Understanding and improving microbial cell factories through Large Scale Data-approaches

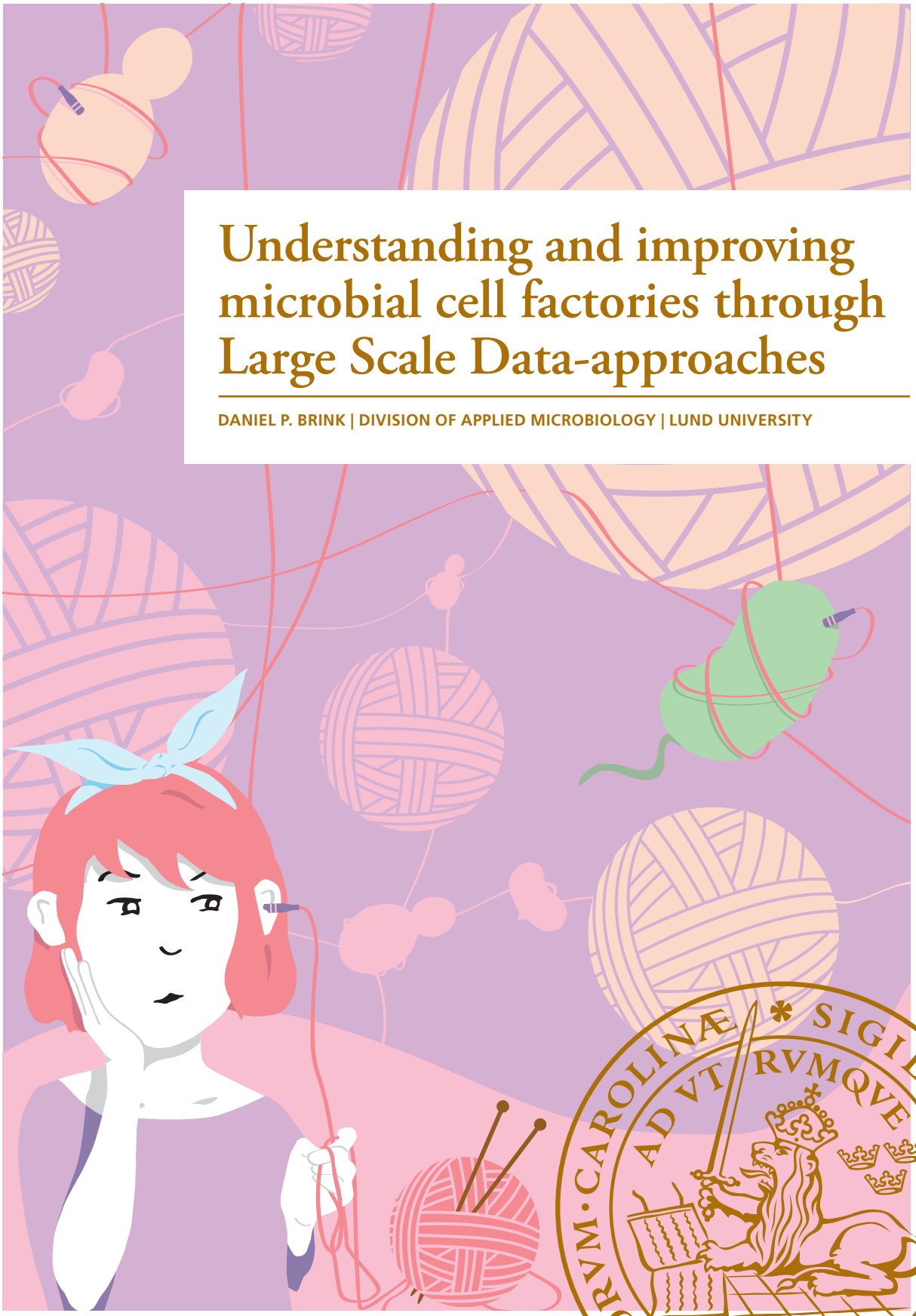
Brink, Daniel

2019

Document Version:
Publisher's PDF, also known as Version of record

Understanding and improving microbial cell factories through Large Scale Data-approaches

DANIEL P. BRINK | DIVISION OF APPLIED MICROBIOLOGY | LUND UNIVERSITY



Understanding and improving microbial cell factories through Large Scale Data-approaches

Daniel P. Brink

Division of Applied Microbiology
Department of Chemistry
Lund University

LUND
UNIVERSITY

DOCTORAL DISSERTATION
by due permission of the Faculty of Engineering, Lund University, Sweden.
To be defended at Kemicentrum, Lecture hall C, Lund
Thursday, 7 th of November 2019 at 10:15.

Faculty opponent
Dr. Kiran Raosaheb Patil
The European Molecular Biology Laboratory (EMBL), Germany

Organization
LUND UNIVERSITY
Division of Applied Microbiology
Department of Chemistry
Faculty of Engineering
P.O Box 124
SE-221 00 Lund, Sweden

Author: Daniel P. Brink

Document name
DOCTORAL DISSERTATION

Date of issue: 14th October 2019

Sponsoring organizations:
Swedish Foundation for Strategic Research
Swedish Energy Agency

Title and subtitle: Understanding and improving microbial cell factories through Large Scale Data-approaches

Abstract

Since the advent of high-throughput genome sequencing methods in the mid-2000s, molecular biology has rapidly transitioned towards data-intensive science. Recent technological developments have increased the accessibility of omics experiments by decreasing the cost, while the concurrent design of new algorithms have improved the computational work-flow needed to analyse the large datasets generated. This has enabled the long standing idea of a *systems* approach to the cell, where molecular phenomena are no longer observed in isolation, but as parts of a tightly regulated cell-wide system. However, large data biology is not without its challenges, many of which are directly related to how to store, handle and analyse ome-wide datasets.

The present thesis examines large data microbiology from a middle ground between metabolic engineering and *in silico* data management. The work was performed in the context of applied microbial lignocellulose valorisation with the end goal of generating improved cell factories for the production of value-added chemicals from renewable plant biomass. Three different challenges related to this feedstock were investigated from a large data-point of view: bacterial catabolism of lignin and its derived aromatic compounds; tolerance of baker's yeast *Saccharomyces cerevisiae* to inhibitory compounds in lignocellulose hydrolysate; and the non-fermentable response to xylose in *S. cerevisiae* engineered for growth on this pentose sugar.

The bibliome of microbial lignin catabolism is vast and consists of a long-standing cohort of fundamental microbiology, and a more recent cohort of applied lignin bio-valorisation. Here, an online database was created with the long-term ambition of closing the gap between the two and make new connections that can fuel the generation of new knowledge. Whole-genome sequencing was used to investigate the genetic basis for observed phenotypes in bacterial isolates capable of growing on different kinds of lignin-derived aromatics. A whole-genome approach was also used to identify key sequence variants in the genotype of an industrial *S. cerevisiae* strain evolved for improved tolerance to inhibitors and high temperature. Finally, assessment of the sugar signalome of *S. cerevisiae* was enabled by the design and validation of a panel of *in vivo* fluorescent biosensors for single-cell cytometric analysis. It was found that xylose triggered a signal similar to that of low glucose in yeast cells engineered with xylose utilization pathways, and that introduction of deletions previously related to improved xylose utilization altered the signal towards that of high glucose.

Taken together, the present thesis illustrates how omics-approaches can aid design of laboratory experiments to increase the knowledge and understanding of microorganisms, and demonstrates the need for a combined knowledge of molecular and computational biology in large-scale data microbiology.

Key words: Lignocellulose, lignin, xylose, bioinformatics, whole-genome sequencing, flow cytometry, signalling pathways, *Saccharomyces cerevisiae, Pseudomonas putida,*

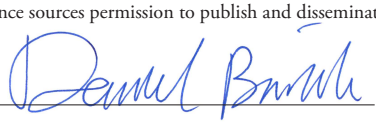
Classification system and/or index terms (if any):

Supplementary bibliographical information: Language: English

ISSN and key title: ISBN: 978-91-7422-684-3

Recipient's notes | Number of pages: 262 | Price

Security classification

Signature  Date: 25th September 2019

Understanding and improving microbial cell factories through Large Scale Data-approaches

Daniel P. Brink

Division of Applied Microbiology
Department of Chemistry
Lund University

LUND
UNIVERSITY

Cover illustration (front):
Hyperthesis: binding the boundaries (digital, 2019); Daniel P. Brink

Cover photo (back):
Lumberhack I (2019); Nikon F2, Nikkor-H Auto 50mm f/2, Tri-X 400; Daniel P. Brink

Division of Applied Microbiology
Department of Chemistry
Faculty of Engineering
P.O Box 124
SE-221 00 Lund
Sweden

ISBN: 978-91-7422-684-3 (print)
ISBN: 978-91-7422-685-0 (digital)

Printed in Sweden by Media-Tryck, Lund University
Lund 2019

Printed matter
3041 0903

Media-Tryck is an environmentally certified and ISO 14001:2015 certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

The only thing I know is that I know nothing

(The Socratic paradox)

Abstract

Since the advent of high-throughput genome sequencing methods in the mid-2000s, molecular biology has rapidly transitioned towards data-intensive science. Recent technological developments have increased the accessibility of omics experiments by decreasing the cost, while the concurrent design of new algorithms have improved the computational work-flow needed to analyse the large datasets generated. This has enabled the long standing idea of a *systems* approach to the cell, where molecular phenomena are no longer observed in isolation, but as parts of a tightly regulated cell-wide system. However, large data biology is not without its challenges, many of which are directly related to how to store, handle and analyse ome-wide datasets.

The present thesis examines large data microbiology from a middle ground between metabolic engineering and *in silico* data management. The work was performed in the context of applied microbial lignocellulose valorisation with the end goal of generating improved cell factories for the production of value-added chemicals from renewable plant biomass. Three different challenges related to this feedstock were investigated from a large data-point of view: bacterial catabolism of lignin and its derived aromatic compounds; tolerance of baker's yeast *Saccharomyces cerevisiae* to inhibitory compounds in lignocellulose hydrolysate; and the non-fermentable response to xylose in *S. cerevisiae* engineered for growth on this pentose sugar.

The bibliome of microbial lignin catabolism is vast and consists of a long-standing cohort of fundamental microbiology, and a more recent cohort of applied lignin biovalorisation. Here, an online database was created with the long-term ambition of closing the gap between the two and make new connections that can fuel the generation of new knowledge. Whole-genome sequencing was used to investigate the genetic basis for observed phenotypes in bacterial isolates capable of growing on different kinds of lignin-derived aromatics. A whole-genome approach was also used to identify key sequence variants in the genotype of an industrial *S. cerevisiae* strain evolved for improved tolerance to inhibitors and high temperature. Finally, assessment of the sugar signalome of *S. cerevisiae* was enabled by the design and validation of a panel of *in vivo* fluorescent biosensors for single-cell cytometric analysis. It was found that xylose triggered a signal similar to that of low glucose in yeast cells engineered with xylose utilization pathways, and that introduction of deletions previously related to improved xylose utilization altered the signal towards that of high glucose.

Taken together, the present thesis illustrates how omics-approaches can aid design of laboratory experiments to increase the knowledge and understanding of microorganisms, and demonstrates the need for a combined knowledge of molecular and computational biology in large-scale data microbiology.

Popular scientific summary

The technological advancements in society continuously change how we live and work. Over the last five decades, computers have helped us organize and process text and numbers, and the internet has given us access to a 24-7 wealth of information and global communication. These developments have also changed how science is performed and disseminated. Specialized instruments can now make hundreds of thousands measurements of a sample in one go, immensely speeding up research outcomes. As a result, some fields in contemporary cell biology are now as much about data handling and -understanding, as they are about the biology itself.

This type of so-called Large Data biology has opened up whole new possibilities on how the microbial cell can be investigated. While traditional molecular microbiology approaches the subject by studying a couple of elements in a cell such as genes and proteins on their own, the new technologies allow to study whole layers (so called *omes*) of the cell at once; for instance, the genome consists of all the genes in a cell, the transcriptome all the mRNA that have been expressed from the genes at a given time, the proteome all the proteins translated from said mRNA at a given time, and the metabolome all the chemical compounds (metabolites) produced by the proteins. The methods used to measure these omes are referred to as *omics*; for instance, the technique to identify the genome (all the genes in the cell) is called genomics.

The sheer size and complexity of the data generated by ome-wide studies calls for scientists to have simultaneous knowledge of the biology (here: the microbial cell) as well as the computational part. The process of handling large biological data is known as bioinformatics, and is together with data management and computer programming an invaluable tool for the modern molecular microbiologist.

In the present thesis, Large Data biology was applied to improve the knowledge and understanding of microbial cells designed for sustainable production of renewable chemicals. Central to the investigation was biological conversion of non-edible plant matter (so called lignocellulose), such as corn stover, wood chips and bagasse, into societally valuable products, e.g. bioethanol. The current work focused on the initial half of the microbial conversion: how lignocellulosic compounds can be better taken up and broken down by the cell.

Three case studies were considered: i) how to better assess the scientific literature; ii) how to determine the genome sequence of complex industrial microorganisms and new isolates (genomics); and iii) how to measure how the cell senses its nutrients (here: different sugars) and controls its breakdown.

In the first case, a web-based database was designed and developed that collects the large and slightly disjointed scientific literature on the microbial breakdown of lignin, one of the major components of lignocellulose. The goal of the database is to collect all current knowledge on lignin biodegradation in a single interactive platform in order to simplify the process of data retrieval for the scientific community.

In the second case, the genomes of lignin-degrading bacteria and a lignocellulose fermenting yeast were determined by whole-genome sequencing methods. This method produces millions of small snippets of DNA that have to be assembled back to the full genome – a process not unlike that of building a jigsaw puzzle, only that the final picture often is unknown at the start. The assembled genomes were then used to determine the presence of genes related to the ability to grow on lignin and its related aromatic compounds. Genomics methods were also used to discover mutations in a yeast strain that had acquired increased tolerance to stressful conditions encountered in industrial lignocellulose fermentation, in order to explain why this yeast had become more robust.

In the third case, the peculiar behavior of baker's yeast *Saccharomyces cerevisiae* to the five-carbon sugar xylose was investigated. This yeast cannot naturally grow on xylose, and has to be genetically modified with genes from other organisms to do so. Still, even after genetic engineering, the yeast grows much slower on xylose than on its preferred sugar glucose, and produces ethanol at a lower rate. To investigate this behavior, a set of green fluorescent markers were constructed that, once installed in the yeast genome, allowed for the measurement of the sugar sensing and signaling network in each cell in real time through fluorescence measurements. It was found that when the cell sensed xylose, it resulted in the same signal as very low concentrations of glucose (i.e. almost starvation) did, and that the modification of previously known key genes for improved use of xylose changed the signal more towards that of regular amounts of glucose.

This thesis illustrates that the use of different forms of Large Data biology allows investigations of the microbial cell in ways that would not be possible or time-wise reasonable with traditional microbial methods. It also shows that the sheer volume of data these approaches generate quickly become a needle-in-the-haystack challenge, where finding the relevant data in the large ocean that is the cellular omes is only possible when molecular biology is combined with computational approaches.

List of papers

This thesis is based on following research papers, which will be referred to by their roman numerals. The papers are found at the end of the thesis.

I. **Mapping the diversity of microbial lignin catabolism: experiences from the eLignin Database**
Brink, D.P., Ravi, K., Lidén, G. & Gorwa-Grauslund, M. F. (2019) *Applied Microbiology and Biotechnology*, 103(10), 3979-4002

II. **Physiological characterization and sequence analysis of a syringate-consuming Actinobacterium**
Ravi, K., García-Hidalgo, J., Brink, D.P., Skywell, M., Gorwa-Grauslund, M.F. & Lidén, G F. (2019) *Bioresource Technology*, 285(1), 121327

III. **Bacterial isolate genome annotation as a driver for improved microbial cell factories: *calA* from *Pseudomonas putida* encodes a vanillin reductase**
García-Hidalgo, J., Brink, D.P., Ravi, K., Paul, C. J., Lidén, G. & Gorwa-Grauslund, M. F. (2019) *Submitted*

IV. **Cell periphery-related proteins as major genomic targets behind the adaptive evolution of an industrial *Saccharomyces cerevisiae* strain to combined heat and hydrolysate stress**
Wallace-Salinas, V., Brink, D.P., Ahrén, D. & Gorwa-Grauslund, M. F. (2015) *BMC genomics*, 16(1), 514

V. **Real-time monitoring of the sugar sensing in *Saccharomyces cerevisiae* indicates endogenous mechanisms for xylose signalling**
Brink, D.P., Borgström, C., Tueros, F.G. & Gorwa-Grauslund, M.F. (2016) *Microbial Cell Factoriess*, 15(1), 183

VI. **Assessing the effect of D-xylose on the sugar signaling pathways of *Saccharomyces cerevisiae* in strains engineered for xylose transport and assimilation**
Osiro, K.O., Brink, D.P., Borgström, C., Wasserstrom, L., Carlquist, M. & Gorwa-Grauslund, M. F. (2018) *FEMS Yeast Research*, 18(1), fox096

VII. **Exploring the xylose paradox in *Saccharomyces cerevisiae* through *in vivo* sugar signalomics of targeted deletants**
Osiro, K.O., Borgström, C., Brink, D.P., Fjölnisdóttir, B.L. & Gorwa-Grauslund, M. F. (2019) *Microbial Cell Factories*, 18(1), 88

I have also contributed to the following review, which is not included in the thesis:

R1. Biological valorization of low molecular weight lignin.
Abdelaziz, O.Y., Brink, D.P., Prothmann, J., Ravi, K. Sun, M., García-Hidalgo, J., Sandahl, M, Hulteberg, C.P., Turner, C., Lidén, G. & Gorwa-Grauslund, M.F. (2016) *Biotechnology Advances*, 34(8), 1318-1346

My contributions to the papers

I. I designed the study from an initial idea of Marie Gorwa-Grauslund, designed and wrote the MySQL database and the web interface (HTML/php), performed the data mining and curated the data. For the paper, I performed the literature review and wrote the manuscript.

II. I designed and performed the in-house bioinformatics and phylogeny analysis and handled the final genome annotation. Together with Krithika Ravi, I analyzed the genome annotation and made the pathway reconstruction.

III. I designed and performed the bioinformatics setup (assembly pipeline, annotation and comparative genomics) and data analysis, and drafted the initial manuscript.

IV. I designed and performed the in-house bioinformatics as well as the viability- and cell wall lysis experiments. I wrote the manuscript together with Valeria Wallace-Salinas

V. I participated in the design of the study, constructed the strains and drafted the initial manuscript. Together with Felipe Tueros I performed the flow cytometry analyses and enzymatic assays, and, together with Celina Borgström, did the molecular biology experiments, wrote the custom scripts and finalized the manuscript.

VI. I did the molecular biology work related to the mutated transporter, constructed eight of the strains and performed the flow cytometry bioinformatics. I wrote the manuscript based on a draft from Karen Ofuji Osiro.

VII. I participated in the design of the study and data analysis, performed the HPLC analysis and wrote the manuscript from a draft by Karen Ofuji Osiro.

Abbreviations

ALE	Adaptive Laboratory Evolution
BLAST	Basic Local Alignment Search Tool
BWA	Burrows-Wheeler Alignment
cAMP	Cyclic adenosine monophosphate
CNV	Copy Number Variations
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
ddNTPs	Dideoxynucleosidetriphosphate
dNTPs	Deoxynucleosidetriphosphate
ER	Ethanol Red (*S. cerevisiae* strain)
FCM	Flow Cytometry
FBA	Flux Balance Analysis
FP	Fluorescent protein
GEM	Genome scale model
GFP	Green Fluorescent Protein
GO	Gene Ontology
HTS	High Throughput Sequencing
INDEL	Insertion and/or Deletion
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAPK	Mitogen-Activated Protein Kinase
MPS	Massive Parallel Sequencing
MS	Mass Spectrometry
NCBI	National Center for Biotechnology Information (US)
NGS	Next Generation Sequencing
OLC	Overlap-Layout-Consensus
ORF	Open Reading Frame
PKA	Protein Kinase A
PTM	Post Translational Modification
QC	Quality Control
ROS	Reactive Oxygen Species
SAM	Sequence Alignment Map
SNP	Single Nucleotide Polymorphism
SQL	Structured Query Language
TGS	Third Generation Sequencing
TOR	Target of Rapamycin
WGS	Whole-genome Sequencing
XI	Xylose isomerase
XR/XDH	Xylose reductase/xylitol dehydrogenase

Table of contents

Preface

There are few buzzwords that describe our computerized, early 21st century world better than the concept of Big Data. The idea that it is possible to measure massive amounts of data points and run it through suitable computer algorithms in order to reveal connections and predictions that were not possible in a "small data world" has infused our society and our behaviour, and is currently a key mechanism in everything from social media to online shopping to science. Big datasets, especially the ones generated in biology, are often complex, messy and noisy – just like the world it tries to describe.

While I have focused the lion's share of the last five years or so on the research that has resulted in this doctoral thesis, my scientific interests has co-inhabited my mind with my long-standing love of art and creativity, such as writing, reading, drawing, designing. I am particularly interested in the interplay of science, literature and art, and their boundaries. To me, science and the arts are two means to the same end: to explore and understand the world that we live in. A 1000-page contemporary novel is also a form of Big Data, in its own way.

These ideas have undeniably coloured this thesis Most notably, I have chosen to preface each of the chapters of this thesis with excerpts from poetry, prose and philosophy that I believe resonate with the content of each section. It is common to see scientific ideas and methods applied to art, but possibly less common in the other direction. It may well be that this approach only serves to make the message of this thesis more messy. Which perhaps makes it not that dissimilar to Big Data?

Sometimes the answers you seek lie between the lines of the dataset. Sometimes the data fails to capture the answer at all. Sometimes Big Data is too Small to answer the question.

September 25th, 2019
Lund, Sweden
Daniel Brink

The world is everything that is the case.
[Die Welt ist alles, was der Fall ist.]

LUDVIG WITTGENSTEIN
The first statement of
Tractatus Logico-Philosophicus (1922)

Chapter 1

Introduction

Modern biology is a data-intensive science. In some regards, this is not a recent phenomenon, as certain sub-fields such as taxonomy and biodiversity, have a long history of reliance on large datasets (Kelling et al., 2009; Leonelli, 2014). Nevertheless, the advent of high-throughput technologies for system-wide assessment of the molecular biology of the cell (such as whole-genome sequencing and liquid chromatography-mass spectrometry) has rapidly changed the stage towards a more computationally demanding biology that needs to handle Big Data as much as it needs to handle biological samples.

Big Data science can in short be said to consist of the capture, curation and analysis of large datasets (Callebaut, 2012), and is often characterized with five V's: volume, velocity, variety, veracity and value (Gudivada et al., 2015; Herschel and Miori, 2017). It has been proposed that Big Data is the *fourth paradigm* in science, with empiricism, theory and computation being the previous three (Bell et al., 2009). However, the concept of Big Data is not stringently defined, and what levels of data quantity, complexity and technology that are needed for a dataset to be considered Big Data may vary considerably between users. In fact, a recent review was able to identify four different groups of definitions of Big Data in literature (De Mauro et al., 2016), and therefore, given how popular as the concept currently is, Big Data will have different meaning depending on the context. This also leads to complications regarding when a dataset can claim to be Big Data (Boyd and Crawford, 2012): is the raw data from the sequencing of the genome of a microbe complex enough to fit the Big Data concept, or does that dataset need to be combined with one or more equally complex datasets (e.g. from transcriptome and proteome studies) before the term even can be considered? Furthermore, Big Data is currently a strong buzzword in many sciences, including biology (Dolinski and Troyanskaya, 2015), and, like other buzzwords, thus tends to be overused. For these reasons, this thesis will instead use *Large Data* in order to avoid getting entangled in the discourse on the semantics of Big Data.

1.1 All cellular layers generate Large Data

The complexity of biology in general, and molecular and cellular biology in particular, makes it so that every attempt of a system wide screening will inevitably lead to generation of Large Data. From a molecular point-of-view, the cell is normally divided into sequential cellular layers according to Crick's theory of the *Central Dogma* (Crick, 1970): the genome (DNA), the transcriptome (mRNA) and the proteome (proteins). In extension, the metabolome (metabolites) is often also considered here despite not being part of Crick's original proposal (Prohaska and Stadler, 2011), see Figure 1. The *-ome* suffix is Latin for "mass" or "many", and *omics* is accordingly defined as the study of a whole ome (e.g. genomics, transcriptomics); due to the nature of the omes, an omics experiment will intrinsically result in a *mass of measurements* per sample (Lay Jr et al., 2006), i.e. Large Data. Omics is sometimes also referred to as *global analysis* (Nielsen and Jewett, 2008), again illustrating their system-wide scope. These methodologies are in fact so closely related to their dataset size that omics data often is seen as the quintessential biological Large Data (Leonelli, 2014). The complexity and temporal resolution increases with each sequential central ome (Figure 1): with the genome being rather stable over time (in terms of e.g. half-life and mutation rate) and the transcriptome, proteome and metabolome being in flux (Lay Jr et al., 2006).

The ome concept has proven to be very useful for describing biological function. Since the word genome was first proposed in 1920 by Hans Winkler (Winkler, 1920)[1], many additional omes outside of the Central Dogma have been defined, from intracellular layers such as the lipidome, epigenome and signalome (the signalling networks of the cell), to extracellular layers such as the secretome, microbiome (e.g. gut flora) and bibliome (the cumulative literature of a scientific discipline) (Grivell, 2002; Prohaska and Stadler, 2011; Topol, 2014), to name a few. In terms of frequency, the three omes of the Central Dogma (genome, transcriptome, proteome) are much more commonly used in literature than the subsequent neologisms, though (Prohaska and Stadler, 2011). The etymology of omics seems to have its root in 1986 when Tom Roderick came up with *Genomics* as the name for the eponymous journal-to-be, with proteomics following suit first in 1995 (Yadav, 2007).

As illustrated in Figure 1, the present thesis work combined methods traditionally regarded as high-throughput (e.g. whole-genome sequencing) with alternative ome assessments, such as single-cell biosensors, and database construction.

[1] It can be noted that a few biological concepts ending in -ome predate genome: e.g. *biome*, *rhizome*, *phyllome*, and that words like these may have been the inspiration for Winkler's proposal (Lederberg and McCray, 2001).

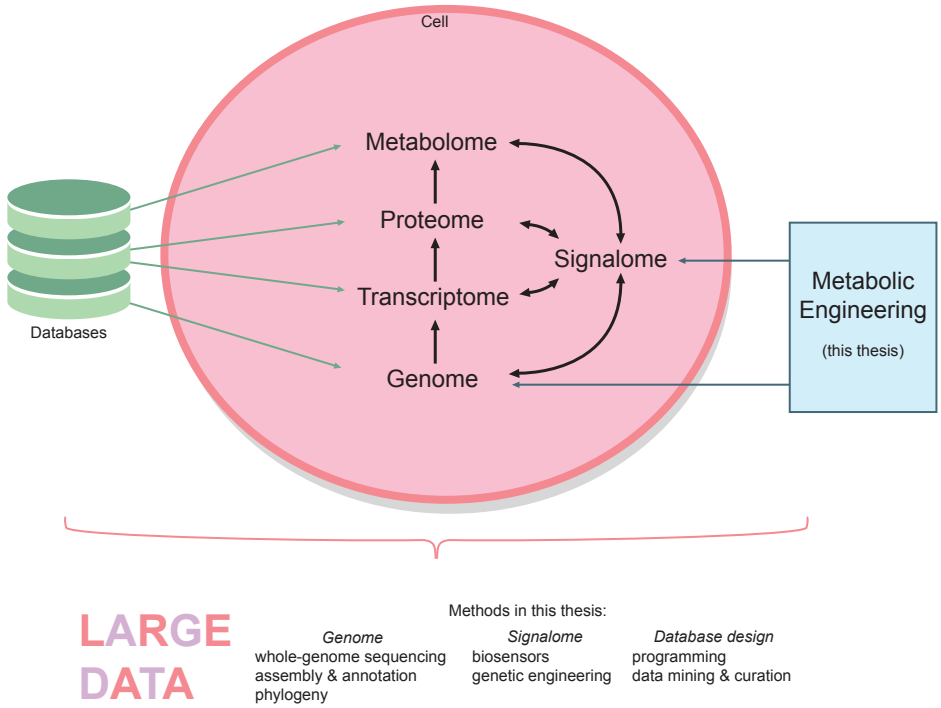


Figure 1: Schematic overview of the main cellular layers of the central dogma (genome, transciptome, proteome, metabolome) and the signalome (all signalling networks in the cell), all of which generate large data. The bottom half illustrates the different methodologies that were used in the thesis work to assess the genome and signalome layer, and how a database was constructed to handle large bibliomes.

1.2 Large Data will always have system boundaries

One of the biggest strengths of Large Data is that it can be used to find new correlations and insights that are not possible or visible in a "small data" world, with a famous example being how Google could predict the spread of the annual flu based on peoples' search queries (Ginsberg et al., 2009). However, every dataset has constraints to what it can predict, which are intrinsically linked to how the data was collected.

A central concern of data-intensive biology is to be able to draw biologically and physiologically relevant conclusions from patterns founds in large datasets (Li and Chen, 2014). For instance, sequencing a genome of an evolved microbe with a novel phenotype will give valuable information of the changes that have occurred in its genetic make-up, but it is not necessarily possible to correlate *which* change in genotype that results in the change in phenotype. Unlike the Google example above, the identification of the underlying causalities of a correlation is much more important in biology, since it is a discipline concerned with understanding *why* something hap-

pens (Mayer-Schönberger and Cukier, 2013). Therefore, when working with Large Data biology and cellular networks (in the present work: metabolic and signalling networks) we have to consider the system boundaries of our data collection methodologies in order to make biologically relevant claims – something that can be easily forgotten among the tempting possibilities promised by the hype surrounding Large Data (Boyd and Crawford, 2012), e.g. the belief that any scientific problem can be solved if a huge enough dataset can be collected and analysed.

To further emphasise this, the thesis is framed by two quotations from Wittgenstein's *Tractatus Logico-Philosophicus*: "*The world is everything that is the case*" and "*Whereof one cannot speak, thereof one must be silent*" (Wittgenstein, 1922). My interpretation of these quotes is that they represent the system boundary of the world – or the world as humans *perceive* it. Likewise, a Large Data biology experiment in itself is *everything that is the case*: it is not possible to draw either systemic or mechanistic conclusions from the assessment of a single of a few omes measured at a limited set of environmental conditions; to that end, better spatio-temporal resolution will be needed. Therefore, it is important to see conclusions from *in silico* biological Large Data experiments as hypotheses until they are verified experimentally, and the Large Data experiments themselves as powerful hypothesis generators.

1.3 Scope of the thesis

As the title implies, the scope of this thesis is to improve the understanding and engineering of microbial cell factories by the means of different data-intensive methodologies. Nevertheless, the sheer width of that statement calls for some system boundaries of its own. As illustrated in Figure 1, the present work will focus on three topics within Large Data microbiology: data- and bibliome handling and its implications (Chapter 2), the genome (Chapter 3), and the signalome (Chapter 4). This will be bookended by a reflection on how the present thesis work relate to and strive to increase the knowledge of said topics (Chapter 5) and an outlook on their future prospects (Chapter 6). Chapters 1-2 will discuss on the current state of large data biology and its benefits and drawbacks, whereas Chapter 3 and 4 will go into the details of the works that are presented in the respective papers.

Being a thesis in Applied Microbiology, all work was made with societal application and impact in mind; in this case within the context of microbial lignocellulose valorisation. The end goal of this field – to which the current work contributes – is construction of improved microbial cell factories for sustainable production of value-added compounds from renewable feedstocks. With the mind-set that Large Data biology is foremost a hypothesis-generator, the present work will demonstrate the benefit of combining *in silico*-approaches with physiological and molecular characterizations.

The *bibliome* studies are represented by **Paper I**, which regards the construction of an online database that indexes the bibliome of microbial catabolism of lignin and lignin-related aromatic compounds. The genome studies are presented in **Papers II-IV**, and addresses different aspects of genome assembly, annotation and detection of mutations, with examples from both bacteria and yeast. Finally, the *signalome* studies are covered by **Papers V-VII**, and demonstrate the development and validation of a panel of *in vivo* single-cell biosensors for real-time monitoring of the sugar signalling networks in baker's yeast *Saccharomyces cerevisiae*. Furthermore, the genomics and signalomics chapters will each conclude with a case study on how these cellular layers were applied for improved microbial utilization of lignocellulosic feedstocks: Chapter 3.4. discusses how comparative genomics was used to correlate the changes in phenotype to changes in genotype in an evolved yeast strain with improved tolerance to the combined inhibition of lignocellulose hydrolysate and elevated temperature; Chapter 4.3. discusses the paradoxical fermentation behaviour of xylose (one of the most abundant sugars in lignocellulose) in *S. cerevisiae* engineered with exogenous xylose catabolism.

apricot trees exist, apricot trees exist

bracken exists; and blackberries, blackberries;
bromine exists; and hydrogen, hydrogen

cicadas exist; chicory, chromium,
citrus trees; cicadas exist;
cicadas, cedars, cypresses, cerebellum

doves exist, dreamers, and dolls;
killers exist, and doves, and doves;
haze, dioxin, and days; days
exist, days and death; and poems
exist; poems, days, death

INGER CHRISTENSEN
Excerpt from *alfabet* (1981)
(an example of alphabetical indexing as a system boundary in poetry)

Chapter 2

How to manage Large Data?

Everything is in a database nowadays. From your email login credentials to your tax return, most information is stored in an electronic database to be accessed online at your convenience. Though they may seem, databases are not by far a new thing, neither in their analogue form – e.g. library index cards, parish registers or national censuses – nor in their digital format – database management systems were invented around the 1960s; (Haigh, 2009)). Nevertheless, with the last decade's developments in Internet connectivity, wireless mobile devices and social media, it is probably safe to assume that there never before have been so many databases that we contact on a daily basis. Digital databases are indeed one of the best ways to organize Large Data, since it not only allows for archiving and indexing (just like an analogue database) but also allows for a whole new level of data connectivity, pattern recognition and synthesis through *in silico* processing. However, as will be discussed throughout the thesis, most biological large datasets are noisy and will require several steps of processing before they can be uploaded to a database.

2.1 Large Data and *in silico*-demanding biology

2.1.1 An explosion of biological data

The rapid developments in computer science and information technology have led to a previously unseen data explosion both in society and in science. In biology, the hitherto biggest data explosion[2] happened in the mid-2000s as a result of the advent of a number of new high-throughput omics methods, especially for nucleotide sequencing (Leonelli, 2014). As the volumes and types of Large Data increases over time with new developments in technology, so does our views on what is large: there was a time where expression data of a single microarray was considered large, which compared

[2] Some disciplines within molecular biology have had data explosions earlier than others due to specific technological developments in their field: e.g. protein crystallography around 1990 (Sussman et al., 1998)

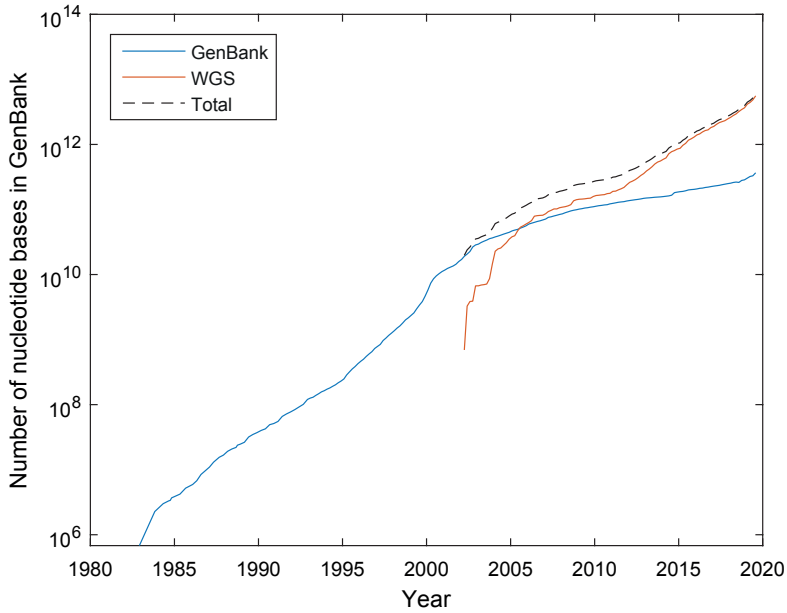


Figure 2: Cumulative number of nucleotide bases uploaded to NCBI GenBank from its launch in 1982 to the latest release in August 2019. A distinction is made between WGS (red), which are the bases in the whole-genome shotgun (WGS) subsection of GenBank introduced in 2002, and GenBank (blue) which does not include the WGS projects. Adapted from publicly available data from NCBI: https://www.ncbi.nlm.nih.gov/genbank/statistics/.

to the throughput of present-day methods seem small in comparison (Dolinski and Troyanskaya, 2015).

The NCBI GenBank database is one of the oldest and largest publically available biological repositories (Benson et al., 2017). Thanks to their open statistics, this repository can be used as a good indicator of how molecular biology has grown since GenBank launch in 1982. Figure 2 illustrates the historical growth of their dataset in terms of number of stored nucleotide bases, which has been exponential since the launch and with a doubling time of approximately 18 months[3]. The whole-genome sequencing subset within GenBank (red line in Figure 2) is a good example of how new technological achievements further contribute to data explosion (further discussed in Chapter 3).

The technological advancements have opened many new possibilities for what can be studied at a reasonably cost and time (a democratization that enables also smaller

[3]See: https://www.ncbi.nlm.nih.gov/genbank/statistics/

labs to do Large Data biology), but the availability of data from published studies has also become an asset in itself. There is an intrinsic value to many Large biological datasets, as its sheer size and molecular complexity makes it possible to actually conduct whole studies based on previously published data without having to generate new data: so called data *re-use* (Marx, 2013; Leonelli, 2014). A few examples include: genome comparisons (Borneman et al., 2011; Vernikos et al., 2015), expression studies (Rung and Brazma, 2013) and computational models of the cell (genome-scale models, GEMs; Price et al. (2004)) – not to mention how database-driven tools such as homology searches by BLAST (Altschul et al., 1990) have enabled and facilitated innumerable amounts of biochemical and metabolic engineering studies. Indeed, there are papers that are cited for their data and not so much for their research findings (Dolinski and Troyanskaya, 2015), just like some papers are primarily cited for their medium recipes (e.g. Verduyn et al. (1992)). Data re-use is however not a trivial problem, since the complex spatio-temporal nature of biological data (what condition, what timespan etc.) complicate re-application and direct comparison. Re-use also introduces new ethical challenges, especially related to authorship (Duke and Porter, 2013) which calls for open data standards and licences (Molloy, 2011).

The benefit of being able to re-use data, perform meta-analysis or integrating multiple individually published datasets to a larger, more systemic analysis is at the end of the day dependent on what raw data is available and the quality of its annotations (Rung and Brazma, 2013). An recent opinion piece phrased the issue thusly: "*Too much published data or too little published data?*"(França and Monserrat, 2019), implying both the issue of handling the large volumes of processed data, and the comparably low amounts of available raw data. This is further complicated by how routines around data sharing differ between disciplines, individual labs and journals. For instance, most journals require raw data and genome assemblies from whole-genome sequencing projects to be uploaded to the NCBI/EBI/DDBJ database consortium prior to submission. Other high-throughput methodologies, such as flow cytometry, do not have established routines for (raw) data sharing, although initiatives have emerged (Spidlen et al., 2012).

While biological data has become simple and cheap to collect, knowledge of data management and -analysis seems to be lagging behind (Peng, 2015). It has for instance been argued that the current "reproducibility crisis" (the fact that very few published studies can be repeated by scientists in other labs) in science (Peng, 2015) is a result of the overwhelming data volumes and of overconfidence in the evidence-power of statistical methods (in particular the commonly used $p<0.05$ threshold in statistical hypothesis testing) (Goodman, 2016; Wasserstein and Lazar, 2016; França and Monserrat, 2019).

There is currently in biology a dichotomy of *data-driven* research and *theory-driven* research (Callebaut, 2012; O'Malley and Soyer, 2012; Dolinski and Troyanskaya,

2015), where, in very general terms, the former uses analyses and modelling of large datasets from cellular phenomena to come up with research ideas after the fact (*a posteriori*), whereas the latter uses *a priori* knowledge from e.g. literature to design experiments (which in themselves can be data-intensive). Nevertheless, it is imperative to remember that

$$data \neq knowledge^{4}$$

and that only thorough experimental design based on previous knowledge and systematic data analysis with suitably large sample sets that are followed by experimental verification can turn large datasets into knowledge. The strength of Large Data is to find correlations, not causalities – but can as such be used as a guidance towards likely causes (Mayer-Schönberger and Cukier, 2013), i.e. new hypotheses and experiments. The present thesis will argue for a theory-driven research complemented by large data-approaches, with the hypothesis generator-aspect of the large data methods operating somewhere in the middle of the two.

2.1.2 Large Data requires large undertakings

In its raw, non-curated form, Large biological Data is often incomprehensible due to its large volumes and varying formats. Managing and analysing the data is therefore not a task suitable to do by hand due the sheer volume and the risk of introducing human errors. This has led to an increased need for programming and heavy-duty bioinformatics in molecular biology (discussed in Section 2.1.3) and researchers well versed in both computer science and biology. In fact, it has become common in data-intensive biology to allocate time on high-performance computing centres (supercomputers) to run more computationally heavy algorithms and pipelines, many of which require programming know-how since there are often no graphical interfaces (Yin et al., 2017). Not only are computations needed, but also methods and infrastructure for dissemination (e.g. public databases). Due to their indispensability in modern molecular biology, Section 2.2 will be dedicated to these types of databases. Like any other methodologies, Large Data biology and their databases come with its benefits and challenges, a few of them being listed in Table 1.

Large Data science is a relatively new field, and some of its potential and accuracy are yet to be confirmed in the long-run. The famous example of how Google Flu Trends could predict the seasonal flu, has, while initially rather accurate, been shown to overestimate the spread of the seasonal flu by a factor two in later years (Lazer et al., 2014). Likewise, the sequencing of the human genome has yet to result in the long-

[4] See also Deming (2018):"...information, no matter how complete and speedy, is not knowledge. Knowledge has temporal speed. Knowledge comes from theory."

standing ambition of a precision medicine tailored towards the individual patient (Coveney et al., 2016). Contrary to what one might first think, these outcomes are probably not caused by the complexity of large data volumes; in fact, the challenge of Large Data biology is that the information volumes contained in contemporary large biological datasets, are *tiny* in comparison to the information complexity of biological systems (Coveney et al., 2016).

This insight aside, the size of e.g. an omics dataset is still massive and difficult to overview. The human tendency of finding patterns where there are none and other cognitive biases such as *confirmation bias* (the tendency to look for results that fits with preconceived expectations) are challenging in science in general (Boyd and Crawford, 2012; Munafò et al., 2017), and in Large Data in particular. The sheer vastness and intrinsic random appearance of Large Data make it more vulnerable to biased and often unconscious analysis. Hypothesis-driven Large Data biology has been suggested as a countermeasure (Lay Jr et al., 2006).

2.1.3 Programming and bioinformatics for microbiologists

Once a Large Data experiment has been suitable designed and the data has been collected, the central challenge of Large Data biology is *in silico* handling. This has thoroughly ushered in a need for biologists to have some level of proficiency in computational biology and programming.

The majority of the state-of-the-art, free-for-academic-use bioinformatics algorithms are implemented in so-called command-line interfaces (text-only terminals where commands are executed by typing, c.f. IBM DOS or *cmd* in Windows), and while this significantly shortens the development time for a new algorithms (no need to develop graphical interfaces), this implementation demands a lot of computer proficiency from the user (Kumar and Dudley, 2007). These command-line software are almost always implemented for use with Unix-systems (e.g. Linux, Mac OS), since this is an environment that is well suited for handling large files (omics data, for instance, is normally gigabytes in size) and has a long tradition of powerful command-line commands for file-manipulation (Bradnam and Korf, 2012). Commercial software tend to have graphical interfaces, but do seldom provide their algorithms (company secrets), leading to a less transparent bioinformatics work-flow. An in-between solution that has proven quite successful is the Galaxy framework (https://galaxyproject.org/; Goecks et al. (2010)) where many of the above-mentioned command-line tools have been implemented in a graphical interface to facilitate for users with less experience in programming. Although the merit of graphical interfaces is clear, as it will decrease the gap between the developers (often bioinformaticians and statisticians) and the end-user scientists (Kumar and Dudley, 2007), a programming knowledge will open many new possibilities for data analysis as custom scripts are often needed to do specific operations, and to combine multiple pre-existing software in an automated

Table 1: Examples of benefits and challenges of biological Large Data experiments and their corresponding databases. Note that the table does not strive to be exhaustive.

Benefits	Challenges
Large Data experiments in biology	
• Enables ome-wide assessments of the cell and can thus give holistic/systems views on cellular phenomena • Can foster discovery of new correlations and insights; generates hypotheses that can be further investigated with complementary experiments • Published datasets can be large enough to be re-used for new research, or as a driver for new hypotheses (Marx, 2013; Peters et al., 2014) • Integration of multi-omics data sets can be used to create *in silico* models of the cell (Heath and Kavraki, 2009) • The high complexity of the datasets may encourace scientists to embrace the complexity of the real world, instead of focusing on isolated observations (Leonelli, 2014) • Current bioinformatics algorithms are mature and established, and improvements follow the technical developments of the field	• Data volume and hetrogenous nature makes processing, analysis and interpretation non-trivial and time-consuming • Typically computationally heavy (due to the above); requires dedicated infrastructures and trained users to process and disseminate data (Yin et al., 2017) • Noisy data (low signal-to-noise ratio); quality pre-processing is therefore needed (De Keersmaecker et al., 2006; Del Fabbro et al., 2013) • Biological large data needs to be annotated to make sense, often using complementary experiments (Prohaska and Stadler, 2011) • Large data volumes are unavoidably prone to inexactitude, compared to "Small Data" (Mayer-Schönberger and Cukier, 2013) • Steep learning-curve for running the algorithms; results may be difficult to reproduce with alternative algorithms (Manzoni et al., 2016)
Biological databases	
• Organization of large data and bibliomes improve data accessibility • Databases are ongoing projects and can in contrast to published litterature reviews grow and be improved over time • Database management systems offers powerful relational tools to connect data; interconnections between databases further simplifies information discovery • Can facilitate data standardization by having quality and format requirements prior to upload • Data sharing increases transparency and collaboration in science	• Curation is needed and is a bottleneck (manual labour intensive) (Howe et al., 2008) • The maturity of the chosen reference databases directly impacts the quality of the bioinformatics analyses (Manzoni et al., 2016) • Heterogeneous nomenclatures and data collection approaches within different disciplines in biology complicates meta-data curation (Leonelli, 2014; Manzoni et al., 2016) • Large amounts of existing data are unavailable (e.g. pre-digital studies and company owned data) (Leonelli, 2014) • Needs continuous maintenance and funding (Bastow and Leonelli, 2010)

work-flow, a so called *pipeline* (Leipzig, 2017). This is complicated by the fact that the output format of a given algorithm is not necessarily compatible with the input format of the next programme in the pipeline (Marx, 2013), meaning that time has to be spent on developing custom scripts for converting between formats within the pipeline.

A few programming languages keep getting recommended for use with biological data (Table 2). Although there are plenty of debate over which language is the best – in a similar manner to how people debate which car or which camera is the best – there is no such thing as an universally superior language; instead they are good at different tasks (Carey and Papin, 2018). Perl and Python do however have a strong tradition within the bioinformatics community, and there is an abundance of documentation, tutorials and previously answered questions available for how to use these languages in general and in biology (Bradnam and Korf, 2012). Both languages are ”general purpose languages”, meaning that they are versatile enough for many different types of applications, and they both handle text well (which is exactly what DNA data is: a string of text). Perl and Python are so called *interpreted languages* (as opposed to *compiled languages*, e.g. Java and C++) which means that there is little need to consider implementation aspects such as CPU and memory allocation, with the drawback that they are slower (Bradnam and Korf, 2012). Memory-intensive algorithms like genome assembly are thus commonly written in compiled languages. The terminology for a program created with an interpreted language is script, and hence scripting is often used as a synonym to programming.

2.2 The importance of biological databases

As has been alluded to throughout this chapter, databases are a necessary infrastructure for handling, storing and sharing large biological data and is as such an important driver for biological discovery (Zhulin, 2015). A database can be defined as a collection of persistent (non-transient) data (Date, 2004) and hence any structured collection of data, like a library catalogue or a set of spreadsheets can be called a database. In the current context the word *database* will be used to imply a computerized *database system*, i.e. the hardware and software that connects the data to the user by structuring it in a systematic way. Benefits of database systems include compactness (no printed papers and filing cabinets), access speed, data sharing, reduced redundancy and inconsistency (e.g. through standardization), data integrity (easy to update data and correct errors) and data independence (can be accessed computationally from different angles and needs) (Date, 2004).

Table 2: List of programming languages and environments that are commonly applied in Large Data biology (and used in the present thesis work). Adapted from Carey and Papin (2018).

Environment	Features/comments
	Scripting/programming languages
bash	Very common Unix shell/command-line interpreter; needed to navigate and execute commands in the Unix-terminal; versatile scripting language, powerful for file manipulation. Essential for work in Unix.
Perl	General-purpose scripting language, good for parsing strings (i.e. DNA sequences, gene annotations, etc.); syntax can be a bit obtuse to read; waning community; in part succeeded by Python, but many bioinformatics script is and have been implemented in Perl, meaning that the language is still very relevant. Dedicated bioinformatics plugins available (BioPerl)
Python	General-purpose scripting language; good for string manipulation; can be used as a scripting languages for webpages; strong community (currently very popular), dedicated plugins for scientific computing (e.g. numpy, matlibplot) and bioinformatics available (BioPython)
	Maths and statistics environments
Matlab	Commercial, but all algorithms are open, large amounts of community deposited scripts and resources are available
R	Open source, community driven development with many bioinformatics plugins ("packages") available; popular alternative to Matlab, especially since there are no licence costs
	Database management
SQL	Relational database language; ISO standard; many database management systems that use SQL are available (e.g. MySQL, a popular open source software). Good for management of large data; the relational model allows for powerful linking of data, and pattern recognitions in datasets

2.2.1 A growing bibliome leads to a growing database demand

According to a recent bibliometric study, the global scientific output has grown exponentially between 1980 and 2012 at a growth rate of circa 3% per year (Bornmann and Mutz, 2015). In addition, the coming of the Internet age has made scientific literature more accessible for reading, assessing and mining. Although large data first need to be structured in the files of individual researchers/labs in order to be analysed in the first place, for data to be shareable and useful, biological databases need to index the data in ways that allow its users to access it in a comprehensible and user-friendly way while annotating each data entry with its meta-data ("information

about information", e.g. data provenance) and with related data that the user may want to consider (e.g. linking the known data on the protein to the gene that it is expressed by). It should also be kept in mind that the bibliome is not only a vessel for scientific data: the bibliome in itself can be analysed for trends and for forecasting of innovation and research directions (Which labs? What type of science? How many citations? How "hot" is a topic?)(Daim et al., 2006; Watatani et al., 2013).

Databases can be classified as *primary* databases, where the data is curated from literature or from direct data submissions from scientists, and as *secondary* (or *meta-*) databases that integrate data from multiple databases into a single platform (Helmy et al., 2016). *Curation* is an essential step towards data sharing, as it regulates how users can find and access the data (Howe et al., 2008), but is a major bottleneck in database development and maintenance, as it is very manual labour intensive (especially for primary databases). Although automation is possible to high degrees and is becoming more advanced (Sehgal et al., 2011), the nature of biological data and difference in tradition and approach between different biological disciplines makes it difficult to implement sufficient automatic curation (Leonelli, 2014). *Minimum Information* initiatives such as the Minimum Information about a Sequencing Experiment (MINSEQE) (Rung and Brazma, 2013) and the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) (Bustin et al., 2009), facilitate meta-data standardization and thus potentially data re-use, but has to be enforced by e.g. journals to reach a higher level of implementation.

Using Leonelli's model of Large Data journeys in biology, data to be deposited in a database goes through three curation-depending stages: *de-contextualisation*, *re-contextualisation* and *re-use* (Leonelli, 2014). De-contextualisation is the process of extracting data from their original context (e.g. a scientific publication) and formatting it to the standards of the database; re-contextualization is the process where the data is becoming available for utilization in new research contexts, which requires good quality meta-data annotations of the data provenance (e.g. experimental procedures, measurements or simulations etc.); finally, re-use is when a dataset has passed through the previous two steps and can be applied to discover new correlations (Leonelli, 2010, 2014). However, most data in biological databases do not reach the re-use phase due to various reasons, e.g. insufficient levels of curation, meta-data and standardization in the technologies used to collect the data (Leonelli, 2014).

Data curation is central for de-contextualization and for the annotation part of re-contextualization. In biology especially, this is complicated by the high degrees of non-standardized nomenclature and naming conventions (e.g. how gene name formats differ between model organisms) and changing classifications over time (e.g. in taxonomy). A countermeasure to this is the implementation of *ontologies*, a shared model or vocabulary for a domain of discourse (Munir and Anjum, 2018), with the seminal one in biology being Gene Ontology (GO) (Ashburner et al., 2000). In

primary databases, the curators will need to extract the meta-data themselves from experimental descriptions, which explains the high manual labour, and underlines that curators need to not only be versed in data science, but also understand the underlying biology. Annotations can also need to be corrected over time when new data becomes available, e.g. functions of predicted genes (cf. *calA* in **Paper III**).

The biological database that was developed in the present thesis (**Paper I**) is a small-scale biological database on microbial lignin valorisation – a growing bibliome that has not been well indexed due to its many pre-digital publications. It was identified that the literature of biological lignin valorization consists of two cohorts: one focusing on the fundamental microbiology of the breakdown of lignin and its related aromatic compounds, with a legacy from at least the 1960s (Ornston and Stanier, 1966), and a second, more recent focused on applied lignin biovalorization that has gained a lot of popularity in the recent decade (Abejón et al., 2018). The vast nature of this bibliome, combined with the many taxonomical re-classifications that have occurred in this niche over more than half a century, and the lack of good pre-existing database functions for lignin-related microbiology makes this field challenging to overview. The eLignin Microbial Database (**Paper I**; www.elignindatabase.com) was therefore designed to facilitate the navigation of this bibliome by creating a searchable, self-contained small scale biological database for use for scientists within the microbial lignin community. Since a majority of the papers in this bibliome are pre-digital, their indexing in eLignin is sometimes their first inclusion in a database system, which means that their curation demanded extra amounts of manual labour.

It is often relatively easy to establish a biological database – e.g. as a part of a bigger research project – but quite difficult is to ensure funding for long-term maintenance (Bastow and Leonelli, 2010). The post-launch period of a database life cycle is therefore likely to be more challenging than the collection and curation of the initial dataset, as it will require continuous maintenance and updates; this a point-of-no-return where a choice has to be made to either "maintain, update or retire" the database (Helmy et al., 2016). In the case of the database discussed in **Paper I**, the publication of the article served as a way to preserve the state of the database in 2018/2019 and its meta-analysis in printed form, should the future of the database become uncertain.

2.2.2 Available biological databases

Given how many specialized databases there are and how new appear and some disappear over time, listing all available biological databases is a near-impossible task. One of the seminal publications on biological databases is the annual Database issue of Nucleic Acids Research that has published papers on biological databases (including human biology) since 1993, with the latest total count being 1613 databases (Rigden and Fernández, 2018). This does however only include databases that have been pub-

lished in this particular journal and within their inclusion criteria, meaning that the actual number is higher. For the sake of orientation, a few examples of some of the more common types of (micro)biological databases are presented in Table 3.

Table 3: A few categories and representative examples within the umbrella concept of biological databases. Partly adapted from Zhulin (2015).

Category	Representative examples	Reference
Genome data	International Nucleotide Sequence Database (GenBank, EMBL, DDBJ)	Cochrane et al. (2015)
	MGnify (EBI Metagenomics)	Mitchell et al. (2017)
Transcriptome data	NCBI GEO (Expression data)	Barrett et al. (2012)
	SILVA (small & large subunit rRNA)	Quast et al. (2012)
Proteome data	Uniprot	UniProt Consortium (2018)
	RCSB Protein Databank	Berman et al. (2000)
	Brenda	Jeske et al. (2018)
	STRING protein-protein associations	Szklarczyk et al. (2018)
Metabolic pathways	KEGG	Kanehisa et al. (2016)
	Metacyc	Caspi et al. (2013)
Signalling pathways	Quorumpeps	Wynendaele et al. (2012)
	MiST (Microbial Signal Transduction database)	Ulrich and Zhulin (2009)
Model organisms	Ecocyc (*E. coli*)	Keseler et al. (2016)
	Pseudomonas genome database	Winsor et al. (2010)
	Saccharomyces genome database	Cherry et al. (2011)
Transporters	TransportDB	Elbourne et al. (2016)
Ontology databases	Gene Ontology	Ashburner et al. (2000)
	ExPASy-Enzyme (enzyme classifications)	Bairoch (2000)
	Transporter classification	Saier Jr et al. (2015)
Bibliome	PubMed Central	Roberts (2001)
	The eLignin Microbial Database	**Paper I**

2.3 Large Data in Metabolic Engineering

It has been proposed that after the human genome project was completed in 2001, biology shifted into a *postgenomics* era where the link between gene and phenotype was no longer considered linear, but branched and multifaceted (Perbal, 2015), and the cell began to be considered not only as a collection of genes and proteins, but as a tightly regulated system that could only be understood when the cellular networks are considered as a whole (Kitano, 2002). In parallel with the developments of ome-level global analysis, genetic engineering also moved towards a more systemic world-view: metabolic engineering. Metabolic engineering has been described as the "*improvement of cellular activities by manipulation of enzymatic, transport and regulatory functions of the cell with the use of recombinant DNA technology*" (Bailey, 1991), and normally see the molecular cell factory as the end-goal (Nielsen and Jewett, 2008). So far we have discussed the philosophical implications of Large Data biology, what type of data it regards and how data has to be handled, stored and annotated. This final section of Chapter 2 will briefly comment on the changes large data has brought to in molecular biology in general, and in metabolic engineering in particular.

2.3.1 Towards a systemic understanding of the cell

With the advent of high throughput techniques came new incentives to integrate different datasets to better describe the cell. Thus, the systems biology discipline emerged, where multi-omics approaches were integrated with the molecular biology needed to understand the cell, the bioinformatics needed to handle the data, and the computer science and mathematics to construct *in silico* models of cellular functions (Heath and Kavraki, 2009). Whereas systems ideas in biology are not new (proposed already in the 1950s, albeit in a slightly different form; von Bertalanffy (1950)), the technological maturation of omics led to a breakthrough for systems biology in the early 2000s (Powell et al., 2007).

A core value of systems biology is holism ("*the whole is larger than the sum of its parts*"), which is in opposition to the traditional reductionist views on molecular biology ("*the whole can be understood by analysis of its parts*") (Fang and Casadevall, 2011). Two different movements have been identified within systems biology: the *localists* who are gene- and pathway-centric and reductionist in their approach, and the *globalists* that are network-centric and use holism (Huang, 2004; Mazzocchi, 2012)[5]. These approaches aside, it should not be interpreted as if molecular biology and physiology has been rendered obsolete by the systems approaches (Gatherer, 2010), since it is a pre-requisite.

[5]There are other characterizations of these two movements (reviewed in O'Malley and Dupré (2005)), but they all seem to agree on that that this dichotomy exists.

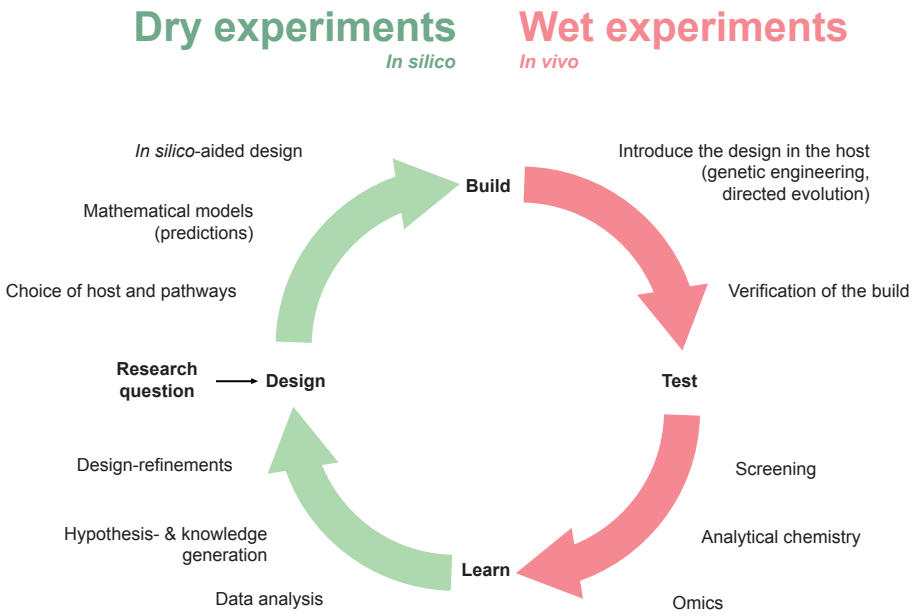


Figure 3: How "dry" and "wet" experiments come together in the iterative design-build-test-learn cycle of metabolic engineering and systems biology. Adapted from Kitano (2002); Petzold et al. (2015); Nielsen and Keasling (2016). Note how a complementary use of *in silico* and *in vivo* methods can generate hypotheses and, eventually, knowledge.

Common to many projects in both systems biology and metabolic engineering is the iterative work flow consisting of four phases: design, build, test, learn – with methods ranging from "wet" experiments to "dry" computer-aided analysis, modelling and design, see Figure 3. The technical challenges of systems biology is largely connected to the challenges of biological large data (Table 1). Notable examples include uneven and unstandardized data quality and need for specialized tools to measure intracellular events at high temporal resolution, preferably at a single-cell level so that population dynamics can be captured (Aderem, 2005).

2.3.2 Data-intensive drivers in systems metabolic engineering

While the scope and ambition of systems biology is grand – e.g. to reach comprehensive systems understanding of the cell that can be demonstrated as a functional *in silico* model of the cell (Powell et al., 2007) – not all systems approaches need to be extensive. For instance, data-intensive systems biology methodologies are often combined with metabolic engineering – sometimes referred to *systems metabolic engineering* – where large scale data are used to drive discoveries of new gene targets (Blazeck and Alper, 2010; Lee et al., 2012) and convey forward momentum to metabolic engi-

neering projects. Although algorithms and bioinformatics speed up the engineering work-flow, they require specialist knowledge and thus call for multi-disciplinary research teams.

Given its foundational role in the central dogma and the maturity of its techniques, the genome is the focal point of systems biology and metabolic engineering. Some genome-based methodologies in this discipline include: metabolic pathway reconstruction, genome scale models, flux balance analysis and reverse engineering of evolved strains. Pathway reconstruction is the process of identifying the nodes of a metabolic pathway (i.e. the enzymes) from assembled genomes and other biochemical data (Schuster et al., 2000; Pinzon et al., 2018). Depending on the size of the pathway and previous knowledge from other organisms (such as homologies and known reactions), reconstruction can be small undertaking that can be done manually using simple tools such as BLAST, or large projects requiring dedicated pipelines such as the KEGG database annotation server (Moriya et al., 2007). Pathway reconstruction was performed in **Papers II-III** in the present thesis work, where two so-called *funnelling pathways* for aromatic degradation in two bacterial species were proposed, based on their assembled and annotated genomes (cf. Chapter 3).

When pathway reconstructions of a given organism reach a critical, systemic level, they can be used to produce genome scale models (GEMs) that can be used to mathematically model the biological functions of that cell (Palsson, 2015). For a GEM to reach this level of sophistication, substantial fundamental knowledge is needed from the system (*knowledge-base*). At the core of a GEM is a matrix that contains the stoichiometry of every reaction of the reconstructed pathway. The large size of a cell-wide stoichiometric matrix makes the system unsolvable unless constraints (i.e. system boundaries) are put on the system, a method known as constraint-based modelling (Palsson, 2015). Luckily, the cell naturally operates under a number of constraints (e.g. environmental, physico-chemical, evolutionary and regulatory) (Covert et al., 2003) and many values that are mathematically possible are invalid in biological systems, such as negative or infinitesimally large concentrations. The models are typically evaluated using Flux Balance Analysis (FBA), where the fluxes in the system are estimated using constraint-based linear algebra (Pinzon et al., 2018). Historically, GEMs have been good at modelling well-known pathways (e.g. aerobic growth on glucose) to the extent that they can be verified by experimental data (Feist et al., 2007; Liao et al., 2011; Lopes and Rocha, 2017), but since all GEMs are reconstructions of the current knowledge of an organism, there will always be pathways that are less known or incomplete and thus result in inaccurate predictions (Orth and Palsson, 2010). GEMs have many potential applications in metabolic engineering, including analysis of deletions and gene up-/down-regulation, engineering target identification and pathway prediction (Kim et al., 2015; Palsson, 2015).

Another genome-centred methodology in metabolic engineering is to subject re-

combinant strains to Adaptive Laboratory Evolution (ALE) to attempt to generate new and improved phenotypes through prolonged exposure to selection pressures (Dragosits and Mattanovich, 2013). Evolution will likely result in mutations related to the new phenotype, but also in an amount of un-related mutations that can potentially be detrimental to the scope of the strain design (e.g. decreased fitness and altered morphology). It is therefore often of interest to identify the relevant mutations of the evolved strain and introduce them in the parent strain (so called *reverse engineering*) to prove that the putative mutations cause the novel phenotypes, and to decrease the burden of all the non-desired mutations (Oud et al., 2012; Dragosits and Mattanovich, 2013).Identification of key mutations correlated to the novel phenotypes is a non-trivial task, but can be approached by the use of comparative genomics and variant calling (cf. Chapter 3.4; **Paper IV**).

There are also Large Data approaches beyond the metabolic pathways and the genome. An example in the present work is network monitoring of the sugar signalome (**Paper V-VI**), where a panel of *in vivo* biosensors were constructed to monitor the cellular response in *S. cerevisiae* to sugars, using a single-cell flow cytometry approach. Commonly, the signals from 10 000 or 100 000 cells were collected, which presents a data management challenge that is very similar to that of omics. The findings of the signal network monitoring can further be applied to improve the response of a cell factory to a desired production, an example of which was reported in **Paper VIII**.

With the current chapter as a theoretical and philosophical springboard, the next two chapters will detail how genomics and signalomics can be used to increase the (systems) level understanding and apply that for building improved microbial cell factories.

The only thing that is constant is change

A common misquote of:

Everything changes and nothing stands still
[*πάντα χωρεῖ καὶ οὐδὲν μένει*]

HERACLITUS OF EPHESUS
(As quoted by Plato in *Cratylus*, 402a)

Chapter 3

A closer look at genomics

The importance of DNA sequencing for biology can hardly be exaggerated. In fact, most of the Large Data biology discussed in the previous chapter would not have existed without sequencing. Like how the genome is the foundational ome of the cell (Figure 1), genomics is a requisite for modern biology. Whereas pre-cloning genetics was concerned with observing phenotypes and looking for the responsible gene, modern genetics often reverse the process by altering genes and observing the phenotypes (Brenner, 2000). Genomics have over the years expanded into different specialized research areas with different levels of integration with data from higher omics, such as metagenomics (Handelsman, 2004), structural genomics (Grabowski et al., 2016), functional genomics (Werner, 2010) and epigenomics (Fazzari and Greally, 2004), to name a few. The present chapter will focus on the "original" aspect of genomics, namely that of whole-genome sequencing (WGS). It will in particular discuss the implications of the recent technological advancements in the field, the different options for assembly of the sequencing data (*reads*) and how WGS can be used to compare the genomes of multiple organisms (*comparative genomics*). Since the terminology of genomics can become rather complex, key concepts will be explained in footnotes, and in a glossary available in Appendix I.

3.1 Timeline of Whole-genome sequencing methods

Just as important as it is to know the provenance of data (cf. Section 2.2), is the provenance of the techniques used to generate it. There are currently three generations of DNA sequencing available, and interestingly enough, all of them are still in use, thanks to their specific strengths and cost requirements.

3.1.1 First generation sequencing

The establishing method for DNA sequencing was the chain termination method (or, more commonly: *Sanger sequencing*) that uses a DNA polymerase, deoxynucleotides

(dNTPs) and elongation-terminating labelled dideoxy-nucleotides (ddNTPs) to determine sequence composition (Sanger et al., 1977b). Incorporation of a ddNTPs inhibits the strand elongation, and the base-pair can be measured thanks to its label, so called *sequencing-by-synthesis* (Heather and Chain, 2016). The first sequenced genome, that of bacteriophage ΦX174 (5386 bp) (Sanger et al., 1977a), was obtained with the *plus and minus* method, which was the precursor to the chain termination method (Sanger and Coulson, 1975). However, Sanger sequencing of larger genomes is laborious, since it has low throughput and each read (raw nucleotide sequence) is at most 1kb in size (Heather and Chain, 2016). The landmark WGS projects that demonstrated the concept of WGS during the 1990s had to utilize variations of the *shotgun sequencing* approach, where the genome was fragmented, the fragments cloned into vectors, sequenced individually, and then assembled[6] *in silico* to longer contiguous sequences (*contigs*[7]) (Oliver et al., 1992; Fleischmann et al., 1995; Fraser et al., 1995; Lander et al., 2001; Venter et al., 2001). These projects notably required the combined efforts of multiple labs and took years to complete, which is evident from the very long lists of authors on the resulting papers (cf. Reference section).

The first fully sequenced organism was *Haemophilus influenza* Rd with a genome size of 1.8 Mbp (Fleischmann et al., 1995), closely followed by *Mycoplasma genitalium* G37, one of the smallest known genomes at 0.58Mbp (Fraser et al., 1995). A year later followed the first sequenced eukaryote genome with baker's yeast *Saccharomyces cerevisiae*; 12Mbp (Goffeau et al., 1996), using strains isogenic to the now standard reference genome S288c (Engel et al., 2014). Around this time, genome sequencing really started to get moving, and a dozen of genomes were released or were underway of release (Goffeau et al., 1996). Famously, the human genome project HUGO was completed in 2001 at a size of 2910 Mbp (Venter et al., 2001).

3.1.2 Second generation sequencing

Soon after the human genome project was completed, high-throughput sequencing methods began to emerge. To differentiate them from (Sanger) Shotgun sequencing, these methods are often referred to as Second Generation Sequencing or by a number of other names: Next Generation Sequencing (NGS), MPS (Massive Parallel Sequencing) and HTS (High-throughput sequencing). Here MPS will be used to refer to the second generation, whereas HTS will be used to refer to post-Sanger methods in general (i.e. both the second and third generation).

[6] *Assembly*: the process of compiling a longer sequence (e.g. a whole-genome sequence) from smaller sequences (reads). Can be done with or without a template genome (reference assembly and *de novo* assembly, respectively).

[7] *Contig*: short for contiguous sequence. A coherent sequence of DNA that is generated in an assembly by piecing together overlapping reads, supported by high confidence levels. Assemblies commonly consists of multiple contigs.

The first HTS method was the today often forgotten Massively Parallel Signature Sequencing (MPSS) method by Lynx Therapeutics (Brenner et al., 2000), that due to high costs never saw a market breakthrough (McPherson, 2014). Instead, the advent of second generation sequencing is normally considered to be five years later in 2005, when the 454 pyrosequencing platform was commercialized (Margulies et al., 2005; Kircher and Kelso, 2010). Many different MPS methods followed suit, and quickly led to a dramatic decrease in the cost of sequencing (Figure 4) and the time required for a WGS run (Kircher and Kelso, 2010). The price drop is fitting with the long-standing dream of the $1000 human genome set up by The National Institutes of Health (Anderson, 2004). Illumina claims to have reached this goal (McPherson, 2014), but as has been pointed out, these estimates only account for the cost of the sequencing run, and not for the costs of data management, storage and downstream bioinformatics, which can be substantial (Sboner et al., 2011). MPS is also the reason for the explosion of WGS-data in the mid-2000s (cf. Figure 2, Chapter 2).

Among the many competing MPS methods (for a review see e.g. Goodwin et al. (2016)), Illumina emerged as the dominating technique, holding 60% of the market share in 2013 (Mohamed and Syed, 2013; McPherson, 2014), which can be explained

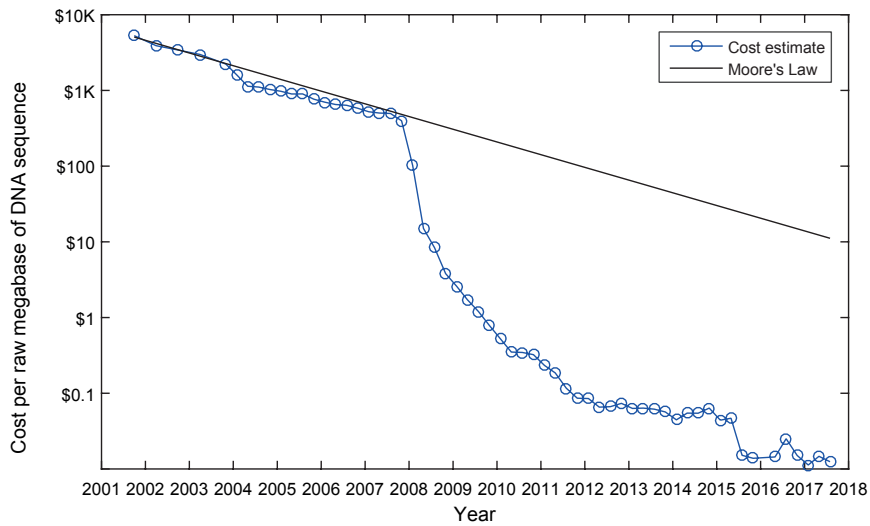


Figure 4: The cost (USD) of sequencing per Raw Megabase since 2001, as determined by the NIH (National Human Genome Research Institute), adapted from (Wetterstrand, 2017). Moore's Law describes how the number of transistors in an integrated circuit historically have doubled at a more or less fixed rate of ~2 years, but can be used to describe other technologies as well (Mack, 2011; Sboner et al., 2011). Technologies that develop according to Moore's law are regarded as well performing (Wetterstrand, 2017), and since the advent of High Throughput Sequencing in the mid-2000s, the cost per mega-basepair has greatly outdone Moore's Law.

by their reasonable (MPS) read length, high throughput and low cost (Schirmer et al., 2015). Given the strong presence in the field and the fact that it was the MPS method of choice for this thesis work (**Papers III-IV**), a detailed description of the Illumina method can be found in Box I.

Typical for MPS methods is that they generate much shorter reads (e.g. Illumina: 150-250bp in paired-end mode[8]) than traditional Sanger sequencing (~1kb) and with higher base calling[9] error rates (Goodwin et al., 2016). The PCR amplification steps in the library preparation and cluster amplification contribute to the increased error rates and amplification bias (Schadt et al., 2010). To compensate for the error rate, MPS assembly requires higher redundancy for each base (*coverage*[10]) in order to produce assemblies with sufficient confidence, meaning that each fragment needs to be sequenced multiple times (so called genome oversampling) (Schadt et al., 2010; Sims et al., 2014). Theoretically, coverage (c) is a the average read length (L) times the number of reads (N) over the haploid (one copy of each chromosome) genome size (G), $c = L * N/G$, meaning that a desired coverage can, to a certain extent, be set by adjusting instrument parameters (Sims et al., 2014)).

3.1.3 Third generation sequencing

The many developments in sequencing instrumentation since the arrival of MPS make it clear that the latest HTS methods belong to a third generation of sequencing (TGS). What criteria that differentiates this from the second generation has been a subject of debate, but a common argument is the method should be capable of single molecule real time sequencing, i.e. inspection of the DNA template without the need for PCR amplification (Schadt et al., 2010; Heather and Chain, 2016).

Two main TGS methodologies have currently emerged, PacBio SMRT (single-molecule real-time) and Oxford Nanopore, that both come with the selling point of being able to produce dramatically longer reads than the generations of sequencers before them (Pacbio: ~10kb; Nanopore: claims to have achieved >150kb reads) (Goodwin et al., 2016; Jain et al., 2016). For in-depth descriptions of these two methods, please see e.g. (Eid et al., 2009; Jain et al., 2016). The longer reads facilitate *de novo* assemblies (Heather and Chain, 2016) and make it possible to resolve regions that are otherwise difficult to determine; but like MPS methods, these TGS methods still have problems with repeats (Liu et al., 2017). Also, these methods have high error rates (~15%, compared to <1% of Illumina) (Berlin et al., 2015; Goodwin et al.,

[8] Paired-end refers to the sequencing of a DNA fragment from both directions (Lander et al., 2001). See Box I for a description of paired-end Illumina sequencing.

[9] *Base calling*: the process of determining the identity and order of nucleotides during the sequencing.

[10] *Coverage*: the average number of sequenced fragments (reads) that support a certain nucleotide position. For example, a coverage of 30x means that on average, each position in the contig was supported by 30 overlapping reads.

2016), which means that special experimental and computational considerations are needed. For instance, since errors in PacBio are random, the DNA fragments can be re-sequenced during the same run to correct for base calling errors (Rhoads and Au, 2015; Liu et al., 2017) – a standard approach for PacBio SMRT, called circular consensus sequencing (Eid et al., 2009; Goodwin et al., 2016) – or combined with MPS data (short-read, but lower error rate) in so-called hybrid assemblies (Bashir et al., 2012; Koren et al., 2012; Goodwin et al., 2015) to compensate for the error rate.

PacBio is the slightly older and more established of the two methods and thus more widely-used (Goodwin et al., 2016) and it was the methodology that was used for the sequencing described in **Paper II**. The principle for *de novo* assembling SMRT data is similar to that of MPS data, although special algorithms are needed to handle the error rates (Berlin et al., 2015).

3.2 Considerations for genomics experiments

The current maturity and low cost of the WGS allows genomics to be more extensively used to investigate specific research questions. WGS can in principle be done in two ways: either by *de novo* sequencing (**Papers II-III**) or *resequencing* (**Paper IV**). As the name implies, *de novo* means that a genome is determined and assembled with sheer computer power without relying on a previous template assembly, whereas resequencing consists in using the reads to determine a sequence (and its variants[11]) relative to a pre-existing reference sequence (Gabaldón and Alioto, 2016). Note that reference sequences can also be used for reference-guided *de novo* assembly (Lischer and Shimizu, 2017), as a final step for reordering *de novo* contigs in biologically relevant order (Del Angel et al. (2018); **Paper III**), or to resolve genetic material not present in the reference (**Paper IV**).

It should be kept in mind that, contrary to what the name *Whole Genome Sequencing* implies, an assembly is always smaller and more fragmented than the biological genome, due to various difficult-to-sequence or difficult-to-assemble regions (Keller and Meese, 2015), including those with very high or low GC content (*GC bias*) and complex repeats such as transposons (Chen et al., 2013; McCoy et al., 2014) ‑in principle: if a repeat is longer than the read length it will be difficult to assemble (Berlin et al., 2015). For most research applications, this incompleteness (often ranging around a few percent; Wetterstrand (2017)) will not be an issue, since the mentioned problematic loci are assumed to be found in non-coding regions. Furthermore, all genome assemblies, both draft and finished, will contain assembly errors, and man-

[11] *(Sequence) variant*: general term for changes in a sequence compared to a control sequence. Semantically similar to mutation, but the term variant is preferred until experimental evidence is in place. Variants can be synonymous (silent; no change in the polypeptide) or non-synonymous (non-silent; changed polypeptide).

Box I: The Illumina MPS method

Illumina uses a sequencing-by-synthesis method known as cyclic reversible termination (Goodwin et al., 2016). The process consists of three steps: library preparation, cluster amplification and sequencing-by-synthesis (Quail et al., 2008), Figure B1.

The library preparation serves to make the DNA compatible with the Illumina flow cell, a glass slide with a high density of covalently bound oligonucleotides (Metzker, 2010). The DNA sample is fragmented (typically in pieces of 0-1200 bp) and adapter oligonucleotides complementary to those on the flow cell – as well as universal sequencing primer binding sites and barcodes – are ligated in the 5' and 3' ends. The adapter-fragment library is filtered by size to select for the optimal template size of the method (200-300bp) and the selected range is PCR amplified (Quail et al., 2008; Kircher and Kelso, 2010; Aird et al., 2011). Transposases capable of simultaneous fragmentation and adapter ligation have been developed to speed up library preparation (Caruccio, 2011). The library is denatured and the single-stranded fragments (ssDNA) are loaded on the flow cell and immobilized by hybridization to the surface oligonucleotides (Kircher and Kelso, 2010).

In the next phase, the cluster amplification, the adapter in the free end of the strand hybridizes to complementary adapters on the flow cell (Fig B1-1). A complementary strand is amplified (Fig B1-2) and the original hybridized fragment is washed away. The adapter end of the complement strand hybridises to another adapter on the flow-cell and forms a bridge (Fig B1-3). The template is amplified (bridge amplification) and denatured (Fig B1-4), resulting in two copies of the fragment that both are attached to the flow cell by their respective adapter (Kircher and Kelso, 2010; Goodwin et al., 2016). This process is repeated to achieve high density clusters with several thousand copies of the template in close proximity (Fig B1-5), and is done for each fragment in parallel (Kircher and Kelso, 2010). The cluster amplification ends with the cleaving and washing out of the reverse strands of the templates (Figure B1-6), forming high density clusters of forward strand ssDNA (Fig B1-6). It has been approximated that Illumina cluster amplification results in 100-200 million template clusters across the flow cell (Metzker, 2010).

The first round of sequencing is done with the forward strand ssDNA clusters on the flow cell and will result in forward direction reads. Like in Sanger Sequencing, primers and fluorescent-labelled nucleotides are added, and each correct nucleotide incorporation results in a nucleotide-specific colour (one for each of the four dNTPs) that is imaged, the fluorophore is removed, and the process is iterated (Kircher and Kelso, 2010; Goodwin et al., 2016). Due to the

formation of multiple clonal clusters on the flow cell, each fragment can be sequenced in parallel, hence the name Massive Parallel Sequencing.

Sequencing reads in general, and MPS reads in particular, are *shorter* than the fragments they amplify. To improve MPS downstream bioinformatics (especially assembly), so called pair-end sequencing can be used to increase the information yielded from each fragment by basically sequencing each fragment from both directions. Therefore, after the first read has been produced, the sequencing process is repeated one more time but with the reverse strand as the template, which results in two sequences per DNA fragment with a known distance from each other (Quail et al., 2008). Paired-end data typically leads to better assemblies as ambiguous regions can be better resolved with the information from paired-reads compared to single reads.

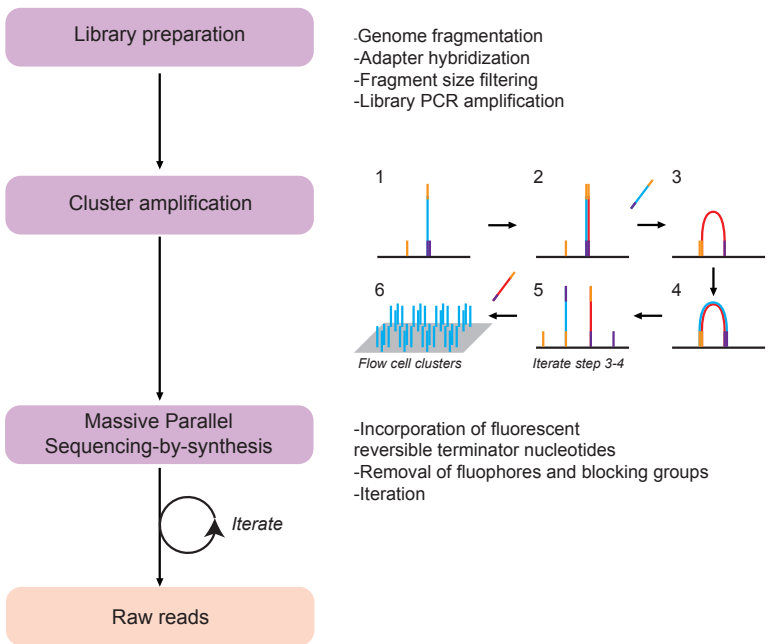


Figure B1: Schematic overview of the Illumina sequencing process. The work-flow consists of three steps: library preparation, cluster amplification and sequencing-by-synthesis. Library preparation serves to ligate sequencing adapters to a fragmented DNA sample and select and amplify an optimal template size. Immobilized clusters are generated across the flow cell in high density by bridge-amplification (Blue = forward strands; Red = reverse strands). The actual sequencing is similar to the chain-termination method in that fluorophore-labeled nucleotides are incorporated by a polymerase one base at a time. However, as the name Massive Parallel Sequencing implies, this is done simultaneously for all clusters on the flow cell, and imaged in real-time.

ual and/or computational validation is very time consuming (Phillippy et al., 2008). Should more completeness be required, hybrid assembly approaches combining short- and long-read methods (e.g. Illumina-PacBio, Illumina-Sanger) will be needed (Bashir et al., 2012; Koren et al., 2012).

Central to a good genomics experimental design is to consider the bioinformatics early in the planning stages. *De novo* sequencing is a difficult scientific and mathematical problem (Pop and Salzberg, 2008; Baker, 2012), and if there is a reference genome available it will facilitate the process and the downstream analyses. If not, *de novo* assembly will have to be used. Resequencing is suitable for most types of comparative approaches such as SNP- and structural variants discovery and genotyping (Olson et al., 2015), and amplicon sequencing (Heyduk et al., 2016). Bacterial assembly is generally more-straightforward than its eukaryotic counterpart due to their haploid nature. With increased ploidy comes the complexity of multiple alleles, which affect both assembly and variant calling[12]. As a rule of thumb, higher ploidy requires higher coverage (Margarido and Heckerman, 2015), e.g. in order to statistically estimate if a sequence variant is found in only some or in all alleles (hetero- and homozygous variants[13]) (Delaneau et al. (2013); **Paper IV**). For WGS applications with MPS data, 30-60x coverage are normally sufficient for most applications, with the exact depth depending on the application (*de novo* assembly and variant calling requires higher coverage) (Bentley et al., 2008; Desai et al., 2013; Fang et al., 2014). It is also important to assess the uniformity of coverage, i.e. the variation in coverage across the genome (Sims et al., 2014). The MPS requirements can be compared to the 6-8x coverage used for the Sanger shotgun WGS of the human genome (Lander et al., 2001), which was possible due to the lower error rate of the method.

Since DNA in general is very stable (in terms of e.g. half-life, mutation rate), only one technical replicate is normally needed for genomics studies, which can be compared to transcriptomics, where the transient nature of mRNA means that a high number of biological replicates are needed to reach statistically appropriate analyses (Schurch et al., 2016). The number of biological replicates needed, will however be dependent on the research question of each project, with comparative approaches and population studies requiring more biological replicates.

The choice of sequencing platform also determines the downstream applications. As could be expected of their different sequencing chemistry, some instruments are better suited to some tasks than others. In short it can be said that data from methods

[12] *Variant calling*: The process of determining variants. Usually done by mapping reads to a reference sequence and identifying statistically supported sequence variants.

[13] *Haplotype estimation*: A method to infer the sequence of alleles in polyploid genomes (more than one copy of each chromosome), e.g. from sequencing reads. Can be used in connection with variant calling in polyploid organisms to determine the allele frequency of the variants (homozygous and heterozygous variants). Also known as haplotyping or haplotype phasing.

that produce longer reads will be easier to process than those with shorter reads (e.g. less coverage needed), but the error prone nature of current long-read methods (HTS) is a drawback that require specific design considerations (Berlin et al., 2015). Also, HTS methods are currently more costly than their MPS counterparts. Sometimes, it can be considered altogether to use alternative DNA strategies instead (microarray, qPCR) instead of HTS (Goodwin et al., 2016; Lavín et al., 2017).

3.3 Assembly, read mapping and annotation

Whereas sequencing itself used to be the bottleneck in genomics, with HTS the bioinformatics needed to assemble the genome has become the new bottleneck (Gabaldón and Alioto, 2016). Assembling and mapping genomes from HTS reads is like solving a jigsaw puzzle, only that each piece comes in multiple, semi-redundant variants and that some pieces are missing altogether. The nature, volume and complexity of the data demand computational support, meaning that genome assembly is done using a number of established algorithms developed by specialist bioinformaticians. Biologists working with Large Data, however, should be more concerned with the *hands-on* work on to assemble a genome, which is to run series of sequential algorithms (a *pipeline*[14]) and tweak their parameters and settings depending on the characteristics of the in-data. This is a complex task in itself that requires knowledge about the expected outcome of each algorithm and the data formats they use. This section will therefore focus on points that the Large Data biologist needs to know, and less on the mathematical basis of the assembly and alignment algorithms.

3.3.1 Pre-processing: data quality control and filtering

The first step in any assembly work-flow (*de novo* or not) is to assess the quality of the data and adjust or filter bad reads. An abbreviated example of typical MPS data is show in Figure 5, stored in the standard FASTQ-format. FASTQ is based on the common FASTA format for DNA sequences (Pearson and Lipman, 1988) but in addition to the sequences of all the reads, FASTQ also stores *read quality* information generated by the sequencing instrument (Figure 5-L4) (Cock et al., 2009), which allows for filtering the data after the run has been completed. MPS read quality is measured in the PHRED-score[15] which gives the probability of an incorrect base call

[14] *Pipeline*: A computational work-flow consisting of a set of software combined in a chain; for instance, the scripts and algortihms needed to run the assembly workflow in Figure 6. Normally consists of a number of different programs that need to be connected, often by adapting the output format of one program to fit with the input format of the next.

[15] The PHRED-score is the base calling quality expressed as the probability of an incorrect base call in a given sequence. A score of 30 corresponds to a probability of 1 incorrect base call in 1000 bases, and

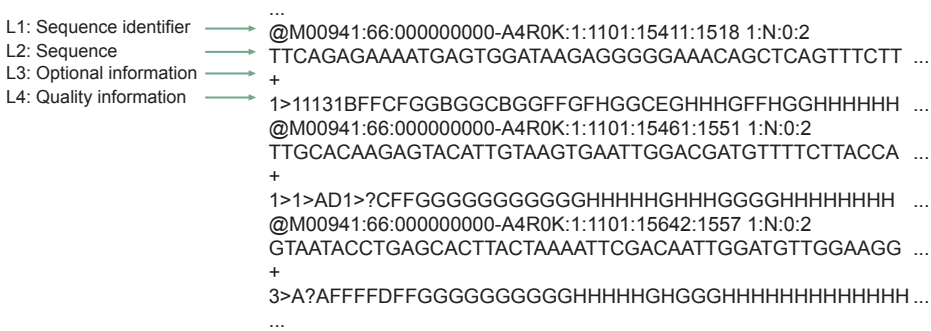


Figure 5: Example of a FASTQ-file containing raw data from an Illumina MPS run (here: **Paper IV** data for *S. cerevisiae* ISO12; SRA accession number: SRR2002960). Each read has four lines of information (L1-L4), with the second and fourth line containing the sequence and its quality data. The figure has been truncated horizontally and vertically as indicated by the red ellipses. This particular file contains the forward reads of a paired-end run that together with the reverse reads sum to over 12 million reads (~4 GB total file size), illustrating that genome assembly is a demanding computational task.

(Cock et al., 2009). Low quality can e.g. be caused by method-dependent amplification biases during the library preparation and the sequencing run (Aird et al., 2011; Nakamura et al., 2011), or by accidental sequencing of the adapter sequences or PCR primers (Bolger et al., 2014). The 3'-ends of the reads tend to accumulate errors as well (Kelley et al., 2010). The removal of low quality bases is known as trimming, and can be done with a number of algorithms such as trimmoatic and sickle (Bolger et al., 2014), and has been shown to be beneficial for assembly and variant calling (Del Fabbro et al., 2013). MPS primers and adapters are normally made public by the instrument manufacturers and can thus easily be found and removed (trimmed). Furthermore, thanks to the coverage redundancy in MPS, read errors can be corrected by substituting low coverage *k*-mers[16] (substrings of a read with size *k*) with those of higher coverage (Kelley et al., 2010). Finally, reads can be filtered out completely if they do not pass the desired quality threshold (Bolger et al., 2014).

3.3.2 *De novo* assembly

At its core, *de novo* assembly algorithms attempt to assemble a series of longer sequences (contigs and scaffolds[17]) from reads by finding redundant overlap. Since this is a non-trivial mathematical problem (Pop and Salzberg, 2008), an assembly pipeline (Figure 6) needs to include a number of iterative quality control (QC) steps.

a score of 40 (preferred threshold) corresponds to 1 in 10000.

[16] See Box II for a short description of *k*-mers and de Bruijn graphs.

[17] *Scaffold*: A colletion of contigs and gaps that together describe a longer portion of a genome sequence.

Early MPS assemblers used *greedy algorithms* to select which reads to merge by always going for the overlap with the highest score (Miller et al., 2010), but was e.g. prone to misassembly (chimeral contigs) in repeat regions (Schatz et al., 2010). Nowadays the methods dedicated for short-read assembly are typically based on so-called de Bruijn graphs[16] for *k*-mers (Schatz et al., 2010; Compeau et al., 2011). The choice of assembly method is thus largely dependent on read length, which in turn is dependent on the method used to generate the reads. For bacterial MPS data, Velvet (Zerbino and Birney, 2008) has long been considered one of the best assemblers (Edwards and Holt, 2013), but have in some regards been surpassed by the newer algorithm SPAdes (Bankevich et al. (2012); used in **Paper III**), as indicated by benchmarking tests (Magoc et al., 2013; Al-okaily, 2016). Later versions of SPAdes are also capable of hybrid assembly of short and long reads (Antipov et al., 2015).

De novo assemblies are contingent on quality control – possibly more so than *resequencing* assemblies – since inaccuracies in the assembly will affect the downstream assessments (Berlin et al., 2015). Quality assessment of an assembly can be done with a number of different metrics, including the number of contigs, how many contigs needed to describe 50% of the assembly when the contigs have been sorted by order of descending length (L50), the size of the smallest of the sorted contigs that describe 50% of the assembly (N50), CG count, and the number of aligned bases and misassemblies compared to a reference genome (if available) (Gurevich et al., 2013). As a rule of thumb: the fewer and longer contigs/scaffolds, the better the assembly. A popular quality control (QC) algorithm for assemblies is QUAST (Gurevich et al., 2013).

Assemblies can potentially be improved computationally be re-running the assembly with altered parameters or by gapfilling algorithms. The latter strives to resolve gaps within contigs to reduce the number of ambiguous bases (Boetzer and Pirovano, 2012), see Figure 6. Algorithms such as Gapfiller, Sealer and AlignGraph use the paired-end reads from the in-data to try to resolve gaps, the latter also capable of using closely related reference genomes (Boetzer and Pirovano, 2012; Bao et al., 2014; Paulino et al., 2015). Finally, once a the user is pleased with the assembly metrics, it is useful to reorder the otherwise randomly ordered contigs/scaffolds in a more biological relevant order with the help of a reference genome (Edwards and Holt, 2013); this can e.g. be done with Mauve (Darling et al., 2004).

3.3.3 Resequencing examples: read mapping and variant calling

Resequencing is the alternative approach to *de novo* assembly and relies on aligning the reads to a previously assembled reference genome (*read mapping*). Following pre-processing, the reads are aligned to the reference genome using an alignment algorithm. Commonly utilized algorithms include bowtie (short reads; Langmead and Salzberg (2012)), BWA (short reads; Li and Durbin (2009)) and blat (long reads; Kent

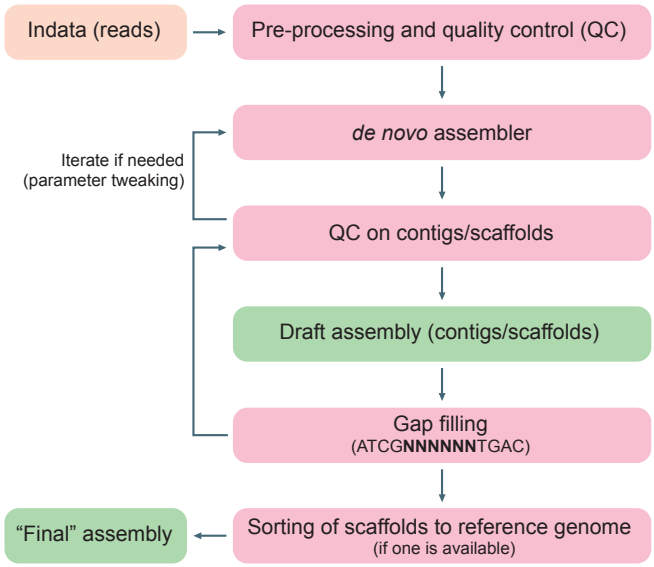


Figure 6: Overview of the work-flow used for the *de novo* assembly in **Paper III**, based on recommendations from (Edwards and Holt, 2013). The final assembly should be considered as "final" in quotation marks, since it can always be improved by e.g. resequencing or hybrid assembly approaches. Gaps and ambiguous bases are represented by N. For this particular assembly, a reference sequence from a related *Psedomonas* species was available, but this is not always the case.

(2002)). Alignments are stored often in SAM format (called BAM when compressed) (Li and Durbin, 2009). Alignment data can be used for many different applications, for instance to calculate phylogeny (**Paper II-III**) and detect sequence variants (**Paper IV**).

Sequence variants can roughly be classified in three types based on their length: Single Nucleotide Polymorphisms (SNPs), insertions and deletions (INDELs) and structural variants (e.g. copy number variations, duplications and translocations), each of which tend to require their own specialized algorithms (Xu, 2018). SNP and INDEL calling is done from SAM alignments, with popular software packages being SAMtools (Li et al., 2009b) and GATK (McKenna et al., 2010). Both are probabilistic variant callers that calculate a likelihood of a genotype at each base (Mielczarek and Szyda, 2016). Human genome data has the benefit of SNP databases that can use known SNP data to help the prediction (Sherry et al., 2001), but that is seldom the case for microbial data. A sequencing error in an isolated read is indistinguishable from a sequence variant, and therefore sufficient depth of coverage is needed in variant calling (Sims et al., 2014). Due to the high occurrences of false positive and false negative calls, the variants needs to be filtered to remove low quality calls, which can

be done with SAMtools and GATK (Altmann et al., 2012). The final step of a variant calling pipeline is to annotate[18] the variants in order to facilitate interpretation and to predict the effects of non-synonymous variants (Altmann et al., 2012). The UCSC Genome browser is a good tool for annotation and variant effect prediction, but it only supports a few model organisms (Rosenbloom et al., 2014). It can also be noted that variant calling can be made from *de novo* assemblies, but since an assembly is a consensus sequence with coverage 1x, it is statistically less strong than read mapping (Olson et al., 2015).

3.3.4 Annotation: predicting and identifying open reading frames

The value of a genome sequence has been associated to the quality of its annotation (Stein, 2001). Since "complete" annotation requires molecular biology evidence of gene expression and function, most annotated genes are putative. This is not a drawback *per se* as tentative ORFs and their potential sequence variants can be a good fuel for hypothesis generation. An example is the draft pathway reconstruction and gene cluster/operon discovery discussed in **Paper II** which could correlate putative ORFs with growth phenotypes, and thus paves the way for future studies on aromatic catabolism in Gram-positives.

Annotation can be divided in two categories: structural (identification of genetic features, e.g. ORFs, ribosomes, CRISPR repeats, transposons) and functional (attachment of meta-data to structural annotations) (Yandell and Ence, 2012). Whole genome annotation is highly dependent on computational approaches, and contrary to e.g. assembly and variant calling where the user commonly builds a custom pipeline suited for the project, genome annotation normally needs to be done with established pipelines (Tatusova et al., 2013) that combine *ab inito* predictors (mathematical approach) and evidence-driven predictions (e.g. alignment with data from related organisms) (Yandell and Ence, 2012; Tatusova et al., 2013).

Prokaryotic automated genome annotation is quite mature and benefits strongly from the high number of sequenced and annotated bacterial genomes that are available today (Tatusova et al., 2013). Examples of common prokaryote pipelines include SEED/RAST (Overbeek et al., 2013), Prokka (Seemann, 2014), and the NCBI prokaryotic genome annotation pipeline (PGAP). PGAP is developed by and integrated within the NCBI databases, and thus has intrinsic access to the largest nucleotide database worldwide (Tatusova et al., 2013). PGAP was the choice of annotation pipeline for **Papers II-III**. A downside of PGAP is that it can only be run on request when uploading a genome to the NCBI, but since it is considered best-practice to upload assemblies to Genbank prior to submission of manuscripts to journals, this

[18] *Annotation*: the process of predicting and identifying features in a nucleotide or amino acid sequence. Genome annotation can be used to identify features such as genes, rRNA, etc.

Box II: *k*-mers and de Bruijn Graphs

Many modern genome assembly algorithms rely on a mathematical concept known as de Bruijn Graphs for *de novo* assembly of short-read data. Here, each read is divided into substrings of length k to facilitate identification of string overlap. Example of k-mers in a short read for different values of k:

Read: ATGGCGTGCA (10bp)
3-mers: ATG, TGG, GGC, GCG, CGT, GTG, TGC, GCA
8-mers: ATGGCGTG, TGGCGTGC, GGCGTGCA
10-mers: ATGGCGTGCA

Let's say that all reads from a MPS sequencing run are 100 bp = 100-mers. For technical reasons, a MPS run cannot capture all 100-mers from the genome. By instead performing the assembly on shorter k-mers from the same reads, commonly ~30-50 bp, the k-mers better represent the composition of the genome. The assembly is calculated by constructing a de Bruijn Graph where each k-mer is connected to two different nodes of size k-1 (one for each end of the k-mer; "left" and "right" k-1-mer) which eventually forms a graph of the relationships of the k-mers in the genome (Figure B2). This circumvents the need for pairwise alignment of each k-mer and significantly decreases the computational burden. For further reading, please see Compeau et al. (2011); Miller et al. (2010).

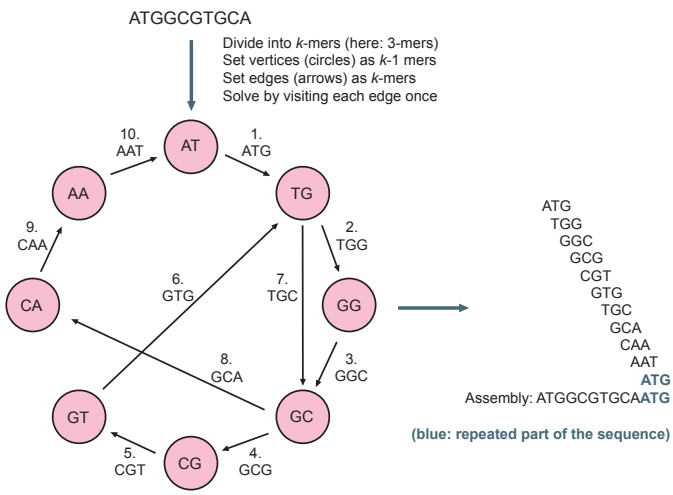


Figure B2: Schematic example of a de Bruijn graph of the example "genome" ATGGCGTGCA. In this example there is only one read for sake of simplicity. In a real case, each read from the sequencing will be divided in its corresponding k-mers, and all unique k-mers will used to attempt the genome assembly. Adapted from Compeau et al. (2011).

could be considered a minor issue.

Eukaryotic annotation poses additional challenges, due to the larger size, introns and high number of repeat-rich regions (Cantarel et al., 2008; Yandell and Ence, 2012). The intron/exon challenge can be assisted by the use of MPS expression data (RNAseq) (Haas et al., 2011; Yandell and Ence, 2012); however, in the case of *S. cerevisiae* – the subject of **Papers IV-VII** – and *Candida* yeasts, only about 5% of the genes require splicing, making structural annotation more similar to that of prokaryotes (Haas et al., 2011). Examples of eukaryotic annotations pipelines include MAKER (Cantarel et al., 2008) and PASA (Haas et al., 2003).

3.4 Comparative genomics for Adaptive Laboratory Evolution

The closing section of this chapter on genomics relies on **Paper IV** as a case study to illustrate on how comparative genomics can be used to identify mutations from evolution experiments. Fermentation of inedible plant matter (lignocellulose) is a sustainable way to produce value-added chemicals from renewable feedstocks (de Jong and Jungmeier, 2015). Lignocellulose pretreatment (here: steam-explosion) results in an hydrolysate rich in five- and six carbon sugars (pentoses and hexoses, respectably), as well as inhibitory compounds such as furaldehydes (here: furfural and hydroxymethylfurfural), weak acids and a number of lignin-derived aromatics (Taherzadeh and Karimi, 2008). Baker's yeast *S. cerevisiae* is commonly applied for lignocellulose fermentation due to its inherent robustness and efficient ethanol production, but it cannot naturally utilize pentose sugars (see Section 4.3; **Papers V-VII**) and it is negatively affected by high concentrations of furaldehydes, aliphatic acids and phenols (Almeida et al., 2007).

In the directly preceding study to **Paper IV**, the already robust industrial *S. cerevisiae* strain Ethanol Red (ER) was subjected to Adaptive Laboratory Evolution (ALE) by growth at elevated temperature (39°C) and in the presence of non-detoxified spruce hydrolysate for ~300 generations, after which a stable clone named ISO12 was isolated with improved tolerance to the two stressors (Wallace-Salinas and Gorwa-Grauslund, 2013). The aim of the **Paper IV** study was to use WGS and variant calling (Section 3.3.3) of the two strains to identify target mutations that could describe the novel phenotype of ISO12. Although both strains were also *de novo* assembled, this project made use of a relative variant calling strategy where the reads of both strains were compared to the gold standard reference genome S288c (Engel et al., 2014) in order to make use of its quality annotations. All variants common between each strain and S288c were discarded, which left the variants that arose between ER and ISO12 (**Paper IV**). Genetic material in ER not present in S288c was extracted from the *de novo* assembly and the reads from ISO12 were used for variant calling of the corresponding region(s). Functional analysis of the coding-region variants revealed 760

Table 4: Candidate genes for reverse engineering of the novel ISO12 phenotype in wild type *S. cerevisiae*.

ORF	Functional annotation	Rationale
MTL1	Cell Wall Integrity sensor	Positive selection in ISO12 (K_a/K_s >1). Belongs to the cell wall integrity MAPK signalling pathways, which is related to heat stress (Verghese et al., 2012)
FLO1/5 /9/11	Flocculation proteins	Positive selection in ISO12 (FLO9/11); Homozygous variants (FLO1/11); Significant CNV increase in ISO12 (FLO1); High variant density: 25 non-syn. calls (FLO5)
CYC3	Cytochrome C heme lyase	Positive selection in ISO12 (K_a/K_s >1)
GPR1	Extracellular glucose sensor (cAMP/PKA signalling)	Premature stop codon in amino acid 251 of 961; heterozygous variant, detected in 59% of the alleles. The cAMP/PKA signalling pathway is related to thermotolerance (Verghese et al., 2012)
ADH7	NADPH-dependent alcohol dehydrogenase	Two-fold CNV increase in ISO12 in non-reference genome regions (*de novo* assemblies)
ENA1/2	ATPase sodium pumps	CNV increase in ISO12 in genetic regions not present in the reference strain; such increases has been associated with acetate and temperature robustness (Gilbert et al., 2009)

non-synonymous variants distributed over 347 ORFs. This illustrates that correlation of genotype-phenotype is not trivial even in short-time ALE, especially when multiple selection pressures are used.

The ISO12 genome was further analysed for copy number variations (CNVs) and evolutionary selection pressure. CNVs are genetic variations where genes change in number rather than in DNA sequence and affect phenotypes through gene dosage effects (Zhang et al., 2013). The selection strength during DNA evolution can be estimated with the ratio between the number of non-synonymous substitutions (K_a) and non-synonymous substitutions (K_s) in an ORF, denoted as ω or K_a/K_s; K_a/K_s > 1 implies a positive selection, K_a/K_s = 1 a neutral selection, and K_a/K_s < 1 a negative selection (Zhang and Yu, 2006). When all these results were taken together, ten genes emerged that seemed extra likely to be correlated to the new phenotype. A summary of key target genes for the new phenotype in ISO12 is listed in Table 4. Although that work focused primarily on non-synonymous mutations, the impact of the synonymous mutations should not be forgotten since they can regulate expression rates (Kudla et al., 2009).

ALE experiments are also likely to result in accumulation of mutations that are not related to improved tolerance to the selection pressure, so called *hitchhiker mutations* (Lang et al., 2013), which make finding the true causative mutations behind novel phenotypes very challenging. One approach is to sequence the genomes of clones isolated during the ALE, and compare the genotype of clones with and without the desired phenotype. In species capable of sexual reproduction (the case of *S. cerevisiae*) it is possible to backcross clones with novel phenotypes to an ancestral strain and sporulate the progeny to generate haploid cells with random allele distribution; variant calling of the different clones, progeny and parent can possibly resolve the causal genotype (Koschwanez et al., 2013). However, neither of these approaches were possible in the ISO12-case because of the genetic instability of the earlier clones, and the loss of sporulation in ISO12.

There have been a number of studies that have used the "evolve and resequence" (Payen et al., 2016) approach to assess improved tolerance to inhibitors from hydrolysate in *S. cerevisiae* (Almario et al., 2013; Wang et al., 2017), and elevated temperature (Caspeta et al., 2014; Satomura et al., 2016). However, it is noteworthy that the suggested driver mutations of these studies were different from those described in **Paper IV**[19], with the exception of the cAMP/PKA pathway that was also a target in one of the studies (Satomura et al., 2016). Although it has been demonstrated that ALE in *S. cerevisiae* can have high repeatability (same key targets being susceptible to mutation across multiple replicates; Lang et al. (2013)), the different outcome of the aforementioned studies show that different experimental approaches and background strains result in different genotypes. The use of temperature and inhibitor co-stressors in the ALE study (Wallace-Salinas and Gorwa-Grauslund, 2013) that was the basis for **Paper IV** also complicated the process of genotype discovery, as it is likely that driver mutations that conveyed simultaneous tolerance to *both* stressors were selected for in ISO12.

[19] Another study worth mentioning in the ISO12 context used TALEN genome editing instead of ALE to generate a library of stress tolerant strains, and found enrichment of sequence variants in proteins in the same Gene Ontology (GO) class (cell-periphery proteins) as in the ISO12 case (Gan et al., 2018).

In the uncertain ebb and flow of time and emotions
much of one's life history is etched in the senses.

「あまりにも不確(ふたし)かな時間や気持ちの流(なが)れの中で、
五感(ごかん)にはいろいろな歴史が刻(きざ)み込(こ)まれている。」

BANANA YOSHIMOTO (吉本ばなな)
Kitchen (キッチン) (1988)
English translation by Megan Backus (1993)

Chapter 4

A closer look at signalomics

Sensing and signal transduction – the means of cell communication – is a necessary property of biological life (Bruni, 2008). The cellular system is regulated by a constant input of signal cascades that form complex signalling networks ranging from molecule-molecule interactions to species level interactions (Weng et al., 1999). Following the systems view of the cell, the term signalome has begun to refer to the entire mass of the signals transmitted in the cumulative signalling networks of a cell. As has been established in previous chapters, Large Data biology is often associated with high-throughput omics methodologies, but, as will be evident in this chapter, is not limited to them. The method of monitoring the *S. cerevisiae* sugar signalome developed in this thesis work (**Papers V-VII**) instead relies of flow cytometry, a data-intensive technology that can measure the fluorescent characteristics of single-cells in quantities of tens of thousands cells per sample.

4.1 What is the signalome?

4.1.1 Towards a definition

In the wake of systems biology, a need to construct new nomenclature for a number of additional system-wide, cellular omes has emerged. Linguistically, it is normally easy to comprehend the meaning of such *ome*-neologisms (e.g. interactome, fluxome and phenome, to name a few more recently coined omes; Baker (2013)), but it can be more challenging to find formal scientific definitions. Compared to the other omes (Figure 1) the concept of the signalome is something that has yet to definitively catch on; it has however begun to make it onto lists of omes (Prohaska and Stadler, 2011). A proposed definition of the signalome is that it is a collection of all the support molecules of all signalling networks active at a given point of time (Bruni, 2008). It is however noteworthy that a very similar concept named *signalsome* is used to describe protein clusters in signalling networks (Wang and Malbon, 2011), which further demonstrates that the naming conventions of the signal network-ome

is not firmly established. A quick literature search reveals that the term signalome is currently mostly used in medicine and cancer research (see e.g. Wicki-Stordeur and Swayne (2014); Dasari et al. (2017); Haqshenas et al. (2017)), but that there are microbial examples as well (Vihinen, 2001; Pitarch et al., 2003; Mhlongo et al., 2018). In the present thesis work, the signalome concept was slightly altered to refer to a smaller set of signalling pathways related to sugar sensing (discussed in Section 4.3.).

4.1.2 Intracellular signalling networks govern cellular functions

The role of signalling networks is to transduce signals that regulate the cellular response to environmental and intracellular cues (Bruni, 2008). Unlike metabolic pathways, signalling pathways do not catalyse enzymatic conversions of a substrate into a product (mass flow), but instead transduce signals through sensors, transducers and actuators (signal flow) (Hyduke and Palsson, 2010). Signals are propagated in cascades (Figure 7) that often start with signalling molecules binding to extracellular receptors, and are transduced by means of post-translational modifications (PTMs), e.g. phosphorylations (Fiedler et al., 2009), ubiquitinations (Woelk et al., 2007), protein-protein interactions (Pawson and Nash, 2000), as well as cellular translocation (Teruel and Meyer, 2000) and second messenger molecules (dedicated signal carriers, e.g. cAMP, calcium; Hofer and Lefkimmiatis (2007)).

It has been proposed that signalling networks should be differentiated from regulatory networks. Although similar in effect they differ in structure: signalling networks are perceptual of the environment and organized as input-intermediate-output systems (Figure 7), while regulatory networks are typically organized as feedback loops (Hyduke and Palsson, 2010). To complicate matters, signalling networks can elicit heterogeneous responses across a cell population (Bruni, 2008), likely due to built-in redundancies (e.g. multiple types of sensors that result in same outcome, usually gene expression modulation) and multiple (sub-)pathways acting in parallel (e.g. two different outcomes to the same signal). While the present work will focus on the signal events inside single cells, the importance of intercellular (cell-cell) signalling in microbiology – with typical examples being bacterial quorum sensing (Waters and Bassler, 2005) and yeast mating pheromones (Dohlman and Thorner, 2001) – should not be underestimated.

There are a number of known signalling networks, many of which occur only in multicellular organisms. For the sake of simplicity, all examples of signalling pathways will here on out be taken from *S. cerevisiae*, unless specified. Being a model single-cell eukaryote, the signalling networks in this yeast are among the most studied signalling systems in microbes, and have also been used as a model for cancer research (e.g. since tumour growth can be related to altered signalling) (Diaz-Ruiz et al., 2011; Cazzanelli et al., 2018).

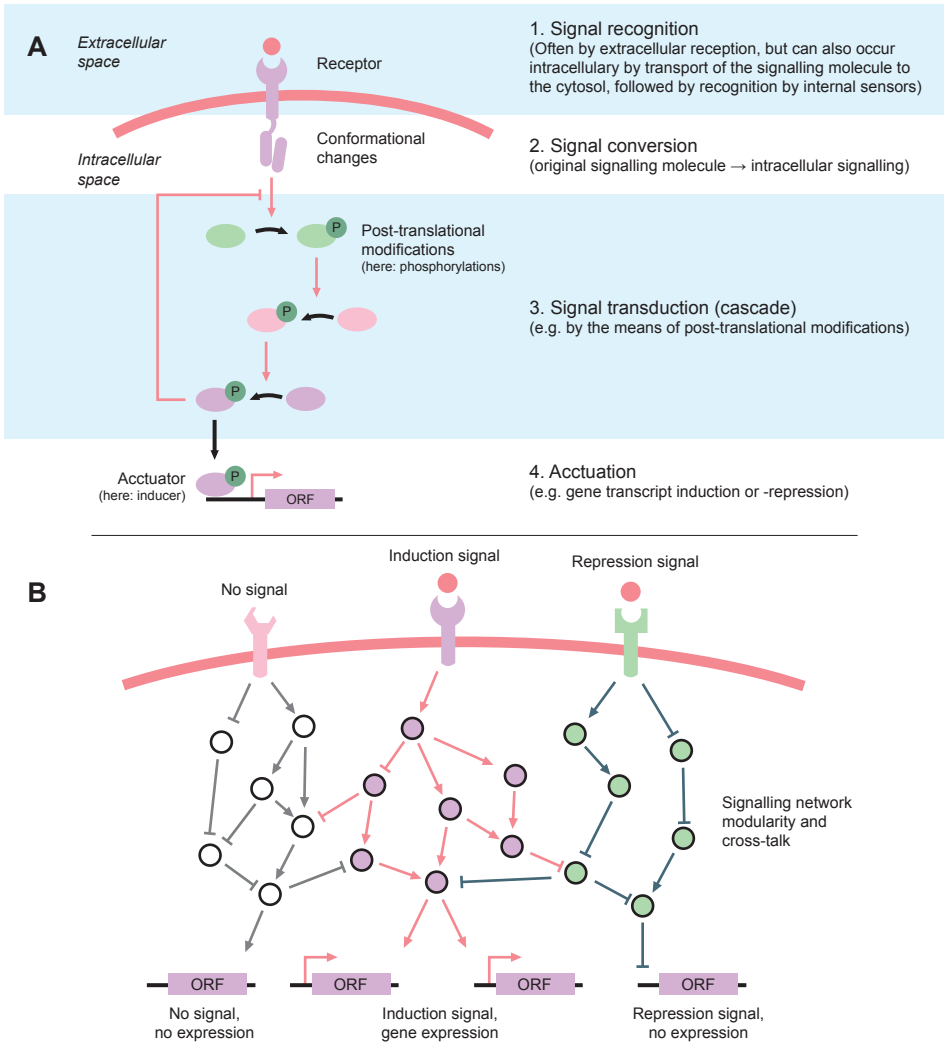


Figure 7: Overview of signalling network function and organization. Black arrows: reactions; Arrowheads: induction; Hammerheads: repression. **A:** Schematic representation of a signalling cascade. In this example, an extracellular signalling molecule (e.g. a nutrient) is sensed by a membrane receptor (1) which passes on a signal to an intracellular signalling network (2). The signal is transduced in a cascade of post-translational modifications and interactions between signal carriers, e.g. signal proteins and second messengers (3) that eventually ends in a signal acctuation, which is often manifested by gene expression regulation (4). As an outcome of the signal transduction, the cellular behaviour changes in response to the original signal (5). **B:** Signalling networks are typically organized in complex, interconnected topologies. A characteristic of signalling networks is their high modularity, i.e. that sets of network components are always expressed together (illustrated by the white, purple and green nodes); modules can however interact with each other as an additional layer of signal modulation (cross-talk). Adapted from Hyduke and Palsson (2010); Yao et al. (2015); Lee and Cho (2018).

According to current knowledge, the *S. cerevisiae* signalome consists of pathways for nutrient, stress, apoptosis, cell growth and mating signals. Nutrients, such as sugars (discussed in Section 4.3), nitrogen, phosphate and other carbon sources are sensed by a number of pathways, including the Snf1/Rgt1, cAMP/PKA (cyclic AMP and Protein Kinase A) and TOR (Target of Rapamycin) pathways (Conrad et al., 2014). Stress signalling is a broad topic that includes the sensing of various kinds of stressors, such as high osmolarity, high temperature, nutrient starvation and cell wall stress. Most stress signalling is regulated by the MAPK (Mitogen-Activated Protein Kinase) and cAMP/PKA pathways; these pathways also cross-talk (send signals to each other) (Thevelein and De Winde, 1999; Chen and Thorner, 2007; Tamaki, 2007). Closely related to these pathways is the ESR (Environmental Stress Response) which is a panel of chaperons and heat shock proteins induced by Msn2p/4p, two proteins which in turn are regulated by e.g. cAMP/PKA, MAPK and TOR (Gasch and Werner-Washburne, 2002; Verghese et al., 2012). Finally, cell growth is controlled by the TOR signalling pathway (Martin and Hall, 2005) and the mating signals are transduced in the pheromone response MAPK pathway (Dohlman and Thorner, 2001). As should be evident from this brief overview, there are a small number of signalling pathways in *S. cerevisiae* that together handle multiple types of environmental cues.

4.2 Methods to analyse the signalome

The fact that signals are transduced with molecules makes it possible to assess the signalome with traditional omics methods. In particular, mass spectrometry (MS)-based techniques such as proteomics and metabolomics have proved applicable to monitor signal transduction (Zhao and Jensen, 2009; Yao et al., 2015). However, the highly transient nature of the signalome (possibly more transient than the transcriptome, which it commonly regulates) has called for real-time methods such as fluorescent imaging and biosensors. As neither omics nor biosensors are currently on their own able to capture the signalome at a simultaneously ome-wide and high temporal resolution (Figure 8), the methods should be seen as complementary.

4.2.1 Omics approaches

Broadly speaking, two commonly used omics-approaches to assess the signalome are molecular profiling (molecules and post-translational modifications (PTMs) that are present at a given time) and molecular perturbation (changes over time, often combined with genetic modifications) (Yao et al., 2015). A non-exhaustive list of proteomic techniques that have been used to assess signal transduction include: chemoprotomics (the interaction of proteins with small molecules), phosphoproteomics (to assess phosphorylation PTMs), protein interactome studies (protein-protein interac-

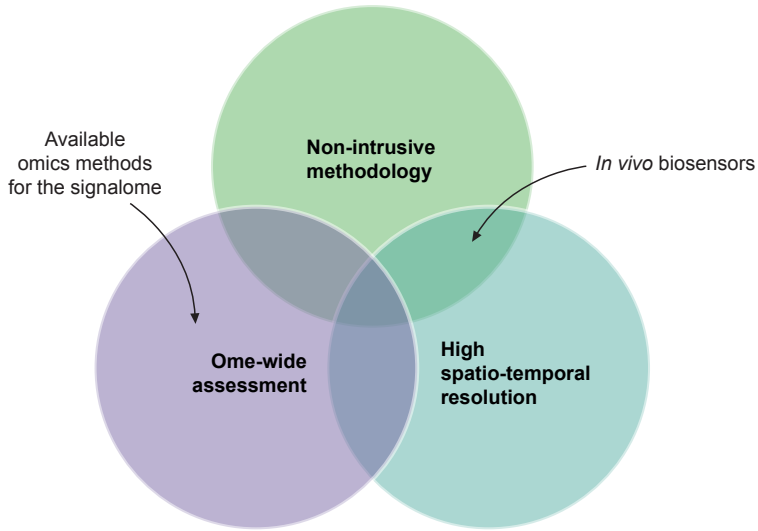


Figure 8: Venn diagram of the desired features of an ideal methodology for signalome assessment. Current methods (omics, biosensors) have their own strenghts and weaknesses. Proteomics and metabolomics can give ome-wide data on signalling, but their very labour-intensive nature (e.g. the need to quench and lyse cells) make them less suitable for sampling at high temporal resolution. Biosensors can be used for close-to-real-time monitoring of signalling processes and allow for subsequent use of the measured cells by cell sorting approaches, but fail to capture the holistic complexity of the signalling networks due to the limited number of sensors that can be simultaneously applied in a single cell.

tions and protein-complex formation), and ubiquitin-remnant profiling (Witze et al., 2007; Xu and Jaffrey, 2013; Yao et al., 2015).

A challenge with PTM proteomics is that the desired proteins exist only in low concentrations (as opposed to e.g. metabolic proteins); enrichment of PTM peptides (e.g. removal of non-relevant peptides) is thus required after sample fragmentation/proteolysis (Witze et al., 2007; Zhao and Jensen, 2009). Other challenges include high rates of false positive PTM-peptide discovery and the fact that some proteins have multiple PTMs and participate in signalling cross-talk, which complicates the mechanistic elucidation (Zhao and Jensen, 2009). However, not all signalling molecules are proteins, and not all of them are intracellular. To give an non-yeast example, bacterial cell-cell and cell-plant communications largely rely on excreted signalling molecules such as volatile organic compounds and quorum sensing molecules that constitute an ”extracellular signalome”, which can be assayed with metabolomics methods (Mhlongo et al., 2018).

Molecular perturbations can e.g. be studied with integrative approaches such as functional genomics. The idea of functional genomics is to correlate the genome with expressed transcripts and proteins at a given time, and thus require multiple omics

datasets, such as genomics, transcriptomics and proteomics data (Werner, 2010). Genetic modifications are introduced and the resulting molecular and phenotypical changes are monitored; the modifications can be done by various methods such as traditional knockout and overexpression, and attenuating approaches such as RNA interference (RNAi) or CRISPR interference (CRISPRi) (Yao et al., 2015).

Chromosome-related signalling mechanisms such as DNA-transcription factor and DNA-protein binding can be assayed with ChIP (chromatin immunoprecipitation) methods (Park, 2009). In the ChIP work-flow, live cells are treated with formaldehyde to fixate all proteins that are bound to DNA; the DNA is then fragmented and antibodies specific to the protein(s) of interest are used to select for the fragments that contain protein binding sites (Kim and Ren, 2006). While originally used with Southern blotting methods, PCR and microarrays (ChIPchip), high-throughput implementation that use MPS have been later developed, such as ChIPSeq (Johnson et al., 2007; Park, 2009). Challenges of ChIP methods include availability of antibodies of suitable specificity and sensitivity, pre-processing artefacts and getting a statistically suitable depth of coverage (Park, 2009).

4.2.2 *In vivo* biosensor approaches

A general drawback of omics methods is that the data is typically an average over many cells and thus cannot resolve single-cell variations and population heterogeneities (Welch et al., 2011). Most omics methods are also intrusive, i.e. the sample preparation requires cell lysis, and highly dynamic omes like the transcriptome, proteome and metabolome furthermore require quenching to ”freeze” the cellular state immediately after sampling (Canelas et al., 2008). This makes attempts at (pseudo)real-time monitoring difficult. An alternative approach to study dynamic signalling processes has instead been to use different forms of fluorescent biosensors for live cell imaging (Newman et al., 2011).

Although fluorescence live-cell imaging methods such as FISH (fluorescence *in situ* hybridization) were developed in the late 1960s (Levsky and Singer, 2003), modern *in vivo* fluorescent biosensors have their origins in a study demonstrating that a gene encoding a green fluorescent protein (GFP) in jellyfish *Aequorea victoria* could be expressed in other host organisms, was cofactor-independent (contrary to other contemporary methods) and did not interfere with cellular functions (Chalfie et al., 1994). Many different fluorescent proteins (FP) with similar properties but different emission spectra (”colours”) have since been discovered and improved by engineering (Newman et al., 2011). A common implementation of fluorescent reporter systems is based on coupling the sequence of a single FP to a gene or its promoter. By adding a FP sequence directly after the gene of interest (*FP-fusion tagging*) a gene product-FP chimera will be formed that can e.g. identify the subcellular localization of the protein and measure protein turnover (Newman et al., 2011). Alternatively, placing the

FP gene under control of a promoter allows for measurement of the expression levels of the promoter during different conditions (*FP-expression*) (**Paper V-VII**). Another type of biosensor design is to make use of FRET (Förster Resonance Energy Transfer), which is the mechanism of energy transfer between two fluorophores (e.g. FPs) in very close proximity, which can be used to monitor molecular interactions such as protein-protein and protein-DNA interactions, and conformational changes (Zadran et al., 2012). The general idea of FRET sensors is to design systems where the two fluorophores either come together or are separated by the molecular event of interest and thus results in a measurable change in signal (Newman et al., 2011; Zadran et al., 2012). A number of studies have demonstrated that *in vivo* biosensors can be used to monitor cellular signalling networks non-intrusively and in more or less real-time[20] (Newman et al., 2011). Many, but not all of these use FRET sensors.

Assessment of fluorescent protein biosensors is based on excitation of the fluorophore with a laser, followed by signal emission, and can thus be measured by a number of different approaches such as fluorescent microscopy fluorimetry and flow cytometry (Shapiro, 2005). The present work (**Paper V-VII**) used flow cytometry, which is a single-cell method where cells are lined up with the help of a fluidics system and measured individually[21] with 10 000-100 000 cells commonly being measured per sample (Shapiro, 2005). Furthermore, cytometry can be combined with Fluorescence-Activated Cell Sorting (FACS) to sort cells of interest based on fluorescent markers (Bonner et al., 1972), and thus used to select for cells with desired phenotypes. This can be beneficial for speeding up the iterative experimental work-flow of metabolic engineering (Figure 3). Sorting can however have negative influence on cell viability (e.g. due to sheath fluid chemistry and pressure) and system cleanliness is crucial to avoid contaminations (Müller and Nebe-von Caron, 2010).

Although biosensors can successfully be used to monitor cellular signalling networks, and flow cytometry can be used for high-throughput assessment of population dynamics, these biosensors are not an ome-wide method, and therefore it is likely that multiple sensors will be needed (preferably within the same cell) to capture a larger picture of the signalling. While the number of possible parallel fluorescent markers per cell are limited by excitation and emission spectra overlap, technological developments have increased the throughput of multi-parameter flow cytometry, with exam-

[20] FPs often require some time to mature (Katranidis et al., 2009), meaning that FP-expression signals are slightly delayed. The extent of the maturation is protein-dependent. Protein half-life is another concern, with e.g. the *S. cerevisiae* yEGFP3 having a half-life of about 7.5h. Engineered alternatives with shorter half-life exist, but tend to rely on degradation by ubiquitination which is an ATP-dependent process and therefore intrusive (Mateus and Avery, 2000).

[21] It is noteworthy that so called *Mass Cytometry* methods have also been developed, where MS is used to detect the signal instead of fluorescence emissions (Bandura et al., 2009; Spitzer and Nolan, 2016). Cells are however destroyed during the process, meaning that this method cannot be used for cell sorting (Saeys et al., 2016).

ples from FCM with up to 18 markers per (immune) cell and from mass cytometry with over 30 simultaneous (isotope-labelled rare-earth metal) markers (Chattopadhyay et al., 2008; Behbehani et al., 2012; O'Neill et al., 2013). Computational approaches to integrate multiple FCM datasets might also be able to increase the width of the signalling information that can be captured by biosensors (Welch et al., 2011).

4.2.3 Computational approaches

Since signalling networks connect the environment with the genome and the metabolism, there is large interest in reconstructing genome-scale *in silico* models of signalling networks (Hyduke and Palsson, 2010). However, signalling pathway reconstructions tend to be less mature than those of metabolic pathways (cf. Section 2.3.2), since signalling entities are difficult to elucidate from genome annotations and because there is a high heterogeneity in types and functions of signalling molecules (Hyduke and Palsson, 2010; Palsson, 2015). Another issue is that the signalling molecule kinetics needed to properly model signalling dynamics are mostly unknown (Imam et al., 2015).

Although there are many mathematical approaches to signalling network modelling (Rother et al., 2013), genome-scale approaches can generally be divided in stoichiometric and Boolean approaches, where the former considers the stoichiometry of the signalling reactions and latter sees the network as a series of switches and logical statements that can be "turned on or off" (Heath and Kavraki, 2009; Hyduke and Palsson, 2010). The different approaches has their strengths and weaknesses, but both seem to model small-scale signalling networks well, although a good method for genome-scale signalling has yet to emerge (Hao et al., 2018).

Like most of the signalling examples so far in this chapter, a majority of the modelling studies come from medicine, with a notable model for T cell signalling being one with the largest signalling reconstructions made (Li et al., 2009a). Examples of microbial signalling reconstructions include e.g. the glucose repression pathways (Christensen et al., 2009; Lubitz et al., 2015) and osmotolerance pathways (Klipp et al., 2005) in *S. cerevisiae* and chemotaxis signalling in *Escherichia coli* (Clausznitzer et al., 2010). However, for microbes, it seems to be more common to see signalling models integrated with reconstructions of metabolism and transcriptional regulation (Hao et al., 2018). Examples include *M. genitalium* (Karr et al., 2012) – one of the smallest known genomes (cf. Section 3.1.1) – and *E. coli* (Covert et al., 2008; Carrera et al., 2014). One of many challenges with integrated models is that signalling and metabolic reconstructions tend to rely on different mathematical approaches that can be challenging to combine in one single model; an example is that molecule concentrations are important in signalling models, but do not matter much for constraint-based FBA models of the metabolism (Imam et al., 2015).

4.3 Monitoring the sensing of xylose in *S. cerevisiae*

In the current thesis work, the subset of the signalome related to sugar sensing, uptake and utilization in *S. cerevisiae* was investigated with a cytometric single-cell *in vivo* biosensor approach (**Papers V-VII**). As such, the concept of the *sugar signalome* will be from now used to describe the three cross-talking signalling pathways that together govern sugar sensing in signalling in *S. cerevisiae*. The challenging nature of ome-wide biosensor assessments of signalling networks discussed above, does necessitate piecemeal-approaches such as this.

4.3.1 The xylose paradox and the *S. cerevisiae* sugar signalome

The present case study was performed in the same lignocellulose valorization context as described in Section 3.4, and concerns the peculiar response to xylose in *S. cerevisiae* strains genetically engineered to grow on this pentose sugar. Despite many successful engineering strategies to improve the fermentation of xylose to ethanol (reviewed in e.g. Moysés et al. (2016); Kim et al. (2013)), the cellular behaviour and previous transcriptomics and metabolomics studies suggests *S. cerevisiae* cells engineered for xylose utilization are tuned towards a non-fermentative response (Salusjärvi et al. (2008); Klimacek et al. (2010); **Paper VI**). The contradictory combination of xylose uptake and non-fermentable behaviour will hereby be referred to as the xylose paradox (**Paper VII**).

Glucose sensing is transduced in three different pathways in *S. cerevisiae*: the Snf3p/Rgt2p, SNF1/Mig1p and cAMP/PKA pathways (Santangelo, 2006); Figure 9. The Snf3p/Rgt2p pathway responds to extracellular glucose and induces expression of high- and low affinity hexose transporters (Ozcan and Johnston, 1995). The SNF1/Mig1p pathway controls carbon catabolite repression and induces genes for utilization of alternative carbon sources in the absence of glucose, the preferred sugar of *S. cerevisiae* (Gancedo, 1998). Finally, the cAMP/PKA pathway is a multifunctional signalling pathway that, in short, can be said to control cell homeostasis (by means of e.g. cell cycle progression control and stress signalling) (Thevelein and De Winde, 1999). These three pathways are subject to involved cross-talk (Kaniak et al., 2004; Gancedo et al., 2015), which emphasises the need to assess all three pathways together. The cAMP/PKA pathway is furthermore interconnected to and/or have overlapping targets with two other main signalling pathways, the TOR and MAPK pathways (Pedruzzi et al. (2003); Tamaki (2007); **Paper VII**).

4.3.2 The xylose signal in wild-type and recombinant *S. cerevisiae*

Glucose sensing is in general well-studied in *S. cerevisiae* (Santangelo, 2006), although some mechanisms are still not fully elucidated. Less is however known about if and

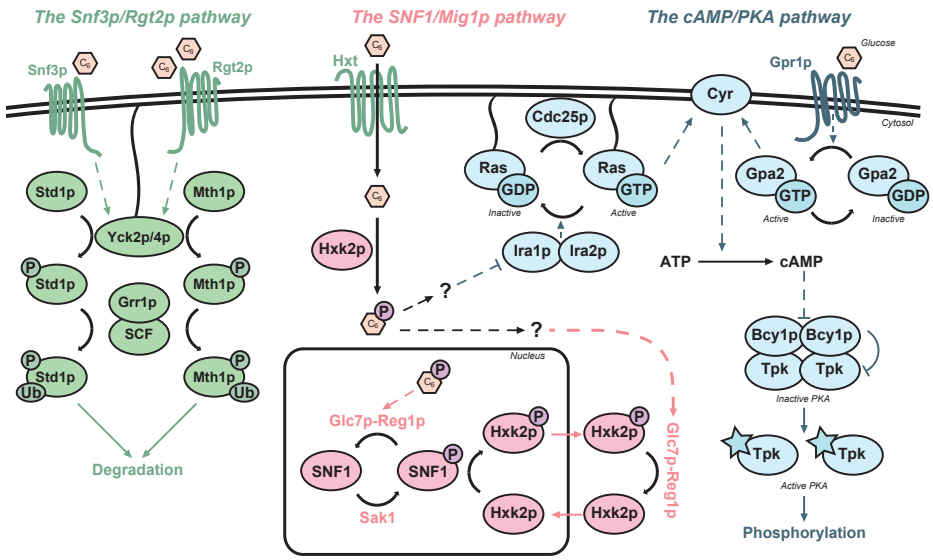


Figure 9: The three main sugar signalling pathways in *S. cerevisiae*, adapted from **Paper VII**. The **Snf3p/Rgt2p** pathway (green) regulates expression of hexose transporters in response to extracellular glucose. The **SNF1/Mig1p** pathway (pink) regulates expression of genes related to alternative (non-glucose) carbon sources in response to intracellular phosphorylated glucose. The **cAMP/PKA** pathway (blue) regulates a variety of responses, such as cellular growth, homeostasis and stress response.

how xylose affects the sugar signalome, since it is not part of the substrate-range of wild type *S. cerevisiae*. This case study therefore proposes a challenging issue in metabolic engineering: sensing and signalling of exogenously enabled substrates.

Sugar signalling in *S. cerevisiae* is rapid. It has for example been shown that, 20 minutes after addition of glucose to glycerol-grown cells, about 40% of the transcriptome has changed expression at least 2-fold (Wang et al., 2004). Because of this, an *in vivo* biosensor approach was chosen to examine the xylose signal in this yeast with a higher temporal resolution than omics (**Paper V**). An integrative single-copy reporter system based on FP-expression was designed where a yeast-enhanced GFP (yEGFP3) was placed under the control of different endogenous yeast promoters known to be regulated by each of the three sugar signalling pathways (**Paper V**).

In *S. cerevisiae* W303-1A laboratory strains lacking recombinant xylose pathways, no induction was found on xylose (**Paper V**), which indicated that *S. cerevisiae* was unable to sense extracellular xylose. However, population heterogeneities in some of the biosensor signals (**Paper V**) led to a hypothesis that *S. cerevisiae* might respond to xylose molecules that has been internalized by the cell with the help of glucose transporters known to transport some levels of xylose (Hamacher et al., 2002).

Following this hypothesis, the biosensor strains were engineered with a mutated

galactose transporter with improved affinity for xylose (Farwick et al., 2014) and an oxidoreductive xylose pathway (XR/XDH; xylose reductase/xylitol dehydrogenase). It was found that the previous ideas had some merit, as the engineered strains did show a signalling response during growth on xylose. However, xylose resulted in the opposite signal of that of glucose abundance (i.e. optimal glucose concentrations for growth), suggesting that signalling was indeed part of the xylose paradox (**Paper VI**). Following these results, a number of recently discovered deletions with positive impact on xylose utilization in strains with a xylose isomerase (XI) pathway (Sato et al., 2016) were introduced in the oxidoreductive pathway strains. It was found that most, but not all, claims of improvements from the XI strains were reproducible in the XR/XDH biosensor background, and that *ira2Δisu2Δ* altered the previous low-glucose signal of xylose to a simultaneous signal of high- and low glucose, reinforcing that signalling engineering is a promising strategy for improved xylose utilization (**Paper VII**).

Other xylose sensors have been developed for *S. cerevisiae* based on bacterial repressor proteins (XylR) (Teo and Chang, 2015; Wang et al., 2016; Hector and Mertens, 2017). They consists of a two-component system where XylR is constitutively expressed and represses the expression of GFP by binding to recognition motifs in a synthetic promoter; the addition of xylose to the cell results in repression of XylR, and induction of GFP (Teo and Chang, 2015). However, rather than monitoring endogenous signalling, these sensors have primarily been used to build synthetic circuits and library screening (Wang et al., 2016). From an applied point-of-view, a combination of the sugar signalome reporter-approach and introduction of synthetic signalling circuits could possibly be a powerful way to elucidate and engineer the signalling control points in xylose-utilizing *S. cerevisiae*.

The apparition of these faces in the crowd;
Petals on a wet, black bough.

EZRA POUND
In a Station of the Metro
in *Lustra* (1917)

(Because metabolism is a subway map waiting for its trains)

Chapter 5
Reflections from this thesis work

The previous chapters have outlined the key aspects of Large Data biology with detailed looks at the genome and signalome. This chapter will attempt to place the thesis work into the larger context and discuss how the challenges of Large Data biology were approached.

5.1 Large Data science and biology

While there is not a lack of examples of the benefits of Large Data science, this field is still surrounded by an intense debate. During the research of the present thesis summary, two main discourses seemed to come up more frequently. On one hand is the debate about the methodology itself, where for instance critical opinions have been raised that the current implementation of the scientific method is not scalable to Large Data (Peters et al., 2014). For instance that, at least in certain biological disciplines, scientists are trained to collect small datasets of high quality rather than applying large data methodologies, that re-use of existing large datasets is still quite uncommon, that there is a lack of *in silico* tools to facilitate a move towards data-intensive applications (Peters et al., 2014), and that a majority of the common statistical methods, such as the *p*-value, were designed with small data in mind (Manzoni et al., 2016). On the other hand is the debate of data-versus-theory (cf. Section 2.2.1). One data-centric argument is that ”more data > better data”; this is based on the idea that the noise in Large Data can be compensated for by increasing the data quantity, and that it is enough to find the general correlations in a large dataset instead of the underlying phenomena (Mayer-Schönberger and Cukier, 2013). This may be a valid mind set for e.g. search engines and online shopping (common examples in Mayer-Schönberger and Cukier (2013)), but is less attractive in biology since the mechanistic understanding is often the end-goal (Leonelli, 2014). On top of this is the fact that the cell operates in multiple dimensions with highly dynamic changes to environmental stimuli, which makes the capture of high spatio-temporal resolution of the cell a massive experimental undertaking that requires interdisciplinary approaches (Figure 10); i.e. ”complete”

data collection from the cell is difficult to achieve. This is in stark contrast to famous cases such as Google's prediction of the seasonal spread of the flu (Ginsberg et al., 2009), which was based on one "ome" – the search queries. This is not to say that search engine data and the human psychology it conveys is a simpler problem than cell biology, but it highlights that the complexity of the cellular network – which we possibly only have started to grasp – makes Large Data biology extra challenging.

5.2 Bibliomes as part of Large Data biology

Biology is full of Large Data, from the different omes within the cells to each individual cell living together in a population. So is also the collected knowledge about these Large Data-generating layers – the bibliome. **Paper I** deals with an important question in Large Data biology: how to handle the literature. Starting out as a smaller microbial section in a review on lignin biovalorization (Abdelaziz et al., 2016), it was quickly realized that it is not possible to present a bibliome in a review article in a way which allows the readers to easily extract the desired information, and that a computerized database which could be queried interactively would instead be preferable. The challenges in database design and curation revolve around data quality and -organization (Helmy et al., 2016). In the case of **Paper I**, the issue was that the database was designed to store multiple type of information about microbial aromatic catabolism, which makes the information more difficult to de-contextualize and standardize than if the database only concerned, say, DNA sequences. The varying level of experimental description in different publications and the sometimes limited access to pre-digital papers also complicated the curation, meaning that some of the data had to be annotated with the caveat that the end-user must refer to the original reference prior to drawing conclusions about the data.

The database discussed in **Paper I** is a useful resource for finding microbes holding certain aromatic pathways, but also shows that it is difficult to get an understanding from literature about how lignin degradation works in nature/in soil, since the majority of the reported studies are on single isolate level and not on community level. Given that only about one percent of all soil bacteria have been estimated to be cultivable (Pham and Kim, 2012), the current knowledge of microbial lignin catabolism comes from a cohort so small that it seems unlikely to be representative of the aromatic processes in microbial soil ecology. The database also illustrates the role of fashions within the scientific community and how certain approaches and topics tend to be favoured over others. Here bibliome studies of microbial ecology reach their limit, and begin to say more about how research is conducted than the natural diversity of microbes. This does not affect the usefulness of the database as a resource for the current knowledge-base on the topic, though.

Although the database has a good potential to make an impact on the commu-

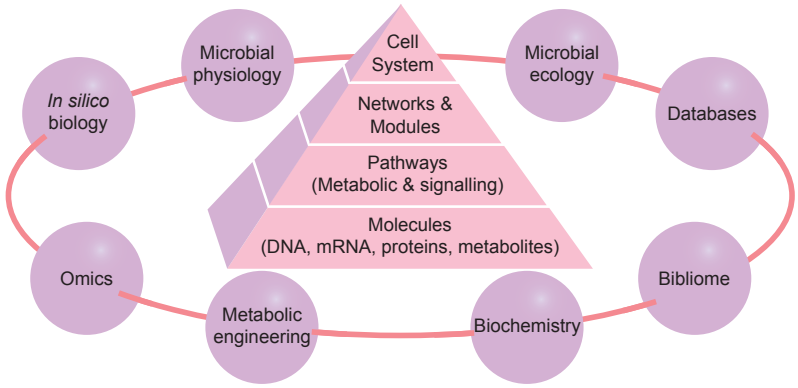


Figure 10: Large data biology from the point-of-view of this thesis. The pyramid was adapted from Oltvai and Barabási (2002) and represents the cellular organization, with increasing complexity from bottom to top. The satellite orbs represent the various techniques that commonly comes together with the content of the pyramid in data-intensive biology.

nity of this field, new challenges related to maintenance and upkeep emerge now that launch-phase has been completed with the publication of **Paper I**. While more automation can be implemented, staff will always be needed for the final curation and technical support. A desired end-goal would be to involve more community activity in the platform, and a step towards that would be to involve other research groups in the database work. Should the database need to be discontinued, the plan is to try to integrate the data with other databases and to deposit the content of the database in an online archive as per recommendation (Helmy et al., 2016). The publication of **Paper I** also serves as an archive of the hitherto work, as it collects the main references and meta-analysis from the database at the time of its publication.

5.3 Whole-Genome Sequencing

The statement that omics is the quintessential Large Data method in biology (Leonelli, 2014) also implies that omics comes with the typical challenges associated with large data biology (Table 1). With the current maturity of HTS techniques, the central issue in whole-genome sequencing is the assembly and read-mapping of the raw reads (Section 3.3). Depending on the DNA source and choice of sequencing method, different degrees of customization of the bioinformatics pipeline may be required, which not only requires a general understanding of the algorithms, but also of programming. In this thesis work, the latter part was approached by learning to program in the scripting languages listed in Table 2, which later enabled the development of the database (**Paper I**), since programming logic is applicable across languages.

Paper II shows that PacBio sequencing data is applicable for assembly of small

genomes as long as it is performed at suitable coverage, and that the outcome of *in silico* annotation pipelines can lead to powerful clues about putative pathways. That study also illustrates the problem of experimentally verifying *in silico* predictions from strains incapable of producing sufficient biomass in the conditions needed to confirm the putative functionalities. In the current study it was not possible to do mRNA analysis (RT-qPCR or RNAseq), knockout studies or enzyme assays since the *Microbacterium* isolate did not produce enough biomass when grown on aromatic compounds. Instead, the function of the proposed candidate genes would need to be confirmed in an exogenous host, which would be a large study on its own. Within this thesis work, another study (**Paper III**) demonstrates how a phenotype can be correlated to a genotype by using custom genome assembly and annotation to find candidates that can be validated experimentally. **Paper III** also shows how putative annotations in databases might contain incorrectly predicted gene function, and that experimental validation can lead to improved reannotations (here: *calA* from *Pseudomonas putida*). Not only were the encoded enzyme activities assessed by overexpression in *E. coli*, but gene deletions in a *P. putida* host strain also confirmed the new annotation.

The WGS described in **Paper IV** enabled us to show that the tolerance towards lignocellulosic inhibitors and elevated temperature in the adapted isolate ISO12 was a function of mutations in multiple genes, many of which related to cell periphery- and stress mechanisms controlled in part by the cAMP/PKA signalling pathway. The fact that this was a two-factor adaptation (temperature and inhibitors) complicated the identification of driver mutations, and it is likely that some of the variants that were selected for resulted in a combined change in phenotype to both selection pressures at once. This could explain why other ALE resequencing projects with similar scope seemed to have found slightly different genes than in this study.

A rather high amount of sequence variants were found between ISO12 and its parental strain ER (**Paper IV**), and it would have been helpful to have sequenced some of the intermediate clones lacking the desired phenotype in order to identify "junk" or hitchhiker mutations. Sadly, the clones isolated throughout the ALE experiment did not have a stable phenotype, and thus this was not possible. Something that *could* be reassessed with the existing data is however the intragenic variants. Granted, it is much more challenging to functionally annotate intragenic variants compared to ORF variants, but the substantial amount of intragenic polymorphisms in ISO12 (**Paper IV**) suggests that regulatory modifications in e.g. promoter and terminator regions might have occurred in addition to the ORF variants.

Tolerance to lignocellulosic inhibitors is an important trait for industrial microbes designed to ferment this feedstock. However, research on the pre-processing of lignocellulose has shown that the levels of furans and phenols can be drastically decreased by modified processing methods (Jönsson and Martín, 2016). It is also unlikely that the adaptive capability of the yeast cell is infinite (despite the high plasticity of its

genome). Therefore, a good approach to this problem would be to combine adapted strains with inhibitor-sparse pre-processing methods. Since stress tolerance is strain background dependent (Zhang et al., 2019), it would be interesting to conduct the ALE again with multiple different parental strains to see whether there are any reproducible driver mutations that would occur independent of strain background.

5.4 *In vivo* biosensors for signalling networks

As established in Chapter 4, omics methods have successfully been used for signalling studies, but fail to capture the temporal dynamics of signalling since they are not real-time methods. To study the overall sugar signalome response to xylose, *in vivo* biosensors based on GFP-coupled endogenous promoters were chosen rather than individual mechanisms in the signal cascade. This presented a number of challenges in sensor design and data analysis.

The choice of promoters was based on known interactions and transcriptomics data from literature. The Snf3p/Rgt2p and SNF1/Mig1p pathways have been well studied (Santangelo, 2006) and thus their target promoters were easily identified. The cAMP/PKA pathway however, proved more complex, since it regulates such a broad range of genes and cellular responses (Thevelein and De Winde, 1999). Finally, two trehalose-6-P synthases (TPS1/2) and one translational elongation factor (TEF4) genes were chosen based on transcription data (Apweiler et al., 2012). However, since the hypotheses about xylose signalling has gone more and more towards the cAMP/PKA pathway (**Papers VII-VII**), it would be interesting to see if the biosensor panel could be complemented with other types of cAMP/PKA sensors to increase resolution. **Paper VI** also showed the importance of performing xylose signalling studies in yeast engineered for xylose utilization, since the addition of a transporter and the oxidoreductive pathway enabled signals from intracellular xylose.

Central to the flow cytometry (FCM) data analysis is the concept of gating: division of measured events (cells) in different groups/populations based on signal intensity in two-dimensional scatter plots (O'Neill et al., 2013). Gating is traditionally a manually-performed graphical method and will thus always contain human bias, subjective selection of target cells and be difficult to reproduce (Saeys et al., 2016). Algorithms are also biased since they are coded by humans, but can be automated, meaning that they result in a systematic bias rather than the more random bias of manual gating. To this end, a cell size regression method (Knijnenburg et al., 2011) was used in **Paper V** to compensate for differences in cell size and morphology when the biosensor strains were grown on different carbons sources. A fully automated pipeline was scripted in Matlab to systematically asses the FI of all samples collected in the project in one run. However, signal normalization of this kind will disguise population heterogeneities, and the data was also run in a pipeline lacking the normalization step,

which revealed the subpopulations when grown on xylose which led to the hypothesis of endogenous xylose sensing that was investigated in **Paper VI**. This illustrates the importance of comparing the outcome of a normalization with non-normalized data. The regression model was discarded in **Paper VI-VII** in favour of overlay histogram plots and subpopulation modelling (Gaussian mixture modelling; **Paper VI**) because it concealed the population dynamics that turned out to be a central finding in these three papers. Another emerging approach in FMC bioinformatics is to use machine learning algorithms to e.g. identify populations in an automated manner, which may be useful for future studies with these biosensors (Saeys et al., 2016). Algorithms for multidimensional FCM data analysis exist, where every parameter measured by the instrument is used instead of the normal step-wise two-parameter-approach of traditional gating (Spear et al., 2017).

Functional as it has proven to be, the reporter system demonstrated in the present thesis work is not without its limitations. The current generation of biosensor strains only has one sensor per strain, meaning that simultaneous measurement of the three pathways is currently not achievable in single-cell. This could be e.g. improved upon by designing a system of multiple fluorophores with non-overlapping emission spectra that could multiplex the sugar signalome, or to use the biosensors to screen for conditions that generate interesting signals that that be further assessed by signalling omics methods. Another limitation is that the system is based on the actuators of the signalling pathway (i.e. the promoters that receive the signal), which means that this reporter system only measure the end-point signalling. This is a fair approach from a biological point-of-view, since the effect of xylose on the cellular behaviour was the end-goal (i.e. protein expression), but it cannot resolve the upstream transduction mechanisms in the pathway. A compliment to the current biosensor setup could therefore be to build FRET sensors for targets in the sugar signallome that are hypothesised to be of extra importance for xylose signalling. Such an approach could possibly elucidate if the signalling networks themselves need to be engineered to enable xylose to be sensed in a manner equal to glucose. Finally, these biosensors are obviously not capable of systems-wide screening (they only look at the sugar signalome), meaning that other systems level effects, including cross-talk with other signalling pathways, cannot be ruled out.

Other valuable information on how xylose signalling functions in yeast could be obtained from applying the current biosensors on natural xylose utilizing yeasts like *Scheffersomyces (Pichia) stipitis* and see how the signal differs. It is known from metabolomics data that the sugar and energy metabolism differs between *S. stipitis* and xylose pathway-engineered *S. cerevisiae* (Shin et al., 2019), which are also likely to be reflected in the signaling. An engineering dream (possibly a pipe dream) would be transplantation of the *S. stipitis* sugar signalome into *S. cerevisiae*, but that would first require mechanistic elucidation of the *S. stipitis* sugar sensing networks to levels

far beyond the current understanding. Hexose-pentose co-consumption is a crucial trait for lignocellulose fermenting cell factories as it will dramatically reduce process times, and signalling engineering is likely to be an important means to this end.

5.5 System boundaries

The division of the cellular activities into omes results in models that do not take the whole system into account. It is therefore imperative to remember that the approaches of this thesis, where cellular activities were investigated from a physiology, genome and signalome point-of-view is a constrained approach to the biology of the cell. This does not mean that said approaches are inadequate – on the contrary they have historically been instrumental in reaching the current level of understanding of the cell – but again shows that omics are not a silver bullet that can resolve the research questions of the cell and that it as a technique has clear system boundaries of its own. The current development of techniques and models for integration of multi-omics data (Macaulay et al., 2017; Hao et al., 2018) is therefore an important move towards being able to make more complete systemic observations of the cell.

Another central question is where the system boundaries for cell factory improvements are. Whereas there is a consensus (this thesis included) that metabolic engineering and adaptive evolution can push the capabilities and capacities of a strain towards more industrially desirable traits, there seems to be less work done on trying to predict of *how far* they can be pushed and at what trade-off[22]. Take the case of **Paper IV** for instance: the aim of isolating a strain with improved tolerance to temperature and inhibitors was reached, but the strain was also highly flocculent and biofilm forming, which are undesired traits for industrial fermentation. It is likely that the latter was related to the former, meaning that it can be difficult to separate desirable phenotypes from the undesired. The boundaries on how far a cell factory can be improved have direct implications on the applicability and technology readiness levels of cell factories. For instance, successful replacement of fossil fuels with renewable alternatives (e.g. the bioethanol of **Papers IV-VII**) is contingent on fermentation processes with minimized costs and high yield and productivity. The theoretical stoichiometric yield is simple to calculate, but the equally important maximum theoretical productivity is much more elusive. However, a recent mathematical framework based on dynamic FBA has demonstrated that maximum theoretical productivity in bacterial succinate production can be predicted (John et al., 2017). Such models that predict the engineering/optimization boundaries of cell factories have large potential to inform the choice of host and product, and estimate the potential of metabolic engineering projects.

[22] Examples towards this end include e.g. thermodynamic-based FBA models (Henry et al., 2007)

Whereof one cannot speak, thereof one must be silent.
[Wovon man nicht sprechen kann, darüber muss man schweigen.]

LUDVIG WITTGENSTEIN
The final statement of
Tractatus Logico-Philosophicus (1922)

Chapter 6
Outlook and concluding remarks

The advent of high-throughput omics methods in the mid-2000s (at the time of writing, about 15 years ago) ushered in a new approach to molecular biology where the cell has begun to be considered from a holistic point-of-view – *begun* in the sense that although the systems view is now widely spread and accepted, the data collection, management and synthesis for systemic assessments of the whole cell (and not just one or two omes) is still in its infancy. Whether this data-intensive approach has changed biological research to the extent that it can be called a new paradigm is a subject of debate (Callebaut, 2012; Leonelli, 2010), but it is clear that Large Data has had big implications on molecular- and cell-biology. Most important of all is probably how technical developments have dramatically cut the price of a high-throughput run (especially for sequencing methods) and thus have democratized Large Data biology by making it affordable to even the smallest labs (McPherson, 2014).

Furthermore, the data-intensive and systemic views of the cell clearly demonstrate the immense complexity of cell biology and stresses that the system boundaries of the current omics methods are more constrained than the system boundaries of the cell – a realization that most scientists working with Large Data biology experience sooner or later. As such, this approach to molecular and cellular biology can be seen as the Socratic paradox in play: the more you learn, the more you understand how little you know (i.e. how much more there is to learn). With new methods for single-cell multi-omics (Macaulay et al., 2017) and initiatives to go from single-genome to pan-genome to study species diversity (Vernikos et al., 2015), the knowledge of cellular biology will only continue to increase in complexity, and so will our need to comprehend the complexity of this research subject.

In terms of outlook, the three Large Data themes discussed in the present thesis (databases, genomics and signalomics) are very likely to remain as central elements of Large Data biology. **Databases** have always been intrinsic to science (irrespective of the form they have taken: encyclopaedia of chemical constants, library catalogues or online data repositories) and will become increasingly important as high-throughput methods are developed and refined. As the scientific community is slowly moving to-

wards increased transparency – with more emphasis from journals and funding agencies to publish manuscripts, programming code and raw and processed data openly accessible – improved database systems are inevitably needed. Throughout the present thesis contributions, raw WGS data, genome assemblies and annotations have for instance been deposited at NCBI and some of the flow cytometry data have also been stored in a dedicated repository (Spidlen et al., 2012).

Genomics is the most standardized of the high-throughput methods (Leonelli, 2014), which is one of the reasons why analysis and storage of genomics data is so mature. During the time of the present thesis work, the third generation of sequencers have been commercially established and started to fill a well needed niche of super long-reads that really complements the short-read MPS methods. Being a new and hot method, TGS is currently about as expensive as MPS was when **Paper IV** was written (>3000€), whereas the more recent MPS study described in **Paper III** was just a couple of hundred Euros. The substantially lower cost and error rates of MPS will likely ensure that MPS will not go out of fashion anywhere soon.

The development of the bioinformatics field is progressing in response and in parallel to the instrument developments. Not only have the algorithms become better, but more and more new software that fills niches that the user previously had to script on their own have emerged. The scene for free-for-academic-use bioinformatics software is thriving, which directly plays into the accessibility, affordability and transparency required for a democratic Large Data biology. These software and algorithms are also often highly validated and state-of-the-art, making many of them well adapted to data-intensive projects.

Signalling studies have also benefited from omics and other high-throughput approaches. The highly transient nature of signal transfer calls for methods such as fluorescent reporter systems that can monitor the signal in higher temporal resolution (preferably in real-time) than what is normally possibly labour-wise by assessment of signalling molecules with proteomics and metabolomics. The drawback of fluorescent reporters is however that they fail to capture the whole signalome, as only a limited number of simultaneous fluorescence proteins are possible in one cell at once due to spectral overlap of the emitted signal. Future systems biology ventures may possibly act as drivers for technological developments of ”signalomics”; for instance, a fully representative *in silico* model of the cell will require signalling modelling (Hao et al., 2018), which implies the high significance of continued studies of the signalome, and here integrative biosensor-omics methods could complement each other. Likewise, the signalling approach to the xylose-paradox in *S. cerevisiae* has good potential to be a substantial step towards the applicability of this yeast as a cell factory for lignocellulose biovalorization; this view is supported by a recent review that called for increased use of a systems regulations approach to xylose engineering in *S. cerevisiae* (Gopinarayanan and Nair, 2019).

In closing, this present thesis has illustrated how Large Data biology can be used as a forward momentum in molecular biology to increase the understanding of microbial cells – especially when combined with the work and knowledge of specialists from different fields (Figure 10). The present work has attempted to establish the importance of using data-intensive biology in concert with traditional biological approaches: theory first; then experimental design, Large Data biology and data analysis; then experimental verification. Large Data biology for the sake of Large Data biology may have been important in the formative years of the respective methodologies to demonstrate their possibilities and accuracies, but today, when the methods are well-established, they should instead be anchored within the knowledge-base of cell biology. This will be imperative for the sustainability of the research in this field in the long-run, especially since one thing is certain: Large data biology is not leaving anytime soon.

No! I am not Prince Hamlet, nor was meant to be;
Am an attendant lord, one that will do
To swell a progress, start a scene or two,
Advise the prince; no doubt, an easy tool,
Deferential, glad to be of use,
Politic, cautious, and meticulous;
Full of high sentence, but a bit obtuse;
At times, indeed, almost ridiculous -
Almost, at times, the Fool.

T.S. ELIOT
The Love Song of J. Alfred Prufrock,
in *Prufrock and Other Observations* (1917)

Acknowledgements

There is a scene towards the end of Jean-Luc Godard's debut film *À bout de souffle* (1960) where Patricia, an aspiring journalist, goes to an open-air terrace of an airport (!) to interview a famous writer. After many attempts of getting her question through, through the dense flock of noisy journalists, she finally catches the writer's attention and this memorable exchange occurs:

"Quelle est votre ambition dans la vie?"
"Devenir immortel et mourir."

[-*What is your greatest ambition in life? -To become immortal and then die.*]

This dialogue about how it is possible to create things with a life and longevity of their own was forever etched into my mind when I first saw this film during my high school years and had just discovered my passion for writing. Although I would personally say it is the creative process itself, rather than a final product or result that gives meaning and satisfaction in life, the impression this scene made has always stayed with me: that writing and publishing texts is a form of non-corporeal immortality. As I write these final pages on my thesis, I think of this scene and the permanence of the written word, but also of another type of immortality that I have come to know during the years of the present work – a much more important one: the memories of the people you have met and worked with. In the process leading up the realization of this thesis, I have met and worked with many people to whom I will be forever grateful.

To Marie (my supervisor), for everything. I am in debt to you. My editor, my sounding board, my benefactor, my mentor, my scientific role model – I who normally consider myself to be somewhat of a wordsmith, struggle to capture the gratitude that I have always wanted to put into words and give to you. But what else is needed to say than that you have changed my life and all for the better. Perhaps words cannot do justice all that you have done for me and given me. Perhaps that is exactly what I wanted to say.

To Gunnar (my co-supervisor), for the support and the state-of-mind. You are probably one of the most thoroughly kind persons I have ever met. You demonstrate that it is possible to have a positive mind-set without being naïve, and your ubiquitous wordplay shows that there is always room for some pun and games even in the most boring of meetings. Thank you for your support and for your world-view.

To Bärbel, for the encouragement and the appreciation. From the beginning of my doctoral studies, you have always given me your heartfelt support, although you

never really were involved in my projects *per se*. Your appreciation for my work has really meant a lot to me – probably more than you can imagine – and I will always remember your kindness.

To Celina, for the discussions. Not only have we known each other since my first day at the engineering programme in biotechnology almost exactly ten years ago (where you were one of the sophomores welcoming the new students), but we have also worked closely together for five years. We are both stubborn people (at least so I have heard), and we have butted heads many times – sometimes gotten on each other's nerves – but I am very happy that we finally found a good wavelength. Thank you for the discussions, for the feedback and for the nerdy small talks.

To Karen, for the hard work. Having been slightly longer at then division than you, I have always seen you as a little sister (and I have sometimes teased you like you would amicably tease a sibling). Your apparition has an air of naïve charm, but underneath show a very skilled knowledge in the science that you do. Though we will likely live in opposite parts of the world from now on, I really hope that we will meet again soon.

To Nina, my previous student and now dear friend. I had the pleasure to supervise you for about eight months during the end of your MSc studies, and was truly happy that the work we did together sparked an interest in research that made you apply to, and get, a PhD student position at the division. It has been amazing to see you grow during the two years since we first met and I hope that we still have a lot of moments ahead of us. I admire your courage in everything you take on, and I am very grateful for the support you have given me when I have needed it.

To Gonzalo and Viktor, my other two students. Gonzalo, your diligent way of working and burning ambition for science are truly admirable and contagious. No one was happier than me when you got your PhD position, and no one was surprised: your passion for research is inspiring. Viktor, you are a very bright person and a very promising scientist; I hope that I was able to light a spark of the passion for research in you.

To Krithika and Javier, for the collaboration we did on the lignin project. Krithika, you truly are a force of nature, and when you want something done, you really go all out; you are very inspiring in that way. Javier, although you don't really make a fuss about it, you are a very skilled molecular biologist and I am happy to have gotten the chance to work on the bacterial lignin projects with you.

To Peter, for the discussions about microbiology and academic careers, and the interim supervision during my second year. To Lisa, for the brainstorming, sound boarding and company in the lab; to Magnus, for the discussions about science in general and flow cytometry in particular; to Jenny, for the care, encouragement and kindness; to Valeria, my Master Thesis supervisor, who taught me the craft of microbiology both during my second and fifth year of the biotechnology programme -

thank you for taking me on as a MSc student during the hectic final year of your PhD studies.

To Catherine, for always being so supportive (despite that I mostly took up your PhD student's time by drinking coffee with her for hours on end...); to Maja, for teaching me a lot about the real world outside of books and novels – I hope I was able to teach you something in return; to Johannes for the calmness and for help with the RT-qPCR; to Anders, for helping with getting started with the molecular biology after coming from the bioinformatics and fermentation work of my Master Thesis.

To Eoin, for the rants, the silly discussions and the *useless-fact-of-the-day* – I really enjoyed sharing the office with you; to Ed, for the discussions on art and popular culture, and for lending me many of your books; to Anette, for all the assistance with paper work, orders and admin; to Christer, for the help with resurrecting dead machines and meeting room silver screens.

To Linda, for your kindness; to Yasmine, for all your efforts towards the social atmosphere at the workplace; to Malin, for the company in the lab (I am still sorry for how long I pestered the lab with the beeps of the colony counter during the work on Paper IV); to Kristjan for the hiking and the film-making session; to Petra, for all the efforts with keeping the labs clean and for helping me with the lab coat laundry; to Paula, for the many insights into the art of typography.

To Omar, Kena, Jens and Bob, Per, Henrik, Lotta, Maggan, and Christian of the lignin group for the good work and fruitful collaborations.

To the past and present members of the division of Applied Microbiology, including my past and present office mates Nikoleta, Alejandro, Jan, Sebastian, Arne, Sofia, Tina and Raquel, and previous members of the yeast group (Diogo, Venkat, Violeta, Kaisa).

To all the students I have had the pleasure to teach and instruct in the classrooms and course labs of Kemicentrum.

To my mother Eva, my father Karl-Johan, and my younger sisters Hanna and Emma. For everything.

To Sandy, my best friend. Who would have known five years ago that the most important take-away of the doctoral studies would be friendship? Who would have known that platonic love could be so strong? For you, I could fill the rest of this book with my gratitude. But words are superfluous. We have already said them all.

Appendix I - Bioinformatics glossary

Table A1: Glossary for some key concepts in the genomics terminology.

Term	Explanation
Annotation	The process of predicting and identifying features in a nucleotide or amino acid sequence. Genome annotation can resolve features such as genes, psueo-genes, rRNA, etc.
Assembly	The process of compiling a longer sequence (e.g. a whole-genome sequence) from smaller sequences (reads). Can be done with or without a template genome (reference assembly and *de novo* assembly, repspectively).
Base calling	The process of determining the identity and order of nucleotides during the sequencing. Base calling quality is commonly measured as the probability of a incorrect base calls in a given sequence length (PHRED-score).
Contig	Short for contiguous sequence. A coherent sequence of DNA that is generated in an assembly by piecing together overlapping reads, supported by high confidence levels. Assemblies commonly consists of multiple contigs.
Coverage	The average number of sequenced fragments (reads) that support a certain nucleotide position. For example, a coverage of 30x means that on average, each position in the contig was supported by 30 overlapping reads. A good assembly has an uniform coverage throughout the contigs. Regions with overly high coverage are indicative of sequencing/assembly errors. Calculated as $c = L * N/G$, where coverage (c) is a the average read length (L) times the number of reads (N) over the haploid (one copy of each chromosome) genome size (G) (Sims et al., 2014).
de Bruijn Graphs	A mathematical concept used in modern short-read genome assembly algorithms. Very simplified (see also Box II): reads are divided into substrings of length *k* (*k*-mers), and alignment is calculated based on constructing a de Bruijn Graph where each *k*-mer is connected to two different nodes of size *k*-1 (one for each end of the *k*-mer; "left" and "right" *k*-1-mer) which eventually forms a graph of the relationships of the *k*-mers in the genome. This cirumvents the need for pairwise alignment of each *k*-mer and signficantly decreases the computational burden. Please refer to Compeau et al. (2011) for an in-depth primer on de Bruijn Graphs in genome assembly. Different implementations of this method in different MPS assemblers have been reviewed in Miller et al. (2010).
Greedy algorithms	An assembly method that finds read overlap based on always selecting the optimal (highest scoring) overlap for each local alignment. Used in older MPS assemblers (Miller et al., 2010)

Haplotype estimation	A method to infer the sequence of alleles in polyploid genomes, e.g. from sequencing reads. Can be used in connection with variant calling in polyploid organisms to determine the allele frequency of the variants (homozygous and heterozygous variants). Also known as haplotyping or haplotype phasing.
K_a/K_s	Estimate of the selection strength during DNA evolution based on the ratio of the number of non-synonymous substitutions (K_a) and synonymous substitutions (K_s) in an ORF. $K_a/K_s > 1$ implies a positive selection strength, $K_a/K_s = 1$ a neutral selection strength, and $K_a/K_s < 1$ a negative selection strength. Sometimes denoted as ω.
***k*-mers**	All possible substrings of length *k* found in a string. In genome sequencing: division of a read into smaller nucleotide sequences of size *k* to facilitate identification of overlap. Used in e.g. de Bruijn Graph assembly (Box II). *k* is commonly around 30-50 bp. Example: Read: ATGGCGTGCA (10bp) 3-mers: ATG, TGG, GGC, GCG, CGT, GTG, TGC, GCA 8-mers: ATGGCGTG, TGGCGTGC, GGCGTGCA 10-mers: ATGGCGTGCA
ORF	Open Reading Frame. Used in genome annotation to descibe putative genes/coding sequences (CDS).
PHRED-score	A base calling quality metric, expressed as the probability of an incorrect base call in a given sequence. A score of 30 corresponds to a probability of 1 incorrect base call in 1000 bases, and a score of 40 (preferred threshold) corresponds to 1 in 10000.
Pipeline	A computational work-flow consisting of a set of software combined in a chain; for instance, the scripts and algortihms needed to run the assembly workflow in Figure 6. Normally consists of a number of different programs that need to be connected, often by adapting the output format of one program to fit with the input format of the next.
Read	(Noun) A sequence of base pairs determined by the sequencer, corrsponing to (a subset of) the DNA fragment that was used as a template.
Scaffold	A colletion of contigs and gaps that together descibe a longer portion of a genome sequence.
Sequencing-by-synthesis	Method that uses a DNA polymerase to incorporate nucleotides based on a DNA template. Labeled-nucleotides allows monioring of the sequence as it is elongated. Examples include Sanger sequencing, Illumina, 454 and IonTorrent.

Variant	General term for a changes in a sequence compared to a control sequence. Semantically similar to *mutation*, but the term variant is preferred until experimental evidence is in place. Variants can be synonymous (silent; no change in the polypeptide) or non-synonymous (non-silent; changed polypeptide). On top of that, variants can be classified as e.g. Single Nucleotide Polymorphism (SNP; point-mutations), INDELS (insertions/deletions) and structural variants (e.g. copy number variations) depending on the length and complexity of the variant. Other examples include frameshift variants and premature stop codons.
Variant calling	The process of determining variants. Usually done by mapping reads to a reference sequence and identifying statistically supported sequence variants.

References

Abdelaziz, O. Y., Brink, D. P., Prothmann, J., Ravi, K., Sun, M., García-Hidalgo, J., Sandahl, M., Hulteberg, C. P., Turner, C., Lidén, G., and Gorwa-Grauslund, M. F. (2016). Biological valorization of low molecular weight lignin. *Biotechnology advances*, 34(8):1318–1346.

Abejón, R., Pérez-Acebo, H., and Clavijo, L. (2018). Alternatives for chemical and biochemical lignin valorization: Hot topics from a bibliometric analysis of the research published during the 2000–2016 period. *Processes*, 6(8):98.

Aderem, A. (2005). Systems biology: its practice and challenges. *Cell*, 121(4):511–513.

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in illumina sequencing libraries. *Genome biology*, 12(2):R18.

Al-okaily, A. A. (2016). HGA: *de novo* genome assembly method for bacterial genomes using high coverage short sequencing reads. *BMC genomics*, 17(1):193.

Almario, M. P., Reyes, L. H., and Kao, K. C. (2013). Evolutionary engineering of *Saccharomyces cerevisiae* for enhanced tolerance to hydrolysates of lignocellulosic biomass. *Biotechnology and bioengineering*, 110(10):2616–2623.

Almeida, J. R., Modig, T., Petersson, A., Hähn-Hägerdal, B., Lidén, G., and Gorwa-Grauslund, M. F. (2007). Increased tolerance and conversion of inhibitors in lignocellulosic hydrolysates by *Saccharomyces cerevisiae*. *Journal of Chemical Technology & Biotechnology: International Research in Process, Environmental & Clean Technology*, 82(4):340–349.

Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E. B., and Müller-Myhsok, B. (2012). A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Human genetics*, 131(10):1541–1554.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Anderson, M. (2004). NIH offers $1000 genome grant. *Genome Biology*, 4(1):spotlight–20040223.

Antipov, D., Korobeynikov, A., McLean, J. S., and Pevzner, P. A. (2015). hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*, 32(7):1009–1015.

Apweiler, E., Sameith, K., Margaritis, T., Brabers, N., van de Pasch, L., Bakker, L. V., van Leenen, D., Holstege, F. C., and Kemmeren, P. (2012). Yeast glucose pathways converge on the transcriptional regulation of trehalose biosynthesis. *BMC genomics*, 13(1):239.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

Bailey, J. E. (1991). Toward a science of metabolic engineering. *Science*, 252(5013):1668–1675.

Bairoch, A. (2000). The enzyme database in 2000. *Nucleic acids research*, 28(1):304–305.

Baker, M. (2012). *De novo* genome assembly: what every biologist should know.

Baker, M. (2013). Big biology: the 'omes puzzle. *Nature News*, 494(7438):416.

Bandura, D. R., Baranov, V. I., Ornatsky, O. I., Antonov, A., Kinach, R., Lou, X., Pavlov, S., Vorobiev, S., Dick, J. E., and Tanner, S. D. (2009). Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Analytical chemistry*, 81(16):6813–6822.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5):455–477.

Bao, E., Jiang, T., and Girke, T. (2014). AlignGraph: algorithm for secondary *de novo* genome assembly guided by closely related references. *Bioinformatics*, 30(12):i319–i328.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2012). NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, 41(D1):D991–D995.

Bashir, A., Klammer, A. A., Robins, W. P., Chin, C.-S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., Peluso, P., Sebra, R., Sorenson, J., Bullard, J., Yen, J., Valdovino, M., Mollova, Emilia nad Luong, K., Lin, S., LaMay, B., Joshi, A., Rowe, L., Frace, M., Tarr, C. L., Turnsek, M., Davis, B. M., Kasarskis, A., Mekalanos, J. J., Waldor, M. K., and Schadt, E. E. (2012). A hybrid approach for the automated finishing of bacterial genomes. *Nature biotechnology*, 30(7):701.

Bastow, R. and Leonelli, S. (2010). Sustainable digital infrastructure: Although databases and other online resources have become a central tool for biological research, their long-term support and maintenance is far from secure. *EMBO reports*, 11(10):730–734.

Behbehani, G. K., Bendall, S. C., Clutter, M. R., Fantl, W. J., and Nolan, G. P. (2012). Single-cell mass cytometry adapted to measurements of the cell cycle. *Cytometry Part A*, 81A(7):552–566.

Bell, G., Hey, T., and Szalay, A. (2009). Beyond the data deluge. *Science*, 323(5919):1297–1298.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2017). Genbank. *Nucleic acids research*, 45(D1):D37–D42.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V.,

Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E. Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O' Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., vandeVondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53.

Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., and Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology*, 33(6):623.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235–242.

Blazeck, J. and Alper, H. (2010). Systems metabolic engineering: Genome-scale models and beyond. *Biotechnology journal*, 5(7):647–659.

Boetzer, M. and Pirovano, W. (2012). Toward almost closed genomes with GapFiller. *Genome biology*, 13(6):R56.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Bonner, W., Hulett, H., Sweet, R., and Herzenberg, L. (1972). Fluorescence activated cell sorting. *Review of Scientific Instruments*, 43(3):404–409.

Borneman, A. R., Desany, B. A., Riches, D., Affourtit, J. P., Forgan, A. H., Pretorius, I. S., Egholm, M., and Chambers, P. J. (2011). Whole-genome comparison reveals novel genetic

elements that characterize the genome of industrial strains of *Saccharomyces cerevisiae*. *PLoS genetics*, 7(2):e1001287.

Bornmann, L. and Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222.

Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679.

Bradnam, K. and Korf, I. (2012). *UNIX and Perl to the rescue! A field guide for the life sciences (and other data-rich pursuits)*. Cambridge University Press.

Brenner, S. (2000). False starts: Inverse genetics. *Current Biology*, 10(18):R649.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J.-i., and Corcoran, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology*, 18(6):630.

Bruni, L. E. (2008). Cellular semiotics and signal transduction. In *Introduction to biosemiotics*, pages 365–408. Springer.

Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W., and Shipley, G. L. (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical chemistry*, 55(4):611–622.

Callebaut, W. (2012). Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1):69–80.

Canelas, A. B., Ras, C., Ten Pierick, A., van Dam, J. C., Heijnen, J. J., and Van Gulik, W. M. (2008). Leakage-free rapid quenching technique for yeast metabolomics. *Metabolomics*, 4(3):226–239.

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A. S., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research*, 18(1):188–196.

Carey, M. A. and Papin, J. A. (2018). Ten simple rules for biologists learning to program.

Carrera, J., Estrela, R., Luo, J., Rai, N., Tsoukalas, A., and Tagkopoulos, I. (2014). An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of *Escherichia coli*. *Molecular systems biology*, 10(7):735.

Caruccio, N. (2011). Preparation of next-generation sequencing libraries using Nextera™ technology: simultaneous DNA fragmentation and adaptor tagging by *in vitro* transposition. In *High-Throughput Next Generation Sequencing*, pages 241–255. Springer.

Caspeta, L., Chen, Y., Ghiaci, P., Feizi, A., Buskov, S., Hallström, B. M., Petranovic, D., and Nielsen, J. (2014). Altered sterol composition renders yeast thermotolerant. *Science*, 346(6205):75–78.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., and Kubo, A. (2013). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research*, 42(D1):D459–D471.

Cazzanelli, G., Pereira, F., Alves, S., Francisco, R., Azevedo, L., Carvalho, P. D., Almeida, A., Corte-Real, M., Oliveira, M. J., Lucas, C., Sousa, M. J., and Preto, A. (2018). The yeast *Saccharomyces cerevisiae* as a model for understanding RAS proteins and their role in human tumorigenesis. *Cells*, 7(2):14.

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W., and Prasher, D. C. (1994). Green fluorescent protein as a marker for gene expression. *Science*, 263(5148):802–805.

Chattopadhyay, P. K., Hogerkorp, C.-M., and Roederer, M. (2008). A chromatic explosion: the development and future of multiparameter flow cytometry. *Immunology*, 125(4):441–449.

Chen, R. E. and Thorner, J. (2007). Function and regulation in MAPK signaling pathways: lessons learned from the yeast *Saccharomyces cerevisiae*. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1773(8):1311–1340.

Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y., and Hwang, C.-C. (2013). Effects of GC bias in next-generation-sequencing data on *de novo* genome assembly. *PloS one*, 8(4):e62856.

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., and Engel, S. R. (2011). *Saccharomyces* genome database: the genomics resource of budding yeast. *Nucleic acids research*, 40(D1):D700–D705.

Christensen, T. S., Oliveira, A. P., and Nielsen, J. (2009). Reconstruction and logical modeling of glucose repression signaling pathways in *Saccharomyces cerevisiae*. *BMC systems biology*, 3(1):7.

Clausznitzer, D., Oleksiuk, O., Løvdok, L., Sourjik, V., and Endres, R. G. (2010). Chemotactic response and adaptation dynamics in *Escherichia coli*. *PLoS computational biology*, 6(5):e1000784.

Cochrane, G., Karsch-Mizrachi, I., Takagi, T., and Sequence Database Collaboration, I. N. (2015). The international nucleotide sequence database collaboration. *Nucleic acids research*, 44(D1):D48–D50.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):1767–1771.

Compeau, P. E., Pevzner, P. A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature biotechnology*, 29(11):987.

Conrad, M., Schothorst, J., Kankipati, H. N., Van Zeebroeck, G., Rubio-Texeira, M., and Thevelein, J. M. (2014). Nutrient sensing and signaling in the yeast *Saccharomyces cerevisiae*. *FEMS microbiology reviews*, 38(2):254–299.

Coveney, P. V., Dougherty, E. R., and Highfield, R. R. (2016). Big data need big theory too. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2080):20160153.

Covert, M. W., Famili, I., and Palsson, B. O. (2003). Identifying constraints that govern cell behavior: a key to converting conceptual to computational models in biology? *Biotechnology and bioengineering*, 84(7):763–772.

Covert, M. W., Xiao, N., Chen, T. J., and Karr, J. R. (2008). Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics*, 24(18):2044–2050.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561.

Daim, T. U., Rueda, G., Martin, H., and Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, 73(8):981–1012.

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, 14(7):1394–1403.

Dasari, S. K., Bialik, S., Levin-Zaidman, S., Levin-Salomon, V., Merrill Jr, A. H., Futerman, A. H., and Kimchi, A. (2017). Signalome-wide RNAi screen identifies GBA1 as a positive mediator of autophagic cell death. *Cell death and differentiation*, 24(7):1288.

Date, C. J. (2004). *An introduction to database systems*. Pearson Education, 8 edition.

de Jong, E. and Jungmeier, G. (2015). Biorefinery concepts in comparison to petrochemical refineries. In *Industrial biorefineries & white biotechnology*, pages 3–33. Elsevier.

De Keersmaecker, S. C., Thijs, I. M., Vanderleyden, J., and Marchal, K. (2006). Integration of omics data: how well does it work for bacteria? *Molecular microbiology*, 62(5):1239–1250.

De Mauro, A., Greco, M., and Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review*, 65(3):122–135.

Del Angel, V. D., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Pettersson, O. V., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J.-F., Vlasova, A., Leskosek, B. L., Soler, L., Binzer-Panchal, M., and Lantz, H. (2018). Ten steps to get started in genome assembly and annotation. *F1000Research*, 7.

Del Fabbro, C., Scalabrin, S., Morgante, M., and Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PloS one*, 8(12):e85024.

Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F., and Marchini, J. (2013). Haplotype estimation using sequencing reads. *The American Journal of Human Genetics*, 93(4):687–696.

Deming, W. E. (2018). *The new economics for industry, government, education*. MIT press.

Desai, A., Marwah, V. S., Yadav, A., Jha, V., Dhaygude, K., Bangar, U., Kulkarni, V., and Jere, A. (2013). Identification of optimum sequencing depth especially for *de novo* genome assembly of small genomes using next generation sequencing data. *PloS one*, 8(4):e60204.

Diaz-Ruiz, R., Rigoulet, M., and Devin, A. (2011). The Warburg and Crabtree effects: On the origin of cancer cell energy metabolism and of yeast glucose repression. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1807(6):568–576.

Dohlman, H. G. and Thorner, J. (2001). Regulation of G protein–initiated signal transduction in yeast: paradigms and principles. *Annual review of biochemistry*, 70(1):703–754.

Dolinski, K. and Troyanskaya, O. G. (2015). Implications of Big Data for cell biology. *Molecular biology of the cell*, 26(14):2575–2578.

Dragosits, M. and Mattanovich, D. (2013). Adaptive laboratory evolution–principles and applications for biotechnology. *Microbial cell factories*, 12(1):64.

Duke, C. S. and Porter, J. H. (2013). The ethics of data sharing and reuse in biology. *BioScience*, 63(6):483–489.

Edwards, D. J. and Holt, K. E. (2013). Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial informatics and experimentation*, 3(1):2.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138.

Elbourne, L. D., Tetu, S. G., Hassan, K. A., and Paulsen, I. T. (2016). TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic acids research*, 45(D1):D320–D324.

Engel, S. R., Dietrich, F. S., Fisk, D. G., Binkley, G., Balakrishnan, R., Costanzo, M. C., Dwight, S. S., Hitz, B. C., Karra, K., Nash, R. S., et al. (2014). The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3: Genes, Genomes, Genetics*, 4(3):389–398.

Fang, F. C. and Casadevall, A. (2011). Reductionistic and holistic science.

Fang, H., Wu, Y., Narzisi, G., ORawe, J. A., Barrón, L. T. J., Rosenbaum, J., Ronemus, M., Iossifov, I., Schatz, M. C., and Lyon, G. J. (2014). Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome medicine*, 6(10):89.

Farwick, A., Bruder, S., Schadeweg, V., Oreb, M., and Boles, E. (2014). Engineering of yeast hexose transporters to transport D-xylose without inhibition by D-glucose. *Proceedings of the National Academy of Sciences*, 111(14):5159–5164.

Fazzari, M. J. and Greally, J. M. (2004). Epigenomics: beyond CpG islands. *Nature Reviews Genetics*, 5(6):446.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., and Palsson, B. Ø. (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular systems biology*, 3(1):121.

Fiedler, D., Braberg, H., Mehta, M., Chechik, G., Cagney, G., Mukherjee, P., Silva, A. C., Shales, M., Collins, S. R., van Wageningen, S., Kemmeren, P., Holstege, F. C., Weissman, J. S., Keogh, M.-C., Koller, D., Shokat, K. M., and Krogan, N. J. (2009). Functional organization of the *S. cerevisiae* phosphorylation network. *Cell*, 136(5):952–963.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocyne, J. D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O., and Venter, J. C. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223):496–512.

França, T. F. and Monserrat, J. M. (2019). To read more papers, or to read papers better? a crucial point for the reproducibility crisis. *BioEssays*, 41(1):1800206.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Furhmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison III, C. A., and Venter, J. C. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, 270(5235):397–404.

Gabaldón, T. and Alioto, T. S. (2016). Whole-Genome Sequencing Recommendations. In *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing*, pages 13–41. Springer.

Gan, Y., Lin, Y., Guo, Y., Qi, X., and Wang, Q. (2018). Metabolic and genomic characterisation of stress-tolerant industrial *Saccharomyces cerevisiae* strains from TALENs-assisted multiplex editing. *FEMS Yeast Research*, 18(5). foy045.

Gancedo, J. M. (1998). Yeast carbon catabolite repression. *Microbiol. Mol. Biol. Rev.*, 62(2):334–361.

Gancedo, J. M., Flores, C.-L., and Gancedo, C. (2015). The repressor Rgt1 and the cAMP-dependent protein kinases control the expression of the SUC2 gene in *Saccharomyces cerevisiae*. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(7):1362–1367.

Gasch, A. P. and Werner-Washburne, M. (2002). The genomics of yeast responses to environmental stress and starvation. *Functional & integrative genomics*, 2(4-5):181–192.

Gatherer, D. (2010). So what do we really mean when we say that systems biology is holistic? *BMC systems biology*, 4(1):22.

Gilbert, A., Sangurdekar, D. P., and Srienc, F. (2009). Rapid strain improvement through optimized evolution in the cytostat. *Biotechnology and bioengineering*, 103(3):500–512.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012.

Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):R86.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen,

P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 genes. *Science*, 274(5287):546–567.

Goodman, S. N. (2016). Aligning statistical and scientific reasoning. *Science*, 352(6290):1180–1181.

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M. C., and McCombie, W. R. (2015). Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. *Genome research*, 25(11):1750–1756.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333.

Gopinarayanan, V. E. and Nair, N. U. (2019). Pentose metabolism in *Saccharomyces cerevisiae*: the need to engineer global regulatory systems. *Biotechnology journal*, 14(1):1800364.

Grabowski, M., Niedzialkowska, E., Zimmerman, M. D., and Minor, W. (2016). The impact of structural genomics: the first quindecennial. *Journal of structural and functional genomics*, 17(1):1–16.

Grivell, L. (2002). Mining the bibliome: searching for a needle in a haystack?: new computing tools are needed to effectively scan the growing amount of scientific literature for useful information. *EMBO reports*, 3(3):200–203.

Gudivada, V. N., Baeza-Yates, R., and Raghavan, V. V. (2015). Big data: Promises and problems. *Computer*, 48(3):20–23.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., Salzberg, S. L., and White, O. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic acids research*, 31(19):5654–5666.

Haas, B. J., Zeng, Q., Pearson, M. D., Cuomo, C. A., and Wortman, J. R. (2011). Approaches to fungal genome annotation. *Mycology*, 2(3):118–141.

Haigh, T. (2009). How data got its base: Information storage software in the 1950s and 1960s. *IEEE Annals of the History of Computing*, 31(4):6–25.

Hamacher, T., Becker, J., Gárdonyi, M., Hahn-Hägerdal, B., and Boles, E. (2002). Characterization of the xylose-transporting properties of yeast hexose transporters and their influence on xylose utilization. *Microbiology*, 148(9):2783–2788.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, 68(4):669–685.

Hao, T., Wu, D., Zhao, L., Wang, Q., Wang, E., and Sun, J. (2018). The genome-scale integrated networks in microorganisms. *Frontiers in microbiology*, 9:296.

Haqshenas, G., Wu, J., Simpson, K. J., Daly, R. J., Netter, H. J., Baumert, T. F., and Doerig, C. (2017). Signalome-wide assessment of host cell response to hepatitis C virus. *Nature communications*, 8:15158.

Heath, A. P. and Kavraki, L. E. (2009). Computational challenges in systems biology. *Computer Science Review*, 3(1):1–17.

Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing dna. *Genomics*, 107(1):1–8.

Hector, R. E. and Mertens, J. A. (2017). A synthetic hybrid promoter for xylose-regulated control of gene expression in *Saccharomyces* yeasts. *Molecular biotechnology*, 59(1):24–33.

Helmy, M., Crits-Christoph, A., and Bader, G. D. (2016). Ten simple rules for developing public biological databases. *PLoS Computational Biology*, 12(11):e1005128.

Henry, C. S., Broadbelt, L. J., and Hatzimanikatis, V. (2007). Thermodynamics-based Metabolic Flux Analysis. *Biophysical Journal*, 92(5):1792 – 1805.

Herschel, R. and Miori, V. M. (2017). Ethics & big data. *Technology in Society*, 49:31–36.

Heyduk, K., Stephens, J. D., Faircloth, B. C., and Glenn, T. C. (2016). Targeted DNA region re-sequencing. In *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing*, pages 43–68. Springer.

Hofer, A. M. and Lefkimmiatis, K. (2007). Extracellular calcium and cAMP: second messengers as "third messengers"? *Physiology*, 22(5):320–327.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D. P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., and Rhee, S. Y. (2008). Big data: The future of biocuration. *Nature*, 455(7209):47.

Huang, S. (2004). Back to the biology in systems biology: what can we learn from biomolecular networks? *Briefings in Functional Genomics*, 2(4):279–297.

Hyduke, D. R. and Palsson, B. Ø. (2010). Towards genome-scale signalling-network reconstructions. *Nature Reviews Genetics*, 11(4):297.

Imam, S., Schäuble, S., Brooks, A. N., Baliga, N. S., and Price, N. D. (2015). Data-driven integration of genome-scale regulatory and metabolic network models. *Frontiers in microbiology*, 6:409.

Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology*, 17(1):239.

Jeske, L., Placzek, S., Schomburg, I., Chang, A., and Schomburg, D. (2018). BRENDA in 2019: a European ELIXIR core data resource. *Nucleic acids research*, 47(D1):D542–D549.

John, P. C. S., Crowley, M. F., and Bomble, Y. J. (2017). Efficient estimation of the maximum metabolic productivity of batch systems. *Biotechnology for biofuels*, 10(1):28.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, 316(5830):1497–1502.

Jönsson, L. J. and Martín, C. (2016). Pretreatment of lignocellulose: formation of inhibitory by-products and strategies for minimizing their effects. *Bioresource technology*, 199:103–112.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361.

Kaniak, A., Xue, Z., Macool, D., Kim, J.-H., and Johnston, M. (2004). Regulatory network connecting two glucose signal transduction pathways in *Saccharomyces cerevisiae*. *Eukaryotic cell*, 3(1):221–231.

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival Jr, B., Assad-Garcia, N., Glass, J. I., and Covert, M. W. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401.

Katranidis, A., Atta, D., Schlesinger, R., Nierhaus, K. H., Choli-Papadopoulou, T., Gregor, I., Gerrits, M., Büldt, G., and Fitter, J. (2009). Fast biosynthesis of GFP molecules: a single-molecule fluorescence study. *Angewandte Chemie International Edition*, 48(10):1758–1761.

Keller, A. and Meese, E. (2015). *Nucleic acids as molecular diagnostics*. John Wiley & Sons.

Kelley, D. R., Schatz, M. C., and Salzberg, S. L. (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome biology*, 11(11):R116.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., and Hooker, G. (2009). Data-intensive science: a new paradigm for biodiversity studies. *BioScience*, 59(7):613–620.

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome research*, 12(4):656–664.

Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velázquez-Ramírez, D. A., Weaver, D., Collado-Vides, J., Paulsen, I., and Karp, P. D. (2016). The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic acids research*, 45(D1):D543–D550.

Kim, B., Kim, W. J., Kim, D. I., and Lee, S. Y. (2015). Applications of genome-scale metabolic network model in metabolic engineering. *Journal of industrial microbiology & biotechnology*, 42(3):339–348.

Kim, S. R., Park, Y.-C., Jin, Y.-S., and Seo, J.-H. (2013). Strain engineering of *Saccharomyces cerevisiae* for enhanced xylose metabolism. *Biotechnology Advances*, 31(6):851 – 861.

Kim, T. H. and Ren, B. (2006). Genome-wide analysis of protein-DNA interactions. *Annu. Rev. Genomics Hum. Genet.*, 7:81–102.

Kircher, M. and Kelso, J. (2010). High-throughput DNA sequencing - concepts and limitations. *Bioessays*, 32(6):524–536.

Kitano, H. (2002). Systems biology: a brief overview. *science*, 295(5560):1662–1664.

Klimacek, M., Krahulec, S., Sauer, U., and Nidetzky, B. (2010). Limitations in xylose-fermenting *Saccharomyces cerevisiae*, made evident through comprehensive metabolite profiling and thermodynamic analysis. *Appl. Environ. Microbiol.*, 76(22):7566–7574.

Klipp, E., Nordlander, B., Krüger, R., Gennemark, P., and Hohmann, S. (2005). Integrative model of the response of yeast to osmotic shock. *Nature biotechnology*, 23(8):975.

Knijnenburg, T. A., Roda, O., Wan, Y., Nolan, G. P., Aitchison, J. D., and Shmulevich, I. (2011). A regression model approach to enable cell morphology correction in high-throughput flow cytometry. *Molecular systems biology*, 7(1):531.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., and Phillippy, A. M. (2012). Hybrid error

correction and *de novo* assembly of single-molecule sequencing reads. *Nature biotechnology*, 30(7):693.

Koschwanez, J. H., Foster, K. R., and Murray, A. W. (2013). Improved use of a public good selects for the evolution of undifferentiated multicellularity. *Elife*, 2:e00367.

Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. *science*, 324(5924):255–258.

Kumar, S. and Dudley, J. (2007). Bioinformatics software for biologists in the genomics era. *Bioinformatics*, 23(14):1713–1717.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu,

N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860.

Lang, G. I., Rice, D. P., Hickman, M. J., Sodergren, E., Weinstock, G. M., Botstein, D., and Desai, M. M. (2013). Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*, 500(7464):571.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357.

Lavín, J. L., Sánchez-Morán, M., Bárcena, L., Cortazar, A. R., Macías-Cámara, N., González, M., and Aransay, A. M. (2017). A fistful of tips for a fruitful high throughput sequencing experiment.

Lay Jr, J. O., Liyanage, R., Borgmann, S., and Wilkins, C. L. (2006). Problems with the "omics". *TrAC Trends in Analytical Chemistry*, 25(11):1046–1056.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science*, 343(6176):1203–1205.

Lederberg, J. and McCray, A. T. (2001). Ome Sweet Omics–a genealogical treasury of words. *The Scientist*, 15(7):8–8.

Lee, D. and Cho, K.-H. (2018). Topological estimation of signal flow in complex signaling networks. *Scientific reports*, 8(1):5262.

Lee, J. W., Na, D., Park, J. M., Lee, J., Choi, S., and Lee, S. Y. (2012). Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nature chemical biology*, 8(6):536.

Leipzig, J. (2017). A review of bioinformatic pipeline frameworks. *Briefings in bioinformatics*, 18(3):530–536.

Leonelli, S. (2010). Packaging data for re-use: Databases in model organism biology. Cambridge University Press.

Leonelli, S. (2014). What difference does quantity make? on the epistemology of Big Data in biology. *Big data & society*, 1(1):2053951714534395.

Levsky, J. M. and Singer, R. H. (2003). Fluorescence *in situ* hybridization: past, present and future. *Journal of cell science*, 116(14):2833–2838.

Li, F., Thiele, I., Jamshidi, N., and Palsson, B. Ø. (2009a). Identification of potential pathway mediation targets in Toll-like receptor signaling. *PLoS computational biology*, 5(2):e1000292.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, 25(14):1754–1760.

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

Li, Y. and Chen, L. (2014). Big biological data: challenges and opportunities. *Genomics, proteomics & bioinformatics*, 12(5):187.

Liao, Y.-C., Huang, T.-W., Chen, F.-C., Charusanti, P., Hong, J. S. J., Chang, H.-Y., Tsai, S.-F., Palsson, B. O., and Hsiung, C. A. (2011). An experimentally validated genome-scale metabolic reconstruction of *Klebsiella pneumoniae* MGH 78578, iYL1228. *Journal of Bacteriology*, 193(7):1710–1717.

Lischer, H. E. and Shimizu, K. K. (2017). Reference-guided *de novo* assembly approach improves genome reconstruction for related species. *BMC bioinformatics*, 18(1):474.

Liu, Q., Zhang, P., Wang, D., Gu, W., and Wang, K. (2017). Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. *Genome medicine*, 9(1):65.

Lopes, H. and Rocha, I. (2017). Genome-scale modeling of yeast: chronology, applications and critical perspectives. *FEMS Yeast Research*, 17(5).

Lubitz, T., Welkenhuysen, N., Shashkova, S., Bendrioua, L., Hohmann, S., Klipp, E., and Krantz, M. (2015). Network reconstruction and validation of the Snf1/AMPK pathway in baker's yeast based on a comprehensive literature review. *Npj Systems Biology And Applications*, 1:15007.

Macaulay, I. C., Ponting, C. P., and Voet, T. (2017). Single-cell multiomics: multiple measurements from single cells. *Trends in Genetics*, 33(2):155–168.

Mack, C. A. (2011). Fifty years of Moore's law. *IEEE Transactions on semiconductor manufacturing*, 24(2):202–207.

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., Tallon, L. J., and Salzberg, S. L. (2013). GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics*, 29(14):1718–1725.

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., and Ferrari, R. (2016). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in bioinformatics*, 19(2):286–302.

Margarido, G. R. and Heckerman, D. (2015). ConPADE: genome assembly ploidy estimation from next-generation sequencing data. *PLoS computational biology*, 11(4):e1004229.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376.

Martin, D. E. and Hall, M. N. (2005). The expanding TOR signaling network. *Current opinion in cell biology*, 17(2):158–166.

Marx, V. (2013). Biology: The big challenges of big data.

Mateus, C. and Avery, S. V. (2000). Destabilized green fluorescent protein for monitoring dynamic changes in yeast gene expression with flow cytometry. *Yeast*, 16(14):1313–1323.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Mazzocchi, F. (2012). Complexity and the reductionism–holism debate in systems biology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 4(5):413–427.

McCoy, R. C., Taylor, R. W., Blauwkamp, T. A., Kelley, J. L., Kertesz, M., Pushkarev, D., Petrov, D. A., and Fiston-Lavier, A.-S. (2014). Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. *PloS one*, 9(9):e106689.

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

McPherson, J. D. (2014). A defining decade in DNA sequencing. *Nature methods*, 11(10):1003.

Metzker, M. L. (2010). Sequencing technologies −the next generation. *Nature reviews genetics*, 11(1):31.

Mhlongo, M. I., Piater, L. A., Madala, N. E., Labuschagne, N., and Dubery, I. A. (2018). The chemistry of plant–microbe interactions in the rhizosphere and the potential for metabolomics to reveal signaling related to defense priming and induced systemic resistance. *Frontiers in plant science*, 9:112.

Mielczarek, M. and Szyda, J. (2016). Review of alignment and SNP calling algorithms for next-generation sequencing data. *Journal of applied genetics*, 57(1):71–79.

Miller, J. R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327.

Mitchell, A. L., Scheremetjew, M., Denise, H., Potter, S., Tarkowska, A., Qureshi, M., Salazar, G. A., Pesseat, S., Boland, M. A., Hunter, F. M. I., ten Hoopen, P., Alako, B., Amid, C., Wilkinson, D. J., Curtis, T. P., Cochrane, G., and D, F. R. (2017). EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic acids research*, 46(D1):D726–D735.

Mohamed, S. and Syed, B. A. (2013). Commercial prospects for genomic sequencing technologies.

Molloy, J. C. (2011). The open knowledge foundation: open data means better science. *PLoS biology*, 9(12):e1001195.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research*, 35(suppl_2):W182–W185.

Moysés, D., Reis, V., Almeida, J., Moraes, L., and Torres, F. (2016). Xylose fermentation by *Saccharomyces cerevisiae*: challenges and prospects. *International journal of molecular sciences*, 17(3):207.

Müller, S. and Nebe-von Caron, G. (2010). Functional single-cell analyses: flow cytometry and cell sorting of microbial populations and communities. *FEMS microbiology reviews*, 34(4):554–587.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature human behaviour*, 1(1):0021.

Munir, K. and Anjum, M. S. (2018). The use of ontologies for effective knowledge modelling and information retrieval. *Applied Computing and Informatics*, 14(2):116–126.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N., and Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic acids research*, 39(13):e90–e90.

Newman, R. H., Fosbrink, M. D., and Zhang, J. (2011). Genetically encodable fluorescent biosensors for tracking signaling dynamics in living cells. *Chemical reviews*, 111(5):3614–3666.

Nielsen, J. and Jewett, M. C. (2008). Impact of systems biology on metabolic engineering of *Saccharomyces cerevisiae*. *FEMS yeast research*, 8(1):122–131.

Nielsen, J. and Keasling, J. D. (2016). Engineering cellular metabolism. *Cell*, 164(6):1185–1197.

Oliver, S. G., van der Aart, Q. J. M., Agostoni-Carbone, M. L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J. P. G., Benit, P., Berben, G., Bergantino, E., Biteau, N., Bolle, P. A., Bolotin-Fukuhara, M., Brown, A., Brown, A. J. P., Buhler, J. M., Carcano, C., Carignani, G., Cederberg, H., Chanet, R., Contreras, R., Crouzet, M., Daignan-Fornier, B., Defoor, E., Delgado, M., Demolder, J., Doira, C., Dubois, E., Dujon, B., Dusterhoft, A., Erdmann, D., Esteban, M., Fabre, F., Fairhead, C., Faye, G., Feldmann, H., Fiers, W., Francingues-Gaillard, M. C., Franco, L., Frontali, L., Fukuhara, H., Fuller, L. J., Galland, P., Gent, M. E., Gigot, D., Gilliquet, V., Glansdorff, N., Goffeau, A., Grenson, M., Grisanti, P., Grivell, L. A., de Haan, M., Haasemann, M., Hatat, D., Hoenicka, J., Hegemann, J., Herbert, C. J., Hilger, F., Hohmann, S., Hollenberg, C. P., Huse, K., Iborra, F., Indje, K. J., Isono, K., Jacq, C., Jacquet, M., James, C. M., Jauniaux, J. C., Jia, Y., Jimenez, A., Kelly, A., Kleinhans, U., Kreisl, P., Lanfranchi, G., Lewis, C., vanderLinden, C. G., Lucchini, G., Lutzenkirchen, K., Maat, M. J., Mallet, L., Mannhaupet, G., Martegani, E., Mathieu, A., Maurer, C. T. C., McConnell, D., McKee, R. A., Messenguy, F., Mewes, H. W., Molemans, F., Montague, M. A., Muzi Falconi, M., Navas, L., Newlon, C. S., Noone, D., Pallier, C., Panzeri, L., Pearson, B. M., Perea, J., Philippsen, P., Pierard, A., Planta, R. J., Plevani, P., Poetsch, B., Pohl, F., Purnelle, B., Ramezani Rad, M., Rasmussen, S. W., Raynal, A., Remacha, M., Richterich, P., Roberts, A. B., Rodriguez, F., Sanz, E., Schaaff-Gerstenschlager, I., Scherens, B., Schweitzer, B., Shu, Y., Skala, J., Slonimski, P. P., Sor, F., Soustelle, C., Spiegelberg, R., Stateva, L. I., Steensma, H. Y., Steiner, S., Thierry, A., Thireos, G., Tzermia, M., Urrestarazu, L. A., Valle, G., Vetter, I., van Vliet-Reedijk, J. C., Voet, M., Volckaert, G., Vreken, P., Wang, H., Warmington, J. R., von Wettstein, D., Wicksteed, B. L., Wilson, C., Wurst, H., Xu, G., Yoshikawa, A., Zimmermann, F. K., and Sgouros, J. G. (1992). The complete DNA sequence of yeast chromosome III. *Nature*, 357(6373):38.

Olson, N. D., Lund, S. P., Colman, R. E., Foster, J. T., Sahl, J. W., Schupp, J. M., Keim, P., Morrow, J. B., Salit, M. L., and Zook, J. M. (2015). Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Frontiers in genetics*, 6:235.

Oltvai, Z. N. and Barabási, A.-L. (2002). Life's complexity pyramid. *Science*, 298(5594):763–764.

O'Malley, M. A. and Dupré, J. (2005). Fundamental issues in systems biology. *BioEssays*, 27(12):1270–1276.

O'Malley, M. A. and Soyer, O. S. (2012). The roles of integration in molecular systems biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1):58–68.

O'Neill, K., Aghaeepour, N., Špidlen, J., and Brinkman, R. (2013). Flow cytometry bioinformatics. *PLoS computational biology*, 9(12):e1003365.

Ornston, L. and Stanier, R. (1966). The conversion of catechol and protocatechuate to β-ketoadipate by *Pseudomonas putida* I. biochemistry. *Journal of Biological Chemistry*, 241(16):3776–3786.

Orth, J. D. and Palsson, B. . (2010). Systematizing the generation of missing metabolic knowledge. *Biotechnology and Bioengineering*, 107(3):403–412.

Oud, B., van Maris, A. J., Daran, J.-M., and Pronk, J. T. (2012). Genome-wide analytical approaches for reverse metabolic engineering of industrially relevant phenotypes in yeast. *FEMS yeast research*, 12(2):183–196.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., and Stevens, R. (2013). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic acids research*, 42(D1):D206–D214.

Ozcan, S. and Johnston, M. (1995). Three different regulatory mechanisms enable yeast hexose transporter (HXT) genes to be induced by different levels of glucose. *Molecular and Cellular Biology*, 15(3):1564–1572.

Palsson, B. (2015). *Systems biology*. Cambridge university press.

Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature reviews genetics*, 10(10):669.

Paulino, D., Warren, R. L., Vandervalk, B. P., Raymond, A., Jackman, S. D., and Birol, I. (2015). Sealer: a scalable gap-closing application for finishing draft genomes. *BMC bioinformatics*, 16(1):230.

Pawson, T. and Nash, P. (2000). Protein–protein interactions define specificity in signal transduction. *Genes & development*, 14(9):1027–1047.

Payen, C., Sunshine, A. B., Ong, G. T., Pogachar, J. L., Zhao, W., and Dunham, M. J. (2016). High-throughput identification of adaptive mutations in experimentally evolved yeast populations. *PLoS genetics*, 12(10):e1006339.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448.

Pedruzzi, I., Dubouloz, F., Cameroni, E., Wanke, V., Roosen, J., Winderickx, J., and De Virgilio, C. (2003). TOR and PKA signaling pathways converge on the protein kinase Rim15 to control entry into G0. *Molecular cell*, 12(6):1607–1613.

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance*, 12(3):30–32.

Perbal, L. (2015). The case of the gene: postgenomics between modernity and postmodernity. *EMBO reports*, 16(7):777–781.

Peters, D. P., Havstad, K. M., Cushing, J., Tweedie, C., Fuentes, O., and Villanueva-Rosales, N. (2014). Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. *Ecosphere*, 5(6):1–15.

Petzold, C. J., Chan, L. J. G., Nhan, M., and Adams, P. D. (2015). Analytics for metabolic engineering. *Frontiers in bioengineering and biotechnology*, 3:135.

Pham, V. H. and Kim, J. (2012). Cultivation of unculturable soil bacteria. *Trends in biotechnology*, 30(9):475–484.

Phillippy, A. M., Schatz, M. C., and Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly. *Genome biology*, 9(3):R55.

Pinzon, W., Vega, H., Gonzalez, J., and Pinzon, A. (2018). Mathematical framework behind the reconstruction and analysis of genome scale metabolic models. *Archives of Computational Methods in Engineering*, pages 1–14.

Pitarch, A., Sánchez, M., Nombela, C., and Gil, C. (2003). Analysis of the *Candida albicans* proteome: II. protein information technology on the Net (update 2002). *Journal of Chromatography B*, 787(1):129–148.

Pop, M. and Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends in genetics*, 24(3):142–149.

Powell, A., O'Malley, M. A., Müller-Wille, S., Calvert, J., and Dupré, J. (2007). Disciplinary baptisms: a comparison of the naming stories of genetics, molecular biology, genomics, and systems biology. *History and philosophy of the life sciences*, pages 5–32.

Price, N. D., Reed, J. L., and Palsson, B. Ø. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11):886.

Prohaska, S. J. and Stadler, P. F. (2011). The use and abuse of-omes. In *Bioinformatics for Omics Data*, pages 173–196. Springer.

Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H., and Turner, D. J. (2008). A large genome center's improvements to the illumina sequencing system. *Nature methods*, 5(12):1005.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(D1):D590–D596.

Rhoads, A. and Au, K. F. (2015). PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289.

Rigden, D. J. and Fernández, X. M. (2018). The 26th annual Nucleic Acids Research database issue and Molecular Biology Database Collection. *Nucleic Acids Research*, 47(D1):D1–D7.

Roberts, R. J. (2001). PubMed central: The GenBank of the published literature.

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S.,

Hickey, G., Hinrichs, A. S., Hubley, R., Karolchik, D., Learned, K., Lee, B. T., Li, C. H., Miga, K. H., Nguyen, N., Paten, B., Raney, B. J., Smit, A. F. A., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M., and Kent, W. J. (2014). The UCSC genome browser database: 2015 update. *Nucleic acids research*, 43(D1):D670–D681.

Rother, M., Münzner, U., Thieme, S., and Krantz, M. (2013). Information content and scalability in signal transduction network reconstruction formats. *Molecular BioSystems*, 9(8):1993–2004.

Rung, J. and Brazma, A. (2013). Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, 14(2):89.

Saeys, Y., Van Gassen, S., and Lambrecht, B. N. (2016). Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, 16(7):449.

Saier Jr, M. H., Reddy, V. S., Tsu, B. V., Ahmed, M. S., Li, C., and Moreno-Hagelsieb, G. (2015). The transporter classification database (TCDB): recent advances. *Nucleic acids research*, 44(D1):D372–D379.

Salusjärvi, L., Kankainen, M., Soliymani, R., Pitkänen, J.-P., Penttilä, M., and Ruohonen, L. (2008). Regulation of xylose metabolism in recombinant *Saccharomyces cerevisiae*. *Microbial cell factories*, 7(1):18.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C., Slocombe, P. M., and Smith, M. (1977a). Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature*, 265(5596):687.

Sanger, F. and Coulson, A. R. (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., and Coulson, A. R. (1977b). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467.

Santangelo, G. M. (2006). Glucose signaling in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.*, 70(1):253–282.

Sato, T. K., Tremaine, M., Parreiras, L. S., Hebert, A. S., Myers, K. S., Higbee, A. J., Sardi, M., McIlwain, S. J., Ong, I. M., Breuer, R. J., Avanasi Narasimhan, R., McGee, M. A., Dickinson, Q., La Reau, A., Xie, D., Tian, M., Reed, J. L., Zhang, Y., Coon, J. J., Hittinger, C. T., Gasch, A. P., and Landick, R. (2016). Directed evolution reveals unexpected epistatic interactions that alter metabolic regulation and enable anaerobic xylose use by *Saccharomyces cerevisiae*. *PLoS genetics*, 12(10):e1006372.

Satomura, A., Miura, N., Kuroda, K., and Ueda, M. (2016). Reconstruction of thermotolerant yeast by one-point mutation identified through whole-genome analyses of adaptively-evolved strains. *Scientific reports*, 6:23157.

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., and Gerstein, M. B. (2011). The real cost of sequencing: higher than you think! *Genome biology*, 12(8):125.

Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. *Human molecular genetics*, 19(R2):R227–R240.

Schatz, M. C., Delcher, A. L., and Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome research*, 20(9):1165–1173.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic acids research*, 43(6):e37–e37.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., and Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, 22(6):839–851.

Schuster, S., Fell, D. A., and Dandekar, T. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature biotechnology*, 18(3):326.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069.

Sehgal, A. K., Das, S., Noto, K., Saier, M., and Elkan, C. (2011). Identifying relevant data for a biological database: Handcrafted rules versus machine learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(3):851–857.

Shapiro, H. M. (2005). *Practical flow cytometry*. John Wiley & Sons.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–311.

Shin, M., Kim, J.-w., Ye, S., Kim, S., Jeong, D., Lee, D. Y., Kim, J. N., Jin, Y.-S., Kim, K. H., and Kim, S. R. (2019). Comparative global metabolite profiling of xylose-fermenting *Saccharomyces cerevisiae* SR8 and *Scheffersomyces stipitis*. *Applied microbiology and biotechnology*, pages 1–12.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121.

Spear, T. T., Nishimura, M. I., and Simms, P. E. (2017). Comparative exploration of multidimensional flow cytometry software: a model approach evaluating T cell polyfunctional behavior. *Journal of leukocyte biology*, 102(2):551–561.

Spidlen, J., Breuer, K., Rosenberg, C., Kotecha, N., and Brinkman, R. R. (2012). Flowrepository: A resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytometry Part A*, 81(9):727–731.

Spitzer, M. H. and Nolan, G. P. (2016). Mass cytometry: single cells, many features. *Cell*, 165(4):780–791.

Stein, L. (2001). Genome annotation: from sequence to biology. *Nature reviews genetics*, 2(7):493.

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., and Abola, E. E. (1998). Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallographica Section D: Biological Crystallography*, 54(6):1078–1084.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., and von Mering, C. (2018).

STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613.

Taherzadeh, M. and Karimi, K. (2008). Pretreatment of lignocellulosic wastes to improve ethanol and biogas production: a review. *International journal of molecular sciences*, 9(9):1621–1651.

Tamaki, H. (2007). Glucose-stimulated cAMP-protein kinase A pathway in yeast *Saccharomyces cerevisiae*. *Journal of bioscience and bioengineering*, 104(4):245–250.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Ciufo, S., and Li, W. (2013). Prokaryotic genome annotation pipeline. In *The NCBI Handbook [Internet]. 2nd edition*. National Center for Biotechnology Information (US).

Teo, W. S. and Chang, M. W. (2015). Bacterial XylRs and synthetic promoters function as genetically encoded xylose biosensors in *Saccharomyces cerevisiae*. *Biotechnology journal*, 10(2):315–322.

Teruel, M. N. and Meyer, T. (2000). Translocation and reversible localization of signaling proteins: a dynamic future for signal transduction. *Cell*, 103(2):181–184.

Thevelein, J. M. and De Winde, J. H. (1999). Novel sensing mechanisms and targets for the cAMP–protein kinase A pathway in the yeast *Saccharomyces cerevisiae*. *Molecular microbiology*, 33(5):904–918.

Topol, E. J. (2014). Individualized medicine from prewomb to tomb. *Cell*, 157(1):241–253.

Ulrich, L. E. and Zhulin, I. B. (2009). The MiST2 database: a comprehensive genomics resource on microbial signal transduction. *Nucleic acids research*, 38(suppl_1):D401–D407.

UniProt Consortium (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets,

R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *science*, 291(5507):1304–1351.

Verduyn, C., Postma, E., Scheffers, W. A., and Van Dijken, J. P. (1992). Effect of benzoic acid on metabolic fluxes in yeasts: a continuous-culture study on the regulation of respiration and alcoholic fermentation. *Yeast*, 8(7):501–517.

Verghese, J., Abrams, J., Wang, Y., and Morano, K. A. (2012). Biology of the heat shock response and protein chaperones: budding yeast (*Saccharomyces cerevisiae*) as a model system. *Microbiol. Mol. Biol. Rev.*, 76(2):115–158.

Vernikos, G., Medini, D., Riley, D. R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Current opinion in microbiology*, 23:148–154.

Vihinen, M. (2001). Bioinformatics in proteomics. *Biomolecular Engineering*, 18(5):241–248.

von Bertalanffy, L. (1950). The theory of open systems in physics and biology. *Science*, 111(2872):23–29.

Wallace-Salinas, V. and Gorwa-Grauslund, M. F. (2013). Adaptive evolution of an industrial strain of *Saccharomyces cerevisiae* for combined tolerance to inhibitors and temperature. *Biotechnology for biofuels*, 6(1):151.

Wang, H.-y. and Malbon, C. C. (2011). Probing the physical nature and composition of signalsomes. *Journal of molecular signaling*, 6(1):1.

Wang, M., Li, S., and Zhao, H. (2016). Design and engineering of intracellular-metabolite-sensing/regulation gene circuits in *Saccharomyces cerevisiae*. *Biotechnology and bioengineering*, 113(1):206–215.

Wang, X., Liang, Z., Hou, J., Shen, Y., and Bao, X. (2017). The absence of the transcription factor Yrr1p, identified from comparative genome profiling, increased vanillin tolerance

due to enhancements of ABC transporters expressing, rRNA processing and ribosome biogenesis in *Saccharomyces cerevisiae*. *Frontiers in microbiology*, 8:367.

Wang, Y., Pierce, M., Schneper, L., Güldal, C. G., Zhang, X., Tavazoie, S., and Broach, J. R. (2004). Ras and Gpa2 mediate one branch of a redundant glucose signaling pathway in yeast. *PLOS Biology*, 2(5).

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on *p*-values: context, process, and purpose. *The American Statistician*, 70(2):129–133.

Watatani, K., Xie, Z., Nakatsuji, N., and Sengoku, S. (2013). Global competencies of regional stem cell research: bibliometrics for investigating and forecasting research trends. *Regenerative medicine*, 8(5):659–668.

Waters, C. M. and Bassler, B. L. (2005). Quorum sensing: cell-to-cell communication in bacteria. *Annu. Rev. Cell Dev. Biol.*, 21:319–346.

Welch, C. M., Elliott, H., Danuser, G., and Hahn, K. M. (2011). Imaging the coordination of multiple signalling activities in living cells. *Nature reviews Molecular cell biology*, 12(11):749.

Weng, G., Bhalla, U. S., and Iyengar, R. (1999). Complexity in biological signaling systems. *Science*, 284(5411):92–96.

Werner, T. (2010). Next generation sequencing in functional genomics. *Briefings in bioinformatics*, 11(5):499–511.

Wetterstrand, K. A. (2017). DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP).

Wicki-Stordeur, L. E. and Swayne, L. A. (2014). The emerging Pannexin 1 signalome: a new nexus revealed? *Frontiers in cellular neuroscience*, 7:287.

Winkler, H. (1920). *Verbreitung und ursache der parthenogenesis im pflanzen-und tierreiche*. G. Fischer.

Winsor, G. L., Lam, D. K., Fleming, L., Lo, R., Whiteside, M. D., Yu, N. Y., Hancock, R. E., and Brinkman, F. S. (2010). *Pseudomonas* Genome Database: improved comparative analysis and population genomics capability for *Pseudomonas* genomes. *Nucleic acids research*, 39(suppl_1):D596–D600.

Wittgenstein, L. (1922). *Tractatus logico-philosophicus*. C. K. Ogden, Trans. Vol. EBook-No. 5740), Project Gutenberg.

Witze, E. S., Old, W. M., Resing, K. A., and Ahn, N. G. (2007). Mapping protein post-translational modifications with mass spectrometry. *Nature methods*, 4(10):798.

Woelk, T., Sigismund, S., Penengo, L., and Polo, S. (2007). The ubiquitination code: a signalling problem. *Cell division*, 2(1):11.

Wynendaele, E., Bronselaer, A., Nielandt, J., D'hondt, M., Stalmans, S., Bracke, N., Verbeke, F., Van De Wiele, C., De Tre, G., and De Spiegeleer, B. (2012). Quorumpeps database: chemical space, microbial origin and functionality of quorum sensing peptides. *Nucleic acids research*, 41(D1):D655–D659.

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and structural biotechnology journal*, 16:15–24.

Xu, G. and Jaffrey, S. R. (2013). Proteomic identification of protein ubiquitination events. *Biotechnology and Genetic Engineering Reviews*, 29(1):73–109.

Yadav, S. P. (2007). The wholeness in suffix *-omics*, *-omes*, and the word *om*. *Journal of biomolecular techniques: JBT*, 18(5):277.

Yandell, M. and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, 13(5):329.

Yao, Z., Petschnigg, J., Ketteler, R., and Stagljar, I. (2015). Application guide for omics approaches to cell signaling. *Nature chemical biology*, 11(6):387.

Yin, Z., Lan, H., Tan, G., Lu, M., Vasilakos, A. V., and Liu, W. (2017). Computing platforms for big biological data analytics: perspectives and challenges. *Computational and structural biotechnology journal*, 15:403–411.

Zadran, S., Standley, S., Wong, K., Otiniano, E., Amighi, A., and Baudry, M. (2012). Fluorescence resonance energy transfer (FRET)-based biosensors: visualizing cellular dynamics and bioenergetics. *Applied microbiology and biotechnology*, 96(4):895–902.

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–829.

Zhang, H., Zeidler, A. F., Song, W., Puccia, C. M., Malc, E., Greenwell, P. W., Mieczkowski, P. A., Petes, T. D., and Argueso, J. L. (2013). Gene copy-number variation in haploid and diploid strains of the yeast *Saccharomyces cerevisiae*. *Genetics*, 193(3):785–801.

Zhang, M.-M., Chen, H.-Q., Ye, P.-L., Wattanachaisaereekul, S., Bai, F.-W., and Zhao, X.-Q. (2019). Development of robust yeast strains for lignocellulosic biorefineries based on genome-wide studies. In *Yeasts in Biotechnology and Human Health*, pages 61–83. Springer.

Zhang, Z. and Yu, J. (2006). Evaluation of six methods for estimating synonymous and nonsynonymous substitution rates. *Genomics, proteomics & bioinformatics*, 4(3):173–181.

Zhao, Y. and Jensen, O. N. (2009). Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics*, 9(20):4632–4641.

Zhulin, I. B. (2015). Databases for microbiologists. *Journal of bacteriology*, 197(15):2458–2467.

Since the advent of high-throughput genome sequencing methods in the mid-2000s, molecular biology has rapidly transitioned towards data-intensive science. Recent technological developments have increased the accessibility of omics experiments by decreasing the cost, while the concurrent design of new algorithms have improved the computational work-flow needed to analyse the large datasets generated. This has enabled the long standing idea of a *systems* approach to the cell, where molecular phenomena are no longer observed in isolation, but as parts of a tightly regulated cell-wide system. However, large data biology is not without its challenges, many of which are directly related to how to store, handle and analyse ome-wide datasets.

The present thesis examines large data microbiology from a middle ground between metabolic engineering and *in silico* data management. The work was performed in the context of applied microbial lignocellulose valorisation with the end goal of generating improved cell factories for the production of value-added chemicals from renewable plant biomass. Three different challenges related to this feedstock were investigated from a large data-point of view: bacterial catabolism of lignin and its derived aromatic compounds; tolerance of baker's yeast *Saccharomyces cerevisiae* to inhibitory compounds in lignocellulose hydrolysate; and the non-fermentable response to xylose in *S. cerevisiae* engineered for growth on this pentose sugar.

Printed by Media-Tryck, Lund 2019 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

ISBN: 978-91-7422-684-3

Applied Microbiology
Department of Chemistry
Faculty of Engineering
Lund University

9 789174 226843