of information available which may allow us to detect patterns and mechanisms that are yet unknown. Better and more accurate algorithms to handle messy data can be built which in turn will also allow us to gain a deeper understanding of systems that now appear extremely messy as we do not understand the underlying mechanisms and the external factors influencing that system. But also in each separate case study described in this dissertation specific opportunities remain.

Transmission of drug resistance continues to pose a huge problem worldwide. The model we introduced in Chapter 2 only focuses on the surveillance drug resistance mutations as these are the ones that confer resistance to the different drug classes. This information can be of potential use in guiding strategic measures such as drug prescription guidelines, research priorities, and knowledge dissemination. First, as already mentioned in the previous section, additional statistical research has to be performed to verify the obtained results. In a next step polymorphisms that are associated with these major mutations and that influence the fitness of the virus in the absence of drug selective pressure can be also included. While this system will be much messier, it is possible

that it leads to a better understanding of the transmission dynamics of these polymorphisms and the resulting effect on transmission of drug resistance.

In case of the embryo selection discussed in Chapter 3, several technical and ethical issues have to be solved before secondary information contained in NGS data can be used in a clinical setting. First of all, the technique of single cell sequencing has to be further improved as the interpretation of this data remains complex and the first cell cycles of an embryo show a large amount of structural instability (Van der Aa et al., 2013; Macaulay & Voet, 2014). Also, more data is needed in order to further elucidate the link between genomic variants and diseases. As the knowledge about the genome increases, it might be possible to create better prediction algorithms. The massive efforts that are going into sequencing hundreds of thousands of genomes will yield a lot of new and important information. Even when all the technical issues are resolved, much has to be done in order to create an ethical and legal basis that protects the privacy and rights of everyone involved, and in this case especially those of the unborn child.

The study on NTDs in Chapter 4 could have benefited from a clear indication of the expected effect size in order to perform a proper power analysis. In this field much research still has to be done to elucidate the mechanisms and distributions of these epigenetic modifications.

For the b-SLIM and MyHealthData projects described in Chapter 5 one of the biggest issue is the initial lack of data. Once more practices will participate in the MyHealthData project, the lack of data will be solved and the project will most likely lead to a better understanding of the data on a population level. For b-SLIM visual analytics can provide a way to reduce the number of dimensions so that enough data is still available compared to the number of dimensions and that prediction algorithms can still be used to accurately predict weight gain or weight loss.

Topics of future research that encompass all these studies are the ethical, legal and societal issues that can arise when dealing with this sort of data. A lot of work on this is already being done by experts in those fields. Although we might not be experts ourselves in those fields, it is nevertheless necessary to always keep these issues in mind when analyzing data and developing models that can have affect other people or a society as a whole.

6.3 Conclusion

In this thesis we presented three case studies and a conceptual framework that can be used to get an indication of how likely it will be to extract accurate and relevant information from biomedical data. The different challenges of each case study were identified and discussed based on the information available before and after the analysis of the data. Contemplating the messiness in the data and the availability of data can help guide a study before conducting the actual analysis thereby avoiding problems that might arise otherwise. In case of highly messy and complex data, visual analytics can provide guidance in order to discover hidden patterns in the data and to gain a deeper understanding of the data.

Bibliography

23andMe Inc (2014). 23andMe's New Custom Chip. Http://blog.23andme.com/news/23andmes-new-custom-chip/ - Accessed 2015-04-05.

Adzhubei Ia, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, & Sunyaev SR (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, **7**(4):248–249.

Ajzen I, Joyce N, Sheikh S, & Cote NG (2011). Knowledge and the Prediction of Behavior: The Role of Information Accuracy in the Theory of Planned Behavior. *Basic and Applied Social Psychology*, **33**(2):101–117.

Ananth P & Koopman C (2003). HIV/AIDS knowledge, beliefs, and behavior among women of childbearing age in India. *AIDS education and prevention : official publication of the International Society for AIDS Education*, **15**(6):529–546.

Andorno R (2004). The right not to know: an autonomy based approach. *Journal of Medical Ethics*, **30**(5):435–439.

Andriani P & McKelvey B (2007). Beyond Gaussian averages: redirecting international business and management research toward extreme events and power laws. *Journal of International Business Studies*, **38**(7):1212–1230.

Angrist J & Krueger A (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, **106**(4):979–1014.

Ardeshirdavani A, Souche E, Dehaspe L, Van Houdt J, Vermeesch J, & Moreau Y (2014). NGS-Logistics: federated analysis of NGS sequence variants across multiple locations. *Genome Medicine*, **6**(9):71.

Arion D, Kaushik N, McCormick S, Borkow G, & Parniak Ma (1998). Phenotypic mechanism of HIV-1 resistance to 3'-azido-3'-deoxythymidine

(AZT): Increased polymerization processivity and enhanced sensitivity to pyrophosphate of the mutant viral reverse transcriptase. *Biochemistry*, **37**(45):15908–15917.

Au KS, Ashley-Koch A, & Northrup H (2010). Epidemiologic and genetic aspects of spina bifida and other neural tube defects. *Developmental Disabilities Research Reviews*, **16**(1):6–15.

Audelin AM, Gerstoft J, Obel N, Mathiesen L, Laursen A, Pedersen C, Nielsen H, Jensen J, Nielsen L, Nielsen C, & Jørgensen LB (2011). Molecular Phylogenetics of Transmitted Drug Resistance in Newly Diagnosed HIV Type 1 Individuals in Denmark, a Nation-Wide Study. *AIDS Research and Human Retroviruses*, **27**(12):1283–1290.

Avraham A, Sandbank J, Yarom N, Shalom A, Karni T, Pappo I, Sella A, Fich A, Walfisch S, Gheber L, & Evron E (2010). A similar cell-specific pattern of HOXA methylation in normal and in cancer tissues. *Epigenetics*, **5**(1):41–46.

Baek SJ, Yang S, Kang TW, Park SM, Kim YS, & Kim SY (2013). MENT: Methylation and expression database of normal and tumor tissues. *Gene*, **518**(1):194–200.

Baguley T (2012). *Serious Stats: A guide to advanced statistics for the behavioral sciences*. Palgrave Macmillan, Hampshire. ISBN 9780230577176.

Bar-Yam Y (2014). Why are public health authorities not concerned about Ebola in the US? Part I. Fat tailed distributions. *eprint arXiv:14111472*, pages 1–16.

Barber Ba & Rastegar M (2010). Epigenetic control of Hox genes during neurogenesis, development, and disease. *Annals of Anatomy*, **192**(5):261–274.

Baril L, Ancelle T, Goulet V, Thulliez P, Tirard-Fleury V, & Carme B (1999). Risk factors for Toxoplasma infection in pregnancy: a case-control study in France. *Scandinavian journal of infectious diseases*, **31**(3):305–9.

Barker D, Eriksson J, Forsen T, & Osmond C (2002). Fetal origins of adult disease: strength of effects and biological basis. *Int J Epidemiol*, **31**(6):1235–1239.

Baron U, Türbachova I, Hellwag A, Eckhardt F, Berlin K, Hoffmuller U, Gardina P, & Olek S (2006). DNA methylation analysis as a tool for cell typing. *Epigenetics*, **1**(1):55–60.

Barré-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J, Dauguet C, Axler-Blin C, Vézinet-Brun F, Rouzioux C, Rozenbaum W,

& Montagnier L (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, **220**(4599):868–871.

Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, Troge J, Ravi K, Esposito D, Lakshmi B, Wigler M, Navin N, & Hicks J (2012). Genome-wide copy number analysis of single cells. *Nature protocols*, **7**(6):1024–41.

Baylin SB & Jones PA (2011). A decade of exploring the cancer epigenome — biological and translational implications. *Nature Reviews Cancer*, **11**(10):726–734.

Bell G, Hey T, & Szalay A (2009). Beyond the Data Deluge. *Science*, **323**(5919):1297–1298.

Bellman RE (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ. ISBN 0486428095.

Bennett DE, Camacho RJ, Otelea D, Kuritzkes DR, Fleury H, Kiuchi M, Heneine W, Kantor R, Jordan MR, Schapiro JM, Vandamme AM, Sandstrom P, Boucher CaB, van de Vijver D, Rhee SY, Liu TF, Pillay D, & Shafer RW (2009). Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PloS one*, **4**(3):e4724.

Bennett DE, Myatt M, Bertagnolio S, Sutherland D, & Gilks CF (2008). Recommendations for surveillance of transmitted HIV drug resistance in countries scaling up antiretroviral treatment. *Antiviral therapy*, **13 Suppl 2**:25–36.

Benzi R, Parisi G, Sutera A, & Vulpiani A (1982). Stochastic resonance in climatic change. *Tellus*, **34**:10–16.

Berg JS, Khoury MJ, & Evans JP (2011). Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. *Genetics in medicine : official journal of the American College of Medical Genetics*, **13**(6):499–504.

Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan JB, & Shen R (2011). High density DNA methylation array with single CpG site resolution. *Genomics*, **98**(4):288–295.

Björkstén B (1999). The environmental influence on childhood asthma. *Allergy*, **54 Suppl 4**:17–23.

Blom HJ, Shaw GM, den Heijer M, & Finnell RH (2006). Neural tube defects and folate: case far from closed. *Nature Reviews Neuroscience*, **7**(9):724–731.

Blommaert J (2013). Language and the study of diversit. *Tilburg Papers in Culture Studies*, **74**:1–14.

Blower SM, Aschenbach AN, Gershengorn HB, & Kahn JO (2001). Predicting the unpredictable: transmission of drug-resistant HIV. *Nature medicine*, **7**(9):1016–20.

Booth CL & Geretti AM (2007). Prevalence and determinants of transmitted antiretroviral drug resistance in HIV-1 infection. *The Journal of antimicrobial chemotherapy*, **59**(6):1047–1056.

Boudadi E, Stower H, Halsall Ja, Rutledge CE, Leeb M, Wutz A, O'Neill LP, Nightingale KP, & Turner BM (2013). The histone deacetylase inhibitor sodium valproate causes limited transcriptional change in mouse embryonic stem cells but selectively overrides Polycomb-mediated Hoxb silencing. *Epigenetics & chromatin*, **6**(1):11.

Bouveyron C & Brunet-Saumard C (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics and Data Analysis*, **71**:52–78.

Boyd D & Crawford K (2012). Critical Question for Big Data. *Information, Communication & Society*, **15**(5):662–679.

Brennan K & Flanagan JM (2008). Cancer Epigenetics 101. *Science*, **322**(5908):1605d–1605d.

Bridge PD & Sawilowsky SS (1999). Increasing Physicians' Awareness of the Impact of Statistics on Research Outcomes. *Journal of Clinical Epidemiology*, **52**(3):229–235.

Broder S (2010). The development of antiretroviral therapy and its impact on the HIV-1/AIDS pandemic. *Antiviral Research*, **85**(1):1–18.

Broekkamp H, van Hout-Wolters BHAM, Rijlaarsdam G, & van den Bergh H (2002). Importance in instructional text: Teachers' and students' perceptions of task demands. *Journal of Educational Psychology*, **94**(2):260–271.

Buckles KS & Hungerman DM (2013). Season of Birth and Later Outcomes: Old Questions, New Answers. *Review of Economics and Statistics*, **95**(3):711–724.

Burke S (2001). Missing values, outliers, robust statistics & non-parametric methods. *LC-GC Europe Online Supplement, Statistics & Data Analysis*, **2**(0):19–24.

Burr RL & Cowan MJ (1992). Autoregressive spectral models of heart rate variability. Practical issues. *Journal of electrocardiology*, **25 Suppl**:224–233.

Cao Y, Li H, & Petzold L (2004). Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *Journal of Chemical Physics*, **121**(9):4059–4067.

Carpenter EM (2002). Hox genes and spinal cord development. *Developmental Neuroscience*, **24**(1):24–34.

Castro H, Pillay D, Cane P, Asboe D, Cambiano V, Phillips A, Dunn DT, Aitken C, Asboe D, Webster D, Cane P, Castro H, Chadwick D, Churchill D, Clark D, Collins S, Delpech V, Geretti AM, Goldberg D, Hale A, Hue S, Kaye S, Kellam P, Lazarus L, Leigh-Brown A, Mackie N, Orkin C, Rice P, Pillay D, Smit E, Templeton K, Tilston P, Tong W, Williams I, Zhang H, Zuckerman M, Greatorex J, Wildfire A, O'Shea S, Mullen J, Mbisa T, Cox A, Tandy R, Hale T, Fawcett T, Hopkins M, Ashton L, Garcia-Diaz A, Shepherd J, Schmid ML, Payne B, Chadwick D, Hay P, Rice P, Paynter M, Clark D, Bibby D, Kaye S, Kirk S, MacLean A, Aitken C, & Gunson R (2013). Persistence of HIV-1 Transmitted Drug Resistance Mutations. *Journal of Infectious Diseases*, **208**(9):1459–1463.

Cavalli P & Copp aJ (2002). Inositol and folate resistant neural tube defects. *Journal of medical genetics*, **39**(2):E5.

Centers for Disease Control and Prevention (1981a). Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men–New York City and California. *MMWR Morbidity and mortality weekly report*, **30**(25):305–308.

Centers for Disease Control and Prevention (1981b). Pneumocystis pneumonia–Los Angeles. *MMWR Morbidity and mortality weekly report*, **30**(21):250–252.

CERN (2015). Animation shows LHC data processing. Http://home.cern/about/updates/2013/04/animation-shows-lhc-data-processing - Accessed 2015-12-27.

Chambeyron S & Bickmore Wa (2004). Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes and Development*, **18**(10):1119–1130.

Chang H, Zhang T, Zhang Z, Bao R, Fu C, Wang Z, Bao Y, Li Y, Wu L, Zheng X, & Wu J (2011). Tissue-specific distribution of aberrant DNA methylation associated with maternal low-folate status in human neural tube defects. *The Journal of Nutritional Biochemistry*, **22**(12):1172–1177.

Chateau M, Swanson MD, & Garcia JV (2013). Inefficient Vaginal Transmission of Tenofovir-Resistant HIV-1. *Journal of Virology*, **87**(2):1274–1277.

Chen X, Guo J, Lei Y, Zou J, Lu X, Bao Y, Wu L, Wu J, Zheng X, Shen Y, Wu BL, & Zhang T (2010). Global DNA hypomethylation is associated with NTD-affected pregnancy: A case-control study. *Birth Defects Research Part A - Clinical and Molecular Teratology*, **88**(7):575–581.

Chi MT (1997). Quantifying Qualitative Analyses of Verbal Data: A Practical Guide. *Journal of the Learning Sciences*, **6**(3):271–315.

Chiou-Tan FY, Magee KN, Tuel SM, Robinson LR, Krouskop TA, Nelson MR, & Moss F (1997). Augmented sensory nerve action potentials during distant muscle contraction. *American journal of physical medicine & rehabilitation*, **76**(1):14–18.

Choi JS, Koren G, & Nulman I (2013). Pregnancy and isotretinoin therapy. *Canadian Medical Association Journal*, **185**(5):411–413.

Chowdhury D & Schadschneider A (1999). Self-organization of traffic jams in cities: Effects of stochastic dynamics and signal periods. *Physical Review E*, **59**(2):R1311–R1314.

Chung EC & Haurin DR (2002). Housing choices and uncertainty: The impact of stochastic events. *Journal of Urban Economics*, **52**(2):193–216.

Cihlar T & Ray AS (2010). Nucleoside and nucleotide HIV reverse transcriptase inhibitors: 25 years after zidovudine. *Antiviral Research*, **85**(1):39–58.

Clapson CD (2015). Privacy Commission takes Facebook to Court. Http://deredactie.be/cm/vrtnieuws.english/News/1.2367528 - Accessed 2015-06-22.

Cluzeau C, Hadj-Rabia S, Jambou M, Mansour S, Guigue P, Masmoudi S, Bal E, Chassaing N, Vincent MC, Viot G, Clauss F, Manière MC, Toupenay S, Le Merrer M, Lyonnet S, Cormier-Daire V, Amiel J, Faivre L, de Prost Y, Munnich A, Bonnefont JP, Bodemer C, & Smahi A (2011). Only four genes (EDA1, EDAR, EDARADD, and WNT10A) account for 90% of hypohidrotic/anhidrotic ectodermal dysplasia cases. *Human mutation*, **32**(1):70–72.

Coffin J, Haase A, Levy JA, Montagnier L, Oroszlan S, Teich N, Temin H, Toyoshima K, Varmus H, & Vogt P (1986). What to call the AIDS virus? *Nature*, **321**(6065):10.

Cohen J (1994). The earth is round ($p < .05$). *American Psychologist*, **49**(12):997–1003.

Colgrove RC, Pitt J, Chung PH, Welles SL, & Japour AJ (1998). Selective vertical transmission of HIV-1 antiretroviral resistance mutations. *AIDS*, **12**(17):2281–2288.

Combescure C, Vallier N, Ledergerber B, Cavassini M, Furrer H, Rauch a, Battegay M, Bernasconi E, Vernazza P, & Hirschel B (2009). How reliable is an undetectable viral load? *HIV Medicine*, **10**(8):470–476.

Conlon CP, Klenerman P, Edwards A, Larder BA, & Phillips RE (1994). Heterosexual Transmission of Human Immunodeficiency Virus Type 1 Variants Associated with Zidovudine Resistance. *Journal of Infectious Diseases*, **169**(2):411–415.

Copp AJ, Stanier P, & Greene ND (2013). Neural tube defects: recent advances, unsolved questions, and controversies. *The Lancet Neurology*, **12**(8):799–810.

Cornwell B (2003). Bonded fatalities: Relational and ecological dimensions of a fire evacuation. *Sociological Quarterly*, **44**(4):617–638.

Cottis RA (2001). Interpretation of electrochemical noise data. *Corrosion*, **57**(3):265–284.

Coupry I, Roudaut C, Stef M, Delrue Ma, Marche M, Burgelin I, Taine L, Cruaud C, Lacombe D, & Arveiler B (2002). Molecular analysis of the CBP gene in 60 patients with Rubinstein-Taybi syndrome. *Journal of medical genetics*, **39**(6):415–421.

Cugell DW & Kamp DW (2004). Asbestos and the pleura: A review. *Chest*, **125**(3):1103–1117.

Czene K, Lichtenstein P, & Hemminki K (2002). Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *International Journal of Cancer*, **99**(2):260–266.

Dalgleish AG, Beverley PC, Clapham PR, Crawford DH, Greaves MF, & Weiss RA (1984). The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature*, **312**(5996):763–767.

Danese S, Sans M, & Fiocchi C (2004). Inflammatory bowel disease: The role of environmental factors. *Autoimmunity Reviews*, **3**(5):394–400.

de Béthune MP (2010). Non-nucleoside reverse transcriptase inhibitors (NNRTIs), their discovery, development, and use in the treatment of HIV-1 infection: a review of the last 20 years (1989-2009). *Antiviral research*, **85**(1):75–90.

de Mendoza C, Rodriguez C, Corral A, del Romero J, Gallego O, & Soriano V (2004). Evidence for Differences in the Sexual Transmission Efficiency of HIV Strains with Distinct Drug Resistance Genotypes. *Clinical Infectious Diseases*, **39**(8):1231–1238.

de Ronde A, Schuurman R, Goudsmit J, van den Hoek A, & Boucher C (1996). First case of new infection with zidovudine-resistant HIV-1 among prospectively studied intravenous drug users and homosexual men in Amsterdam, The Netherlands. *AIDS*, **10**(2):231–232.

de Wert G (2009). Preimplantation genetic testing: normative reflections. In Harper J, editor, *Preimplantation Genetic Diagnosis*, pages 259–273. Cambridge University Press, second edition. ISBN 9780511581571. Cambridge Books Online.

Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, & Fuks F (2014). A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in Bioinformatics*, **15**(6):929–941.

Deforche K, Camacho R, Grossman Z, Silander T, Soares M, Moreau Y, Shafer R, Van Laethem K, Carvalho A, & Wynhoven B (2007). Bayesian network analysis of resistance pathways against HIV-1 protease inhibitors. *Infection, Genetics and Evolution*, **7**(3):382–390.

Demchenko Y, Grosso P, de Laat C, & Membrey P (2013). Addressing big data issues in Scientific Data Infrastructure. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pages 48–55. IEEE. ISBN 978-1-4673-6404-1.

Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhart M, Di Marzio P, Marmon S, Sutton RE, Hill CM, Davis CB, Peiper SC, Schall TJ, Littman DR, & Landau NR (1996). Identification of a major co-receptor for primary isolates of HIV-1. *Nature*, **381**(6584):661–666.

Di YM, Chan E, Wei MQ, Liu JP, & Zhou SF (2009). Prediction of Deleterious Non-synonymous Single-Nucleotide Polymorphisms of Human Uridine Diphosphate Glucuronosyltransferase Genes. *The AAPS Journal*, **11**(3):469–480.

Dierynck I, De Wit M, Gustin E, Keuleers I, Vandersmissen J, Hallenberger S, & Hertogs K (2007). Binding kinetics of darunavir to human immunodeficiency virus type 1 protease explain the potent antiviral activity and high genetic barrier. *Journal of virology*, **81**(24):13845–13851.

Dohm JC, Lottaz C, Borodina T, & Himmelbauer H (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, **36**(16):e105–e105.

Dolling D, Sabin C, Delpech V, Smit E, Pozniak A, Asboe D, Brown AL, Churchill D, Williams I, Geretti AM, Phillips A, Mackie N, Murphy G, Castro H, Pillay D, Cane P, & Dunn D (2012). Time trends in drug resistant HIV-1 infections in the United Kingdom up to 2009: multicentre observational study. *BMJ*, **345**:e5253–e5253.

Dondorp W, de Wert G, & Cornel MC (2010). The Quality of Genetic Screening: An Integral Approach. In Kristoffersson U, Schmidtke J, & Cassiman JJ, editors, *Quality Issues in Clinical Genetic Services*, pages 165–172. Springer Netherlands, Dordrecht. ISBN 978-90-481-3918-7.

Doniach T (1995). Basic FGF as an inducer of anteroposterior neural pattern. *Cell*, **83**(7):1067–1070.

Drescher SM, von Wyl V, Yang WL, Böni J, Yerly S, Shah C, Aubert V, Klimkait T, Taffé P, Furrer H, Battegay M, Ambrosioni J, Cavassini M, Bernasconi E, Vernazza PL, Ledergerber B, Günthard HF, & Kouyos RD (2014). Treatment-naive individuals are the major source of transmitted HIV-1 drug resistance in men who have sex with men in the Swiss HIV Cohort Study. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, **58**(2):285–294.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, Dahl F, Fernandez A, Staker B, Pant KP, Baccash J, Borcherding AP, Brownley A, Cedeno R, Chen L, Chernikoff D, Cheung A, Chirita R, Curson B, Ebert JC, Hacker CR, Hartlage R, Hauser B, Huang S, Jiang Y, Karpinchyk V, Koenig M, Kong C, Landers T, Le C, Liu J, McBride CE, Morenzoni M, Morey RE, Mutch K, Perazich H, Perry K, Peters Ba, Peterson J, Pethiyagoda CL, Pothuraju K, Richter C, Rosenbaum AM, Roy S, Shafto J, Sharanhovich U, Shannon KW, Sheppy CG, Sun M, Thakuria JV, Tran A, Vu D, Zaranek AW, Wu X, Drmanac S, Oliphant AR, Banyai WC, Martin B, Ballinger DG, Church GM, & Reid Ca (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**(5961):78–81.

du Bois RM, Branthwaite MA, Mikhail JR, & Batten JC (1981). Primary Pneumocystis carinii and cytomegalovirus infections. *Lancet*, **2**(8259):1339.

Duro JA & Teixidó-Figueras J (2013). Ecological footprint inequality across countries: The role of environment intensity, income and interaction effects. *Ecological Economics*, **93**:34–41.

Eckert Ka & Kunkel Ta (1991). DNA polymerase fidelity and the polymerase chain reaction. *Genome Research*, **1**(1):17–24.

Ellis J, Riordan O, Cameron J, & Jordan A (2006). Resource ? New Literature on the Precautionary Principle. *European Journal of International Law*, **17**(2):445–462.

Erice A, Mayers DL, Strike DG, Sannerud KJ, McCutchan FE, Henry K, & Balfour HH (1993). Brief report: primary infection with zidovudine-resistant human immunodeficiency virus type 1. *The New England journal of medicine*, **328**(16):1163–1165.

European Commission (2015). Dual-use export controls. Http://ec.europa.eu/trade/import-and-export-rules/export-from-eu/dual-use-controls/index_en.htm - Accessed 2015-12-25.

Evans JSBT (2006). The heuristic-analytic theory of reasoning: extension and evaluation. *Psychonomic bulletin & review*, **13**(3):378–395.

Fan J, Han F, & Liu H (2014). Challenges of Big Data analysis. *National Science Review*, **1**(2):293–314.

Feng Y, Broder CC, Kennedy PE, & Berger EA (1996). HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science*, **272**(5263):872–877.

Firth HV & Wright CF (2011). The Deciphering Developmental Disorders (DDD) study. *Developmental medicine and child neurology*, **53**(8):702–3.

Fisher D, DeLine R, Czerwinski M, & Drucker S (2012). Interactions with big data analytics. *Interactions*, **19**(3):50.

Fisher, Elizabeth Charlotte Jones JS & von Schomberg R (2006). *Implementing the Precautionary Principle: Perspectives and Prospects*. Edward Elgar Publishing, Cheltenham, UK.

Fitbit Inc (2015). Fitbit Official Site for Activity Trackers & More. Https://www.fitbit.com/ - Accessed 2015-08-26.

Fitbug Ltd (2015). Fitness, Sleep & Activity Tracker. Https://www.fitbug.com - Accessed 2015-08-20.

Fitzgerald TW, Gerety SS, Jones WD, van Kogelenberg M, King Da, McRae J, Morley KI, Parthiban V, Al-Turki S, Ambridge K, Barrett DM, Bayzetinova T, Clayton S, Coomber EL, Gribble S, Jones P, Krishnappa N, Mason LE, Middleton A, Miller R, Prigmore E, Rajan D, Sifrim A, Tivey aR, Ahmed M, Akawi N, Andrews R, Anjum U, Archer H, Armstrong R, Balasubramanian M, Banerjee R, Baralle D, Batstone P, Baty D, Bennett C, Berg J, Bernhard B, Bevan aP, Blair E, Blyth M, Bohanna D, Bourdon L, Bourn D, Brady A,

Bragin E, Brewer C, Brueton L, Brunstrom K, Bumpstead SJ, Bunyan DJ, Burn J, Burton J, Canham N, Castle B, Chandler K, Clasper S, Clayton-Smith J, Cole T, Collins A, Collinson MN, Connell F, Cooper N, Cox H, Cresswell L, Cross G, Crow Y, D'Alessandro M, Dabir T, Davidson R, Davies S, Dean J, Deshpande C, Devlin G, Dixit A, Dominiczak A, Donnelly C, Donnelly D, Douglas A, Duncan A, Eason J, Edkins S, Ellard S, Ellis P, Elmslie F, Evans K, Everest S, Fendick T, Fisher R, Flinter F, Foulds N, Fryer A, Fu B, Gardiner C, Gaunt L, Ghali N, Gibbons R, Gomes Pereira SL, Goodship J, Goudie D, Gray E, Greene P, Greenhalgh L, Harrison L, Hawkins R, Hellens S, Henderson A, Hobson E, Holden S, Holder S, Hollingsworth G, Homfray T, Humphreys M, Hurst J, Ingram S, Irving M, Jarvis J, Jenkins L, Johnson D, Jones D, Jones E, Josifova D, Joss S, Kaemba B, Kazembe S, Kerr B, Kini U, Kinning E, Kirby G, Kirk C, Kivuva E, Kraus A, Kumar D, Lachlan K, Lam W, Lampe A, Langman C, Lees M, Lim D, Lowther G, Lynch Sa, Magee A, Maher E, Mansour S, Marks K, Martin K, Maye U, McCann E, McConnell V, McEntagart M, McGowan R, McKay K, McKee S, McMullan DJ, McNerlan S, Mehta S, Metcalfe K, Miles E, Mohammed S, Montgomery T, Moore D, Morgan S, Morris A, Morton J, Mugalaasi H, Murday V, Nevitt L, Newbury-Ecob R, Norman A, O'Shea R, Ogilvie C, Park S, Parker MJ, Patel C, Paterson J, Payne S, Phipps J, Pilz DT, Porteous D, Pratt N, Prescott K, Price S, Pridham A, Procter A, Purnell H, Ragge N, Rankin J, Raymond L, Rice D, Robert L, Roberts E, Roberts G, Roberts J, Roberts P, Ross A, Rosser E, Saggar A, Samant S, Sandford R, Sarkar A, Schweiger S, Scott C, Scott R, Selby A, Seller A, Sequeira C, Shannon N, Sharif S, Shaw-Smith C, Shearing E, Shears D, Simonic I, Simpkin D, Singzon R, Skitt Z, Smith A, Smith B, Smith K, Smithson S, Sneddon L, Splitt M, Squires M, Stewart F, Stewart H, Suri M, Sutton V, Swaminathan GJ, Sweeney E, Tatton-Brown K, Taylor C, Taylor R, Tein M, Temple IK, Thomson J, Tolmie J, Torokwa A, Treacy B, Turner C, Turnpenny P, Tysoe C, Vandersteen A, Vasudevan P, Vogt J, Wakeling E, Walker D, Waters J, Weber A, Wellesley D, Whiteford M, Widaa S, Wilcox S, Williams D, Williams N, Woods G, Wragg C, Wright M, Yang F, Yau M, Carter NP, Parker M, Firth HV, FitzPatrick DR, Wright CF, Barrett JC, & Hurles ME (2014). Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, **519**(7542):223–228.

Flemming W (1882). *Zellsubstanz, Kern und Zelltheilung.* FCW Vogel, Leipzig.

Ford CE, Jones KW, Polani PE, De Almeida JC, & Briggs JH (1959). A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). *Lancet*, **1**(7075):711–713.

Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suñer D, Cigudosa JC, Urioste M, Benitez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu YZ,

Plass C, & Esteller M (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(30):10604–10609.

Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, Anderson CA, Bis JC, Bumpstead S, Ellinghaus D, Festen EM, Georges M, Green T, Haritunians T, Jostins L, Latiano A, Mathew CG, Montgomery GW, Prescott NJ, Raychaudhuri S, Rotter JI, Schumm P, Sharma Y, Simms LA, Taylor KD, Whiteman D, Wijmenga C, Baldassano RN, Barclay M, Bayless TM, Brand S, Büning C, Cohen A, Colombel JF, Cottone M, Stronati L, Denson T, De Vos M, D'Inca R, Dubinsky M, Edwards C, Florin T, Franchimont D, Gearry R, Glas J, Van Gossum A, Guthery SL, Halfvarson J, Verspaget HW, Hugot JP, Karban A, Laukens D, Lawrance I, Lemann M, Levine A, Libioulle C, Louis E, Mowat C, Newman W, Panés J, Phillips A, Proctor DD, Regueiro M, Russell R, Rutgeerts P, Sanderson J, Sans M, Seibold F, Steinhart AH, Stokkers PCF, Torkvist L, Kullak-Ublick G, Wilson D, Walters T, Targan SR, Brant SR, Rioux JD, D'Amato M, Weersma RK, Kugathasan S, Griffiths AM, Mansfield JC, Vermeire S, Duerr RH, Silverberg MS, Satsangi J, Schreiber S, Cho JH, Annese V, Hakonarson H, Daly MJ, & Parkes M (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics*, **42**(12):1118–1125.

Franken N (2009). Visual exploration of algorithm parameter space. In *Proceedings of the Eleventh conference on Congress on Evolutionary Computation*, CEC'09, pages 389–398. IEEE Press, Piscataway, NJ, USA. ISBN 978-1-4244-2958-5.

Frentz D, Boucher CaB, & van de Vijver DaMC (2012). Temporal changes in the epidemiology of transmission of drug-resistant HIV-1 across the world. *AIDS reviews*, **14**(1):17–27.

Freund JA & Pöschel T (2000). *Stochastic processes in physics, chemistry, and biology*, volume 557. Springer Science & Business Media, Berlin. ISBN 3540410740.

Friso S, Choi SW, Girelli D, Mason JB, Dolnikowski GG, Bagley PJ, Olivieri O, Jacques PF, Rosenberg IH, Corrocher R, & Selhub J (2002). A common mutation in the 5,10-methylenetetrahydrofolate reductase gene affects genomic DNA methylation through an interaction with folate status. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(8):5606–5611.

Frosst P, Blom HJ, Milos R, Goyette P, Sheppard Ca, Matthews RG, Boers GJ, den Heijer M, Kluijtmans La, & van den Heuvel LP (1995). A

candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. *Nature genetics*, **10**(1):111–113.

Funk S, Salathé M, & Jansen Vaa (2010). Modelling the influence of human behaviour on the spread of infectious diseases: a review. *Journal of the Royal Society Interface*, **7**(50):1247–1256.

Gallo RC, Salahuddin SZ, Popovic M, Shearer GM, Kaplan M, Haynes BF, Palker TJ, Redfield R, Oleske J, & Safai B (1984). Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science*, **224**(4648):500–503.

Garland-Thomson R (2012). The Case for Conserving Disability. *Journal of Bioethical Inquiry*, **9**(3):339–355.

Garrett W (2013). Why More Data and Simple Algorithms Beat Complex Analytics Models. Http://data-informed.com/why-more-data-and-simple-algorithms-beat-complex-analytics-models/ - Accessed 2015-06-03.

Gautier and Rpy2 contributors (2014). Rpy2 (version 2.4). Http://rpy.sourceforge.net/rpy2/doc-2.4/html/index.html - Accessed 2014-08-10.

Genomics England (2015). The 100,000 Genomes Project. Http://www.genomicsengland.co.uk/the-100000-genomes-project/ - Accessed 2015-12-30.

Gerke J, Lorenz K, Ramnarine S, & Cohen B (2010). Gene–Environment Interactions at Nucleotide Resolution. *PLoS Genetics*, **6**(9):e1001144.

Gillespie DT (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, **22**(4):403–434.

Gillespie DT (2007). Stochastic simulation of chemical kinetics. *Annual review of physical chemistry*, **58**:35–55.

Gilliland FD, Li YF, & Peters JM (2001). Effects of maternal smoking during pregnancy and environmental tobacco smoke on asthma and wheezing in children. *American journal of respiratory and critical care medicine*, **163**(2):429–436.

Goel NS & Richter-Dyn N (2013). *Stochastic models in biology*. Elsevier. ISBN 1483278107.

Goff SP (2001). Intracellular trafficking of retroviral genomes during the early phase of infection: Viral exploitation of cellular pathways. *Journal of Gene Medicine*, **3**(6):517–528.

Google Inc (2015). What Data Does Google Collect? Https://privacy.google.com/data-we-collect.html - Accessed 2015-06-22.

Gottlieb MS, Schroff R, Schanker HM, Weisman JD, Fan PT, Wolf RA, & Saxon A (1981). Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *The New England journal of medicine*, **305**(24):1425–1431.

Gould R, Kreuter F, & Palmer C (2006). Towards statistical thinking: Making real data real. *Proceedings of the Seventh International Conference on the Teaching of Statistics, July 2006*, pages 1–6.

Graham JW (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, **60**(1):549–576.

Gray VE, Kukurba KR, & Kumar S (2012). Performance of computational tools in evaluating the functional impact of laboratory-induced amino acid mutations. *Bioinformatics*, **28**(16):2093–2096.

Greenberg A (2013). Meet The 'Liberator': Test-Firing The World's First Fully 3D-Printed Gun. Http://www.forbes.com/sites/andygreenberg/2013/05/05/meet-the-liberator-test-firing-the-worlds-first-fully-3d-printed-gun/ - Accessed 2015-08-19.

Greene ND, Stanier P, & Moore GE (2011). The emerging role of epigenetic mechanisms in the aetiology of neural tube defects. *Epigenetics*, **6**(7):875–883.

Guderian a, Dechert G, Zeyer KP, & Schneider FW (1996). Stochastic Resonance in Chemistry. 1. The Belousov-Zhabotinsky Reaction. *The Journal of Physical Chemistry*, **100**(11):4437–4441.

Gymrek M, McGuire AL, Golan D, Halperin E, & Erlich Y (2013). Identifying Personal Genomes by Surname Inference. *Science*, **339**(6117):321–324.

Hadfield J & Loman N (2015). High-Throughput Sequencing Maps. Http://omicsmaps.com/stats - Accessed 2015-06-26.

Hafley WL & Lewis JS (1963). Analyzing Messy Data. *Industrial & Engineering Chemistry*, **55**(4):37–39.

Hagendorens MM, Ebo DG, Bridts CH, de Water LV, De Clerck LS, & Stevens WJ (2004). Prenatal exposure to house dust mite allergen (Der p 1), cord blood T cell phenotype and cytokine production and atopic dermatitis during the first year of life. *Pediatric Allergy and Immunology*, **15**(4):308–315.

Hales CN & Barker DJP (2013). Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. *International Journal of Epidemiology*, **42**(5):1215–1222.

Hall P, Marron JS, & Neeman A (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **67**(3):427–444.

Hallowell N, Hall A, Alberg C, & Zimmern R (2015). Revealing the results of whole-genome sequencing and whole-exome sequencing in research and clinical investigations: some ethical issues. *Journal of Medical Ethics*, **41**(4):317–321.

Hänggi P (2002). Stochastic resonance in biology: How noise can enhance detection of weak signals and help improve biological information processing. *ChemPhysChem*, **3**(3):285–290.

Harford T (2014). Big data: are we making a big mistake? Http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2xSTI6uzS - Accessed 2015-12-28.

Harper JC & Sengupta SB (2012). Preimplantation genetic diagnosis: state of the art 2011. *Human genetics*, **131**(2):175–186.

Harper PS (2008). *A Short History of Medical Genetics*. Oxford University Press, New York, NY. ISBN 9780195187502.

Harris MJ & Juriloff DM (2007). Mouse mutants with neural tube closure defects and their role in understanding human neural tube defects. *Birth Defects Research Part A - Clinical and Molecular Teratology*, **79**(3):187–210.

Harris MJ & Juriloff DM (2010). An update to the list of mouse mutants with neural tube closure defects and advances toward a complete genetic perspective of neural tube closure. *Birth Defects Research Part A - Clinical and Molecular Teratology*, **88**(8):653–669.

Harvey A, Jan Koopman S, & Penzer J (1999). Messy Time Series. In Fomby TB & Hill RC, editors, *Messy Data (Advances in Econometrics, Volume 13)*, pages 103–143. Emerald Group Publishing Limited, Bradford.

Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, & Ullah Khan S (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, **47**:98–115.

Hauser MW (2012). Messy Data, Ordered Questions. *American Anthropologist*, **114**(2):184–195.

Hecht FM, Grant RM, Petropoulos CJ, Dillon B, Chesney MA, Tian H, Hellmann NS, Bandrapalli NI, Digilio L, Branson B, & Kahn JO (1998). Sexual Transmission of an HIV-1 Variant Resistant to Multiple Reverse-Transcriptase and Protease Inhibitors. *New England Journal of Medicine*, **339**(5):307–311.

Helberg C (1995). Pitfalls of Data Pitfalls of Data Analysis. Http://www.execpc.com/˜helberg/pitfalls/ - Accessed 2015-12-28.

Helweg-Larsen M & Collins BE (1997). A Social Psychological Perspective on the Role of Knowledge About AIDS in AIDS Prevention. *Current Directions in Psychological Science*, **6**(2):23–26.

Hens K, Dondorp W, & de Wert G (2013a). Embryos without secrets: an expert panel study on comprehensive embryo testing and the responsibility of the clinician. *European journal of medical genetics*, **56**(2):67–71.

Hens K, Dondorp W, Handyside AH, Harper J, Newson AJ, Pennings G, Rehmann-Sutter C, & de Wert G (2013b). Dynamics and ethics of comprehensive preimplantation genetic testing: a review of the challenges. *Human reproduction update*, **19**(4):366–375.

Hert DG, Fredlake CP, & Barron AE (2008). Advantages and limitations of next-generation sequencing technologies: A Comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis*, **29**(23):4618–4626.

Hey T, Tansley S, & Tolle K, editors (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, WA. ISBN 9780982544204.

Hodkinson BP & Grice Ea (2015). Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. *Advances in Wound Care*, **4**(1):50–58.

Hoff E (2003). The specificity of environmental influence: socioeconomic status affects early vocabulary development via maternal speech. *Child development*, **74**(5):1368–1378.

Hofman PL, Regan F, Jackson WE, Jefferies C, Knight DB, Robinson EM, & Cutfield WS (2004). Premature birth and later insulin resistance. *The New England journal of medicine*, **351**(21):2179–2186.

Holland JJ, De La Torre JC, & Steinhauer DA (1992). RNA Virus Populations as Quasispecies. In Holland J, editor, *Genetic Diversity of RNA Viruses SE - 1*, volume 176 of *Current Topics in Microbiology and Immunology*, pages 1–20. Springer Berlin Heidelberg. ISBN 978-3-642-77013-5.

Holliday R (2006). Epigenetics: A historical overview. *Epigenetics*, **1**(2):76–80.

Holliday R & Pugh J (1975). DNA modification mechanisms and gene activity during development. *Science*, (187):226–232.

Hotelling H (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**(6):417–441.

Hughes G (2011). How big is 'big data' in healthcare? Http://blogs.sas.com/content/hls/2011/10/21/how-big-is-big-data-in-healthcare/ - Accessed 2015-12-27.

Hwu F & Sun S (2012). The aptitude-treatment interaction effects on the learning of grammar rules. *System*, **40**(4):505–521.

IBM (2015). IBM - SPSS SamplePower. Http://www-03.ibm.com/software/products/en/spss-samplepower - Accessed 2015-08-19.

Ikegame T, Bundo M, Murata Y, Kasai K, Kato T, & Iwamoto K (2013). DNA methylation of the BDNF gene and its relevance to psychiatric disorders. *Journal of human genetics*, **58**(7):434–438.

Illumina Inc (2015). HiSeq X Ten and HiSeq X Five Systems. Http://www.illumina.com/systems/hiseq-x-sequencing-system.html - Accessed 2015-04-05.

INSERM (1997). Orphanet: an online database of rare diseases and orphan drugs. Http://www.orpha.net - Accessed 2013-01-23.

Institute of Medicine (1994). *Assessing Genetic Risks*. National Academies Press, Washington, D.C. ISBN 978-0-309-04798-2.

Ito S (2004). Novel Complex GUCY2D Mutation in Japanese Family with Cone-Rod Dystrophy. *Investigative Ophthalmology & Visual Science*, **45**(5):1480–1485.

Izzi B, aM Binder, & Michels K (2014). Pyrosequencing Evaluation of Widely Available Bisulfite Conversion Methods: Considerations for Application. *Medical Epigenetics*, **2**(1):28–36.

Izzi B, Decallonne B, Devriendt K, Bouillon R, Vanderschueren D, Levtchenko E, de Zegher F, Van den Bruel A, Lambrechts D, Van Geet C, & Freson K (2010). A new approach to imprinting mutation detection in GNAS by Sequenom EpiTYPER system. *Clinica Chimica Acta*, **411**(23-24):2033–2039.

Izzi B, Francois I, Labarque V, Thys C, Wittevrongel C, Devriendt K, Legius E, Van den Bruel A, D'Hooghe M, Lambrechts D, de Zegher F, Van Geet C, & Freson K (2012). Methylation Defect in Imprinted Genes Detected in Patients with an Albright's Hereditary Osteodystrophy Like Phenotype and Platelet Gs Hypofunction. *PLoS ONE*, **7**(6):e38579.

Jacobs PA & Strong JA (1959). A Case of Human Intersexuality Having a Possible XXY Sex-Determining Mechanism. *Nature*, **183**(4657):302–303.

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, & von Mering C (2009). STRING 8 - A global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, **37**(SUPPL. 1):412–416.

Jensen R (2004). Daggers, Rifles and Dynamite: Anarchist Terrorism in Nineteenth Century Europe. *Terrorism and Political Violence*, **16**(1):116–153.

Jernigan TL, Gamst AC, Archibald SL, Fennema-Notestine C, Mindt MR, Marcotte TL, Heaton RK, Ellis RJ, & Grant I (2005). Effects of methamphetamine dependence and HIV infection on cerebral morphology. *American Journal of Psychiatry*, **162**(8):1461–1472.

Jessell TM (2000). Neuronal specification in the spinal cord: inductive signals and transcriptional codes. *Nature reviews Genetics*, **1**(1):20–29.

Jimenez-Gomez JM, Corwin Ja, Joseph B, Maloof JN, & Kliebenstein DJ (2011). Genomic Analysis of QTLs and Genes Altering Natural Variation in Stochastic Noise. *PLoS Genetics*, **7**(9):e1002295.

Johnson CW (2005). Lessons from the evacuation of the world trade centre, 9/11 2001 for the development of computer-based simulations. *Cognition, Technology and Work*, **7**(4):214–240.

Johnson PLF & Slatkin M (2006). Inference of population genetic parameters in metagenomics: A clean look at messy data. *Genome Research*, **16**(10):1320–1327.

Johnson Va, Calvez V, Gunthard HF, Paredes R, Pillay D, Shafer RW, Wensing AM, & Richman DD (2013). Update of the drug resistance mutations in HIV-1: March 2013. *Topics in antiviral medicine*, **21**(1):6–14.

Jordan A & O'Riordan T (2004). The precautionary principle: a legal and policy history. In Martuzzi M & Tickner JA, editors, *The precautionary principle: protecting public health, the environment and the future of our children*, page 219. WHO, Copenhagen. ISBN 92 890 1098 3.

Kalff-Suske M, Wild A, Topp J, Wessling M, Jacobsen EM, Bornholdt D, Engel H, Heuer H, Aalfs CM, Ausems MG, Barone R, Herzog A, Heutink P, Homfray T, Gillessen-Kaesbach G, König R, Kunze J, Meinecke P, Müller D, Rizzo R, Strenge S, Superti-Furga A, & Grzeschik KH (1999). Point mutations throughout the GLI3 gene cause Greig cephalopolysyndactyly syndrome. *Human molecular genetics*, **8**(9):1769–1777.

Keim D, Kohlhammer J, Ellis G, & Mansmann F (2010). *Mastering the Information Age - Solving Problems with Visual Analytics*. ISBN 978-3-905673-77-7.

Kempf DJ, Marsh KC, Kumar G, Rodrigues aD, Denissen JF, McDonald E, Kukulka MJ, Hsu A, Granneman GR, Baroldi Pa, Sun E, Pizzuti D, Plattner JJ, Norbeck DW, & Leonard JM (1997). Pharmacokinetic enhancement of inhibitors of the human immunodeficiency virus protease by coadministration with ritonavir. *Antimicrobial Agents and Chemotherapy*, **41**(3):654–660.

Kennedy W (2012). Modelling Human Behaviour in Agent-Based Models. In Heppenstall AJ, Crooks AT, See LM, & Batty M, editors, *Agent-Based Models of Geographical Systems SE - 9*, pages 167–179. Springer Netherlands. ISBN 978-90-481-8926-7.

Kerner BS (2013). Criticism of generally accepted fundamentals and methodologies of traffic and transportation theory: A brief review. *Physica A: Statistical Mechanics and its Applications*, **392**(21):5261–5282.

Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, & Schilling TF (1995). Stages of embryonic development of the zebrafish. *Developmental dynamics*, **203**(3):253–310.

Kirke PN, Molloy AM, Daly LE, Burke H, Weir DG, & Scott JM (1993). Maternal plasma folate and vitamin B12 are independent risk factors for neural tube defects. *The Quarterly journal of medicine*, **86**(11):703–708.

Knipl DH, Pilarczyk P, & Röst G (2015). Rich Bifurcation Structure in a Two-Patch Vaccination Model. *SIAM Journal on Applied Dynamical Systems*, **14**(2):980–1017.

Koenekoop RK, Fishman Ga, Iannaccone A, Ezzeldin H, Ciccarelli ML, Baldi A, Sunness JS, Lotery AJ, Jablonski MM, Pittler SJ, & Maumenee I (2002). Electroretinographic abnormalities in parents of patients with Leber congenital amaurosis who have heterozygous GUCY2D mutations. *Archives of ophthalmology*, **120**(10):1325–1330.

Koenig M, Hoffman EP, Bertelson CJ, Monaco aP, Feener C, & Kunkel LM (1987). Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell*, **50**(3):509–517.

Koren S & Phillippy AM (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, **23**:110–120.

Krauss S, Johansen T, Korzh V, Moens U, Ericson JU, & Fjose a (1991). Zebrafish pax[zf-a]: a paired box-containing gene expressed in the neural tube. *The EMBO journal*, **10**(12):3609–3619.

Krauss S, So J, Hambrock M, Köhler A, Kunath M, Scharff C, Wessling M, Grzeschik KH, Schneider R, & Schweiger S (2009). Point mutations in GLI3 lead to misregulation of its subcellular localization. *PloS one*, **4**(10):e7471.

Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CMT, & Richards JB (2012). The Empirical Power of Rare Variant Association Methods: Results from Sanger Sequencing in 1,998 Individuals. *PLoS Genetics*, **8**(2):e1002496.

Laird PW (2010). Principles and challenges of genomewide DNA methylation analysis. *Nature reviews Genetics*, **11**(3):191–203.

Langevin SM & Kelsey KT (2013). The fate is not always written in the genes: Epigenomics in epidemiologic studies. *Environmental and Molecular Mutagenesis*, **54**(7):533–541.

Laskey SB & Siliciano RF (2014). A mechanistic theory to explain the efficacy of antiretroviral therapy. *Nature Reviews Microbiology*, **12**(11):772–780.

Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Sung KWK, Rigoutsos I, Loring J, & Wei CL (2010). Dynamic changes in the human methylome during differentiation. *Genome Research*, **20**(3):320–331.

Lazer D, Pentland A, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, & Van Alstyne M (2009). Social science. Computational social science. *Science*, **323**(5915):721–723.

Lees-Murdock DJ & Walsh CP (2008). DNA Methylation Reprogramming in the Germ Line. pages 1–15. Springer New York. ISBN 9780387775753.

Legendre P (1993). Real data are messy. *Statistics and Computing*, **3**(4):197–199.

Leigh Brown AJ, Frost SDW, Mathews WC, Dawson K, Hellmann NS, Daar ES, Richman DD, & Little SJ (2003). Transmission fitness of drug-resistant human immunodeficiency virus and the prevalence of resistance in the antiretroviral-treated population. *The Journal of infectious diseases*, **187**(4):683–686.

Lejeune J, Gautier M, & Turpin R (1959). Etude des chromosomes somatiques de neuf enfants mongoliens. *Comptes rendus hebdomadaires des seances de l'Academie des sciences*, **248**(11):1721–1722.

Lem S, Onghena P, Verschaffel L, & Van Dooren W (2013). The heuristic interpretation of box plots. *Learning and Instruction*, **26**:22–35.

Li D, Xie Z, Pape ML, & Dye T (2015). An evaluation of statistical methods for DNA methylation microarray data analysis. *BMC Bioinformatics*, **16**(1):217.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, & Durbin R (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16):2078–2079.

Lin CDJ (2012). Sample Size Calculation and Power Analysis for Design of Experiments Using Proc GLMPOWER. *Western Users of SAS Software Proceedings*, pages 1–10.

Lindquist S (2009). Protein folding sculpting evolutionary change. *Cold Spring Harbor Symposia on Quantitative Biology*, **74**:103–108.

Longford NT (2001). Multilevel analysis with messy data. *Statistical methods in medical research*, **10**(6):429–444.

Lopes MC, Joyce C, Ritchie GRS, John SL, Cunningham F, Asimit J, & Zeggini E (2012). A combined functional annotation score for non-synonymous variants. *Human heredity*, **73**(1):47–51.

Los Alamos National Security LLC (2014). Landmarks of the HIV genome. Http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html - Accessed 2015-06-02.

Ludovici DW, De Corte BL, Kukla MJ, Ye H, Ho CY, Lichtenstein Ma, Kavash RW, Andries K, De Béthune MP, Azijn H, Pauwels R, Lewi PJ, Heeres J, Koymans LMH, De Jonge MR, Van Aken KJa, Daeyaert FFD, Das K, Arnold E, & Janssen PaJ (2001a). Evolution of anti-HIV drug candidates. Part 3: Diarylpyrimidine (DAPY) analogues. *Bioorganic and Medicinal Chemistry Letters*, **11**(17):2235–2239.

Ludovici DW, Kavash RW, Kukla MJ, Ho CY, Ye H, De Corte BL, Andries K, De Béthune MP, Azijn H, Pauwels R, Moereels HEL, Heeres J, Koymans LMH, De Jonge MR, Van Aken KJa, Daeyaert FFD, J Lewi P, Das K, Arnold E, & Janssen PaJ (2001b). Evolution of anti-HIV drug candidates. Part 2: Diaryltriazine (DATA) analogues. *Bioorganic and Medicinal Chemistry Letters*, **11**(17):2229–2234.

Ludovici DW, Kukla MJ, Grous PG, Krishnan S, Andries K, De Béthune MP, Azijn H, Pauwels R, De Clercq E, Arnold E, & Janssen PAJ (2001c). Evolution of anti-HIV drug candidates. Part 1: From α-Anilinophenylacetamide (α-APA) to imidoyl thiourea (ITU). *Bioorganic and Medicinal Chemistry Letters*, **11**(17):2225–2228.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers Ca, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo Ma, Banks E, Hu M, Handsaker RE, Rosenfeld Ja, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IHa, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, Wang J, Li Y, Gibbs Ra, McCarroll Sa, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, & Tyler-Smith C (2012). A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science*, **335**(6070):823–828.

MacArthur DG & Tyler-Smith C (2010). Loss-of-function variants in the genomes of healthy humans. *Human Molecular Genetics*, **19**(R2):R125–R130.

Macaulay IC & Voet T (2014). Single Cell Genomics: Advances and Future Perspectives. *PLoS Genetics*, **10**(1):e1004126.

Magolda MBB & Astin AW (1993). What "Doesn't" Matter in College? *Educational Researcher*, **22**(8):32.

Malashkevich VN, Chan DC, Chutkowski CT, & Kim PS (1998). Crystal structure of the simian immunodeficiency virus (SIV) gp41 core: conserved helical interactions underlie the broad inhibitory activity of gp41 peptides. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(16):9134–9139.

Malpas PJ (2006). Why tell asymptomatic children of the risk of an adult-onset disease in the family but not test them for it? *Journal of Medical Ethics*, **32**(11):639–642.

Mansky LM & Bernard LC (2000). 3'-Azido-3'-deoxythymidine (AZT) and AZT-resistant reverse transcriptase can increase the in vivo mutation rate of human immunodeficiency virus type 1. *Journal of virology*, **74**(20):9532–9539.

Mansky LM & Temin HM (1995). Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of virology*, **69**(8):5087–5094.

Marcus G & Davis E (2014). Eight (No, Nine!) Problems With Big Data. Http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html?_r=1 - Accessed 2015-12-28.

Mardis ER (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, **24**(3):133–141.

Marjoram P, Zubair a, & Nuzhdin SV (2014). Post-GWAS: where next? More samples, more SNPs or more biology? *Heredity*, **112**(1):79–88.

Martens R, Bastiaens T, & Kirschner Pa (2007). New Learning Design in Distance Education: The impact on student perception and motivation. *Distance Education*, **28**(1):81–93.

Martin AM, Nolan D, Gaudieri S, Almeida CA, Nolan R, James I, Carvalho F, Phillips E, Christiansen FT, Purcell AW, McCluskey J, & Mallal S (2004). Predisposition to abacavir hypersensitivity conferred by HLA-B*5701 and a haplotypic Hsp70-Hom variant. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(12):4180–4185.

Marx JL (1982). New disease baffles medical community. *Science*, **217**(4560):618–621.

Masliah E, Dumaop W, Galasko D, & Desplats P (2013). Distinctive patterns of DNA methylation associated with Parkinson disease: Identification of concordant epigenetic changes in brain and peripheral blood leukocytes. *Epigenetics*, **8**(10):1030–1038.

Masur H, Michelis MA, Greene JB, Onorato I, Stouwe RA, Holzman RS, Wormser G, Brettman L, Lange M, Murray HW, & Cunningham-Rundles S (1981). An outbreak of community-acquired Pneumocystis carinii pneumonia: initial manifestation of cellular immune dysfunction. *The New England journal of medicine*, **305**(24):1431–1438.

Mayer AN, Dimmock DP, Arca MJ, Bick DP, Verbsky JW, Worthey Ea, Jacob HJ, & Margolis Da (2011). A timely arrival for genomic medicine. *Genetics in Medicine*, **13**(3):195–196.

McDonald E & Brown CT (2013). khmer: Working with Big Data in Bioinformatics. *arXiv preprint arXiv:13032223*, (February):1–18.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, & DePristo Ma (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**(9):1297–1303.

McKusick-Nathans Institute of Genetic Medicine JHU (2012). Online Mendelian Inheritance in Man, OMIM. Http://www.omim.org - Accessed 2012-10-28.

McLean GA, George MH, Funkhouser GE, & Chittum CB (1996). Aircraft Evacuations Onto Escape Slides and Platforms I: Effects of Passenger Motivation. Technical report.

Meaburn E & Schulz R (2012). Next generation sequencing in epigenetics: Insights and challenges. *Seminars in Cell and Developmental Biology*, **23**(2):192–199.

Mearian L (2015). CERN's data stores soar to 530M gigabytes. Http://www.computerworld.com/article/2960642/cloud-storage/cerns-data-stores-soar-to-530m-gigabytes.html - Accessed 2015-12-30.

Mertes H & Hens K (2015). The Diversity of Genetic Perfection. *The American Journal of Bioethics*, **15**(6):34–36.

Metzker ML (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, **11**(1):31–46.

Milliken GA & Johnson DE (1993). *Analysis of Messy Data: Designed Experiments*. v. 1. Chapman & Hall, London. ISBN 9780412990816.

Miyoshi T, Ueno Y, Nakagawa M, & Nakayasu H (2010). Urgent evacuation simulation in aircraft accident using AAMAS. In *40th International Conference on Computers and Industrial Engineering: Soft Computing Techniques for Advanced Manufacturing and Service Systems, CIE40 2010*. ISBN 9781424472956.

Monkhorst K, Jonkers I, Rentmeester E, Grosveld F, & Gribnau J (2008). X Inactivation Counting and Choice Is a Stochastic Process: Evidence for Involvement of an X-Linked Activator. *Cell*, **132**(3):410–421.

Morse RP & Evans EF (1996). Enhancement of vowel coding for cochlear implants by addition of noise. *Nature medicine*, **2**(8):928–932.

Muir HC, Bottomley DM, & Marrison C (1996). Effects of Motivation and Cabin Configuration on Emergency Aircraft Evacuation Behavior and Rates of Egress. *The International Journal of Aviation Psychology*, **6**(1):57–77.

Murphy K, Travers P, & Walport M (2008). *Janeway's Immunobiology*. Garland Science, Geneva. ISBN 978-0-8153-4123-9.

Murtagh F (2009). The remarkable simplicity of very high dimensional data: Application of model-based clustering. *Journal of Classification*, **26**(3):249–277.

Myers Ta (2011). Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data. *Communication Methods and Measures*, **5**(4):297–310.

National Human Genome Research Institute (2010). Genome Project Completion: Frequently Asked Questions. Https://www.genome.gov/11006943 - Accessed 2015-04-05.

National Institutes of Health (1993). U.S. National Library of Medicine. Http://www.nlm.nih.gov - Accessed 2013-01-24.

National Institutes of Health (2015). Precision Medicine Initiative Cohort Program. Https://www.nih.gov/precision-medicine-initiative-cohort-program - Accessed 2015-12-30.

Navas-Acien A, Guallar E, Silbergeld EK, & Rothenberg SJ (2007). Lead Exposure and Cardiovascular Disease: A Systematic Review. *Environmental Health Perspectives*, **115**(3):472–482.

Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J, & Wigler M (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, **472**(7341):90–94.

NCSS (2015). NCSS | Sample Size Software | Power Analysis Software | PASS | NCSS.com. Http://www.ncss.com/software/pass/ - Accessed 2015-08-19.

Neggers Y (2014). The Relationship between Folic Acid and Risk of Autism Spectrum Disorders. *Healthcare*, **2**(4):429–444.

Neuhauser M & Senske R (2004). The Baumgartner-Wei -Schindler test for the detection of differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **20**(18):3553–3564.

Ng PC (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, **31**(13):3812–3814.

NIH (2011). Press Statement on the NSABB Review of H5N1 Research. Http://www.nih.gov/news/health/dec2011/od-20.htm - Accessed 2015-08-21.

O'Neil C (2015). Big Data Is The New Phrenology. Http://mathbabe.org/2015/02/26/big-data-is-the-new-phrenology/ - Accessed 2015-12-28.

Ormond KE, Wheeler MT, Hudgins L, Klein TE, Butte AJ, Altman RB, Ashley Ea, & Greely HT (2010). Challenges in the clinical application of whole-genome sequencing. *The Lancet*, **375**(9727):1749–1751.

Palma aC, Araújo F, Duque V, Borges F, Paixão MT, & Camacho R (2007). Molecular epidemiology and prevalence of drug resistance-associated mutations in newly diagnosed HIV-1 patients in Portugal. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, **7**(3):391–398.

Panel on Antiretroviral Guidelines for Adults and Adolescents (2014). Guidelines for the use of antiretroviral agents in HIV-infected adults and adolescents. Department of Health and Human Services. Available at

http://aidsinfo.nih.gov/ContentFiles/Adultand AdolescentGL.pdf. Section accessed 2014-08-09.

Perelson AS, Neumann AU, Markowitz M, Leonard JM, & Ho DD (1996). HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, **271**(5255):1582–1586.

Petronis A (2010). Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature*, **465**(7299):721–727.

Philippidou P & Dasen JS (2013). Hox Genes: Choreographers in Neural Development, Architects of Circuit Organization. *Neuron*, **80**(1):12–34.

Pirinen M, Donnelly P, & Spencer CCa (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics*, **44**(8):848–851.

Pollard VW & Malim MH (1998). The HIV-1 Rev protein. *Annual review of microbiology*, **52**:491–532.

Prabu-Jeyabalan M, Nalivaika E, & Schiffer Ca (2002). Substrate shape determines specificity of recognition for HIV-1 protease: Analysis of crystal structures of six substrate complexes. *Structure*, **10**(3):369–381.

Prince VE, Joly L, Ekker M, & Ho RK (1998). Zebrafish hox genes: genomic organization and modified colinear expression patterns in the trunk. *Development*, **125**(3):407–420.

Priplata Aa, Niemi JB, Harry JD, Lipsitz La, & Collins JJ (2003). Vibrating insoles and balance control in elderly people. *Lancet*, **362**(9390):1123–1124.

Proulx G (2001). Occupant behaviour and evacuation. *Proceedings of the 9th International Fire Protection Symposium*, pages 219–232.

Pruitt KD, Harrow J, Harte Ra, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Maidak BL, Mudge J, Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber Ja, Wilming L, Wu W, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, & Lipman D (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome research*, **19**(7):1316–1323.

Quinn TC, Wawer MJ, Sewankambo N, Serwadda D, Li C, Wabwire-Mangen F, Meehan MO, Lutalo T, & Gray RH (2000). Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group. *The New England journal of medicine*, **342**(13):921–929.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Http://www.r-project.org/ - Accessed 2014-08-10.

Rajan U, Seru A, & Vig V (2015). The failure of models that predict failure: distance, incentives and defaults. *Journal of Financial Economics*, **115**(2):237–260.

Rajaraman A (2008). More data usually beats better algorithms. Http://anand.typepad.com/datawocky/2008/03/more-data-usual.html - Accessed 2015-06-03.

Rakyan VK, Down Ta, Balding DJ, & Beck S (2011). Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, **12**(8):529–541.

Rambaut A, Posada D, Crandall Ka, & Holmes EC (2004). The causes and consequences of HIV evolution. *Nature Reviews Genetics*, **5**(1):52–61.

Ramdohr P (2009). HIV-1 Replication Cycle. Adapted from https://www.flickr.com/photos/opiado/4191226676 - Licensed under CC BY-NC 2.0: https://creativecommons.org/licenses/by-nc/2.0/.

Ramos FJ, Kaplan BS, Bellah RD, Zackai EH, & Kaplan P (1998). Further evidence that the Hajdu-Cheney syndrome and the "serpentine fibula-polycystic kidney syndrome" are a single entity. *American journal of medical genetics*, **78**(5):474–481.

Raz R, Roberts AL, Lyall K, Hart JE, Just AC, Laden F, & Weisskopf MG (2014). Autism Spectrum Disorder and Particulate Matter Air Pollution before, during, and after Pregnancy: A Nested Case–Control Analysis within the Nurses' Health Study II Cohort. *Environmental Health Perspectives*, **123**(3):264–70.

Reed JF & Stark DB (1996). Hinge estimators of location: Robust to asymmetry. *Computer Methods and Programs in Biomedicine*, **49**(1):11–17.

Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén SE, Greco D, Söderhäll C, Scheynius A, & Kere J (2012). Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *PLoS ONE*, **7**(7):e41361.

Ren J & Stammers DK (2008). Structural basis for drug resistance mechanisms for non-nucleoside inhibitors of HIV reverse transcriptase. *Virus Research*, **134**(1-2):157–170.

Reynolds A, McDearmid JR, Lachance S, Marco PD, Merello E, Capra V, Gros P, Drapeau P, & Kibar Z (2010). VANGL1 rare variants associated with neural tube defects affect convergent extension in zebrafish. *Mechanisms of Development*, **127**(7-8):385–392.

Rhee SY, Blanco JL, Jordan MR, Taylor J, Lemey P, Varghese V, Hamers RL, Bertagnolio S, de Wit TFR, Aghokeng AF, Albert J, Avi R, Avila-Rios S, Bessong PO, Brooks JI, Boucher CaB, Brumme ZL, Busch MP, Bussmann H, Chaix ML, Chin BS, D'Aquin TT, De Gascun CF, Derache A, Descamps D, Deshpande AK, Djoko CF, Eshleman SH, Fleury H, Frange P, Fujisaki S, Harrigan PR, Hattori J, Holguin A, Hunt GM, Ichimura H, Kaleebu P, Katzenstein D, Kiertiburanakul S, Kim JH, Kim SS, Li Y, Lutsar I, Morris L, Ndembi N, Ng KP, Paranjape RS, Peeters M, Poljak M, Price Ma, Ragonnet-Cronin ML, Reyes-Terán G, Rolland M, Sirivichayakul S, Smith DM, Soares Ma, Soriano VV, Ssemwanga D, Stanojevic M, Stefani Ma, Sugiura W, Sungkanuparph S, Tanuri A, Tee KK, Truong HHM, van de Vijver DaMC, Vidal N, Yang C, Yang R, Yebra G, Ioannidis JPa, Vandamme AM, & Shafer RW (2015). Geographic and Temporal Trends in the Molecular Epidemiology and Genetic Mechanisms of Transmitted HIV-1 Drug Resistance: An Individual-Patient- and Sequence-Level Meta-Analysis. *PLOS Medicine*, **12**(4):e1001810.

Ricci F, Rokach L, & Shapira B (2011). Introduction to Recommender Systems Handbook. In Ricci F, Rokach L, Shapira B, & Kantor PB, editors, *Recommender Systems Handbook SE - 1*, pages 1–35. Springer US, New York, NY. ISBN 978-0-387-85819-7.

Ringwood JV & Malpas SC (2001). Slow oscillations in blood pressure via a nonlinear feedback model. *American journal of physiology Regulatory, integrative and comparative physiology*, **280**(4):R1105–R1115.

Riordan, Rommens J, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou J, & Et A (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*, **245**(4922):1066–1073.

Robinson PN & Mundlos S (2010). The human phenotype ontology. *Clinical genetics*, **77**(6):525–534.

Rogers Y, Sharp H, & Preece J (2011). *Interaction Design: Beyond Human - Computer Interaction, 3rd Edition*. Wiley. ISBN 0471492787.

Rommens J, Iannuzzi M, Kerem Bs, Drumm M, Melmer G, Dean M, Rozmahel R, Cole J, Kennedy D, Hidaka N, & Et A (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science*, **245**(4922):1059–1065.

Rowe DC, Jacobson KC, & Van den Oord EJCG (1999). Genetic and environmental influences on vocabulary IQ: parental education level as moderator. *Child development*, **70**(5):1151–1162.

Rozek LS, Dolinoy DC, Sartor Ma, & Omenn GS (2014). Epigenetics: Relevance and Implications for Public Health. *Annual Review of Public Health*, **35**(1):105–122.

Rubin LH, Witkiewitz K, Andre JS, & Reilly S (2007). Methods for Handling Missing Data in the Behavioral Neurosciences: Don't Throw the Baby Rat out with the Bath Water. *Journal of Undergraduate Neuroscience Education*, **5**(2):A71–A77.

Sanderson S, Zimmern R, Kroese M, Higgins J, Patch C, & Emery J (2005). How can the evaluation of genetic tests be enhanced? Lessons learned from the ACCE framework and evaluating genetic tests in the United Kingdom. *Genetics in Medicine*, **7**(7):495–500.

Sanger F & Coulson AR (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, **94**(3):441–448.

Sanger F, Nicklen S, & Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**(12):5463–5467.

Sarafianos SG, Marchand B, Das K, Himmel DM, Parniak Ma, Hughes SH, & Arnold E (2009). Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition. *Journal of Molecular Biology*, **385**(3):693–713.

Sarngadharan MG, Popovic M, Bruch L, Schüpbach J, & Gallo RC (1984). Antibodies reactive with human T-lymphotropic retroviruses (HTLV-III) in the serum of patients with AIDS. *Science*, **224**(4648):506–508.

SAS Institute Inc (2014). Business Analytics and Business Intelligence Software - SAS. Http://www.sas.com - Accessed 2015-06-01.

Sasco aJ (2001). Epidemiology of breast cancer: an environmental disease? *APMIS : acta pathologica, microbiologica, et immunologica Scandinavica*, **109**(5):321–332.

Saudi Human Genome Program (2013). Saudi Human Genome Program. Http://shgp.kacst.edu.sa/site/ - Accessed 2015-12-30.

Schadschneider A, Klingsch W, Kluepfel H, Kretz T, Rogsch C, & Seyfried A (2011). Evacuation Dynamics: Empirical Results, Modeling and Applications. In Meyers RA, editor, *Extreme Environmental Events*, pages 517–550. Springer, New York, NY. ISBN 978-1-4419-7695-6.

Schafer JL & Graham JW (2002). Missing data: Our view of the state of the art. *Psychological Methods*, **7**(2):147–177.

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, & Quince C (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, **43**(6):e37–e37.

Schmidt D, Kollan C, Fätkenheuer G, Schülter E, Stellbrink HJ, Noah C, Jensen BEO, Stoll M, Bogner JR, Eberle J, Meixenberger K, Kücherer C, Hamouda O, & Bartmeyer B (2014). Estimating trends in the proportion of transmitted and acquired HIV drug resistance in a long term observational cohort in Germany. *PloS one*, **9**(8):e104474.

Schwarz JM, Rödelsperger C, Schuelke M, & Seelow D (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods*, **7**(8):575–6.

Sciascia S, Mazzola P, Astrachan JH, & Pieper TM (2012). The role of family ownership in international entrepreneurship: Exploring nonlinear effects. *Small Business Economics*, **38**(1):15–31.

Semenov Ma & Barrow EM (1997). Use of a stochastic weather generator in the development of climate change scenarios. *Climatic Change*, **35**(4):397–414.

Seraphim Sense Ltd (2015). Angel open sensor wristband | The future of health and well being. Http://www.angelsensor.com/ - Accessed 2015-08-26.

Shendure J & Ji H (2008). Next-generation DNA sequencing. *Nature Biotechnology*, **26**(10):1135–1145.

Sifrim A, Popovic D, Tranchevent LC, Ardeshirdavani A, Sakai R, Konings P, Vermeesch JR, Aerts J, De Moor B, & Moreau Y (2013). eXtasy: variant prioritization by genomic data fusion. *Nature Methods*, **10**(11):1083–1084.

Sifrim A, Van Houdt JK, Tranchevent LC, Nowakowska B, Sakai R, Pavlopoulos Ga, Devriendt K, Vermeesch JR, Moreau Y, & Aerts J (2012). Annotate-it: a Swiss-knife approach to annotation, analysis and interpretation of single nucleotide variation in human disease. *Genome medicine*, **4**(9):73.

Simonotto E, Riani M, Seife C, Roberts M, Twitty J, & Moss F (1997). Visual Perception of Stochastic Resonance. *Physical Review Letters*, **78**(6):1186–1189.

Simpson JL, Rechitsky S, & Kuliev A (2013). Next-generation sequencing for preimplantation genetic diagnosis. *Fertility and sterility*, **99**(5):1203–1204.

Sims D, Sudbery I, Ilott NE, Heger A, & Ponting CP (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, **15**(2):121–132.

Skar H, Hedskog C, & Albert J (2011). HIV-1 evolution in relation to molecular epidemiology and antiretroviral resistance. *Annals of the New York Academy of Sciences*, **1230**(1):108–118.

Sluis-Cremer N, Temiz NA, & Bahar I (2004). Conformational changes in HIV-1 reverse transcriptase induced by nonnucleoside reverse transcriptase inhibitor binding. *Current HIV research*, **2**(4):323–332.

Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, Siragusa E, Zemojtel T, Buske OJ, Washington NL, Bone WP, Haendel MA, & Robinson PN (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature Protocols*, **10**(12):2004–2015.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH, & Hood LE (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, **321**(6071):674–679.

Soshnikova N, Dewaele R, Janvier P, Krumlauf R, & Duboule D (2013). Duplications of hox gene clusters and the emergence of vertebrates. *Developmental Biology*, **378**(2):194–199.

Sparrow R (2015). Imposing Genetic Diversity. *The American Journal of Bioethics*, **15**(6):2–10.

Srikanthan R & McMahon Ta (2001). Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and Earth System Sciences Discussions*, **5**(4):653–670.

Stadeli KM & Richman DD (2013). Rates of emergence of HIV drug resistance in resource-limited settings: a systematic review. *Antiviral therapy*, **18**(1):115–23.

Stark C (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, **34**(90001):D535–D539.

Stein Ra (2011). Super-spreaders in infectious diseases. *International Journal of Infectious Diseases*, **15**(8):e510–e513.

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel Ja, Thomas NST, Abeysinghe S, Krawczak M, & Cooper DN (2003). Human Gene Mutation Database (HGMD): 2003 update. *Human mutation*, **21**(6):577–581.

Strozzi F & Aerts J (2011). A Ruby API to query the Ensembl database for genomic features. *Bioinformatics*, **27**(7):1013–1014.

Suñé C, Brennan L, Stover DR, & Klimkait T (2004). Effect of polymorphisms on the replicative capacity of protease inhibitor-resistant HIV-1 variants under drug pressure. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, **10**(2):119–126.

Sutton WS (1903). The Chromosomes in Heredity. *Biological Bulletin*, **4**(5):231.

Taleb NN (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House Publishing Group. ISBN 9781588365835.

Taleb NN (2015). *Silent risk, Lectures on Probability Vol. 1*. DesCartes Publishing.

Taleb NN, Read R, Douady R, Norman J, & Bar-yam Y (2014). The precautionary principle (with Application to the Genetic Modification of Organisms). *eprint arXiv:14105787*, pages 1–17.

Tang aC, Yang Z, Reeb-Sutherland BC, Romeo RD, & McEwen BS (2012). Maternal modulation of novelty effects on physical development. *Proceedings of the National Academy of Sciences*, **109**(6):2120–2125.

Taylor JR (1997). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books, Sausalito, CA. ISBN 0935702423.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319):1061–1073.

The Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, **72**(6):971–983.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, pages 267–288.

Tjio JH & Levan A (1956). The Chromosome Number of Man. *Hereditas*, **42**(1-2):1–6.

Tory M & Möller T (2004). Human factors in visualization research. *IEEE Transactions on Visualization and Computer Graphics*, **10**(1):72–84.

Tost J (2008). *Epigenetics*. Horizon Scientific Press, Norfolk, UK. ISBN 9781904455233.

Tost J (2010). DNA methylation: An introduction to the biology and the disease-associated changes of a promising biomarker. *Molecular Biotechnology*, **44**(1):71–81.

Trenberth KE (2010). Global change: The ocean is warming, isn't it? *Nature*, **465**(7296):304–304.

Tsai CW, Lai CF, Chao HC, & Vasilakos AV (2015). Big data analytics: a survey. *Journal of Big Data*, **2**(1):21.

Tsai PC & Bell JT (2015). Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *International Journal of Epidemiology*, **44**(4):1429–1441.

Tufte E (1991). *The Visual Display of Quantitative Information, Second Edition*. Graphic Press, Cheshire, CT.

Tylee DS, Kawaguchi DM, & Glatt SJ (2013). On the outside, looking in: A review and evaluation of the comparability of blood and brain "-omes". *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, **162**(7):595–603.

UCLA: Statistical Consulting Group (2015). Statistical Computing Seminars: Introduction to Power Analysis. Http://www.ats.ucla.edu/stat/seminars/Intro_power/ - Accessed 2015-08-19.

UNAIDS (2013). *GLOBAL REPORT: UNAIDS report on the global AIDS epidemic 2013*. UNAIDS, Geneva. ISBN 9789292530327.

Urquhart NS & Weeks DL (1978). Linear models in messy data: some problems and alternatives. *Biometrics*, **34**(4):696–705.

US Department of Health and Human Services (2015). FDA-Approved HIV Medicines. Https://aidsinfo.nih.gov/education-materials/fact-sheets/21/58/fda-approved-hiv-medicines - Accessed 2015-06-23.

Vagata P & Wilfong K (2014). Scaling the Facebook data warehouse to 300 PB. Https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/ - Accessed 2015-12-30.

Vahlne A (2009). A historical reflection on the discovery of human retroviruses. *Retrovirology*, **6**:40.

Van der Aa N, Esteki M, Vermeesch JR, & Voet T (2013). Preimplantation genetic diagnosis guided by single-cell genomics. *Genome Medicine*, **5**(8):71.

van Dijk EL, Jaszczyszyn Y, & Thermes C (2014). Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research*, **322**(1):12–20.

Van Kampen NG (1992). *Stochastic processes in physics and chemistry*. Elsevier, Amsterdam. ISBN 0444893490.

van Rossum G & Drake F (2001). Python Software Foundation. Http://www.python.org/ - Accessed 2015-06-02.

Van Vliet J, Oates Na, & Whitelaw E (2007). Epigenetic mechanisms in the context of complex diseases. *Cellular and Molecular Life Sciences*, **64**(12):1531–1538.

Vandamme AM, Camacho RJ, Ceccherini-Silberstein F, de Luca A, Palmisano L, Paraskevis D, Paredes R, Poljak M, Schmit JC, Soriano V, Walter H, & Sönnerborg A (2011). European recommendations for the clinical use of HIV drug resistance testing: 2011 update. *AIDS reviews*, **13**(2):77–108.

Veenstra J, Schuurman R, Cornelissen M, van 't Wout AB, Boucher CAB, Schuitemaker H, Coudsmit J, & Coutinho RA (1995). Transmission of Zidovudine-Resistant Human Immunodeficiency Virus Type 1 Variants Following Deliberate Injection of Blood from a Patient with AIDS: Characteristics and Natural History of the Virus. *Clinical Infectious Diseases*, **21**(3):556–560.

Veraksa A, Del Campo M, & McGinnis W (2000). Developmental Patterning Genes and Their Conserved Functions: From Model Organisms to Humans. *Molecular Genetics and Metabolism*, **69**(2):85–100.

Verbeeck N (2014). *Datamining of Imaging Mass Spectrometry Data for Biomedical Tissue Exploration*. KU Leuven, Leuven. ISBN 9789460189371.

Vercauteren J, Wensing AMJ, van de Vijver DaMC, Albert J, Balotta C, Hamouda O, Kücherer C, Struck D, Schmit JC, Asjö B, Bruckova M, Camacho RJ, Clotet B, Coughlan S, Grossman Z, Horban A, Korn K, Kostrikis L, Nielsen C, Paraskevis D, Poljak M, Puchhammer-Stöckl E, Riva C, Ruiz L, Salminen M, Schuurman R, Sonnerborg A, Stanekova D, Stanojevic M, Vandamme AM, & Boucher CaB (2009). Transmission of drug-resistant HIV-1 is stabilizing in Europe. *The Journal of infectious diseases*, **200**(10):1503–8.

Vermetten YJ, Vermunt JD, & Lodewijks HG (2002). Powerful learning environments? How university students differ in their response to instructional measures. *Learning and Instruction*, **12**(3):263–284.

Vernazza P, Hirschel B, Bernasconi E, & Flepp M (2008). Les personnes seropositives ne souffrant d'aucune autre MST et suivant un traitment antiretroviral efficace ne transmettent pas le VIH par voie sexuelle. *Bulletinde medecins suisses*, **89**(5):165–169.

Vieux-Rochas M, Mascrez B, Krumlauf R, & Duboule D (2013). Combined function of HoxA and HoxB clusters in neural crest cells. *Developmental Biology*, **382**(1):293–301.

Vigen T (2015a). *Spurious Correlations*. Hachette Books, New York, NY. ISBN 978-0316339438.

Vigen T (2015b). Spurious Correlations. From: http://www.tylervigen.com/spurious-correlations - Accessed 2015-12-28 - Licensed under CC BY 4.0 - https://creativecommons.org/licenses/by/4.0/legalcode.

Villanova University (2015). What is big data? Http://www.villanovau.com/resources/bi/what-is-big-data/#.Vn–NMrOlXY - Accessed 2015-12-27.

Viti F, Hoogendoorn S, Immers L, Tampère C, & Lanser S (2008). National Data Warehouse: How the Netherlands Is Creating a Reliable, Widespread, Accessible Data Bank for Traffic Information, Monitoring, and Road Network Control. *Transportation Research Record: Journal of the Transportation Research Board*, **2049**:176–185.

Voelkerding KV, Dames Sa, & Durtschi JD (2009). Next-generation sequencing:from basic research to diagnostics. *Clinical Chemistry*, **55**(4):641–658.

Voet T, Kumar P, Van Loo P, Cooke SL, Marshall J, Lin ML, Zamani Esteki M, Van der Aa N, Mateiu L, McBride DJ, Bignell GR, McLaren S, Teague J, Butler A, Raine K, Stebbings La, Quail Ma, D'Hooghe T, Moreau Y, Futreal PA, Stratton MR, Vermeesch JR, & Campbell PJ (2013). Single-cell paired-end genome sequencing reveals structural variation per cell cycle. *Nucleic acids research*, **41**(12):6119–38.

Volcik Ka, Blanton SH, Kruzel MC, Townsend IT, Tyerman GH, Mier RJ, & Northrup H (2002). Testing for genetic associations in a spina bifida population: Analysis of the HOX gene family and human candidate gene regions implicated by mouse models of neural tube defects. *American Journal of Medical Genetics*, **110**(3):203–207.

von Kanel T, Stanke F, Weber M, Schaller A, Racine J, Kraemer R, Chanson M, Tümmler B, & Gallati S (2013). Clinical and molecular characterization of

the potential CF disease modifier syntaxin 1A. *European Journal of Human Genetics*, **21**(12):1462–1466.

Waddington CH (1942). The epigenotype. *Endeavour*, (1):18–20.

Wainberg Ma & Friedland G (1998). Public health implications of antiretroviral therapy and HIV drug resistance. *JAMA : the journal of the American Medical Association*, **279**(24):1977–1983.

Wallingford JJB, Niswander LaL, Shaw GGM, & Finnell RRH (2013). The continuing challenge of understanding, preventing, and treating neural tube defects. *Science*, **339**(6123):1–14.

Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, Ambrosone CB, Johnson CS, Smiraglia DJ, & Liu S (2012). IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics*, **28**(5):729–730.

Wang L, Wang F, Guan J, Le J, Wu L, Zou J, Zhao H, Pei L, Zheng X, & Zhang T (2010). Relation between hypomethylation of long interspersed nucleotide elements and risk of neural tube defects. *American Journal of Clinical Nutrition*, **91**(5):1359–1367.

Wang T, Guan W, Lin J, Boutaoui N, Canino G, Luo J, Celedón JC, & Chen W (2015). A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data. *Epigenetics*, **10**(7):662–669.

Wang WYS, Barratt BJ, Clayton DG, & Todd Ja (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, **6**(2):109–118.

Watson J & Crick FHF (1953). Molecular structure of nucleic acids. *Nature*, **171**(4356):737–738.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, & Parkinson H (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, **42**(D1):1001–1006.

Wensing AMJ, van de Vijver Da, Angarano G, Asjö B, Balotta C, Boeri E, Camacho R, Chaix ML, Costagliola D, De Luca A, Derdelinckx I, Grossman Z, Hamouda O, Hatzakis A, Hemmer R, Hoepelman A, Horban A, Korn K, Kücherer C, Leitner T, Loveday C, MacRae E, Maljkovic I, de Mendoza C, Meyer L, Nielsen C, Op de Coul EL, Ormaasen V, Paraskevis D, Perrin L, Puchhammer-Stöckl E, Ruiz L, Salminen M, Schmit JC, Schneider F, Schuurman R, Soriano V, Stanczak G, Stanojevic M, Vandamme AM, Van

Laethem K, Violin M, Wilbe K, Yerly S, Zazzi M, & Boucher Ca (2005). Prevalence of drug-resistant HIV-1 variants in untreated individuals in Europe: implications for clinical management. *The Journal of infectious diseases*, **192**(6):958–966.

Wensing aMJ, van Maarseveen NM, & Nijhuis M (2010). Fifteen years of HIV Protease Inhibitors: raising the barrier to resistance. *Antiviral Research*, **85**(1):59–74.

Westerfield M (1995). *The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish (Danio rerio), 3rd Edition*. University of Oregon Press, Eugene, OR. ISBN 9994860577.

WHO (2015). Obesity and overweight. Http://www.who.int/mediacentre/factsheets/fs311/en/ - Accessed 2015-08-25.

Wild a, Kalff-Suske M, Vortkamp A, Bornholdt D, König R, & Grzeschik KH (1997). Point mutations in human GLI3 cause Greig syndrome. *Human molecular genetics*, **6**(11):1979–1984.

Wilks D & Wilby R (1999). The weather generation game: a review of stochastic weather models. *Progress in Physical Geography*, **23**(3):329–357.

Wilson JMG, Jungner G, & World Health Organization (1968). Principles and practice of screening for disease. *WHO Chronicle: Public Health Papers*, **22**:473.

WIV-ISP (2013). Gezondheidsenquete 2013. Technical report. Available at: https://his.wiv-isp.be/nl/Gedeelde documenten/Summ_HC_NL_2013.pdf - Accessed 2015-08-25.

Wong aHC, Gottesman II, & Petronis A (2005). Phenotypic differences in genetically identical organisms: The epigenetic perspective. *Human Molecular Genetics*, **14**(SPEC. ISS. 1):11–18.

Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, Li F, Tsang S, Wu K, Wu H, He W, Zeng L, Xing M, Wu R, Jiang H, Liu X, Cao D, Guo G, Hu X, Gui Y, Li Z, Xie W, Sun X, Shi M, Cai Z, Wang B, Zhong M, Li J, Lu Z, Gu N, Zhang X, Goodman L, Bolund L, Wang J, Yang H, Kristiansen K, Dean M, Li Y, & Wang J (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, **148**(5):886–895.

Xue Y, Chen Y, Ayub Q, Huang N, Ball E, Mort M, Phillips A, Shaw K, Stenson P, Cooper D, & Tyler-Smith C (2012). Deleterious- and Disease-Allele Prevalence in Healthy Individuals: Insights from Current Predictions,

Mutation Databases, and Population-Scale Resequencing. *The American Journal of Human Genetics*, **91**(6):1022–1032.

Yaffe K, Haan M, Byers a, Tangen C, & Kuller L (2000). cognitive decline Evidence of gene – environment interaction. *Neurology*, **54**:1949–1953.

Yang WL, Scherrer A, Yerly S, Böni J, Klimkait T, Aubert V, Kouyos R, Günthard H, & the Swiss HIV cohort study (SHCS) (2013). Prevalence of transmitted drug resistance and relation to mean population viral load of treatment failing patients: a 16-year analysis within the Swiss HIV Cohort Study (SHCS). *Antiviral therapy*, **18**(Suppl 1):A21.

Yerly S, Junier T, Gayet-Ageron A, Amari EBE, von Wyl V, Günthard HF, Hirschel B, Zdobnov E, & Kaiser L (2009). The impact of transmission clusters on primary drug resistance in newly diagnosed HIV-1 infection. *AIDS*, **23**(11):1415–1423.

Yilmaz E, Uzuntarla M, Ozer M, & Perc M (2013). Stochastic resonance in hybrid scale-free neuronal networks. *Physica A: Statistical Mechanics and its Applications*, **392**(22):5735–5741.

Zernant J, Külm M, Dharmaraj S, den Hollander AI, Perrault I, Preising MN, Lorenz B, Kaplan J, Cremers FPM, Maumenee I, Koenekoop RK, & Allikmets R (2005). Genotyping microarray (disease chip) for Leber congenital amaurosis: detection of modifier alleles. *Investigative ophthalmology & visual science*, **46**(9):3052–3059.

Zhuang J, Jetzt AE, Sun G, Yu H, Klarmann G, Ron Y, Preston BD, & Dougherty JP (2002). Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *Journal of virology*, **76**(22):11273–11282.

Zou M, Baitei EY, Alzahrani AS, Parhar RS, Al-Mohanna FA, Meyer BF, & Shi Y (2011). Mutation prediction by PolyPhen or functional assay, a detailed comparison of CYP27B1 missense mutations. *Endocrine*, **40**(1):14–20.

Zuk O, Hechter E, Sunyaev SR, & Lander ES (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, **109**(4):1193–1198.

List of publications

IT (Articles in internationally reviewed academic journals)

Winand R, Theys K, Eusébio M, Aerts J, Camacho R, Gomes P, Suchard M, Vandamme A-M, Abecasis A (2015). Assessing transmissibility of HIV-1 drug resistance mutations from treated and from drug naïve individuals. *AIDS*, **29**(15):2045-2052.

Rochtus A, Izzi B, Vangeel E, Louwette S, Wittevrongel C, Lambrechts D, Moreau Y, Winand R, Verpoorten C, Jansen K, Van Geet C, Freson K (2015) DNA methylation analysis of Homeobox genes implicates HOXB7 hypomethylation as risk factor for neural tube defects. *Epigenetics*, **10**(1):92-101.

Winand R, Hens K, Dondorp W, de Wert G, Moreau Y, Vermeesch J, Liebaers I, Aerts J (2014) In vitro screening of embryos by whole-genome sequencing: now, in the future or never?. *Human Reproduction*, **29**(4):842-851.

Sakai R, Winand R, Verbeiren T, Vande Moere A, Aerts J (2014) Dendsort: modular leaf ordering methods for dendogram representations in R. *F1000Research*, **3**(177): 1-10.

IMa (Meeting abstracts, presented at international scientific conferences and symposia, published or not published in proceedings or journals)

Winand R, Theys K, Eusébio M, Aerts J, Camacho R, Gomes P, Vandamme A-M, Abecasis A. HIV-1 transmitted drug resistance: new insights into the transmissibility of SDRMs (presentation). *22nd International HIV Dynamics & Evolution*, Budapest, Hungary, May 2015.

Winand R, Eusébio M, Li G, Gomes P, Camacho R, Aerts J, Vandamme A-M, Abecasis A. Transmissibility of HIV-1 subtype B Drug Resistance Mutations: A Phylogenetic Framework. *19th International BioInformatics Workshop on*

Virus Evolution and Molecular Epidemiology (19th VEME), Rome, Italy, Sep. 2014.

Hens K, Winand R, Dondorp W, de Wert G, Moreau Y, Vermeesch J, Liebaers I, Aerts J. In vitro screening and selection of embryos by whole-genome sequencing and analysis - assessment of its usefulness and ethical implications. *30th Annual Meeting of ESHRE*, Munich, Germany, Jun-Jul. 2014.

Winand R, Theys K, Eusébio M, Aerts J, Camacho R, Gomes P, Vandamme A-M, Abecasis A. HIV-1 transmitted drug resistance : new insights into the transmissibility of SDRMs. *9th Benelux Bioinformatics Conference. BBC 2014*, Luxemburg, Dec. 2014.

Eusébio M, Winand R, Pineda Pena A, Li G, Camacho R, Vandamme A-M, Abecasis A. TDR levels and mutation patterns in Portugal differ significantly in HIV-1 subtype G compared to subtype B. *14th. European AIDS conference*, Brussels, Belgium, Oct 2013.

Winand R, Hens K, Sifrim A, Liebaers I, Vermeesch J, Moreau Y, Aerts J. Prediction accuracy for deleterious and disease causing mutations in healthy individuals. *Benelux Bioinformatics Conference. BBC 2013*, Brussels, Belgium, Dec. 2013.

Eusébio M, Winand R, Camacho R, Vandamme A-M, Abecasis A. HIV-1 drug resistance transmission chains in subtype B in Portugal. *Controlling the HIV Epidemic with Antiretrovirals: From Consensus to Implementation*, London, UK, Sep. 2013.

Winand R, Eusébio M, Camacho R, Vandamme A-M, Abecasis A. Development of a phylogenetic framework for the detection of HIV-1 drug resistance transmission chains. *18th International Bioinformatics. Workshop on Virus Evolution*, Gainesville, Florida, USA, Aug. 2013.

Winand R, Hens K, Liebaers I, Vermeesch J, Moreau Y, Aerts J. Accuracy of algorithms and databases for the prediction of deleterious and disease causing mutations in healthy individuals. *Proc. of ISMB/ECCB 2013*, Berlin, Germany, Jul. 2013.

Winand R, Hens K, Moreau Y, Aerts J. Paving the way for preventive genomics - The clinical validity of secondary use of genomic information. *Benelux Bioinformatics conference. BBC 2012*, Nijmegen, Nederland, Dec. 2012.

Winand R, Hens K, Dondorp W, de Wert G, Liebaers I, Vermeesch J, Aerts J. Healthy embryos, healthy people ? Assessing the predictive value and clinical utility of whole genome analysis of in vitro embryos. *Personal genomes and medical genomics*, New York, USA, Nov. 2012.

Winand R, Borry P, d'Haenens L, Derboven J, Dumortier J, Geerts D, Verhenneman G, Vermeesch J, Moreau Y, Aerts J. Extension of Healthy Life Using Preventive Genomics. *Capita selecta in Complex Disease Analysis 2012*, Liège, Belgium, May-Jun. 2012.

Winand R, Borry P, d'Haenens L, Derboven J, Dumortier J, Geerts D, Verhenneman G, Vermeesch J, Moreau Y, Aerts J. Extension of Healthy Life Using Preventive Genomics. *Genomic Disorders 2012 – The Genomics of Rare Diseases*, Hinxton, UK, Mar. 2012.

Winand R, Borry P, d'Haenens L, Derboven J, Dumortier J, Geerts D, Verhenneman G, Vermeesch J, Moreau Y, Aerts J. Extension of Healthy Life Using Preventive Genomics. *EMBL Conference: Omics and Personalised Health*, Heidelberg, Germany, Feb. 2012.

AMa (Meeting abstracts, presented at other scientific conferences and symposia, published or not published in proceedings or journals)

Winand R, Hens K, Vermeesch J, Moreau Y, Aerts J. Secondary use of genomic information: deleterious and disease causing mutations in healthy individuals. *Belgian Society of Human Genetics (BeSHG) 13th Annual Meeting "Genetics of Human Development EXPOsed"*. Brussels, Belgium, Mar. 2013.

Winand R, Borry P, d'Haenens L, Derboven J, Dumortier J, Geerts D, Verhenneman G, Vermeesch J, Moreau Y, Aerts J. Extension of Healthy Life Using Preventive Genomics. *Belgian Society of Human Genetics (BeSHG) 12th Annual Meeting "Next Generation Sequencing and Recent Advances in Genetics"*, Liège, Belgium, Mar. 2012.



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
STADIUS CENTER FOR DYNAMICAL SYSTEMS, SIGNAL PROCESSING AND DATA ANALYTICS
Kasteelpark Arenberg 10 box 2446
B-3001 Leuven