# Executive Summary

## Analysis

Moaraj Hasan's thesis was analyzed for plagiarism with a combination of i) a manual search and of ii) a software tool[1].

All instances of plagiarism were grouped into the following three categories, depending on the infringement of Basic principles as stated in the **Citation Etiquette** guidelines set by ETH Zürich

1. **Missing citations**: Text is copied verbatim from another source or mildly rephrased. There is no citation to help distinguish between the student's own contribution and the work of others.
2. **Misleading citations**: Text is copied verbatim from another source or mildly rephrased. There is a citation but it does not reflect the true source of the borrowed text.
3. **Incorrect citation**: Text is copied word-by-word from another source. A citation can be found in the neighborhood of the borrowed text, but the text does not follow the above mentioned **Citation Etiquette** guidelines, e.g. it is not quoted in inverted commas.

## Format of the report

Each of the problems detected in the thesis are reported in the following way:

**Type of infringement** (color coded, see below)

[MH] Page 99, Section 99.99

Text1 text2 text3

[Ref 99]

Text1 some Text3

Where

- **Type of infringement**: is a header indicating the type of problem reported. It is one of the three types listed above and is color coded as shown in the next subsection
- [MH] Page 99, Section 99.99: Marks the beginning of Mr. Moaraj Hasan's

[1] Docol©c. URL: http://www.docoloc.de/

(MH) text, followed by the page number and section in the document where the quote can be found.

- [Ref 99]: Indicates the publication/reference to which MH's text is associated. The numbered references can be found at the end of this report.
- Text1 text2 text3: Is the text extracted from MH's thesis. Highlighted in orange, are words copied verbatim from the reference. Highlighted in yellow, are words rephrased from the reference. No highlight indicates words that were ignored in the quote.
- Text1 some Text3: Is the text extracted from the reference. As before, the colors are used to highlight portions of the text that were copied word-by-word (in orange) versus text that was rephrased (yellow).

### Type of infringement. Color coding

For each of the issues listed above, this report highlights their occurrence with a header color-coded in the following way:

**Missing citation**

**Misleading citation**

**Incorrect citation**

## Summary of findings

| Type of infringement | Count |
|---|---|
| **Missing citation** | 17 |
| **Misleading citation** | 11 |
| **Incorrect citation** | 3 |
| Total | 31 |

# Report

## Chapter 1, Introduction

---

### Missing citation

[MH] Page 9, General introduction

The late 19th and early 20th centuries were a golden era of health innovation. Breakthroughs like germ theory, antibiotics, and widespread vaccination, as well as major public-health advances in sanitation and regulation, neutralized many long-leading causes of death

[Ref 1]

The late 19th and early 20th centuries were a golden era of American health innovation. Breakthroughs like germ theory, antibiotics, and widespread vaccination, as well as major public-health advances in sanitation and regulation, neutralized many long-leading causes of death.

---

### Missing citation

[MH] Page 9, General introduction

With life expectancy increasing rapidly in the focus of medical innovation shifted away from the eradicating of infection disease and more towards managing chronic age-related conditions.

[Ref 1]

Life expectancy skyrocketed as a result, but brought with it new demons. For the past 50 years, medical innovation has focused less on eradicating disease and more on managing chronic conditions.

---

**Missing citation**

[MH] Page 9, General introduction

The development of longitudinal assessments of aging phenotypes in multiple model organisms, with attention to differences in sex, strain and diet composition, could accelerate our ability to better screen for interventions that would lead to people living longer and healthier lives.

[Ref 2]

The development of longitudinal assessments of aging phenotypes in multiple model organisms, with attention to differences in sex, strain and diet composition, could accelerate our ability to better screen for interventions that would lead to people living longer and healthier lives.

**Missing citation**

[MH] Page 10, General introduction

The ideal interventions in aging would be able to preserve all the faculties of a young person in their prime in an old person and maintain them for as long as possible, ideally well into a person's golden years.

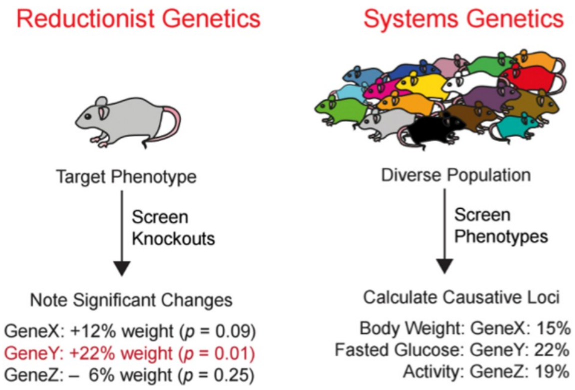
[Ref 2]

The ideal would be to preserve all the faculties of a young person in their prime in an old person and maintain them for as long as possible, ideally well into a person's golden years.

## Missing citation

[MH] Page 12, Section 1.1 Project Outline

Figure 1.2



[Ref 3]

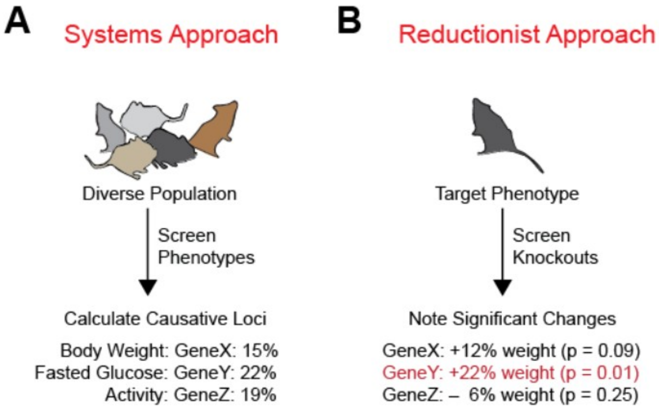


Figure 1:2 **Principal Study Designs**

## Missing citation

[MH] Pages 13-14, Section 1.3 Reverse and Forward genetics

The process of disruption or alteration can either be targeted specifically as in the case of gene silencing or homologous recombination or can rely on non-targeted random disruptions (e.g., chemical mutagenesis, transposon mediated mutagenesis) followed by screening a library of individuals for lesions at a specific location.

[Ref 4]

The process of disruption or alteration can either be targeted specifically as in the case of gene silencing or homologous recombination or can rely on non-targeted random disruptions (e.g. chemical mutagenesis, transposon mediated mutagenesis) followed by screening a library of individuals for lesions at a specific location.

---

**Missing citation**

[MH] Page 14, Section 1.3 Reverse and Forward genetics

Variants that deviate from the average trait presentation in heterogeneous population can be measured at many scales from macro (body size, morphology) to different levels of micro variation (protein profiles and DNA sequence variation).

[Ref 4]

Variation can be measured at many scales – from macro (body size, morphology) to different levels of micro variation (crude protein profiles to DNA sequence variation).

---

**Missing citation**

[MH] Page 14, Section 1.3 Reverse and Forward genetics

In many cases the observable variation has been induced through controlled out breeding(in the case of model organisms) but also may be naturally occurring as in the case of subsets of wild populations that have been sequenced.

[Ref 4]

In many cases the observable variation has been induced using a DNA damaging agent (mutagen) but also may be naturally occurring.

---

**Missing citation**

[MH] Pages 14-15, Section 1.3 Reverse and Forward genetics

Instead of going from phenotype to sequence as in forward genetics, we work in the opposite direction a gene sequence is known and hypothesize to be driving an age-related phenotype, but its exact function is uncertain.

[Ref 5]

So, instead of going from phenotype to sequence as in forward genetics, reverse genetics works in the opposite direction – a gene sequence is known, but its exact function is uncertain.

---

**Incorrect citation, text copied word-by-word**

[MH] Page 15, Section 1.4 Systems Approach to Complex Trait Analysis

Systems Genetics is an approach to understand the flow of biological information that underlies complex traits. It uses a range of experimental and statistical methods to quantitative and integrate intermediate phenotypes, such as transcript, protein or metabolite levels, in populations that vary for traits of interest.

[Ref 6]

Systems genetics is an approach to understand the flow of biological information that underlies complex traits. It uses a range of experimental and statistical methods to quantitate and integrate intermediate phenotypes, such as transcript, protein or metabolite levels, in populations that vary for traits of interest

---

**Missing citation**

[MH] Page 19, Section 1.6 Studying Heredity in Model Organism Populations

What this means is that while there is an influence exerted by environmental factors (nutritional intake) in the weight of an individual, the major portion of the variation in mouse body weight can be explained by genetics. More importantly, it really tells us that the major influence in explaining the differences between individuals is accounted for in this fashion. Note that it tells us nothing about what gave rise to the particular weight for any particular individual, but rather what explains the differences between individuals within a particular population.

[Ref 7]

What this means is that while there is an influence exerted by environmental factors (such as nutrients) in the height of an individual, the major portion of the influence is exerted by the genes. More importantly, it really tells us that the major influence in explaining the DIFFERENCES between individuals is accounted for in this fashion. Note that it tells us nothing about what gave rise to the particular height for any particular individual, but rather what explains the differences between individuals within a particular population.

---

**Missing citation**

[MH] Page 20, Section 1.6 Studying Heredity in Model Organism Populations

This is because the variation between the groups is not accounted for solely by the genes, but also by the diets that produce the weight.

[Ref 7]

...indicated previously because the variation between the groups is not accounted for solely by the genes, but also by the environmental effects that produce the stunting.

---

**Missing citation**

[MH] Page 20, Section 1.6 Studying Heredity in Model Organism Populations

In both cases, the environment is essentially held constant, so the variation in weight is solely due to the genes, hence 71% and 74% heritability are calculated.

[Ref 7]

In both cases, the environment is essentially held constant, so the variation in height is solely due to the genes, hence 100% heritability.

---

## Misleading citation

[MH] Page 20, Section 1.7 What is QTL Analysis?

A Quantitative Trait Locus (QTL) analysis is a statistical method that is aimed at linking measurements of continuous phenotypic trait and genotypic molecular markers, in attempt to explain the phenotypic variation with genetic variation.

[Ref 8]

Quantitative trait locus (QTL) analysis is a statistical method that links two types of information, phenotypic data (trait measurements) and genotypic data (usually molecular markers), in an attempt to explain the genetic basis of variation in complex traits.

## Misleading citation

[MH] Page 24, Section 1.8 QTL vs. GWAS

QTL mapping suffers from two limitations; only allelic diversity that segregates between the parents of the particular F2 cross or within the RI population can be assayed ,and second, the amount of recombination that occurs during the creation of the RI population places a limit of the resolution of the mapping.

[Ref 9]

Despite this success, QTL mapping suffers from two fundamental limitations; only allelic diversity that segregates between the parents of the particular F2 cross or within the RIL population can be assayed [5], and second, the amount of recombination that occurs during the creation of the RIL population places a limit on the mapping resolution.

## Misleading citation

[MH] Page 24, Section 1.9 RI Mouse Population for GxE

Mice and other inbred and isogenic model organisms are extremely well suited to evaluate complex experimental effects in the context of QTL mapping. The ability to impose well-controlled perturbations across large cohorts is among the strongest motivations to use model organisms. This kind of design is already the most common and critical in agricultural genetics.

[Ref 10]

Mice and other inbred and isogenic model organisms are extremely well suited to evaluate complex experimental effects in the context of QTL mapping. The ability to impose well-controlled perturbations across large cohorts is among the strongest motivations to use model organisms. This kind of design is already the most common and critical in agricultural genetics.

---

## Misleading citation

[MH] Page 25, Section 1.9 RI Mouse Population for GxE

As discussed about the most important disadvantage of conventional RI strains and other standard two-parent crosses is that they segregate for only a fraction of all known polymorphisms. For example, the BXD family segregates for a total of ˜5.2 million sequence variants about 44% of common variants among standard inbred strains.

[Ref 10]

The most important disadvantage of conventional RI strains and other standard two-parent crosses is that they segregate for only a fraction of all known polymorphisms. For example, the BXD family segregates for a total of ~5.2 million sequence variants — about 44% of common variants among standard inbred strains.

---

## Missing citation

[MH] Page 26, Section 1.10.2 Mouse Diets

It is composed of agricultural byproducts, such as ground wheat, corn, or oats, alfalfa and soybean meals, a protein source such as fish, and vegetable oil and is supplemented with minerals and vitamins. Thus, chow is a high fiber diet containing complex carbohydrates, with fats from a variety of vegetable sources. In contrast, defined high- fat(HF) diet (Teklad20180614) consist of amino acid supplemented casein, corn- starch, maltodextrose or sucrose, and soybean oil or lard, also supplemented with minerals and vitamins. Fiber is often provided by cellulose.

[Ref 11]

Regular chow is composed of agricultural byproducts such as ground wheat, corn, or oats, alfalfa and soybean meal, a protein source such as fish, and vegetable oil and is supplemented with minerals and vitamins. Thus, chow is a high-fiber diet containing complex carbohydrates, with fats from a variety of vegetable sources. Chow is inexpensive to manufacture and is palatable to rodents. In contrast, defined high-fat diets consist of amino acid-supplemented casein, corn starch, maltodextrose or sucrose, and soybean oil or lard, also supplemented with minerals and vitamins. Fiber is often provided by cellulose.

**Missing citation**

[MH] Pages 26-27, Section 1.10.2 Mouse Diets

CD and HF diets may exert significant separate and independent unintended effects on the measured metabolites, proteins and transcripts.

[Ref 11]

Chow and defined diets may exert significant separate and independent unintended effects on the measured phenotypes in any research protocol.

**Misleading citation**

[MH] Page 27, Section 1.10.3 Mouse Sex

In humans, the reproductive cycle, called the menstrual cycle, lasts approximately 28 days, in rodents this cycle, called the estrous cycle, lasts approximately 4-5 days.

[Ref 12]

In humans, the reproductive cycle, called the menstrual cycle, lasts approximately 28 days, in rodents this cycle, called the estrous cycle, lasts approximately 4-5 days.

**Misleading citation**

[MH] Page 28, Section 1.10.3 Mouse Sex

The dominant female mice physically nibble or pluck fur and whiskers from their cage mates.

[Ref 13]

The dominant mice physically nibble or pluck fur and whiskers from their cage mates.

---

**Incorrect citation, text copied word-by-word**

[MH] Page 28, Section 1.11 Why Multiple Omics

Large-scale initiatives toward personalized medicine are driving a massive expansion in the number of human genomes being sequenced.

[Ref 14]

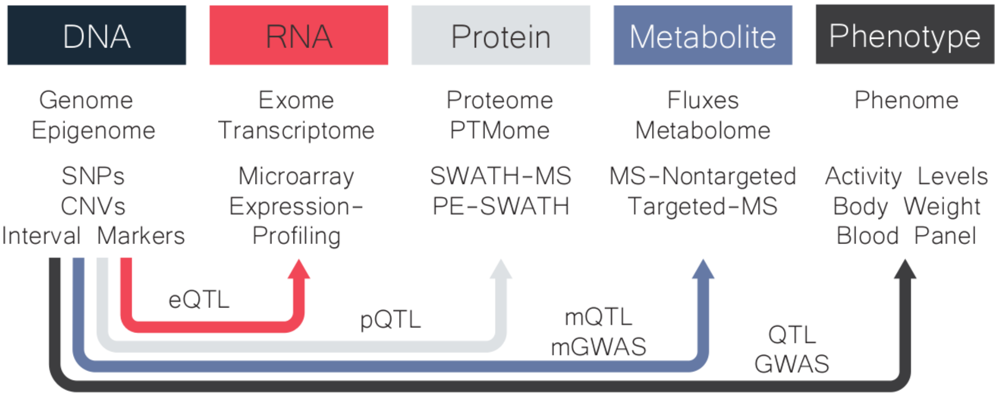
Large-scale initiatives toward personalized medicine are driving a massive expansion in the number of human genomes being sequenced.

---

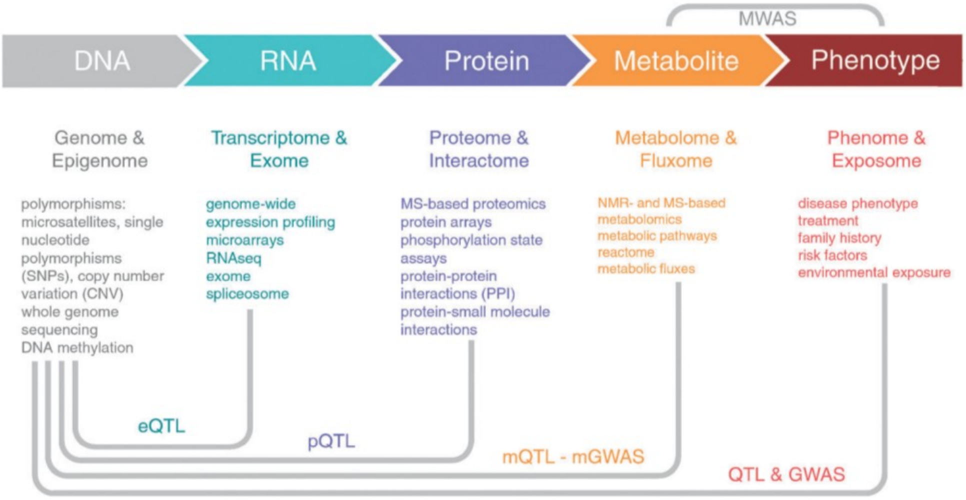
## Missing citation

[MH] Page 29, Section 1.11 Why Multiple Omics

Figure 1.8



[Ref 15]



## Misleading citation

[MH] Page 29, Section 1.11 Why Multiple Omics

Owing to the fact that transcript, protein and metabolites levels have only a modest correlation with each other (Williams et al., 2016), and that metabolites can be further modified by enzymatic processes, good experimental design...

[Ref 16]

Owing to the fact that transcript and protein levels have only a modest correlation with each other, and that metabolites can be further modified by enzymatic processes and can originate from...

## Chapter 2, Metabolomics

### Misleading citation

[MH] Page 32, Section 2.1 Introduction to Non-Targeted Metabolomics

Therefore, metabolomics analyses are always focused on a part of a given metabolome, for small polar molecules which are the focus of this study or the lipids extracted from a biological sample

[Ref 17]

Therefore, metabolomics analyses are always focused on a part of a given metabolome, e.g., the small polar molecules or the lipids extracted from a biological sample.

### Incorrect citation, text copied word-by-word

[MH] Page 32, Section 2.1 Introduction to Non-Targeted Metabolomics

Such an inevitable focus is the reason for the rather independently evolving disciplines of small polar molecules metabolomics, lipidomics, glycomics and other metabolome related approaches.

[Ref 17]

Such an inevitable focus is the reason for the rather independently evolving disciplines of “small polar molecules” metabolomics, lipidomics, glycomics and other metabolome related approaches.

### Missing citation

[MH] Page 33, Section 2.1.1 Metabolomics Methods

It involves the application of advanced analytical tools to profile the diverse metabolic complement of BXD mouse livers.

[Ref 18]

Metabonomics involves the application of advanced analytical tools to profile the diverse metabolic complement of a given biofluid or tissue.

# Chapter 3, Proteomics

## Misleading citation

[MH] Page 90, Section 3.7 Random Forest

Random Forests is an ensemble classifier which uses many decision tree models to predict a classification. A different subset training data is selected, with replacement to train each tree(Goel and Abhilasha, 2017). We can get an idea the mechanism from the name itself-"Random Forests". A collection trees is a forest, and the trees are being trained on subsets which are being selected at random, hence random forests(Goel and Abhilasha, 2017). This can be used for classification and old and young age group or regression against the 4 age cohorts.

[Ref 19]

Random Forests is an ensemble classifier which uses many decision tree models to predict the result. A different subset of training data is selected, with replacement to train each tree. We can get an idea of the mechanism from the name itself-"Random Forests". A collection of trees is a forest, and the trees are being trained on subsets which are being selected at random, hence random forests. This can be used for classification and regression problems. Class assignment is made by the number of votes from all the trees and for regression the average of the results is used.

**Misleading citation**

[MH] Page 96, Section 4.1.2 Data Extraction

The CDF (Chip Description File) contains information about the layout the chip. There is one for each chip type. In this experiment, we are looking for the same cohort of transcripts in all of the samples so there is only one CDF file. A CDF file allows you to link between probes and probesets Identify and determine which probes are perfect matches and which are control mismatch probes.

[Ref 20]

The CDF (Chip Description File) contains information about the layout of the chip. There is one for each chip type. So for an experiment you normally have only one. A CDF file allows you to

- Link between probes and probesets
- Identify which probes are PM and which are MM
- Identify control probes.

**Misleading citation**

[MH] Page 96, Section 4.1.2 Data Extraction

a CEL file stores the results the intensity calculations on the pixel values that are extracted from a DAT file. This includes an intensity value, standard deviation of the intensity, the number pixels used to calculate the intensity value, a flag to indicate an outlier as calculated by the algorithm and a user defined flag indicating the feature should be excluded from future analysis(Miller and Tang, 2009). The file stores this aforementioned data for each feature on the probe array.

[Ref 21]

The CEL file stores the results of the intensity calculations on the pixel values of the DAT file. This includes an intensity value, standard deviation of the intensity, the number of pixels used to calculate the intensity value, a flag to indicate an outlier as calculated by the algorithm and a user defined flag indicating the feature should be excluded from future analysis. The file stores the previously stated data for each feature on the probe array.

---

**Missing citation**

[MH] Pages 96-97, Section 4.1.3 Normalization

When running experiments that involve multiple high density oligonucleotide arrays, it is important to remove sources variation between arrays non-biological origin. Normalization is a process for reducing this variation. It is common to see non-linear relations between arrays and the standard normalization provided by Affymetrix does not perform well in these situations.

[Ref 22]

When running experiments that involve multiple high density oligonucleotide arrays, it is important to remove sources of variation between arrays of non-biological origin. Normalization is a process for reducing this variation. It is common to see non-linear relations between arrays and the standard normalization provided by Affymetrix does not perform well in these situations.

---

# References

[1] Joe Pinsker**.** Why We Live 40 Years Longer Today Than We Did in 1880. The golden age of medicine—in one chart. The Atlantic. 11.2013 Issue.

[2] Mary Armanios, Rafael de Cabo, Joan Mannick, Linda Partridge, Jan van Deursen and Saul Villeda. Translational strategies in aging and age-related disease. Nature Medicine 21, 1395–1399 (2015)

[3] Evan Williams. A Systems Approach to Identify Genetic and Environmental Regulators of Metabolism. THÈSE NO 6486 (2015). ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE.

[4] Tierney, Melinda Beth, "Reverse Genetics with TILLING in Phytophthora sojae. " Master's Thesis, University of Tennessee, 2005.
URL: http://trace.tennessee.edu/utk_gradthes/2552

[5] Tierney, M.B. and Lamour, K.H. 2005. An Introduction to Reverse Genetic Tools for Investigating Gene Function. *The Plant Health Instructor.* DOI: 10.1094/PHI-A-2005-1025-01.

[6] Mete Civelek and Aldons J. Lusis. Systems genetics approaches to understand complex traits. Nature Reviews Genetics 15, 34–48 (2014) doi:10.1038/nrg3575

[7] What Is Heritability? Science 2.0. Evolution. Gerhard Adam. Sep 06, 2012.
http://www.science20.com/gerhard_adam/what_heritability-93424

[8] Wilson Thau Lym Yong, Grace Joy Wei Lie Chin and Kenneth Francis Rodrigues. Genetic Identification and Mass Propagation of Economically Important Seaweeds. http://dx.doi.org/10.5772/62802
URL: https://cdn.intechopen.com/pdfs-wm/50406.pdf

[9] Arthur Korte and Ashley Farlow. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 2013, 9:29

[10] Williams, R. W. and Williams, E. G. (2017). Resources for systems genetics. In Methods in Molecular Biology (Vol. 1488, pp. 1-29). (Methods in Molecular Biology; Vol. 1488). Humana Press Inc.. DOI: 10.1007/978-1-4939-6427-7_1
**Note: This reference is also found in MH's thesis**

[11] Warden, Craig H. et al. Comparisons of Diets Used in Animal Models of High-Fat Feeding. Cell Metabolism , Volume 7 , Issue 4 , 277

[12] Claudia Caligioni. Assessing Reproductive Status/Stages in Mice. Curr Protoc Neurosci. 2009 July; APPENDIX: Appendix–4I. doi: 10.1002/0471142301.nsa04is48

[13] Peter Kelmenson. OH NO, MY MICE ARE BALDING! Blog Post June 06, 2012
https://www.jax.org/news-and-insights/jax-blog/2012/june/oh-no-my-mice-are-balding20150422t160237

[14] Amalio Telenti, Levi C. T. Pierce, William H. Biggs, Julia di Iulio, Emily H. M. Wong, Martin M. Fabani, Ewen F. Kirkness, Ahmed Moustafa, Naisha Shah, Chao Xie, Suzanne C. Brewerton, Nadeem Bulsara, Chad Garner, Gary Metzker, Efren Sandoval, Brad A. Perkins, Franz J. Och, Yaron Turpaz, and J. Craig Venter. Deep sequencing of 10,000 human genomes. PNAS 2016 113 (42) 11901-11906; doi:10.1073/pnas.1613365113
**Note: This reference is also found in MH's thesis**

[15] Marc-Emmanuel Dumas. Metabolome 2.0: quantitative genetics and network biology of metabolic phenotypes. Mol. BioSyst., 2012, 8, 2494-2502. DOI: 10.1039/C2MB25167A

[16] Caroline H. Johnson, Julijana Ivanisevic and Gary Siuzdak. Metabolomics: beyond biomarkers and towards mechanisms. Nature Reviews Molecular Cell Biology 17, 451–459 (2016) doi:10.1038/nrm.2016.25

[17] ETH Zürich. Functional Genomics Center Zurich.
URL: http://www.fgcz.ch/omics_areas/met.html
**Note: This reference is also found in MH's thesis**

[18] Muireann Coen. A metabonomic approach for mechanistic exploration of pre-clinical toxicology. Toxicology. Volume 278, Issue 3, 30 December 2010, Pages 326-340

[19] Quora.com; What is the difference between random forest and decision tress?
URL: https://www.quora.com/What-is-the-difference-between-random-forest-and-decision-tress#

[20] Ben Bolstad. (Presentation slides) Software for affy data analysis Bioconductor, AffyExtensions and RMAExpress
URL: http://bmbolstad.com/stuff/SoftwareTalk.pdf

[21] Leigh Wilson and David Chambers. Transcriptomic analysis of midbrain and individual hindbrain rhombomeres in the chick embryo. Scientific Data 1, Article number: 140014 (2014) doi:10.1038/sdata.2014.14
Table 1 (caption) URL: https://www.nature.com/articles/sdata201414/tables/1

[22] Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19(2):185–93.
**Note: This reference is also found in MH's thesis**