Algorithms for sequence-based reverse metabolic engineering

Proefschrift

ter verkrijging van de graad van doctor aan de Technische Universiteit Delft, op gezag van Rector Magnificus prof. ir. K.Ch.A.M. Luyben, voorzitter van het College voor Promoties, in het openbaar te verdedigen op 18 december 2013 om 12:30 door Jurgen Franciscus NIJKAMP, ingenieur in Media & Knowledge Engineering, geboren te Houten
Dit proefschrift is goedgekeurd door de promotor:
prof. dr. ir. M.J.T. Reinders

Copromotor:
dr. ir. D. de Ridder

Samenstelling promotiecommissie:
Rector Magnificus, Voorzitter
prof. dr. ir. M.J.T. Reinders, Technische Universiteit Delft, promotor
dr. D. de Ridder, Technische Universiteit Delft, copromotor
prof. dr. R.A.L. Bovenberg, DSM, Delft
prof. dr. E.P.J.G. Cuppen, Hubrecht Institute, Utrecht
dr. S.A.F.T. van Hijum, Radboud Universitair Medisch Centrum, Nijmegen
dr. M. Pop, University of Maryland, MD, USA
prof. dr. J.T. Pronk, Technische Universiteit Delft
prof. dr. ir. J.J. Heijnen, Technische Universiteit Delft, reservelid

Dit onderzoek werd financieel ondersteund door het Kluyver Centre for Genomics of Industrial Fermentation, dat wordt ondersteund door het Nederlands Genomics Initiatief (NGI). Het onderzoek is uitgevoerd in de sectie Patroon Herkenning & Bioinformatica van de Technische Universiteit Delft, in samenwerking met de sectie Industriële Microbiologie. Een gedeelte van dit onderzoek is uitgevoerd bij het Center for Bioinformatics and Computational Biology, University of Maryland, USA, ondersteund met een EMBO short-term fellowship. Het drukken van het proefschrift werd financieel ondersteund door Heineken.

ISBN: 978-94-6191-994-6
## Contents

1 Introduction.............................................................................................................. 1
2 Integrating genome assemblies with MAIA............................................................. 13
3 The *S. cerevisiae* CEN.PK113-7D genome.............................................................. 29
4 Laboratory evolution of lactate transport............................................................... 51
5 *De novo* detection of copy number variation......................................................... 75
6 Laboratory evolution of multicellular yeast............................................................ 93
7 Exploring variation aware contig graphs................................................................. 113
8 Discussion.................................................................................................................. 129
References................................................................................................................... 135
Summary....................................................................................................................... 143
Samenvatting................................................................................................................. 144
Curriculum Vitae......................................................................................................... 145
Acknowledgements...................................................................................................... 146
Publications.................................................................................................................. 148
Introduction

1.1 Microbial cell factories

1.1.1 Application of microorganisms in industrial fermentation

In industrial biotechnology, microorganisms are applied in a wide variety of processes for production of food and ingredients, beverages, pharmaceutical compounds, nutraceuticals, and fine and bulk chemicals. Industrial biotechnology contributes to a sustainable society by for example relieving the dependence on fossil fuels, and contributes to a higher quality of life and to the economy. Various microorganisms are used for specific products, such as *Penicillium chrysogenum* for antibiotics, *Escherichia coli* for human insulin and *Lactobacillus* for bioplastics.

The eukaryotic yeast *Saccharomyces cerevisiae* has historically been one of the key species in industrial fermentation, even before its discovery. Yeast fermentation has been applied for thousands of years in the production of alcoholic beverages and bread and, more recently, in large scale production of biofuels. From a more fundamental perspective, yeast is interesting due to the easy accessibility to classical and molecular genetics techniques, which have made it popular in fundamental biological studies and as a model organism for more complex eukaryotes.

*S. cerevisiae* was in 1996 the first eukaryotic organism for which the genome sequence was determined and published (81). The sequencing project was a large and costly endeavour at that time, performed by an international consortium. The genetic material originated from the strain S288C, which was isolated from a rotting fig in Central California in 1938 (156). Although several important phenotypic differences between S288C, industrial and additional laboratory yeast strains were known, it took until 2005 for other *S. cerevisiae* genomes to be sequenced, such as RM11-1a (Broad institute, 2005) and YJM789 (240). Currently, at least 82 wild and domestic yeast genomes have been sequenced, which are available as draft genome or raw sequencing data (25).

The haploid MATa *S. cerevisiae* strain CEN.PK113-7D used throughout this thesis is a popular strain for systems biology studies and is used as a model organism for industrial microbiology (29, 123). It was created in the ’90s of the previous century by crossing
several laboratory strains to obtain a strain with good genetic accessibility and stable growth characteristics in a laboratory environment (64).

1.1.2 Reverse engineering cell factories in industrial microbiology

Understanding the behaviour of yeast and other microorganisms is of crucial importance to design efficient processes, which can be achieved by improving the characteristics of the cell factories, such as yield and robustness. Metabolic engineering, i.e. the modification of cellular processes by genetic modification, is one of the key tools used for such strain improvements. This rational design method is often complemented with non-rational approaches, such as random mutagenesis, evolutionary engineering or combinations of these. In random mutagenesis, strains are treated with a chemical or physical (e.g. UV) mutagen to infer mutations in the DNA, followed by a screening process that selects the rare mutants with an improved phenotype. Evolutionary engineering is rapidly gaining popularity as a tool to analyse genome function and to select for microorganisms with industrially relevant properties (168). Strains are cultivated in specifically designed conditions in which strains that acquire a desired trait have a selective advantage. Additionally, genes conferring desired traits can be obtained from natural sources and integrated into the genome, followed by mutagenesis and evolutionary engineering for fine-tuning (175).

The genetic makeup that underlies the improved phenotype generated using nontargeted approaches can be deduced in a process called reverse engineering, which is a cyclic process to obtain novel genotype-phenotype links that can be used as input for strain generation (Fig. 1.1). Reverse engineering is essential to increase our fundamental understanding of molecular mechanisms, but in a more applied setting it also allows us to use the obtained knowledge in forward engineering approaches transferring the phenotypic traits by transferring the corresponding genetic elements.

Phenotypes have previously been reverse engineered using measurements on the transcriptome, metabolome and proteome. Transcriptome profiling in particular provides an economically feasible overview of changes in gene regulation. In a study by Oud et al. (169) transcriptome profiling of an evolved Pdc- mutant was used to identify a mutation in MTH1, a repressor of RGT1, which is involved in the regulation of several glucose transporter (HXT) genes. Transcriptome profiling of the evolved mutants revealed the altered expression of HXTs compared to a wild type strain, therefore a mutation in the transcriptional regulator could be hypothesised and validated by targeted sequencing.

However, it is not possible to reverse engineer a genotype using transcriptome profiling when it does not exert its phenotype by modifying gene expression. In such cases directly comparing the genome of a wild type and mutant strain is necessary. Direct measurement of DNA has become economically feasible since the advent of next-generation sequencing. By sequencing and analysing the DNA of the strains that are improved by nontargeted approaches we can now learn what genomic characteristics encode features of the organisms that we observe.
1.2 DNA sequencing and its computational challenges

1.2.1 DNA sequencing technology

The invention of sequencing in 1977 enabled mankind to read the DNA sequence of living organisms. While in the 1990’s the cost of reading a complete DNA sequence was still in the order of millions of euros, recent technological developments made this multiple orders of magnitude cheaper. As a result, a 3 Gbp human genome can now be ‘re-sequenced’ for approximately €4k, whereas the cost in 2001 were €71M (http://www.genome.gov/sequencingcosts/, January 2013). The sequencing technologies that succeeded Sanger sequencing in the beginning of the 21st century are collectively referred to as Next-Generation Sequencing (NGS). Several NGS platforms came onto the market since 2004, such as the 454 pyrosequencer (Roche Applied Science; Basel, Switzerland), Illumina’s platforms based on Solexa’s technology (Illumina; San Diego, CA), the SOLiD platform (Applied Biosystems; Foster City, CA) and later the HeliScope (Heliocos; Cambridge, MA) and SMRT (Pacific Biosciences; Menlo Park, CA) single molecule sequencers. In this thesis data from both the 454 and the Illumina platform were used.

The 454 platform relies on random fragmentation of the DNA followed by a size selection step to generate single-stranded fragments as starting material for sequencing-by-synthesis. The fragments are amplified in a process called emulsion PCR, which binds single adaptor-ligated fragments to beads and captures single beads in small water-in-oil reactors. A PCR reaction is performed inside each of these pico-reactors to fill the bead with identical fragments. The DNA on each bead is then sequenced by flowing the four types of nucleotides (A,T,C,G) one type at a time over a PicoTiter-Plate filled with the beads, each cell in the plate containing a single bead. The complementary strand of the
1. INTRODUCTION

fragments is elongated by a polymerase at each cycle, starting from a generic primer. Complementary nucleotides are incorporated when flown over the cell. This incorporation emits light through the release of a pyrophosphates (PPi). The emitted signals are translated into a sequence of \{A,T,G,C\} letters, called reads. The advantage of the 454 platform is that it generates relatively long read lengths (up to 1kb). However, it suffers from a high indel (insertions/deletions) error rate in particular in homopolymeric regions; if multiple identical nucleotides are incorporated in a single flow, the exact number has to be inferred from the intensity of the signal, which can be inaccurate.

The Illumina platform is similar to the 454 platform, i.e. it uses fragmentation, size selection, fragment amplification and sequencing-by-synthesis. Amplification is performed on a plate instead of beads in emulsion. A plate is sparsely populated with adapter-ligated single stranded DNA. Bridge PCR amplifies the fragments into clusters of identical fragments. Sequencing is performed with reversible terminator chemistry, which incorporates a single nucleotide at each cycle, thereby preventing the homopolymer problem of the 454 platform. The Illumina platform has a higher throughput than the 454 platform and is currently cheaper per sequenced base, but offers shorter read lengths (50-150 bp).

The DNA fragments that have been hybridised to plates in the Illumina sequencing protocol have different adapters at their 5’ and 3’ ends. Paired-end reads are obtained by first sequencing from one adapter, and once finished, sequencing from the other adapter (Fig. 1.2F). As such, two reads from both sides of the fragment are obtained, one from the Watson and one from the Crick strand, which are at a known physical distance (200-500 bp) separated on the genome. The mate-pair protocol brings genome fragments that were originally separated 3-5 kbp’s together into smaller (400-600 bp) fragments, which are again sequenced from both sides on the Illumina sequencer (Fig. 1.2A-D). The mate-pair sequencing protocol on the 454 machine is similar to Illumina’s protocol, but here a long single sequencing run reads the biotin-purified fragment all the way through the linker that was used to generate the circular fragments (Fig. 1.2E). The paired sequences are then obtained by splicing-out the linker sequence in silico.

1.2.2 Whole genome reconstruction

Whole genome shotgun sequencing (WGS) is the commonly used approach for sequencing genomic DNA. The NGS technology can only read a limited number of continuous nucleotides in a sequence, whereas genome lengths are in the range of millions (bacteria) to billions (higher eukaryotes) of nucleotides. Reconstructing the original genome from these short reads, called de novo assembly, is done by computers and poses many challenges to be solved in the field of bioinformatics. The complications that arise in the assembly process are often compared to making a jigsaw puzzle (178). First, a genome has many (almost) identical regions, similar to large patches in a jigsaw puzzle with just a single colour. There can be many ways to reconstruct these patches and connect them to the rest of the puzzle. Second, during the reading of the DNA errors are made, which result in pieces of the puzzle with damaged edges that do not fit anywhere or, worse, fit where they do not belong. Third, the complexity of a jigsaw puzzle does not increase linearly.
1.2 DNA sequencing and its computational challenges

**Figure 1.2: Mate-pair (A-E) and paired-end (F) sequencing** - A) Fragments of 2-20 kbp are selected and labeled with biotin. B) Fragments are circularised with a linker and randomly sheared into 400-600 bp fragments. C) Sheared fragments with biotin labels are affinity purified. D) In Illumina’s mate-pair protocol, purified fragments are sequenced from both ends, which were originally separated by several kbp’s. E) In 454’s mate-pair protocol, purified fragments are sequenced from one end through the linker. F) In Illumina’s paired-end protocol 200-500bp fragments are selected and sequenced from both ends.

with the number of pieces, e.g. a puzzle with thousand pieces is more than twice as hard as a puzzle with five hundred pieces. The DNA puzzle consist of millions to billions of pieces, many of which are identical or damaged, which leads to a range of computational challenges.

To reconstruct a genome from sequencing reads it has to be oversampled to ensure sufficient overlap between them to be merged. The number of times a genome is oversampled is referred to as the depth of coverage, e.g. a genome that has a length of 20 Mbp is sequenced at 3X depth of coverage if the lengths of the reads sum to 60 Mbp. The depth of coverage is usually not uniformly distributed across a genome, as it is influenced by genomic features such as the GC-content. Therefore, high coverage is necessary to ensure every base of the genome is seen.

**De novo assembly**

Automatic procedures that attempt to merge reads into longer contiguous sequences, called contigs, are known as de novo assembly algorithms. We can distinguish three main de novo assembly paradigms:

**Greedy** assembly was one of the first paradigms to be used in several assemblers, such as PHRAP (51) and the TIGR assembler (216), and is quite intuitive. It starts by assigning each read to a contig, which are then iteratively merged. In each iteration the two contigs with the best overlap (according to some heuristic) are merged. The main disadvantage of greedy assembly is that it creates chimeric contigs when repeats are present in a genome. Detecting contigs that are repetitive in the genome can avoid making these misassemblies,
1. INTRODUCTION

although this in turn leads to highly fragmented assemblies. Greedy assembly has lost most of its popularity since the advent of overlap-layout-consensus assembly.

Overlap-layout-consensus (OLC) assembly, like greedy assembly, is based on the calculation of pairwise overlaps between reads, but represents reads and overlaps in a fragment assembly string graph \(^{158}\). String graphs are simplified by transitive reduction and subsequent unitigging, a process of collapsing simple paths without branches. The result is the bidirected contig string graph, in which nodes represent contigs and edges represent reads spanning two contigs (Fig. 1.3B). The contigs cannot be unitigged further, since for each pair of contigs connected by an edge the in- or out-degree (dependent on strand) is larger than 1. String graph assemblers were first developed in the Sanger era, to assemble relatively long sequencing reads, but they are also suitable for 454 reads. Popular string graph assemblers are Minimus \(^{211}\) and Newbler (Roche).

De novo assembly using OLC assemblers of next-generation sequencing reads from Illumina machines was for a long time infeasible, since the short read length required such high depth of coverage to have enough overlap between the reads that calculating the pairwise alignments and the resulting graph became intractable. Recently though, the development of assemblers using efficient indexing and compression made it possible to use string graphs for the assembly of such short reads \(^{208}\).

De Bruijn graph assemblers, such as Velvet \(^{258}\) and Abyss \(^{209}\), were specifically developed for short next-generation sequencing reads, such as the produced by Illumina. These assemblers construct a de Bruijn graph instead of a string graph (Fig. 1.3A). First all unique \(k\)-mers, i.e. all substrings of length \(k\) found in the reads are stored together with the number of occurrences of this \(k\)-mer. In the de Bruijn graph, nodes represent \((k-1)\)mers and edges connect two nodes if there is a \(k\)-mer in the data that has one such \((k-1)\)mer as a prefix and the the other as a suffix. The main advantage is that the number of nodes in the de Bruijn graph is proportional to the genome size instead of to the number of reads (as in the string graph). The original genome can be reconstructed by finding a Eulerian path (or Eulerian cycle in case of a circular genome) through the graph, visiting each edge exactly once. In practice the Eulerian path cannot be found, since this is only possible under the condition that all \(k\)-mers are unique in the genome, all \(k\)-mers in the genome are in the set of reads and the \(k\)-mers are error free. In a typical genome sequencing project these assumptions do no hold and the result of an assembly is a set of contigs, as in OLC assembly.

Comparative assembly

In the Sanger sequencing era only one whole genome sequencing project per species was performed, i.e. the genome of a single individual was chosen to be representative for the species. These reference genomes generated a wealth of information on the number of genes a species has and their sequences. The sequencing of multiple genomes per species,
1.2 DNA sequencing and its computational challenges

TATCTTAGCTTCA

Figure 1.3: Assembly graphs - A sequenced toy genome (top) with four reads (bars below genome). Left: A de Bruijn graph generated from the reads with $k = 3$. Nodes in the de Bruijn graph are $(k-1)$-mers. Nodes are connected with an edge if there exists a $k$-mer that has one node as prefix and the other as suffix. Right: A fragment assembly string graph, or overlap graph, where nodes are reads that are connected by an edge if they overlap. The dashed edge is removed by the transitive reduction algorithm [158], since there is a path that includes read AGCTT that spells the same sequence.

such as individual yeast strains, became economically feasible only with the introduction of next-generation sequencing.

De novo assembly is hard with short reads and generates fragmented assemblies. The availability of a closely related (finished) genome can help improve the assembly quality by using a process called comparative genome assembly that involves mapping of the reads to a reference genome before assembling them into a genome [179]. Mapping against a closely related genome will only yield those parts that are identical in target and template genome. To get the unique components in the target genome, a de novo assembly will always be required.

Scaffolding

Contigs are typically shorter than the chromosomes that are targeted in a genome assembly project. Genomic repeats longer than the read length and sequencing errors cannot be unambiguously assembled. Scaffolding is a post-processing step of an assembly, in which contigs are placed in the right order, orientation and at approximately the right distance. Additional data than just the read data is required, such as paired-end reads (Fig 1.2) that contain information on relative orientation, order and approximate physical distance. Additionally, other sources providing information on distance and orientation can be used, such as reference genomes and restriction maps.

In 2010 few scaffolders were available that could operate stand-alone and make use of additional data sources, to our knowledge only Bambus [180] and SOPRA [48], although the latter was not originally designed for this purpose. In Gritsenko et al. [85] we described
1. INTRODUCTION

a novel scaffolding algorithm called GRASS that can be applied to any type of scaffolding information.

1.2.3 Comparative genomics using next-generation sequencing

Classes of genomic variation

Genomes can differ in many ways, i.e. from small local differences such as single nucleotide variations (SNVs) and short insertions and deletions (indels) to the larger structural variations (SVs). SVs are traditionally defined as regions larger than 1kbp, but with the increasing resolution to detect SVs currently also regions between 50 bp and 1 kbp are included.

SVs can be divided into balanced and unbalanced mutations. Balanced mutations, such as inversions and translocations, preserve the copy number of a given allele, whereas unbalanced mutations, such as indels and duplications, change the number of copies of the involved allele. Differences between genomes in the latter respect are referred to as copy number variation (CNV).

Detection of variation using read mapping

Given DNA sequencing datasets of two strains, the most intuitive way of discovering genomic differences between them is by assembling both genomes from the reads and comparing the reconstructed genomes by aligning them. Unfortunately, given the read length it is impossible to reconstruct genomes accurately enough. Therefore, the scientific community resorted to read-mapping approaches, which require reconstructing the genome of a single individual to serve as a template. The reads from other genomes are then mapped to this reference genome and differences between the target and reference genome are inferred from aberrant read mappings, read-depth variation, split-read mappings, etc. The patterns of the mapped reads that indicate a class of variation is referred to as the sequence signature for a particular SV class.\(^5\)

Reference-free variation detection

Many situations in biological and medical research arise where the reference genome is not the most logical sequence to compare to. For example, in matched tumor-normal pairs one seeks to find differences between samples from healthy and tumor tissue. Although comparing these paired samples directly to each other is the most intuitive, the samples are currently compared via a reference genome, since few tools are available for direct comparison. The indirect mapping-based comparison requires extensive post-processing of the detected variants to eliminate the variants between the sequenced individual and the reference genome, as we demonstrated in chapter\(^3\).

Matched pairs arise from laboratory evolution similarly as in matched tumour-normal studies. Microorganisms are evolved under certain selective pressure in the laboratory for weeks to months, and in some cases even for years to select for a certain phenotype.
Sequencing both the evolved strain and its pre-evolution ancestor can elucidate the mutations that have occurred in the DNA during the laboratory evolution period (Fig 1.4), thereby giving insight in genotype-phenotype relations. These matched pre- and post-evolution sequencing samples are much closer related to each other than to an available reference genome. The *S. cerevisiae* laboratory strains used in our evolution studies diverge up to 0.5% (~60,000 nt) from the reference genome, while during a lab evolution only few changes are expected. For example, laboratory strain CEN.PK113-7D has approximately 20,000 single nucleotide variations compared to the reference genome of strain S288C, whereas two parallel laboratory evolutions of two of these strains resulted in only 7 resp. 10 single nucleotide variations compared to their ancestor.

Reference-free genome variation detection methods are particularly useful for metagenome sequencing projects. The DNA of metagenomes is directly sampled from the environment, without purification of specific microbes. To get the complete picture of all DNA present, its biological functions and the genetic variation, whole-metagenome sequencing is performed. Sequence level variations in metagenomes have until now hardly been analysed. Conversely, in isolated genomes genetic variation is a major research theme, which is usually performed with reference-based methods. Reference-based variation detection works for single genomes, but these tools are less suitable for metagenomes, since reference genomes are usually not available for all species in the sample, and additionally the species present are unknown. Cultivation of many bacterial species in the lab is not possible, which complicates construction of reference genomes. Reference-free methods are therefore the intuitive choice for metagenomes.

![Reverse engineering workflow](image.png)

**Figure 1.4: Reverse engineering workflow** - A. *S. cerevisiae* strains are evolved under selective pressure in serial batch reactors. B. The DNA of the evolved strain is sequenced, generating single-end, paired-end or mate-pair read libraries. C. The sequencing reads are processed by either mapping them to a reference genome or assembling them *de novo*. D. DNA variants are detected from the read mapping or by analysing the *de novo* assemblies. E. The list of found mutations is filtered, prioritised and associated to the phenotype. F. The associated mutations are validated by introducing them in the starting strain and testing it for the phenotype.
1. INTRODUCTION

1.3 Contributions of this thesis

This thesis contributes to several aspects of the reverse metabolic engineering cycle (Fig. 1.1). It describes the development of graph-based and statistical analysis methods for next-generation sequencing data and the application to biological challenges encountered in industrial microbiology research. The developed algorithms and the assembled and annotated CEN.PK113-7D genome relieve bottlenecks in the reverse engineering workflow (Fig. 1.4) and help close the cycle (Fig. 1.1), as exemplified by the two reverse engineering projects in chapters 4 and 6.

We generated a high quality reference genome for the laboratory yeast strain *S. cerevisiae* CEN.PK113-7D, which is used as starting strain for several evolutionary engineering projects (Fig. 1.4A). We started with 50 bp reads generated by the Illumina platform, which did not yield a satisfactory assembly. Therefore, we complemented our sequencing data with longer reads from the 454 platform. It was not trivial to efficiently deal with these two sequencing types using a single available assembler. Additionally, there was a high quality reference genome available (strain S288C) that could be used as template for comparative assembly. Making optimal use of the data available to assemble the CEN.PK113-7D genome required the application of de Bruijn graph, OLC and comparative assemblers. In chapter 2 we present MAIA (Multiple Assembly IntegrAtor), an overlap graph-based integrator, to integrate the resulting *de novo* and comparative assemblies (Fig. 1.4C).

In chapter 3 we describe the reconstruction and analysis of the CEN.PK113-7D genome. Here, we chose to construct a genome not biased w.r.t. any reference genome and therefore integrated only assemblies generated by *de novo* assemblers. Comparison to the genome of S288C uncovered several differences, such as the remarkable discovery that CEN.PK113-7D is, unlike S288C, a biotin prototroph having the full set of BIO genes in its genome required for biotin biosynthesis.

The CEN.PK113-7D is useful as resource for the community using this laboratory strain. However, the genome sequence is still in draft status and we found such a draft genome to be suboptimal for variant discovery. Additionally, we worked with other strains such as two-species hybrid lager brewing yeasts with aneuploid genomes for which no good reference genomes were available. For this reason we ventured into developing reference-free variation detection methods (Fig. 1.4D). In chapter 5 we present a coassembly-based algorithm, called Magnolya, for CNV detection between matched sequencing pairs from a contig string graph using a Poisson mixture model. We used Magnolya to analyse copy number variation of the aneuploid genomes of lager brewing yeasts.

In chapter 7 we present the MaryGold algorithm to reference-free detect sequence variation within metagenomes by analysing contig graphs (Fig. 1.4D). Finding variation within a single metagenome also allowed us to improve assemblies. Additionally, MaryGold can detect variation between multiple metagenomes through co-assembly of all reads. To our knowledge no tool was previously available to perform reference-free genome variation discovery, which is set to become a common analysis as more metagenomics datasets will become available.
Using the the sequenced CEN.PK113-7D genome and our developed algorithms to detect DNA variants, such as aneuploidy, we could employ next-generation sequencing to fully reverse-engineer two laboratory evolved yeast strain (Fig. 1.4A-F). First, in chapter 4 we show that a single base pair mutation transformed an acetate transporter $ADY2$ in an efficient lactate transporter. The part of chromosome III that is occupied by $ADY2$ was found to be triplicated, thereby increasing the gene dosage of this, given the imposed selective pressure, important allele. Second, in chapter 6 we establish that a combination of aneuploidy and a frame shift mutation in the $ACE2$ gene were responsible for the formation of multi-cellular aggregating yeast. Although the $ACE2$ mutation could be related functionally related to the phenotype, polyploidy was found to be essential for the yeast’s aggregating behaviour.
Integrating genome assemblies with MAIA

Jurgen Nijkamp, Wynand Winterbach, Marcel van den Broek, Jean-Marc Daran, Marcel Reinders and Dick de Ridder

Published in *Bioinformatics*
ECCB 2010, Volume 26, pages i433-i439
Abstract

Motivation: De novo assembly of a eukaryotic genome with NGS data is still a challenging task. Over the past few years several assemblers have been developed, often suitable for one specific type of sequencing data. The number of known genomes is expanding rapidly, therefore it becomes possible to use multiple reference genomes for assembly projects. We introduce an assembly integrator that makes use of all available data, i.e. multiple de novo assemblies and mappings against multiple related genomes, by optimizing a weighted combination of criteria.

Results: The developed algorithm was applied on the de novo sequencing of the *Saccharomyces cerevisiae* CEN.PK 113-7D strain. Using Solexa and 454 read data, two de novo and three comparative assemblies were constructed and subsequently integrated, yielding 29 contigs, covering more than 12 Mbp; a drastic improvement compared to the single assemblies.
2.1 Introduction

Next generation sequencing (NGS) platforms, such as Roche/454, AB/Solid and Illumina/Solexa, allow for gigabytes of data generation at an affordable cost. The third generation sequencing platforms (Helicos, Pacific Biosciences) may even let the cost per megabase drop under $1 per megabase. Considering the relatively low cost of these platforms, compared to classical Sanger sequencing, it becomes possible to use them for de novo sequencing projects. However, the millions of short DNA sequences generated by NGS platforms, called reads, are still relatively small. Given this limited read length and the many repetitive regions in a eukaryotic genome, de novo assembly is still a challenging task. To alleviate this problem it is essential to design algorithms that make full use of all available data.

Over the past few years, several assemblers have been developed for NGS data. Assemblers pull millions of reads together into larger contiguous sequences, called contigs. A typical assembly of a eukaryotic genome is a set with thousands of contigs. These contigs are unordered as well as unoriented, i.e. it is unknown whether they come from the forward or reverse strand. The process to determine orientation and relative ordering of contigs is called scaffolding. Some assemblers have built-in scaffolders; otherwise, an external scaffolder can be used, such as Bambus. An alternative to de novo assembly is mapping the reads against a finished or draft genome from a close relative (a template). From such a mapping a consensus can be called, generating a comparative assembly. As the number of known genomes is growing rapidly, in the future, it will be more often the case that multiple close relative genomes are available to create such assemblies. However, mapping against a closely related genome will only yield those parts that are identical in target and template genome. To get the unique components in the target genome, a de novo assembly will always be required.

Assemblers are often specialized for a specific type of reads. De Bruijn graph-based assemblers, such as Velvet, Abyss and ALLPATHS are most suitable for short reads (Solid, Solexa), whereas overlap-layout-consensus algorithms, such as Newbler (Roche) and CABOG, are more suitable for the longer 454 sequences. It is not trivial to deal efficiently with different read types simultaneously or to combine outputs of different assemblers. Hybrid strategies (using two types of sequencing data) mostly work by altering the output of a first assembler to make it suitable for application on a second. Reinhardt et al. generated contigs with VCAKE using Solexa data, which were subsequently used as input to the Newbler assembler together with 454 data. Goldberg et al. simulated Sanger reads from a set of contigs assembled by Newbler with 454 data. These reads were subsequently used as input to the Celera assembler combined with true Sanger reads. We are aware of only one de novo assembler designed to integrate Sanger and NGS data, called Forge. However, Forge does not allow for integration of comparative assemblies. Other hybrid strategies combine assemblies using Minimus. Minimus is restricted to only two assemblies, so to combine three or more assemblies it has to be applied iteratively. Minimus also does not allow for weighted combinations of contigs.
2. INTEGRATING GENOME ASSEMBLIES WITH MAIA

A. Perform de novo and comparative assembly

De novo assembly 1 (Abyss)
De novo assembly 2 (Celera)
Mapping against related genome 1 (S288c)
Mapping against related genome 3 (RM11-1A)

B. Calculate pairwise overlaps between contigs

C. Construct overlap graph, determine start and end node and weigh edges with Z-scores

D. Determine orientation by depth-first traversing the graph in order of weights

E. Edge direction follows from end-to-end alignments

F. Find the highest scoring path using a Tabu search and call consensus

Reference genome
Integrated assembly

Figure 2.1: An overview of the process of integrating several assemblies with MAIA. A. Multiple de novo and comparative assemblies are created using specialized assemblers. B. The resulting contigs are pairwise aligned to each other to find end-to-end overlaps. C. An overlap graph is constructed, in which nodes represent contigs and edges represent overlaps. A forward and a reverse edge is added between the pairs of nodes, but these are indicated by an undirected edge for simplicity. A start node and an end node is determined using a reference genome. Edges are assigned weights based on several properties of the alignments and contigs, combined using weighted Z-scores. D. An orientation is assigned to the contigs by traversing the graph depth-first in order of weight (indicated by the numbers). Edge 9 (dashed box in panel B) assigns reverse orientation to the blue node, while a forward orientation has already been assigned via edge 1, therefore it is recognized as conflicting and it is removed. E. Oriented contigs and end-to-end overlaps form a directed graph. F. The highest scoring path is found using a Tabu search procedure, which leads to the assembly of a chromosome.

In this paper, we describe MAIA (Multiple Assembly IntegrAtor), a graph-based algorithm for integration of several de novo and comparative assemblies. Assembly integration is related to both de novo assembly and scaffolding, but differs in its input. An assembler deals with short sequences (reads) and high genome coverage to account for read errors and repeats in the genome. A scaffolder tries to determine the relative ordering and orientation of large sequences (contigs) of a single de novo assembly, assuming the target genome is covered once. An integrator is a hybrid of these, dealing both with contigs and manifold genome coverage, allowing a number of assemblies to be considered simultaneously. MAIA is not restricted in the number of assemblies and uses the full contigs produced, not requiring these to be broken into reads or k-mers of any type. Pairwise alignments of contigs are calculated to generate an overlap graph. In this graph nodes represent contigs and edges represent alignments. These edges are weighted with several properties of the contigs and alignments, which are combined using weighted Z-scores. Assemblies are integrated at chromosome level by finding the combination of contigs which yields the highest score. This is achieved by optimizing a path in the overlap graph between the contigs that align closest to the 5’ and 3’ ends of a reference genome. The assembled chromosome directly follows from this path.

The MAIA approach has two main advantages. First, multiple known related genomes
2.2 Methods

MAIA is an assembly integrator using the overlap-layout-consensus paradigm, known from genome assembly algorithms, to combine several assemblies into a single integrated assembly. The algorithm takes as input sets of contigs, each set originating from either a de novo or a comparative assembly, i.e. from mapping against a related genome (Fig. 2.1A). Overlap between contigs is detected by pair-wise aligning the contigs in an all-vs-all fashion among the sets (Fig. 2.1B). An undirected overlap graph is then constructed, with nodes representing contigs and edges representing overlaps. Using a reference genome, i.e. the evolutionary closest of the related genomes available that is of high quality, a 5' start and 3' end node is determined to guide the integration. (Fig. 2.1C). The edges are weighted, reflecting the likelihood that the alignment represents an actual overlap in the genome (Fig. 2.1D). The graph is then directed by assigning orientation to the contigs, i.e. forward or reverse (Fig. 2.1E). Assembly integration is finally achieved by finding a highest scoring path between the start and end nodes in the overlap graph and calling the consensus (Fig. 2.1F). These steps are described in detail below.

2.2.1 Constructing an overlap graph from pairwise alignments of contigs

A graph $G = (V, E)$ is created in which $V = \{v_1, v_2, ..., v_n\}$ is the set of nodes and $E = \{\{v_i, v_j\}|v_i, v_j \in V\}$ is the set of edges. Each contig $c \in C$ is assigned to a node. Overlapping regions in contigs are detected by pair-wise aligning all contig pairs in different sets. For every aligned pair of contigs, two filters are applied. First, only the longest mutually consistent set of alignments is selected. Second, if there still is more than one match between two contigs, only the longest is retained. For these steps we used Nucmer and Delta-filter, both part of the Mummer package [53], although other
tools could be used. The resulting alignments \( a \in A \) are used to generate the edges in \( G \). Contigs \( v_i \) and \( v_j \) that overlap end-to-end with a minimum alignment length \( l^{A,\text{min}} \) and maximum length \( l^{O,\text{max}} \) of non-aligned overhang (Fig. 2.2), i.e. the part of the contig that will be clipped when merging the two, are then joined by a forward and a reverse edge, \((v_i, v_j)\) and \((v_j, v_i)\).

### 2.2.2 Weighting the edges in the overlap graph

A score \( Z(e) \) is assigned to each edge \( e \in E \) to reflect the quality of the alignment and the quality of the contig to which the edge leads. Edge weights differ between forward and reverse edges. For three properties of contigs and alignments (Fig. 2.2), a \( p \)-value is calculated. Null distributions for these properties are inferred using all contigs and possible contig pairs. These distributions reflect the probability that the property occurs by chance in a pair of contigs, which do not overlap in the target genome. The \( p \)-values are transformed into \( Z \)-scores using the inverse of the cumulative density function \( N^{-1} \), and combined into Liptak-Stouffer’s weighted \( Z \)-score, where the weights \( w_i \) are user-specified per property (101):

\[
Z(e) = \frac{1}{\sqrt{\sum w_i^2}} \sum_{i=1}^{3} w_i N^{-1}(1 - p_i(e)) + w_4 Z_4(e) \tag{2.1}
\]

For each edge \( e \in E \) the following four properties (illustrated in Fig. 2.2) are calculated:

1. The length of the contig; longer contigs are preferred over smaller contigs. A \( p \)-value for a particular contig \( c \) is estimated as \( p(l^C \geq l^C_c) = \frac{|\{c' \in C | l^C_c \geq l^C_c\}|}{|C|} \), where \( l^C_c \) is the length of the contig \( c \) to which edge \( e \) points and \( |C| \) is the total number of contigs.

2. The length of the alignment; longer overlap between contigs is preferred. For the calculation of the \( p \)-value, only the number of correctly aligned nucleotides \( l^A = l^A \cdot f^A \) are considered, where \( l^A \) is the full length of the alignment and \( f^A \) is the fraction of aligned nucleotides that are identical. The \( p \)-value for a particular alignment \( a \) is estimated as \( p(l^A \geq l^A_a) = \frac{|\{a' \in A | l^A_{a'} \geq l^A_a\}|}{|A|} \) where \( |A| = \prod_{x=1}^{n} |C_x| \) is the total number of possible contig pairs, \( x \) is the assembly number and \( n \) is the total number of assemblies.

3. The percentage of non-aligned overhang; the length of the non-aligned overhang \( l^O \) should ideally be zero. For a particular alignment \( a \) a \( p \)-value is calculated as \( p(l^O \leq l^O_a) = \frac{|\{a' \in A | q_{a'} \leq q_a\}|}{|A|} \), where \( q_a = \frac{l^O_a}{l^O_a + l^O} \) is the fraction of non-aligned overhang and \( l^O \) is the number of nucleotides that have to be clipped if the two contigs would be merged. We consider the number of nucleotides overhang relative to the contig length to avoid connecting small contigs with large overhangs.
Finally, a manually assigned score $Z_4$ is added for the quality of the assembly, which can differ per assembly source. This score attribute is used to have MAIA prefer high quality assemblies.

### 2.2.3 Directing the overlap graph

All contig alignments are end-to-end and can be represented as directed edges in the overlap graph. The direction of each edge depends on the orientation of the contigs it connects. If the upstream end of node $v_i$ aligns to the downstream end of node $v_j$, the edge in the graph would be $e = (v_i, v_j)$. Since the orientation of the contigs is unknown, taking the reverse complement of the two contigs flips the edge to $e = (v_j, v_i)$. These two edges represent the forward and reverse strands of the DNA. Since only one strand needs to be assembled, the orientation of the contigs is fixed.

Assigning an orientation to the contigs can cause problems, by introducing cycles in the graph that disagree on orientation. These cycles are caused by alignments of contigs that are not actually overlapping in the genome. Edges causing these conflicts have to be removed. An optimal solution would be to assign an orientation to the contigs which minimizes the number of conflicting edges. Since this problem is NP-hard, a greedy approach is used, similar to the contig orientation method in Bambus [180]. This approach starts by fixing the orientation of the start node to forward. Next, the graph is traversed depth-first in order of descending weights. For every node an orientation is assigned based on the alignment and orientation of the previously visited node.

The contig orientation is illustrated by the example in Fig. 2.2, in which arrows represents contigs and dashes their alignments. Node $v_1$ is the start node and has a fixed forward orientation. The graph is traversed to $v_2$. Since the reverse complement of $v_2$ aligns to $v_1$ (opposing arrows), a reverse orientation is assigned to $v_2$. Subsequently, using the same reasoning, a forward and reverse orientation will be assigned to $v_3$ and $v_4$, respectively. If $v_4$ had already been visited and was assigned a forward orientation, the edge between $v_3$ and $v_4$ conflicts with the previously assigned orientation and will be removed from the graph. After all nodes have been oriented it is known for each end of a contig whether it is the up- or downstream end. The end-to-end alignments can now be used to direct the graph, e.g. as node $v_2$ aligns to the downstream end of node $v_1$, the directed edge will be $e = (v_1, v_2)$.

### 2.2.4 Finding the highest scoring path

A chromosome can be assembled by finding a simple path $P = v_0v_1v_2...v_k$ connected by edges $e_1e_2...e_k$ in the overlap graph, visiting no node more than once (Fig 2.1F). A start node $v_0$ and end node $v_k$ are determined to avoid having to evaluate paths between all possible node pairs in the graph. These nodes are set to be those contigs that originate from the 5’ and 3’ ends of comparative assembly against the reference genome (Fig. 2.1C). The combination of contigs connecting $v_0$ and $v_k$ is optimized by maximizing the sum of
2. INTEGRATING GENOME ASSEMBLIES WITH MAIA

the edge scores \( S(P) = \max(\sum_{e \in P} Z(e)) \). This optimization can be shown to be \( \mathcal{NP} \)-complete by taking an instance of \( G \) with only positively weighted edges, thereby reducing the maximization to a search for a Hamiltonian path, which is known to be \( \mathcal{NP} \)-complete. This makes finding the global optimum computationally expensive; therefore we search for the highest-scoring path using a Tabu procedure (80).

The Tabu search starts by finding an initial solution for \( P \) by performing a Dijkstra shortest path search on the graph with inverted edge weights \( \tilde{Z}(e) \). These inverted weights are calculated for each edge \( e \) as \( \tilde{Z}(e) = \max_e (Z(e)) - Z(e) + 1 \). The Tabu search proceeds by systematically applying the change to the path that yields the most improvement in terms of \( S(P) \). All pairs of adjacent edges in \( P \) are considered for modification (2-opt).

Four modifications are possible to a set of two edges. Fig. 2.3 shows an example for a set of three nodes \( v_i, v_j \) and \( v_k \), connected by the edges \((v_i, v_j)\) and \((v_j, v_k)\). The possible modifications are: (1) \( v_j \) is bypassed by directly connecting \( v_i \) and \( v_k \) with edge \((v_j, v_k)\); (2) \( v_i \) and \( v_k \) are connected via a fourth node \( v_l \); (3) \( v_i \) or \( v_j \) are connected via \( v_l \) and (4) \( v_i \) and \( v_j \) are connected via \( v_l \). After the change has been applied, the inverse of the change (the "undo") is stored in the Tabu list. Changes in the Tabu list are not allowed to be applied to avoid entrapment in cycles of repeated identical changes. After a certain number of cycles (here, 3) the change is removed from the Tabu list. The algorithm proceeds until for a certain number of changes (15) no improvement is seen compared to the best path found so far. As the initial solution is often close to the final one, convergence is often fast. In practice, the algorithm is limited by memory size (to hold the overlap graph) rather than computational complexity.

2.2.5 Connecting unconnected subgraphs

If no path exists between start and end node, contigs from one or more assemblies are aligned to the reference. If a region of the reference genome is not covered by an aligned contig, a pseudo node \( \tilde{v} \) is created, containing the DNA sequence of this non-covered region (Fig. 2.4). Edges with a low score (i.e. a penalty) are inserted between \( \tilde{v} \) and the pair

**Figure 2.3:** Four types of modifications which are applied iteratively to the path by the Tabu search procedure.
of nodes that align on both sides of \( \tilde{v} \). The number of pseudo nodes in the graph is kept minimal by gradually increasing the allowed pseudo node size until a path between start and end node is found.

### 2.2.6 Post processing of the path

The start and end node determined using the reference genome are not necessarily the ends of the target chromosome. Therefore, the path \( P \) is greedily extended towards the 5’ and 3’ extremes of the target genome. \( P \) is iteratively extended from the current last node to the node connected with the highest edge weight, provided that is has a specified minimum alignment length \( l_{A,\text{min}}^{C,\text{min}} \) and minimum percentage alignment identity \( f_{A,\text{min}}^{C,\text{max}} \).

The maximum size of a pseudo node is set to \( \tilde{l}_{c,\text{max}}^{C} \). \( P \) is split at pseudo nodes exceeding \( \tilde{l}_{c,\text{max}}^{C} \) and backtracked to the nearest branchpoint on both sides of the pseudo node. From there on the paths are greedily extended until no extension is possible, similar to the end extension described above, resulting in multiple contigs per chromosome. Fig. 2.4 gives an example of splitting \( P \) at a large pseudo node.

### 2.2.7 Calling the consensus

Finally, the integrated contigs follow from the path found by the Tabu search. The contigs and their associated pairwise alignments are transformed into an alignment matrix with one row for every source assembly. The consensus is called by taking for each column the base or gap (arising from the nucmer gapped alignment) of the highest quality assembly present in that column. In the resulting consensus, gaps are removed and the bases are tagged with the assembly from which they originate. This information can be used in further analyses of the assembly.

![Figure 2.4: A pseudo node \( \tilde{v}_4 \) is inserted in the graph if no path exists between start and end node. An initial path \( P = v_1 v_2 v_3 \tilde{v}_1 v_5 v_6 v_7 \) has been found. If \( \tilde{v}_4 \) exceeds the maximum size \( l_{c,\text{max}}^{C,\text{max}} \), the path is subsequently split into \( P_1 = v_1 v_2 v_3 \) and \( P_2 = v_5 v_6 v_7 \). \( P_1 \) and \( P_2 \) are backtracked (red arrows) to the nearest branchpoints (\( v_2 \) and \( v_5 \), respectively) and greedily extended to \( v_9 \) and \( v_{10} \) yielding \( P_3 = v_1 v_2 v_8 v_9 \) and \( P_4 = v_{10} v_{11} v_5 v_6 v_7 \). Finally, the highest scoring paths are chosen from \( \{P_1, P_3\} \) and \( \{P_2, P_4\} \).

21
Table 2.1: CEN.PK assembly statistics of single input and hybrid assemblies Only contigs ≥ 200 bp were used to generate statistics (two rightmost columns).

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Assembly</th>
<th>Package</th>
<th># contigs</th>
<th>Total size (Mb)</th>
<th>N50 (kb)</th>
<th>Mapped reads (%)</th>
<th>Supporting pairs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>de novo</td>
<td>Abyss</td>
<td>1,223</td>
<td>11.64</td>
<td>20</td>
<td>84.8</td>
<td>97.6</td>
<td></td>
</tr>
<tr>
<td>de novo</td>
<td>Celera</td>
<td>4,148</td>
<td>9.03</td>
<td>3</td>
<td>62.8</td>
<td>98.5</td>
<td></td>
</tr>
<tr>
<td>single input</td>
<td>comparative (S288c)</td>
<td>Maq</td>
<td>375</td>
<td>12.06</td>
<td>162</td>
<td>96.9</td>
<td>99.0</td>
</tr>
<tr>
<td>comparative (YJM789)</td>
<td>Maq</td>
<td>907</td>
<td>11.77</td>
<td>44</td>
<td>90.8</td>
<td>98.3</td>
<td></td>
</tr>
<tr>
<td>comparative (RM11-1A)</td>
<td>Maq</td>
<td>795</td>
<td>11.54</td>
<td>41</td>
<td>78.2</td>
<td>98.5</td>
<td></td>
</tr>
<tr>
<td>hybrid</td>
<td>de novo + comparative</td>
<td>Minimus</td>
<td>634</td>
<td>11.40</td>
<td>72</td>
<td>75.5</td>
<td>97.7</td>
</tr>
<tr>
<td>hybrid</td>
<td>de novo + comparative</td>
<td>MAIA</td>
<td>71</td>
<td>12.21</td>
<td>290</td>
<td>92.1</td>
<td>99.3</td>
</tr>
</tbody>
</table>

2.2.8 Assembly validation

To assess the quality of both the individual assemblies and the MAIA integrated assembly, the paired-end Solexa reads were mapped onto the assemblies using BWA (136). Two statistics were extracted from the mappings using Samtools (138). First, to assess the completeness of an assembly, the percentage of reads that mapped on the assembly was calculated. Second, to assess the accuracy of an assembly, the percentage supporting read pairs was calculated. This is calculated as the percentage of the total number of mapped pairs that map at a proper distance from each other on a contig. The insert size distribution $N(208, 13)$ and the maximum allowed insert size ($\sim 6\sigma$) were estimated by BWA.

2.2.9 Experimental setup

DNA of the S. cerevisiae lab strain CEN.PK 113-7D (MATa MAL2-8c SUC2) was prepared (28). A library of 200 bp fragments was created and sequenced paired-end using the Illumina Solexa system, generating $\sim 56$ million paired reads. A second library with mate-pair reads with an insert size of 8 kb was prepared sequenced on the Roche 454 Titanium. Both libraries were prepared according to manufacturer recommendations (Illumina and Roche). The pairing rate for the 454 mate pair library was 19%, yielding 149,900 paired reads.

De novo unscaffolded assemblies were performed with Abyss (209) and the Celera assembler (154) on the Solexa and 454 reads, respectively. Abyss was tested for all combinations of $k$-mer size $\in \{23, \ldots, 33\}$ and coverage cut-off $\in \{0, \ldots, 12\}$; the combination yielding the best N50 was chosen. The Celera assembler was used with standard settings as described in Lee (134). Comparative assemblies were made by mapping the Solexa reads to the (draft) genomes of the S. cerevisiae strains S288c, YJM789 and RM11-1A using MAQ (137). These genomes are 99.3%, 98.4% and 98.0% identical to CEN.PK, calculated by dividing the number of identical bases by the length of the genome. The consensus sequences were split into contigs at every occurrence of an ’N’. Contigs smaller than 200 bp have been discarded.

Integration of the assemblies with MAIA has been performed per chromosome. From the S288c comparative assembly, only contigs originating from the chromosome being assembled were used. A minimum alignment length $l_{A,\text{min}}$ of 20 nucleotides is used for
finding pairwise alignments. MAIA finds all contigs that align end-to-end. The maximum allowed non aligned overhang $l^{O,max}$ was set to 10 nucleotides. Scores for the assembly qualities were set to $Z = 3, 2.5, 2, 1, 0.5$ for the Abyss, Celera, S288c, YJM789 and RM11-1A assembly, respectively, reflecting our beliefs concerning the relative quality of the assemblies. De novo assemblies received the highest $Z$-scores, since these may contain structural variants unique to the target genome. The weights in the combined $Z$-score for the contig length, alignment length, non-aligned overlap and assembly quality were rather arbitrarily set to be 0.35, 0.25, 0.15 and 0.25, respectively, corresponding to the relative importance of the forms of evidence for merging contigs. Pseudo nodes are iteratively added with increasing sizes until a path from start to end node is found. Only contigs from the S288c assembly were used to create pseudo nodes. The edge weight of a pseudo node is set to -10 and maximum pseudo node size $l^{C,max}$ was set to 250.

Two other hybrid methods were applied as a comparison. First, a de novo assembly with Velvet was performed on a combination of the Solexa reads and 454 data-based contigs pre-assembled by the Celera assembler. The parameters ($k$-mer size and coverage cut-off) were optimized w.r.t. the N50. The paired-end information of the Solexa reads was then used for scaffolding. Second, an assembly integration with Minimus was performed by merging two assemblies and iteratively applying Minimus to the merged result and a next assembly, whereby the singletons were discarded in every step. The order of combination was: S288c + Abyss, + Celera, + YJM789, + RM11-1A.

## 2.3 Results and Discussion

MAIA has been developed to integrate multiple assemblies. An integrated assembly of the *S. cerevisiae* lab strain CEN.PK 113-7D, from hereon called CEN.PK, has been constructed to demonstrate the algorithm.

### 2.3.1 Individual assemblies are fragmented and vary in error rates

Two de novo and three comparative assemblies have been made for CEN.PK. The results for the individual assemblies are shown in Table 2.1. Despite the high genome coverage ($\sim 160X$ for the Solexa and $\sim 20X$ for the 454 data), the Abyss and Celera de novo assemblers generated fragmented assemblies, with an N50 of 20.3 and 2.7 kb, respectively. The N50 is the smallest possible contig length, such that the sum of lengths of all contigs $c' \in C$ with $l^{c'} \geq N50$ is at least 50% of the total assembly size.

The level of fragmentation of the comparative assemblies depends on evolutionary closeness and quality of the genome. The comparative assembly using S288c as template yields the best individual assembly, covering 12.06 million nucleotides with only 375 contigs. The available S288c genome is of high quality and evolutionary closer to CEN.PK than the other strains [201]. Most reads could be mapped to the S288c comparative assembly, which is therefore the most complete; only 3.1% of the reads could not be mapped. The least number of reads mapped to the Celera and Velvet hybrid assemblies. Running Velvet to assemble only Solexa reads (results not shown) allowed 10% more reads to be
mapped. That the use of more 454 reads lowers the percentage of mapped Solexa reads, hints at 454 data quality problems.

2.3.2 MAIA drastically lowers the number of contigs

The number of contigs larger than 200 bp in the individual source assemblies range from 375 to 1,223. MAIA reduces this to 29. Most chromosomes have been assembled in a single contig, except for chromosomes 1, 3, 8, 10, 12 and the mitochondrial DNA, which consist of 5, 4, 3, 2, 2 and 2 contigs, respectively. These chromosomes are known to be relatively divergent among *S. cerevisiae* strains. Schacherer *et al.* (201) showed deleted regions in every one of these chromosomes using a whole-genome tiling array. The most apparent of these deletions is the 10 Kbp deleted region at the extreme of the left arm of chromosome 1, which is also seen in the MAIA assembly (Fig. 2.5). The splits in the chromosomes assembled by MAIA are generally observed near their ends, which are known to be divergent regions in yeast (7). Divergent regions can benefit less from comparative assemblies and therefore MAIA cannot not fully close the genome.

The final integrated CEN.PK genome is compiled of five source assemblies; two *de novo* and three comparative assemblies. Four additional MAIA runs were performed where in each step one of the assemblies was incrementally added to its input, starting with only the S288c comparative assembly. Fig. 2.4 shows each individual input assembly positively contributes to the final result. Table 2.1 and Fig. 2.5 show the assemblies and their use for integration. The usage differs from only 0.8% for the comparative assembly with RM11-1A as template to 80% with S288c as template. The S288c genome is fully finished and of high quality. S288c and CEN.PK are both laboratory strains, known to be evolutionary close (201); therefore mapping yields large contigs. Both contig quality and contig length are reflected in the Z-scores on the edges in the overlap graph. Therefore, MAIA has a preference for the S288c contigs and often selects them for integration. On the contrary, the RM11-1A genome is a draft genome composed of a set of supercontigs. RM11-1A is a phylogenetically more distant wine strain and therefore contributed to a much lower extent in the final assembly. The MAIA assembly contains 0.04% sequences from 102 pseudo nodes, which are 4340 nucleotides in stretches individually not larger than 250 bp. These sequences do no originate from read data, but from the reference genome.

As an illustration the overlap graph of chromosome 9 is shown in Fig. 2.8. All five input assemblies are used to construct chromosome 9 of CEN.PK. The contigs of the YJM789 comparative assembly have been grouped by the chromosome from which they originate and divided among three levels in the layout, indicated by the arrows in Fig. 2.8. Contigs originating from YJM789’s chromosomes 14 and 15 appear in this graph because of repeat sequences that are present in both these chromosomes and CEN.PK’s chromosome 9. Although these repeat induced connections are present, the Tabu search does not include them in the path. Only the contigs originating from YJM789’s chromosome 9 are incorporated in the MAIA integrated chromosome 9 of CEN.PK.
2.3.3 MAIA integrates assemblies at low error rate

The quality of both the integrated and single assemblies has been assessed using the percentage of mapped pairs that map at a proper distance from each other (Table 2.1). These supporting pairs reflect the accuracy of the assembly algorithms. In both the MAIA and the Minimus assembly 99.3% of the mapped pairs can be mapped at their proper distance, showing that these assemblies are of the highest quality in the list. However, only 92.1% of the reads mapped on the Minimus assembly.

The S288c comparative assembly is 50 Kbp longer than the MAIA assembly. This is also reflected in the percentage of reads that map to the assemblies; 96.9% of the reads map to the S288c comparative assembly compared to 96.5% to the MAIA assembly (Table 2.1). The length difference can be partially attributed to the relaxed comparative assembly settings that were used for Maq; no minimum read depth and mapping quality was used. Analysis of the reads mapped by BWA on both the MAIA integrated assembly and the S288c comparative assembly showed that a far larger part of the latter is covered by only few reads (Fig. 2.7). In particular, 15 Kbp of the nucleotides were covered by five reads or less, whereas for the MAIA assembly this was the case for only 3 Kbp.

2.3.4 Other hybrid strategies are less complete

The Minimus assembly contains more than 12 million base pairs, but only 92.1% of the reads mapped to it. This indicates the iterative approach taken with Minimus results in overlapping information within the integrated assembly. The hybrid de novo assembly generated with Velvet is less complete than the MAIA assembly; only 75.5% of the reads can be mapped to it.

2.4 Conclusions

We developed MAIA, an integrator for assembly information. Our work extends previously developed algorithms for de novo and comparative assembly, enabling integration of multiple assemblies at once. MAIA makes it possible to use specific assemblers for different NGS data sources, to use multiple reference genomes for comparative assemblies or to combine outputs of different runs of an assembler. The number of known genomes is currently increasing rapidly. In the future it will be more common that multiple closely related genomes are available, as is the case already for *S. cerevisiae*. These genomes can be leveraged by using MAIA in combination with a comparative assembler.

MAIA improves genome assemblies by making use of all available information. The algorithm integrates single assemblies from different sources into longer contigs, up to chromosomal length as shown in the *S. cerevisiae* assembly integration. Five single assemblies were integrated into 29 contigs covering 12.01 Mbp. In the MAIA integrated assembly, 99.3% of the mapped read pairs mapped at a correct distance from each other. This percentage is higher than for each of the single assemblies, indicating that the integrated assembly is of higher quality than the single assemblies.
The edge weighting system in MAIA can be extended. Integration is achieved by building an overlap graph from pairwise aligned contigs and subsequently finding the highest scoring path. Edges in this graph are weighted by properties of the involved contigs and alignments. Currently, four properties for the edge weighting are implemented. These can be extended by calculating $p$-values for additional properties such as alignment Overspanning mate pair data, distances of contigs on the related genomes, or physical or genetic map information.

Acknowledgements: We would like to thank David Adams from the Sanger institute for his help with obtaining the 454 reads.
2.4 Conclusions

Figure 2.5: Usage of the different assemblies in the input per chromosome.

Figure 2.6: MAIA results for the incremental addition of input assemblies.
Figure 2.7: Histograms of coverage of reads mapped on the MAIA integrated assembly and the S288c comparative assembly.

Figure 2.8: The overlap graph for chromosome 9. The highest scoring path found by the Tabu search is indicated by the red arrows.
De novo sequencing, assembly and analysis of the genome of the laboratory strain *Saccharomyces cerevisiae* CEN.PK113-7D, a model for modern industrial biotechnology

Jurgen Nijkamp, Marcel van den Broek, Erwin Datema, Stefan de Kok, Lizanne Bosman, Marijke A. Luttik, Pascale Daran-Lapujade, Wanwipa Vongsangnak, Jens Nielsen, Wilbert Heijne, Paul Klaassen, Chris Paddon, Darren Platt, Peter Kötter, Roeland van Ham, Marcel Reinders, Jack Pronk, Dick de Ridder and Jean-Marc Daran

Published in *Microbial Cell Factories*
26 March 2012, Volume 11, issue 1
Abstract

Saccharomyces cerevisiae CEN.PK 113-7D is widely used for metabolic engineering and systems biology research in industry and academia. We sequenced, assembled, annotated and analyzed its genome. Single-nucleotide variations (SNV), insertions/deletions (indels) and differences in genome organization compared to the reference strain *S. cerevisiae* S288C were analyzed. In addition to a few large deletions and duplications, nearly 3000 indels were identified in the CEN.PK113-7D genome relative to S288C. These differences were overrepresented in genes whose functions are related to transcriptional regulation and chromatin remodelling. Some of these variations were caused by unstable tandem repeats, suggesting an innate evolvability of the corresponding genes. Besides a previously characterised mutation in adenylate cyclase, the CEN.PK113-7D genome sequence revealed a significant enrichment of non-synonymous mutations in genes encoding for components of the cAMP signalling pathway. Some phenotypic characteristics of the CEN.PK113-7D strains were explained by the presence of additional specific metabolic genes relative to S288C. In particular, the presence of the *BIO1* and *BIO6* genes correlated with a biotin prototrophy of CEN.PK113-7D. Furthermore, the copy number, chromosomal location and sequences of the MAL loci were resolved. The assembled sequence reveals that CEN.PK113-7D has a mosaic genome that combines characteristics of laboratory strains and wild-industrial strains.
3.1 Background

The 1000-dollar genome, an iconic goal in human genomics, is already a reality for the yeast Saccharomyces cerevisiae (based on September 2011 quotes from several sequencing companies for sequencing a 12 Mb genome via paired-end short-read sequencing, at over 40-fold coverage).

Although a high quality reference genome of the laboratory strain S. cerevisiae S288C has been available since 1996 [81], there are four main reasons to (re)sequence the genomes of other S. cerevisiae strains. First, the considerable sequence divergence among S. cerevisiae species may cause practical complications, for example, the design of oligonucleotide arrays and cassettes for gene disruption in non-S288C strains. The discovery of >250,000 polymorphisms in 71 S. cerevisiae strains sequenced at low coverage [142] illustrates that this is not a trivial problem. Secondly, although the genomes of S. cerevisiae strains appears to be much more strongly conserved than those of other organisms, such as E. coli [144], S. cerevisiae strains do show physiologically relevant differences in their gene complement. For example, the absence of a functional MALx3 gene in S. cerevisiae S288C leads to a maltose-negative phenotype, while an atypical ENA gene complement renders the laboratory strain CEN.PK113-7D more sensitive to lithium ions [45]. The possible importance of strain-specific genes is illustrated by the identification of a probable horizontal gene transfer event in the S. cerevisiae wine strain EC1118, that led to the acquisition of genes from the spoilage yeast Zygosaccharomyces bailii [164]. Third, in addition to the presence or absence of coding regions, differences can occur in non-coding regions, such as promoter regions. Knowledge of such differences is essential for the analysis and modeling of regulatory networks in systems biology [68]. Finally, laboratory evolution is rapidly gaining popularity as a tool to analyse genome function and to select for yeast strains with industrially relevant properties [41, 95, 168, 243, 248]. Genome comparisons based on mapping short-read data to a distant relative may overlook structural changes. Hence availability of a well-annotated, high-quality reference genome is essential to interpret the changes that occur during laboratory evolution. Several wild and domestic yeast strains have been sequenced. At the moment, forty-seven genome projects for S. cerevisiae have been registered at GenBank from which twenty-eight contain a de novo assembled (draft) genome [4, 7, 22, 23, 57, 59, 67, 81, 102, 164, 240].

The isogenic family of CEN.PK strains was developed by crossing of different laboratory strains of S. cerevisiae in the 1990s by a consortium of German yeast researchers [64]. A subsequent multi-laboratory study in which four S. cerevisiae strains were compared, confirmed that the CEN.PK strains combine good accessibility to classical and molecular genetics techniques with excellent growth characteristics under controlled, industrially relevant conditions [226]. These strains, and in particular the haploid MATa strain CEN.PK113-7D, have since become extremely popular for studies in systems biology [29, 123]. Moreover, the excellent growth characteristics of the CEN.PK strains have resulted in their broad application in metabolic and evolutionary engineering studies, for example for the fermentation of pentose sugars [15, 118, 128, 249], production of ethanol.
3. THE S. CEREVISIAE CEN.PK113-7D GENOME

and spirits (202) production of lactate and pyruvate (229, 230), production of C4-dicarboxylic acids (256), isoprenoids (149, 167), and fungal polyketide (6-methylsalicylic acid) (239).

Genomic differences between S. cerevisiae CEN.PK113-7D and the S288C strain have been the subject of several studies. Daran-Lapujade et al. (43) performed a comparative genotyping of the two strains by hybridization of genomic DNA to oligonucleotide gene-expression arrays. This work led to the identification of several genes that were absent in CEN.PK113-7D, but present in S288C. Schacherer et al. (200) employed an oligonucleotide tiling microarray (Affymetrix S. cerevisiae Tiling 1.0R array) based on the S288C genome to detect locations of single nucleotide variation (SNV) in order to narrow down the amount of sequencing needed using traditional sequencing approaches and to find genes absent in CEN.PK113-7D such as RDS1 and EHD3. SNVs in CEN.PK113-7D compared to S288C have previously been characterized by mapping next-generation DNA sequencing data to the S288C reference genome followed by SNV calling (167). The use of short read (35-bp) sequences and a limited coverage, prohibited detection of insertions and deletions (indels), unique CEN.PK113-7D sequences and structural variations.

The goal of the present study was to make a high-quality assembled and annotated reference genome of S. cerevisiae CEN.PK113-7D sequence available to the academic and industrial yeast research communities. Additionally, we aim to compare the CEN.PK113-7D sequence to that of strain S288C and other previously sequenced S. cerevisiae strains. To this end, we performed high-coverage sequencing, de novo genome assembly, scaffolding and annotation of S. cerevisiae CEN.PK113-7D strain. We explored differences with the S288C genome, including single nucleotide variations, small insertions and deletions (indels) and larger structural variation, copy number variation (CNV) and strain-specific sequences and ORFs.

3.2 Results and Discussion

3.2.1 Genome assembly, scaffolding and annotation

The genome assembly of the CEN.PK113-7D strain sequence was performed by combining Illumina (36M reads, 51 bp, paired-end) and 454 (0.6M reads, mean length 280 bp) sequencing datasets (see Methods and Supplementary methods) that together represented more than 150-fold coverage of the genome. A hybrid assembly strategy followed by scaffolding using paired-end read information resulted in 565 scaffolds with a total size of 11.6 Mbp (GenBank BioProject PRJNA52955; [http://cenpk.bt.tudelft.nl]) (Table 3.1), which were subsequently placed into chromosomal scaffolds based on homology to S288C. Genes in the CEN.PK113-7D genome were predicted using a combination of ab initio and alignment based gene predictors. Combination of predictions by Jigsaw (6) resulted in a total of 5472 ORFs that were predicted with high confidence, comparable to the 5538 genes annotated in S288C (112). The difference could be attributed to imperfect gene predictions, to missing sequence in the CEN.PK113-7D genome mostly due to repetitive sequences and to genomic content missing from the sequence data due to sequencing bias (e.g. due
to nucleotide composition of specific regions). Analysis of the 0.5 Mbp present in S288C but absent in the CEN.PK113-7D draft genome showed that 26% was in genomic regions with low read mapping probably caused by extreme GC content. Most missing sequence was due to sequence repeats. Out of the 0.5 Mbp absent in the CEN.PK assembly more than 90% was in repetitive regions. Moreover, about 24 kbp in the CEN.PK assembly was found absent in the S288C genome. This additional sequence will be discussed below.

Table 3.1: Assembly statistics. The 454 and Illumina reads were assembled with four approaches (column 1-4). The resulting contig sets were combined into a final assembly with the MAIA algorithm [75]. Ordering and orienting the contigs using homology to the S288C chromosomes resulted in chromosomal scaffolds. The N50 is the length-weighted median of the contig lengths. Coverage on S288C indicates the number of base pairs from the S288C genome covered in the assembly.

<table>
<thead>
<tr>
<th>454 Newbler contigs</th>
<th>Hybrid Mira contigs</th>
<th>Illumina velvet PE scaffolds</th>
<th>Hybrid velvet PE scaffolds</th>
<th>MAIA integrated PE scaffolds</th>
<th>Chromosomal scaffolds</th>
<th>Contigs not in chromosomal scaffolds</th>
</tr>
</thead>
<tbody>
<tr>
<td>454</td>
<td>Hybrid</td>
<td>Illumina</td>
<td>Hybrid</td>
<td>MAIA</td>
<td>Chromosomal</td>
<td>Contigs not in chromosomal scaffolds</td>
</tr>
<tr>
<td>727</td>
<td>2237</td>
<td>647</td>
<td>617</td>
<td>565</td>
<td>16-1</td>
<td>55</td>
</tr>
<tr>
<td>Largest contig (kbp)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>149</td>
<td>284</td>
<td>263</td>
<td>271</td>
<td>353</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N50 (kbp)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>62</td>
<td>71</td>
<td>73</td>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total size (Mbp)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.6</td>
<td>13.0</td>
<td>11.6</td>
<td>11.6</td>
<td>11.6</td>
<td>11.5</td>
<td>0.06</td>
</tr>
<tr>
<td>Total size contigs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.5</td>
<td>12.8</td>
<td>11.5</td>
<td>11.5</td>
<td>11.5</td>
<td>11.5</td>
<td>0.06</td>
</tr>
<tr>
<td>Tiling on S288C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.01</td>
<td>9.7</td>
<td>10.8</td>
<td>10.6</td>
<td>10.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage on S288C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.5</td>
<td>11.8</td>
<td>11.6</td>
<td>11.6</td>
<td>11.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.2.2 Reassembly and analysis of replicated sequences

A major challenge of de novo genome assembly with the current read length of next generation sequencing methods is still the identification and reconstruction of repetitive regions. Regions with sequence similarity within the genome are therefore often left unresolved and cause assembly fragmentation. The approximately 38 Long Terminal Repeat (LTR) transposons in the CEN.PK113-7D genome with a length of about 6 kbp each, as well as the whole genome duplication that occurred in the evolutionary history of *S. cerevisiae* [251] contribute to the presence of repetitive sequences throughout the genome.

*S. cerevisiae* has five retrotransposon families (Ty1y5) (Table S1). Each Ty element is flanked by LTR sequences [113]. A total of 50 LTR flanked retrotransposons have been characterized in the S288C strain [81], whereas only 17 retrotransposons were identified in the strain YJM789, a strain isolated from the lung of an AIDS patient with pneumonia [240]. Of the 50 retrotransposons found in S288C [81], evidence of the presence of 40 was obtained in CEN.PK113-7D genome assembly. We found that 39 retrotransposon sequences were indeed located on contig breaks, which suggests an intact transposon or a remnant of one (Figure S1). A 40th transposon, *YCLWTy5-1*, was fully de novo assembled. *YCLWTy5-1* is the only Ty5-type retrotransposon in *S. cerevisiae* (SGD). Analysis of the depth-of-coverage ratio between S288C and CEN.PK113-7D gave a rough estimate of the transposon content and suggested that CEN.PK113-7D had 38 Ty retrotransposons, two fewer than the 40 possible retrotransposons derived from the assembly. This difference
might be caused by two contig break locations not containing full-length transposons, possibly due to Ty excision by internal recombination, leaving remnant LTRs.

In addition to transposons, paralogs resulting from duplication events form a class of sequences with high similarities that are difficult to assemble de novo. To identify duplicated genes, read depths of the CEN.PK113-7D and S288C genomes were compared. Mapping sequencing reads of both S288C and CEN.PK113-7D to the S288C genome allowed the calculation of a depth of coverage ratio at every location on the genome (Figure 3.1A), which corrects for systematic biases in mapping depth of individual read sets, such as those caused by differences in GC content. This analysis enabled the identification of six replicated regions in the strain CEN.PK113-7D relative to S288C (Table 3.2). Five groups of genes comprised these regions including i) BIO2, IMA1, ii) RDL1, RDL2, iii) MAL33, MAL31, MAL32, YBR298C-A, YBR300C, iv) LEU2, NFS1 and v) PHO12, IMD2. To confirm the sequencing results, electrophoretic karyotypes and Southern blots with BIO2, IMA1, MAL32, RDL1 and PHO12 probes were carried out (Figures 3.1C and S2). Each probing experiment confirmed the duplication of the tested genes. The hybridization pattern with the RDL1 probe confirmed the presence of an extra copy on chromosome VI (CHRVI) in addition to the expected copy on CHRXV (Figure S2). The PHO12 hybridization pattern in S288C revealed three chromosomes (I, II and VIII) of which only CHRVIII corresponds to PHO12 (Figure S2). Hybridization with chromosomes I and II resulted from cross-hybridization with PHO11 (CHR1), PHO13 and PHO5 (both on CHRII). In addition to the three bands in S288C, CEN.PK113-7D exhibited a fourth hybridization on CHRXI (Figure S2), which confirms the extra copy found in CEN.PK113-7D sequencing data.

Table 3.2: Copy number variation between the S288C and CEN.PK113-7D genomes was estimated by mapping CEN.PK and S288C reads to the S288C genome using BWA (136).

<table>
<thead>
<tr>
<th>CNV S288C coordinates</th>
<th>Log2 (S288C/S288C)</th>
<th>CEN.PK CHR Amplified Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>VII 1064291—1068431</td>
<td>1.01 × 10⁻¹⁰</td>
<td>VII, III, XI BIO2, IMA1</td>
</tr>
<tr>
<td>XV 849425—850253</td>
<td>1.63 × 10⁻²¹</td>
<td>XV, VI RDL1, RDL2</td>
</tr>
<tr>
<td>II 805955—806853</td>
<td>3.92 × 10⁻⁷⁶</td>
<td>H, III MAL32, YBR300C</td>
</tr>
<tr>
<td>II 801609—805541</td>
<td>6.79 × 10⁻¹¹⁰</td>
<td>H, III MAL33, MAL31, MAL32, YBR298C-A</td>
</tr>
<tr>
<td>III 90771—93253</td>
<td>2.15 × 10⁻⁸⁷</td>
<td>? LEU2, NFS1</td>
</tr>
<tr>
<td>VIII 553415—556313</td>
<td>2.81 × 10⁻¹³³</td>
<td>VIII, XI PHO12, IMD2</td>
</tr>
</tbody>
</table>

Read depth ratios and p-values were calculated using CNV-seq (253). Regions with a significantly higher read depth in the CEN.PK113-7D genome are reported in this table.

In contrast to S288C, strains from the CEN.PK lineage are able to grow on maltose as sole carbon source. Both the MAL3 locus and the BIO2/IMA1 locus have double sequence coverage in CEN.PK113-7D (Figure 3.1A). In S288C the contiguous gene pair BIO2/IMA1 is located upstream of the MAL1 locus on chromosome VII (Figure 3.1C). As already observed for PHO12 the presence of paralogs complicated the hybridization pattern. IMA1, encoding for an isomaltase, has four paralogs in S288C. Therefore, in
3.2 Results and Discussion

Figure 3.1: Overview of the reconstruction of the MAL2 locus in *S. cerevisiae* CEN.PK113-7D. - A) Depth of coverage analysis of MAL loci. Sequencing reads from the CEN.PK113-7D and S288C have been mapped onto the S288C genome sequence. Log2-ratio’s of the average sequencing depth in 414-bp windows of CEN.PK113-7D over S288C have been plotted versus genomic position. Log2-ratios were capped at -4. B) Schematic representation of the paired-end linkage analysis that showed anomalous read pairs (red arrows) on the end of chromosome III that mapped to BIO2 on CHRVII. C) Southern blots and karyotypes, denoted with ‘b’ and ‘k’, respectively, of S288C and CEN.PK113-7D, denoted with ‘S’ and ‘C’, respectively. Probes for IMA1, BIO2 and MAL32 were amplified with the primers listed in Table S8. Square boxes indicate hybridized chromosomes. Chromosomes marked with an asterisk (*) depicts cross-hybridization of the IMA1 probe to paralogs of IMA1 (see text). D) Schematic organization of the MAL loci in CEN.PK113-7D. Loci marked with (#) were individually amplified, sequenced and assembled.
addition to the hybridization band on CHRVII corresponding to *IMA1*, we observed three more hybridized chromosomes, CHRIX (*IMA3*), CHRX (*IMA4, IMA5*) and CHRXV (*IMA2*). In CEN.PK113-7D, the *BIO2* and *IMA1* hybridization patterns consistently revealed not one but two additional copies of the paired genes on CHRXI and CHRIII. In S288C, *BIO2* and *IMA1* are physically linked to the locus *MAL1* (CHRVII). However, CEN.PK paired-end read mapping to the S288C genome established that the *BIO2/IMA1* duplication was located on the right arm subtelomeric region of chromosome III (Figure 3.1B). Hybridization chromosomal profiles with the *MAL32* probes confirmed the co-localization of *BIO2/IMA1* and *MAL* loci on three chromosomes CHRVII, and CHRIII but also on CHRXI as extra copies of *BIO2* and *IMA1*. It was therefore hypothesized that the chromosome III *BIO2/IMA1* genes preceded the *MAL2* locus as well, given the gene order also observed on the *MAL1* locus. This hypothesis was validated by long-range PCR amplification of *MAL1, MAL2*, and *MAL3* in CEN.PK113-7D. The PCR products were sequenced and assembled individually, which was not possible using whole-genome shotgun data because of the high sequence homology. The sequence assembly results confirmed that the additional copy located on chromosome III linked the additional *MAL2* locus to *BIO2/IMA1*. We were not able to reconstruct the *MAL4* locus located on CHRXI that seems to share high similarity with CHRIII *MAL2*.

The maltase *MAL22* and maltose permease *MAL21* genes are very similar to their *MAL3* paralogs, but the regulator gene *MAL23* is very different from both the *MAL1* and the *MAL3* locus. This *MAL23* mutant allele also known as *MAL2-8C* (117) is responsible for the partial de-repression of the *MAL* genes in CEN.PK strains in the absence of maltose (118). Interestingly, the two additional *MAL* (2 and 4) loci found in CEN.PK113-7D presented features of the two common loci in S288C (*MAL1* and *MAL3*). They probably originated from a two-step event with an initial recombination between *MAL1* and 3 at *MALx3* genes, placing *BIO2* and *IMA1* in front of *MAL2* or 4, followed by a duplication of newly recombined *MAL* locus (Figure 3.1D).

The last region that was called duplicated contained two genes, *LEU2* and *NFS1*. However, this duplication could not be confirmed by Southern blotting of karyotype, nor by direct resequencing of the *LEU2* locus.

*Single nucleotide variation* Single Nucleotide Variation (SNV) and insertions/deletion (indel) analysis was performed by alignment of the CEN.PK113-7D assembly and the reference sequence of S288C. This analysis identified a total of 21,889 SNVs (Table S2), of which 13,235 were located in 1,843 open reading frames. 4,677 SNVs (35%) resulted in amino acid changes and affected a total number of 1,406 proteins.

An earlier study by Otero et al (2010), restricting their analysis to genes involved in metabolism, revealed that galactose uptake and ergosterol biosynthesis pathways were enriched for non-synonymous SNVs. In addition to confirming these results, our sequencing data enabled the identification of new variations (Table S4).

An emblematic mutation in the CEN.PK strain family is located in the *CYR1* gene, which encodes for adenylate cyclase, a key enzyme for cAMP production and cAMP-dependent protein kinase signalling. The CEN.PK113-7D *CYR1* gene carries a non-
synonymous mutation that results in an amino acid substitution (Lys1876Met) (232). The metabolic repercussions of this mutation are very limited. It was accompanied by the absence of a cAMP peak and a reduced trehalase activation after sudden exposure to high glucose concentration and a delayed mobilization of storage carbohydrates. While we confirmed the occurrence of this SNV in CYR1, several other mutations were found in genes encoding for components of cAMP signalling pathway including PLC1, GPA2, GPB2, IRA2 (Figure 3.2 Table S5). The IRA2 gene exhibits a very early frameshift that disrupts the coding sequence and most importantly the C-terminus of the protein that is essential for its activity (218). The repercussion of these mutations might be suppressed by the presence of the lcr1 (cyr1met1876) mutation, especially for genes encoding for components of the cAMP signalling pathway located upstream to CYR1 (i.e. GPA2 and IRA2) (232).

3.2.3 Small insertions and deletions

In total 2,859 small indels (<90 bp) were called with an average length of 3 bp. Indels often reside within low-complexity regions, mostly in tandem repeats (77). 420 indels were found inside a gene, together affecting 297 genes. In 132 genes with indels the start and stop codons usage was not affected (Table S3).

The DAVID functional annotation tool (98) was applied on the genes containing indels to find enrichment of Gene Ontology (GO) terms using the GO Fat subset (8) (Figure 3.3). The biological process term ‘regulation of transcription’ (44 genes of 297 genes with indels)
tandem repeats have been shown to predominantly occur in promoter regions and in genes was significantly overrepresented ($p = 9.3 \times 10^{-3}$, Bonferroni corrected). The percentage of gene products involved in 'regulation of transcription' even increased to nearly a quarter of all genes with indels (44 of 220, 20%) when genes classified as uncharacterized or dubious in the Saccharomyces Genome Database (SGD) were not considered. Eventually, in 72% (32 of the 44) of the cases, the indel did not cause a frameshift in the transcription factor genes (i.e. start and stop codon usage was conserved). For example, the CEN.PK113-7D SWI1 allele carries an in-frame 69 bp deletion, conversely CEN.PK113-7D SNF11 contains a 12 bp extension relative to the S288C SNF11 allele. Swi1 and Snf1 are both subunits of the Swi/Snf ATP-dependent chromatin remodelling complex and are part of the RNA polymerase II holoenzyme complex. Sequence alignment of the CEN.PK113-7D SWI1 and SNF11 genes with indels (44 of 220, 20%) when genes classified as uncharacterized or dubious in gene products involved in 'regulation of transcription' even increased to nearly a quarter of strains showed a high diversity in gene and protein length (Figure 3.4 and Figure S5).

Figure 3.3: Enrichment analysis - Enrichment analysis of non-synonymous SNVs ($n = 1406$) and indels ($n = 297$) performed on the Gene Ontology Fat subset and their Bonferroni corrected $p$-values are shown. All significantly enriched GO terms are shown ($\alpha = 0.05$).

Figure 3.4: Tandem repeats - Tandem repeat variation in SNF11 amongst sequenced S. cerevisiae strains. Multiple sequence alignment of SNF11 was performed with ClustalW. The cladogram was created using whole genome distances and the UPGMA algorithm. Edge lengths in the cladogram are meaningless. Genbank accessions are listed in Table S7.

Interestingly, the indels in SWI1 and SNF11 occur in tandem repeat regions. Such tandem repeats have been shown to predominantly occur in promoter regions and in genes
encoding for transcription factors and cell-surface proteins (77). In *S. cerevisiae* recombination of intragenic repeats of cell surface proteins has been suggested to be a reversible mechanism of evolutionary adaptation to the environment by creating cell surface diversity (234). Although 128 genes that encode for transmembrane and cell wall proteins carried at least one indel, this number was not deemed significant (*p* < 0.05, Bonferroni corrected). However, similarly to analysis of human genome sequences (77), category enrichment analysis of genes with indels identified overrepresentation of the functional categories ‘regulation of transcription’ and ‘DNA binding’ (Figure 3.3). These repetitive sequences induce contraction and expansion of the gene length by recombination and slippage of the DNA polymerase (77). In humans, repeat expansion in coding sequences has been implicated in several degenerative diseases as Huntington disease and spinocerebellar ataxias. The genetic cause of Huntington disease is a trinucleotide expansion in exon 1 of the *IT15* gene that encodes for huntingtin, a protein involved in several cellular functions including transcription (76).

The severe effects of tandem repeat length variation in humans raises the question of the effect of this phenomenon in *S. cerevisiae*. Intriguingly, the physiological implications of these polymorphisms in transcription factors have not yet been studied. Dedicated research in the future should address this to understand the role of such variation.

### 3.2.4 Absent and specific genes in CENPK113-7D relative to the reference S288C

A substantial number of *S. cerevisiae* genes are redundant as a result of genome duplication, thus allowing the loss of one copy of each duplicated gene. To systematically analyze the genes that were absent in CEN.PK113-7D sequence relative to S288C, a methodical search of S288C homologous genes with at least 95% identity in CEN.PK113-7D was performed. To prevent false negatives (i.e. genes called absent because they were not assembled), genes with a lower copy number (log2 ratio < -0.6) in CEN.PK113-7D than in S288C using the read mapping analysis were listed and compared to those S288C genes that did not present a homolog in CEN.PK113-7D. This analysis identified 83 genes absent in CEN.PK113-7D relative to S288C (Table S6). Absence of 62 of these genes can be explained by the absence of several large fragments that, in strain S288C, are located in subtelomeric regions (Table S6), such as on the left arm of CHRI (25 Kbp), on the left arm of CHRVIII (12 kb), on the left arm of CHRXII (17 kb) and on the right arm of CHRXIV (14 kb).

Only 21 deletions were found outside subtelomeric regions. Our assembly confirmed the size difference of the *PMR2* locus in the CEN.PK113-7D strain (15). In S288C the *PMR2* locus on chromosome IV harbors five copies of *ENA* genes (*ENA1-ENA5*) that encode plasma membrane sodium pumping ATPases. Sequence analysis of the *PMR2* locus in *S. cerevisiae* CEN.PK113-7D revealed the presence of a single and new *ENA6* allele that showed substantial sequence differences, both at the nucleotide level and at the predicted amino acid sequence level, with previously described *ENA* genes (Figure 3.5). The presence of this single and atypical *ENA* gene correlated with hypersensitivity
3. THE S. CEREVISIAE CEN.PK113-7D GENOME

to sodium and, in particular, to lithium ions. This CEN.PK113-7D locus was also found in several industrial S. cerevisiae strains and natural isolates. The closest homolog was found in S. cerevisiae YJM269 (strain isolated from Blauer Portugieser grapes (GenBank: AEWN00000000.1)), which exhibited a complete coverage of the locus with an identity of 99% at the nucleotide level (Figure 3.5A).

![Diagram showing genomic regions](http://www.genome.wustl.edu/genomes/view/saccharomyces_cerevisiae_pw5)

**Figure 3.5: Unique genomic regions** - Occurrence analysis of three regions present in the CEN.PK113-7D genome, but not in the S288C genome. A) Venn diagram that represents the occurrence of the three regions over the available S. cerevisiae sequenced strains in Genbank (Table S7). B and C) Annotation of the regions and RNA-seq expression profiles. RNA-seq data from glucose- and nitrogen limited anaerobic chemostat cultures (red and blue, respectively) were plotted (one bar every 10th base) for the CEN.PK113-7D specific ENA locus (B) and the two specific contigs (C). Expression data, expressed as the number of times a base is covered by a read, are ranged from are [0-750] for contig379 and contig151 and [0-250] for contig596.

The assembled CEN.PK113-7D genome sequence contained several sequences that are absent in S288C. The CEN.PK specific genes were organized in two large blocks: contig151 and contig596 (Figure 3.5C). The first block contig151, showed similarity with sequence data of S. cerevisiae strains YJM269 (79%) and PW5 (isolated from Nigerian Raphia palm wine) (http://www.genome.wustl.edu/genomes/view/saccharomyces_cerevisiae_pw5 (63%)) (Figure 3.5A). Of the five adjacent genes on contig151, the three new genes could not be functionally annotated. However, their expression could be confirmed by RNA-seq on samples taken from anaerobic carbon- and nitrogen-limited chemostats, commonly used cultivation conditions to study yeast physiology. Interestingly, uncharacterized genes 430, 431 and 432 on contig151 exhibited differential expression and were significantly up-regulated under nitrogen-limitation relative to carbon-limitation (46, 217) (Figure 3.5C). Furthermore, Southern blotting of chromosomal separation gels established that contig151 could be placed on chromosome XVI (Figure S3). These observations support the existence and functionality of these open reading frames in CEN.PK113-7D (Figure 3.5). Functional analysis of these new genes should be part of a larger, emerging challenge: the functional
analysis of the growing pangenome of the species *S. cerevisiae* that arises from the advent of next-generation sequencing.

### 3.2.5 CEN.PK113-7D specific sequences lead to biotin prototrophy

The second sequence block present in CEN.PK113-7D but not in S288C is contig596 (Figure 3.5C). This contig contained three predicted open reading frames. The first of the three shared strong similarity (95% identity) with *PHO12* that encodes an acid phosphatase. As mentioned earlier, *PHO12* had three paralogs in the CEN.PK113-7D genome, which preclude the exact localization of this *PHO12* family member found on contig596. The second and third open reading frames identified on this contig shared 72% identity with *BIO1* (*YJR154W*) that encodes for a pimeloyl-CoA synthase and 95% identity with *BIO6* that encodes for the following step in biotin biosynthesis, a 7-keto-8-amino pelargonic acid (KAPA) synthase (Figure 3.6A). Southern blotting of S288C and CEN.PK113-7D chromosomal separation gels indicated that these two open reading frames alongside the newly identified *PHO12* family member were located on CHRI (Figure 3.6B).

The CEN.PK113-7D *BIO1* and *BIO6* did exhibit similarity within the *S. cerevisiae* pangenome. The genes are conserved (over 99%) in several other *S. cerevisiae* strains, such as Sigma1278b (laboratory strain used in pseudohyphal studies, GenBank: ACVY00000000.1), YJM269 and UC5 (isolated from Sene sake in Kurashi, Japan, GenBank: AFDD00000000.1) (Figure 3.5A). Although, the systematic sequence of S288C as reported in SGD indicates that these genes were absent, Hall and Dietrich (2007) identified two pairs of pseudo genes on CHRI ([*YAR069W-A* (*BIO6*) and *YHR214W-F* (*BIO8*)]) and CHRVIII [*YAR070W-A* (*BIO1*) and *YHR214W-G* (*BIO7*)] in S288C. Southern blot chromosomal localization of *BIO1* and *BIO6* genes in S288C and CEN.PK113-7D established that these two genes were indeed exclusively present in CEN.PK113-7D and were located on CHRI. If the pseudogene pairs were indeed present in S288C as previously proposed, they displayed very low degree of similarity. In addition *BIO1* and *BIO6* transcripts were detected in glucose- and nitrogen-limited conditions (Figure 3.5C).

To verify whether these differences in genotype between S288C and CEN.PK113-7D were reflected in phenotype, we tested their growth in chemically defined medium with and without added biotin. After two consecutive shake-flask cultures in biotin-free medium to exhaust biotin reserves, while S288C strain failed to grow ($\mu < 0.0005 h^{-1}$ in 100 hours), the CEN.PK113-7D strain successfully proliferated in the absence of biotin, at a specific growth rate of 0.045 $h^{-1}$ (compared to 0.38 $h^{-1}$ in presence of biotin). Biotin is an indispensable water-soluble vitamin that is mainly used as a cofactor by biotin-dependent carboxylases. In the literature, most *S. cerevisiae* strains are considered to have lost the ability to grow in absence of biotin. To complement this auxotrophy, biotin is routinely included in chemical defined fermentation media. Biotin needs to be dissolved at high pH at which it is not stable. This step is critical in medium preparation and can cause batch-to-batch variation. Strains prototrophic for biotin might offer a solution to this, and CEN.PK113-7D is an excellent model to evaluate the use of biotin prototrophic *S.
3. THE S. CEREVISIAE CEN.PK113-7D GENOME

Figure 3.6: Biotin biosynthesis - Biotin biosynthesis characterization in CEN.PK113-7D and S288C. A) Biotin biosynthesis pathway in CEN.PK113-7D. B) Southern blotting of S288C and CEN.PK113-7D chromosomes using specific probes for BIO1 and BIO6 genes. C) Maximal specific growth rate of CEN.PK113-7D and S288C in shake flask cultivations with synthetic medium in absence (white bar), and presence (grey bar) of biotin. Growth rates were calculated from biomass measurements expressed as optical density measured at 660 nm.
3.2 Results and Discussion

cerevisiae strains for industrial application. Such biotin prototrophy would be economically interesting. A quick estimate based on the amount of biotin used in chemically defined medium designed to grow around 3 g l\(^{-1}\) biomass, an average catalog price for biotin of 100\(\text{€}.g^{-1}\) and a fermentation volume of 150 m\(^3\) would reduce the medium cost by 1000\(\text{€}\). This cost saving might be increased by 5 to 20-fold with media elaborated to grow higher biomass densities (e.g. fed-batch for protein production) (75).

3.2.6 CEN.PK113-7D shares genomic material with both laboratory and industrial strains

Our analysis so far focused on the genome comparison of CEN.PK113-7D and the main reference sequence of the laboratory strain S288C. Phylogenetic analysis that aimed at comparing CEN.PK113-7D genome to the other S. cerevisiae genomes available in GenBank (Table S7), showed the CEN.PK113-7D genome was located in between the laboratory strains (S288C and Sigma1278b) and industrial strains, such as the bioethanol (JAY291), wine (EC1118, RM11-1a, AWRI1631, AWRI1796, Vin13, VL3, T73 and Lalvin QA23) and beer (CLIB382, FostersO and FostersB) strains (Figure 3.7). Phylogenetic analyses performed previously showed similar clustering wine and laboratory strains. CEN.PK113-7D is not a typical laboratory strain, but has both features typically found in industrial S. cerevisiae strains and features that make it a suitable laboratory strain. Regions more closely related to either industrial strains or laboratory strains were distributed in a mosaic structure across the CEN.PK113-7D genome (Figure 3.8, Table S7). The industrial background of CEN.PK113-7D may further be elucidated by sequencing its parental strains (ENY-WA-1A and MC996A), which have different phenotypes but are of unclear origin.

![Figure 3.7: Phylogeny](image-url) - Phylogenetic tree generated using UPGMA on genomic distances in the SplitsTree4 package using the available S. cerevisiae genomes in Genbank (Table S7).
Figure 3.8: Mosaic CEN.PK113-7D genome - CEN.PK chromosomes colored by their identity to *S. cerevisiae* genomes, which were divided into the groups: laboratory (lab) (FL100:PRJNA60147, S288C:PRJNA128, Sigma1278b:PRJNA39317), industrial (e.g. wine, beer, bio-ethanol) (AWRI1631:PRJNA30553, AWRI796:PRJNA48559, CBS7960:PRJNA60391, CLIB382:PRJNA60145, EC1118:PRJEA37863, Foster:B:PRJNA48569, FosterO:PRJNA48567, JAY291:PRJNA32809, Kyokai no. 7:PRJNA45827, Lalvin QA23:PRJNA48561, M22:PRJNA28815, PW5:PRJNA60181, RM11-1a:PRJNA13674, T73:PRJNA60195, UC5:PRJNA60197, Vin13:PRJNA48563, VL3:PRJNA48565, YJM269:PRJNA60389) and other (CLIB215:PRJNA60143, CLIB324:PRJNA60415, EC9-8:PRJNA73985, T7:PRJNA60387, Y10:PRJNA60201, YJM789:PRJNA13304, YPS163:PRJNA28813) (Table S7). Each group was assigned one of the color channels of the RGB figure (red: lab, green: other and blue: industrial). The genome was divided in non-overlapping fragments of 1000 base pairs, represented by one pixel in the figure, which were aligned to the available *S. cerevisiae* genomes in GenBank (Table S7). The identity of the best alignment in a group was set to be the value of the corresponding color channel. These values were scaled between 0 and 1; 0 meaning a maximal identity of 97% or lower, 1 meaning a maximal identity of 100%. For example, a white pixel color means 100% conservation of the fragment in all three groups. Blue and cyan mean conservation in industrial strains, but not in laboratory strains.
3.3 Conclusion

We have sequenced, assembled and annotated the genome of *S. cerevisiae* CEN.PK113-7D. Complementary to previous re-sequencing efforts (167), our de novo assembly provides additional insight in still unexplored differences between *S. cerevisiae* strains, such as large indels. Comparisons of the de novo assembled genome of CEN.PK113-7D to other *S. cerevisiae* genomes revealed many unstable tandem repeats in genes involved in transcriptional regulation. Repeat expansion and contraction in genes coding for regulatory proteins may be a mechanism of evolving transcription regulation, reconfiguring binding affinity and specificity of transcription factors. From a metabolic engineering perspective, varying tandem repeat lengths may be a promising approach to tune transcription regulation networks. De novo assembly of the genome allowed us to uncover genes in CEN.PK113-7D that are absent in S288C and relate to phenotypic characteristics specific to CEN.PK113-7D such as maltose metabolism and biotin biosynthesis. These findings led to the surprising discovery that CEN.PK113-7D is biotin prototroph, a phenotypic trait potentially interesting for industrial application. Phylogenetic analysis based on whole-genome sequence established that CEN.PK113-7D is not solely related to laboratory strains like S288C, but also to industrial strains. These genetic specificities are the basis of the robust physiological performance of this strain, which is used as a yeast model for industrial applications. Together, these results contribute to the elucidation of the genetic diversity within the *S. cerevisiae* pangenome. In addition, a high-quality annotated genome of this popular model strain is now available as a resource to the yeast research community (GenBank BioProject PRJNA52955; http://cenpk.tudelft.nl).

3.4 Materials & Methods

3.4.1 DNA purification and sequencing

The genomic DNA of *S. cerevisiae* CEN.PK 113-7D (64) was purified as previously described (50). High-throughput sequencing DNA libraries for Illumina (Illumina, San Diego, CA) and 454 (454 Life Sciences, Branford, CT) sequencing were prepared according the manufacturers recommendations. The Illumina paired-end library was sequenced on Genome Analyzer IIx (Illumina) with a read length of 50 bp at ServiceXS (Leiden, The Netherlands). The 454 library was sequenced on a GS FLX + system (454 Life Sciences) with an average read length of 350 bp at GATC-Biotech (Konstanz, Germany).

3.4.2 Assembly, scaffolding, annotation and RNAseq

The two sequencing methods used in this study yielded data with different properties in terms of error models and read length. Different assemblers, i.e. de Bruijn graph and overlap-layout-consensus assemblers, yield optimal results given these data types. We therefore chose a hybrid assembly strategy using the Velvet (258) assembler for the Illumina data and the Newbler assembler for the 454 data. The output of both assemblers was combined using the Maia algorithm (161). Paired-end scaffolding was then
performed using Velvets scaffoldor (258). The resulting contigs were placed into chromosomal scaffolds based on homology with the S288C genome (Saccharomyces Genome Database, http://www.yeastgenome.org/, March 3rd, 2011) using MUMmer (53). The chromosomal scaffolds were generated for visualization purposes. Variation analysis was performed on contigs rather than on scaffolds, as the bear no bias towards S288C. ORF finding was performed using several tools combined in the Cyrille2 pipeline (70). Ab initio gene prediction was performed using Augustus (213), SNAP (120) and GeneMark-S (16). Comparative gene prediction was performed by mapping S288C ORFs, S288C proteins and UniProt fungal proteins with GenomeThreader (127).

The output of the above predictors was combined by Jigsaw (6) to predict the final ORFs for CEN.PK. These ORFs were aligned using BLAST (259) to the Saccharomyces Genome Database (SGD) to assign the name of the closest S288C homolog.

3.4.2.1 RNAseq data

cDNAs from anaerobic carbon-limited, and anaerobic nitrogen-limited chemostat samples (217) were prepared as previously described (18). The Illumina cDNA libraries were prepared according the manufacturers recommendations. The libraries were sequenced on Genome Analyzer IIx with read length of 50 bp at Baseclear (Leiden, The Netherlands). The resulting reads were mapped onto the CEN.PK113-7D genome sequence using the Burrows-Wheeler Alignment tool BWA (136). Visualization of the mapped reads was performed using the Integrated genomics viewer (189).

3.4.2.2 Gap and missing sequence analysis

We investigated the causes of fragmentation and potentially missing sequence between two CEN.PK113-7D contigs. Every CEN.PK113-7D contig was assigned a location in the S288C genome using alignment by MUMmer. Regions in S288C not covered by one of the CEN.PK113-7D contigs, i.e. the missing sequences or gaps, were analyzed. There are several possible causes for missing sequence: (1) the sequence is unique to S288C; (2) the regions were not well amplified in the PCR reaction in the sequencing instrument, because of too high or too low GC content; or (3) the reads are present but not assembled, possibly because of repeats. If the reads are present and the region is not repetitive, a contig break is unlikely. We therefore investigated the latter two causes of gaps: too low read coverage or repeats. The Illumina reads were mapped to the S288C genome using BWA (136) and analysed with Samtools (138). First we counted the number of regions in the S288C genome between consecutively aligned CEN.PK113-7D contigs that contained at least one base not covered by any read. Such gaps we call missing sequence. Second, we determined if a repetitive sequence could be located in the gap. To this we aligned the S288C genome to itself using MUMmer. Regions were called repetitive if they had >95% identity with sequence elsewhere in the genome. If a gap overlapped with one of these repetitive regions in S288C, it was labeled unassembled. The total size of missing sequence gaps and unassembled gaps was then tallied, providing insight in the causes of the fragmentation of the assembly.
3.4.3 Variation analysis

3.4.3.1 SNV indels and novel DNA

CEN.PK113-7D contigs were aligned to S288C with MUMmer with an identity threshold of at least 97%. SNVs and indels were called from alignments with a non-ambiguous mapping. Small indels were detected as gapped alignments, maximized to a length of 90 bp by MUMmers default settings.

We define a region in CEN.PK113-7D to be unique if it shows less than 95% identity to any region in S288C. To list these unique regions, first all non-unique regions were found by whole genome alignment of CEN.PK113-7D and S288C, keeping only alignments with more than 95% identity. For each position in the S288C genome only the best aligned CEN.PK113-7D region was considered; if other regions aligned, this was considered to be a duplication in CEN.PK113-7D, hence unique. All resulting unaligned regions were then called unique to CEN.PK113-7D and analyzed for genes. Conservation of these unique regions in other yeast strains was investigated by aligning them to the NCBI Whole-Genome-Shotgun Sequences database using blastn. Species which contained a similar sequence were selected by sorting the hits on score and subsequently selecting the top hits with query coverage >60% and identity >90%.

Enrichment of Gene Ontology (GO) terms in the GO Fat subset in the sets of SNVs and indels was performed in the DAVID functional annotation tool. P-values were calculated using the EASE score, a modified version of the Fisher exact test. Bonferroni correction was applied by DAVID to account for multiple testing.

3.4.3.2 Deleted and duplicated genes

Copy number variation of regions in CEN.PK113-7D compared to S288C was investigated by mapping both S288C and CEN.PK113-7D reads to the S288C reference genome. The log2-ratio of the CEN.PK113-7D depth of coverage over that of S288C was calculated using CNV-seq version 0.2-6 and normalized to an average of 1 per chromosome. For this analysis, 36 bp reads for both strains were obtained from. The default CNV-seq threshold of 0.6 was used as a cut-off for the log2-ratio.

S288C genes that could not be mapped to the CEN.PK113-7D assembly by GenomeThreader with an identity score of at least 95% were labeled not present. An additional validation was performed to prevent false positive deletion calls of genes that should be present in the genome but are not assembled; using copy number variation between CEN.PK113-7D and S288C. Genes with a significantly lower copy number value in CEN.PK113-7D were considered not to be present in the data. The intersection of the set of genes not assembled and the set of genes not present in the data resulted in a high confidence list of deleted genes.

3.4.3.3 Transposon analysis

Ty retrotransposons mostly cause contig breaks in the assembly process, because their sequences are repetitive in the genome. The presence of the S288C transposons in the
3. THE *S. CEREVISIAE* CEN.PK113-7D GENOME

CEN.PK113-7D genome can be investigated by using these contig breaks. When inspecting the alignment of CEN.PK113-7D and S288C on locations these Ty elements in the S288C genome, three situations can be observed, the transposon is not present, therefore a gapped alignment of one contig spans the transposon location (GA) with a gap of size similar to the Ty retrotransposon size (∼6-kbp); the transposon is present in CEN.PK113-7D, but unassembled, therefore two contigs align to both sides of the transposon location (CB); or the transposon is present in CEN.PK and assembled, therefore a gapless alignment is observed (AS). The first situation yields a not present call for the transposon, the second and third situation yield a present call. The presence of transposons unique to CEN.PK113-7D was not investigated, since this requires long insert mate-pair libraries. An estimate of the number of Ty elements in the CEN.PK113-7D genome for each member of the Ty family (Ty1-Ty5) was obtained by calculating the log2-ratio of the depth of coverage for each Ty element in S288C.

3.4.3.4 Southern blotting

The chromosomes of the CEN.PK113-7D strain were prepared and separated by clamped homogeneous electrical field (CHEF) electrophoresis as previously described in (43). Transfer of the DNA from the gel to a nylon membrane (Amersham HybondTM-N+, GE Healthcare Europe GmbH, Diegem, Belgium), was performed as previously described in (194). Genomic DNA probe fragments were amplified from genomic DNA of *S. cerevisiae* CEN.PK113-7D using Phusion Hot-Start Polymerase (Finnzymes, Landsmeer, The Netherlands) and the oligonucleotides listed in Table S8. The probes were labelled according to the manufacturers instructions (Amersham Gene Images AlkPhos Direct Labelling and Detection System, Buckinghamshire, UK). Hybridization was done overnight at 55C in hybridization buffer (50% formamide, 5x Saline-Sodium Citrate (SSC) buffer, 2% blocking reagent (Roche), 0.1% Na-lauroylsarcosyl, 0.02% Sodium Dodecyl Sulfate (SDS)). Membranes were washed twice with primary wash buffer (2M urea, 0.1% SDS, 50mM sodium phosphate, 50M sodium chloride, 0.2% blocking reagent) for 15 minutes at 55C and twice with secondary wash buffer (50mM Tris base, 0.1M sodium chloride, 2mM magnesium chloride) for 5 minutes at room temperature. Digoxigenin-labelled probes were detected by chemiluminescence using CPD-star (Roche, Paris, France).

3.4.3.5 PCR amplification of the *MAL* loci

The genomic *MAL1*, 2 and 3 loci were amplified from genomic DNA of *S. cerevisiae* CEN.PK113-7D using the Qiagen LongRange PCR kit (Qiagen, Hilden, Germany) according to manufacturers instructions in a Biometra TGradient Thermocycler (Biometra, Gottingen, Germany). The following primer combinations were used for amplification: ZUO1 Fw/MALx2 Rv (*MAL1* locus), YCR102W-A Fw/MALx2 Rv (*MAL2* locus) and PHO89 Fw/MALx2 Rv (*MAL3* locus) (Table S9). Sequencing DNA libraries of each of MAL PCR locus for Illumina were prepared and the PCR products were sequenced on Genome Analyzer IIx with read length of 50 bp at Baseclear (Leiden, The Netherlands).
The subsequent sequences were assembled using Newbler (454 Life Sciences, Brandford, CT).

3.4.3.6 Phylogenetic tree construction

The *S. cerevisiae* genomes were downloaded from GenBank (accession date: 21 Nov. 2011) (Table S7). Similarities between all pairs of genomes were determined using MUMmer with the settings as recommended in (10). Pairwise distances were calculated using the coverage distance function (91). The phylogenetic tree was created by performing hierarchical clustering (UPGMA) with the SplitsTree4 package (100).

3.4.3.7 Strains and cultivation conditions

The prototrophic *S. cerevisiae* strains CEN.PK113-7D (64) and S288C (ATCC 204508) (156) were grown in liquid cultures. The shake-flask cultivations with biotin were performed in 500 ml flasks containing 100 ml of medium, which were incubated at 30 °C on an orbital shaker set at 200-rpm. The composition of the synthetic medium (SM) was as follows: glucose (20 g.l\(^{-1}\)), (NH\(_4\))\(_2\)SO\(_4\) (5 g.l\(^{-1}\)) KH\(_2\)PO\(_4\) (3 g.l\(^{-1}\)), MgSO\(_4\) (0.5 g.l\(^{-1}\)), trace elements and vitamin solutions (233). A separate biotin-free vitamin solution was used for growth in absence of biotin. The pH of the medium was adjusted to 5.0 and sterilized by autoclaving. Glucose was autoclaved separately. Vitamins were filter-sterilized and added to the medium. Growth of the various strains was monitored by OD measurements at 660 nm.

3.5 Supplemental data

Supplemental methods and data are available online at http://www.microbialcellfactories.com/content/11/1/36
3. **THE *S. CEREVISIAE* CEN.PK113-7D GENOME**
Laboratory evolution of new lactate transporter genes in a jen1Δ mutant of *Saccharomyces cerevisiae* and their identification as *ADY2* alleles by whole-genome resequencing and transcriptome analysis

Stefan de Kok, Jurgen Nijkamp, Bart Oud, Filipa Roque, Dick de Ridder, Jean-Marc Daran, Jack Pronk, Ton van Maris

Published in *FEMS Yeast Research*
23 January 2012, Volume 12, issue 3, pages 359-374
Abstract

Laboratory evolution is a powerful approach in applied and fundamental yeast research, but complete elucidation of the molecular basis of evolved phenotypes remains a challenge. In this study, DNA microarray-based transcriptome analysis and whole-genome resequencing were used to investigate evolution of novel lactate transporters in \textit{Saccharomyces cerevisiae} that can replace Jen1p, the only documented \textit{S.cerevisiae} lactate transporter. To this end, a \textit{jen1}∆ mutant was evolved for growth on lactate in serial batch cultures. Two independent evolution experiments yielded growth on lactate as sole carbon source (0.14 and 0.18 h$^{-1}$, respectively). Transcriptome analysis did not provide leads, but whole-genome resequencing showed different single-nucleotide changes (C755G/Leu219Val and C655G/Ala252Gly) in the acetate transporter gene \textit{ADY2}. Introduction of these \textit{ADY2} alleles in a \textit{jen1}∆ \textit{ady2}∆ strain enabled growth on lactate (0.14 h$^{-1}$ for Ady2p\textit{Leu219Val} and 0.12 h$^{-1}$ for Ady2p\textit{Ala252Gly}), demonstrating that these alleles of \textit{ADY2} encode efficient lactate transporters. Depth of coverage of DNA sequencing, combined with karyotyping, gene deletions and diagnostic PCR, showed that an isochromosome III (c. 475 kb) with two additional copies of \textit{ADY2}C755G had been formed via crossover between retrotransposons YCLW∆15 and YCRC∆6. The isochromosome formation shows how even short periods of selective pressure can cause substantial karyotype changes.
4.1 Introduction

Laboratory evolution (referred to as ‘evolutionary engineering’ in applied contexts) is a powerful strategy to obtain microorganisms with improved characteristics that are difficult to engineer through rational approaches, especially when the genetic determinants for a phenotype are unknown or complex (197). Subsequent discovery of the underlying mutation(s) is of key importance, both for functional gene analysis and to enable transfer of the acquired traits to non-evolved industrially relevant strain backgrounds (‘reverse metabolic engineering’; (11)). In addition to the mutation(s) of interest, random mutations can accumulate during evolution. Combined with our incomplete understanding of metabolic and regulatory networks, this makes elucidation of the genetic basis of acquired phenotypes the main challenge of reverse metabolic engineering. During evolution, different types of mutations can occur: (1) small local changes (substitutions, insertions, deletions and duplications) that result in either differential gene expression or changes in the amino acid sequence of the encoded protein and/or (2) rearrangements of larger DNA fragments, potentially resulting in a different gene dosage or expression (197).

Several genome-wide techniques have been described in literature to analyse mutations that occur during laboratory evolution (for reviews, see (27, 79, 83). Classical techniques include construction and analysis of plasmid-based genomic libraries and quantitative trait locus mapping via crossing. More recent developments have enabled whole-genome DNA sequencing, DNA hybridization to DNA microarrays, transcriptome analysis using DNA microarrays, proteome, metabolome and fluxome profiling (27, 79, 84). Analysis with DNA microarrays was, until recently, the only affordable genome-wide method. Microarray-based transcriptome analysis only indirectly indicates changes on the DNA level. Several factors, such as complex regulatory networks, transcription factor binding, translation efficiency and stability of mRNA and proteins, complicate the interpretation of the exact genetic basis of changes observed at mRNA and protein level. In contrast, comparative genome hybridization using DNA microarrays can identify genetic changes directly (43, 200, 247). However, limitations in the maximum probe density imply that either (1) the resolution of these DNA microarrays is limited to several base pairs, necessitating sequencing of selected regions or (2) multiple DNA microarrays are required to cover larger (eukaryotic) genomes. Furthermore, DNA microarrays can only be used for a comprehensive analysis when the strains under investigation are congenic to the strain on whose genome sequence the microarray design was based. In contrast, whole-genome DNA sequencing, which because of recent price developments has become an interesting technique for analysis of evolved strains (165, 199), has the potential to directly reveal all mutations accumulated in an evolved strain. However, discriminating between random and relevant mutations remains a challenging task. Combining DNA sequencing with other genome-wide techniques that provide insight in the underlying physiology can help to elucidate the genetic basis of a phenotype acquired via evolution.

Transport of carboxylic acids plays a key role in weak organic acid stress (173) and in metabolic engineering strategies for organic acid production with S.cerevisiae (108, 230, 231). However, the responsible membrane transporters are often poorly studied and
4. LABORATORY EVOLUTION OF LACTATE TRANSPORT

encoded by multiple redundant genes (30). The goal of this study was to investigate whether laboratory evolution and subsequent elucidation of the underlying mutations can lead to the identification of alternative lactate transporters in *S. cerevisiae*. In *S. cerevisiae*, Jen1p was previously identified as the only efficient lactate importer by generation of a UV mutant unable to grow on lactate and subsequent functional complementation with a genomic library (30). The lactate uptake rate of the resulting *jen1* null mutant was close to the detection limit (30). Interestingly, export of lactate in engineered lactate producing *S. cerevisiae* is unaffected by deletion of *JEN1* (our unpublished results). Even though mechanisms for import or export can differ, these combined observations suggest the presence of at least one alternative lactate transporter. To explore the presence of other lactate transporters in *S. cerevisiae*, a *jen1Δ* strain was constructed and evolved for growth on lactate as the sole carbon and energy source. The evolved strains were subjected to a combination of transcriptome analysis and whole-genome DNA (re)sequencing to identify the relevant mutations. The resulting lead genes were tested for their lactate transport activity via knockout studies in the evolved strains and introduction into nonevolved strains.

4.2 Materials and methods

4.2.1 Strains and maintenance

The *S. cerevisiae* strains used and constructed in this study (Table 4.1) are all derived from CEN.PK113-7D (64, 226). Stock cultures were grown at 30 °C in shake flasks containing either 100 mL synthetic medium (233) with 20 g L⁻¹ glucose as carbon source or 100 mL complex medium containing 20 g L⁻¹ glucose, 10 g L⁻¹ Bacto yeast extract and 20 g L⁻¹ Bacto peptone. After overnight growth, 20% (v/v) glycerol was added and 1-mL aliquots were stored at -80 °C.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Relevant genotype</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEN.PK113-7D</td>
<td>MATa URA3 ADY2 JEN1</td>
<td>P. Köttler, Frankfurt</td>
</tr>
<tr>
<td>IMK302</td>
<td>MATa URA3 ADY2 jen1::loxP-KanMX4-loxP</td>
<td>This study</td>
</tr>
<tr>
<td>IMW004</td>
<td>IMK302 evolved for growth on lactate</td>
<td>This study</td>
</tr>
<tr>
<td>IMW005</td>
<td>IMK302 evolved for growth on lactate</td>
<td>This study</td>
</tr>
<tr>
<td>IMW032</td>
<td>IMW004 ady2::loxP-hphNT1-loxP</td>
<td>This study</td>
</tr>
<tr>
<td>IMW040</td>
<td>IMW004 ady2::loxP-hphNT1-loxP ady2::loxP-ura3::loxP-loxP</td>
<td>This study</td>
</tr>
<tr>
<td>IMW041</td>
<td>IMW004 ady2::loxP-hphNT1-loxP ady2::loxP-ura3::loxP-loxP</td>
<td>This study</td>
</tr>
<tr>
<td>IMW033</td>
<td>IMW005 ady2::loxP-hphNT1-loxP</td>
<td>This study</td>
</tr>
<tr>
<td>IMK322</td>
<td>MATa ura3::loxP-hphNT1-loxP ADY2 jen1::loxP-KanMX4-loxP</td>
<td>This study</td>
</tr>
<tr>
<td>IMK338</td>
<td>MATa ura3::loxP ADY2 jen1::loxP</td>
<td>This study</td>
</tr>
<tr>
<td>IMK341</td>
<td>MATa ura3::loxP ady2::loxP-hphNT1-loxP jen1::loxP</td>
<td>This study</td>
</tr>
<tr>
<td>IMZ272</td>
<td>IMK341 pUDC011 (CEN6/ARS4, URA3, ADY2-ADY2-TADY2)</td>
<td>This study</td>
</tr>
<tr>
<td>IMZ273</td>
<td>IMK341 pUDC012 (CEN6/ARS4, URA3, ADY2-ADY2-TADY2)</td>
<td>This study</td>
</tr>
<tr>
<td>IMZ276</td>
<td>IMK341 pUDC015 (CEN6/ARS4, URA3, ADY2-ADY2-TADY2)</td>
<td>This study</td>
</tr>
</tbody>
</table>
4.3 Plasmid and strain construction

The plasmids and primers used in this study are listed in Tables 4.2 and 4.3, respectively. Transformations of *S. cerevisiae* were carried out according to the LiAc/ssDNA method (78). Gene deletions were performed using the loxP-marker-loxP/Cre recombinase system, using pUG6, pUG66, pUG-natNT2 and pUG-hphNT1 as templates for PCR amplification of the knockout cassettes (88). Selection of knockout mutants was performed on agar plates containing 20 g L$^{-1}$ glucose, 10 g L$^{-1}$ Bacto yeast extract, 20 g L$^{-1}$ peptone and either 200 mg L$^{-1}$ G418, 10 mg L$^{-1}$ phleomycin, 100 mg L$^{-1}$ nourseothricin or 200 mg L$^{-1}$ hygromycin B. The natNT2 marker was cloned from pFA6a-natNT2 (104) to pUG6 via the SacI and BglII restriction sites, thereby replacing the KanMX4 marker, resulting in pUG-natNT2. The JEN1 knockout cassette was amplified from pUG6 using primers JEN1 KO Fw and JEN1 KO Rv and transformed to CEN.PK113-7D, resulting in strain IMK302. Correct integration of the knockout cassette was confirmed using primer combinations JEN1 Ctrl Fw/KanMX4 Ctrl Rv and KanMX4 Ctrl FW/JEN1 Ctrl Rv. The URA3 knockout cassette was amplified from pUG-hphNT using primers URA3 KO Fw and URA3 KO Rv and transformed to IMK302, resulting in IMK322. Correct integration of the knockout cassette was confirmed using primer combinations URA3 Ctrl Fw/hphNT1 Ctrl Rv and hphNT1 Ctrl FW/URA3 Ctrl Rv. The KanMX4 and hphNT1 markers in IMK322 were removed using the Cre/loxP system with pSH47 (88), resulting in IMK338. Correct marker removal was confirmed using primer combinations JEN1 Ctrl Fw/URA3 Ctrl Rv and URA3 Ctrl Fw/URA3 Ctrl Rv.

Table 4.2: Plasmids used in this study

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Characteristic</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>pCR-BLUNT II-TOPO</td>
<td>Gateway entry plasmid</td>
<td>Invitrogen</td>
</tr>
<tr>
<td>pRS416</td>
<td>Centromeric plasmid, URA3</td>
<td>Sikorski and Hieter (207)</td>
</tr>
<tr>
<td>pUG66</td>
<td>PCR template for loxP-KanMX4-loxP cassette</td>
<td>Gueldener et al. (88)</td>
</tr>
<tr>
<td>pUG66</td>
<td>PCR template for loxP-ble-loxP cassette</td>
<td>Gueldener et al. (88)</td>
</tr>
<tr>
<td>pSH47</td>
<td>Centromeric plasmid, URA3, GAL1-cre-T CY C1</td>
<td>Gueldener et al. (88)</td>
</tr>
<tr>
<td>pUG-hphNT1</td>
<td>PCR template for loxP-hphNT1-loxP cassette</td>
<td>de Kok et al. (49)</td>
</tr>
<tr>
<td>pFA6a-natNT2</td>
<td>Plasmid with natNT2 marker</td>
<td>Jankes et al. (104)</td>
</tr>
<tr>
<td>pUG-natNT2</td>
<td>PCR template for loxP-natNT2-loxP cassette</td>
<td>This study</td>
</tr>
<tr>
<td>pUD151</td>
<td>Gateway entry clone, ADY2-ADY2C655G-TADY2</td>
<td>This study</td>
</tr>
<tr>
<td>pUD152</td>
<td>Gateway entry clone, ADY2-ADY2C755G-TADY2</td>
<td>This study</td>
</tr>
<tr>
<td>pUD152</td>
<td>Gateway entry clone, ADY2-ADY2C755G-TADY2</td>
<td>This study</td>
</tr>
<tr>
<td>pUD152</td>
<td>Gateway entry clone, ADY2-ADY2C755G-TADY2</td>
<td>This study</td>
</tr>
<tr>
<td>pUD152</td>
<td>Gateway entry clone, ADY2-ADY2C755G-TADY2</td>
<td>This study</td>
</tr>
<tr>
<td>pUD155</td>
<td>Gateway entry clone, URA3, ADY2-ADY2C655G-TADY2</td>
<td>This study</td>
</tr>
<tr>
<td>pUD155</td>
<td>Gateway entry clone, URA3, ADY2-ADY2C755G-TADY2</td>
<td>This study</td>
</tr>
<tr>
<td>pUD155</td>
<td>Gateway entry clone, URA3, ADY2-ADY2C755G-TADY2</td>
<td>This study</td>
</tr>
</tbody>
</table>

The first ADY2 knockout cassette was amplified from pUG-hphNT1 using primers ADY2 KO Fw A and ADY2 KO Rv B and transformed to IMK338, IMW004 and IMW005, resulting in IMK341, IMW032 and IMW033, respectively. Correct integration of the knockout cassette was confirmed using primer combinations ADY2 Ctrl Fw/hphNT1 Ctrl Rv and hphNT1 Ctrl FW/ADY2 Ctrl Rv. The second ADY2 knockout cassette was amplified from pUG-natNT2 using primers ADY2 KO Fw B and ADY2 KO Rv B and transformed to IMW032, resulting in IMW040. Correct integration of the knockout cassette was confirmed using primer combinations ADY2 Ctrl Fw/natNT2 Ctrl Rv and natNT2 Ctrl FW/ADY2 Ctrl Rv. The third ADY2 knockout cassette was amplified from pUG66.
### Table 4.3: Primers used in this study

<table>
<thead>
<tr>
<th>Name Sequence (5′ → 3′)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>JEN1 KO Fw</strong></td>
</tr>
<tr>
<td><strong>JEN1 KO Rv</strong></td>
</tr>
<tr>
<td><strong>JEN1 Ctrl Fw</strong></td>
</tr>
<tr>
<td><strong>JEN1 Ctrl Rv</strong></td>
</tr>
<tr>
<td><strong>ADY2 KO Fw</strong></td>
</tr>
<tr>
<td><strong>ADY2 KO Rv</strong></td>
</tr>
<tr>
<td><strong>ADY2 Ctrl Fw</strong></td>
</tr>
<tr>
<td><strong>ADY2 Ctrl Rv</strong></td>
</tr>
<tr>
<td><strong>URA3 KO Fw</strong></td>
</tr>
<tr>
<td><strong>URA3 KO Rv</strong></td>
</tr>
<tr>
<td><strong>URA3 Ctrl Fw</strong></td>
</tr>
<tr>
<td><strong>URA3 Ctrl Rv</strong></td>
</tr>
<tr>
<td><strong>KanMX4 Ctrl Fw</strong></td>
</tr>
<tr>
<td><strong>KanMX4 Ctrl Rv</strong></td>
</tr>
<tr>
<td><strong>hphNT1 Ctrl Fw</strong></td>
</tr>
<tr>
<td><strong>hphNT1 Ctrl Rv</strong></td>
</tr>
<tr>
<td><strong>natNT2 Ctrl Fw</strong></td>
</tr>
<tr>
<td><strong>natNT2 Ctrl Rv</strong></td>
</tr>
<tr>
<td><strong>ble Ctrl Fw</strong></td>
</tr>
<tr>
<td><strong>ble Ctrl Rv</strong></td>
</tr>
<tr>
<td><strong>ADY2 inside Fw</strong></td>
</tr>
<tr>
<td><strong>ADY2 inside Rv</strong></td>
</tr>
<tr>
<td><strong>ADY2p Fw</strong></td>
</tr>
<tr>
<td><strong>ADY2p Rv</strong></td>
</tr>
<tr>
<td><strong>P5</strong></td>
</tr>
<tr>
<td><strong>P6</strong></td>
</tr>
</tbody>
</table>
using primers ADY2 KO Fw C and ADY2 KO Rv C and transformed to IMW040, resulting in IMW041. Correct integration of the knockout cassette was confirmed using primer combinations ADY2 Ctrl Fw/ble Ctrl Rv and ble Ctrl FW/ADY2 Ctrl Rv. Presence or absence of the ADY2 gene was confirmed using primers ADY2 inside Fw and ADY2 inside Rv. The ADY2 gene including promoter and terminator was amplified from genomic DNA of CEN.PK113-7D, IMW004 and IMW005 using primers ADY2p Fw and ADY2t Rv, and the resulting PCR products were cloned into pCR-BLUNT II-TOPO using Gateway Technology (Invitrogen, Carlsbad, NM), resulting in pUD151, pUD153 and pUD152, respectively. The ADY2 alleles including promoter and terminator were subsequently cloned into pRS416 via the BamHI and XbaI sites, resulting in pUDC011, pUDC015 and pUDC012, respectively. pRS416, pUDC011, pUDC012 and pUDC015 were transformed into IMK341 and selected on a synthetic medium without uracil, resulting in IMZ271, IMZ272, IMZ273 and IMZ276.

4.3.1 Molecular biology techniques

PCR amplification was performed using Phusion Hot Start High Fidelity Polymerase (Finnzymes, Espoo, Finland) according to manufacturer’s instructions in a Biometra TGradient Thermocycler (Biometra, Gottingen, Germany). DNA fragments were separated on a 1% (w/v) agarose (Sigma, St. Louis, MO) gel in 1xTAE (40 mM Trisacetate pH 8.0 and 1 mM EDTA). Isolation of fragments from gel was performed with the Zymoclean Gel DNA Recovery kit (Zymo Research, Orange, CA). Restriction endonucleases (New England Biolabs, Beverly, MA) and DNA ligases (Roche, Basel, Switzerland) were used according to manufacturer’s instructions. Transformation and amplification of plasmids was performed in Escherichia coli One Shot TOP10 competent cells (Invitrogen) according to manufacturer’s instructions. Plasmids were isolated from E. coli with the Sigma GenElute Plasmid Miniprep Kit (Sigma). DNA constructs were routinely sequenced by Baseclear BV (Baseclear, Leiden, The Netherlands).

4.3.2 Laboratory evolution of strains IMW004 and IMW005

Glycerol stocks of IMK302 were used to inoculate two 500-mL shake flasks containing 100 mL of synthetic medium (Verduyn et al., 1992) with 5 g L$^{-1}$ l-lactic acid (Fluka 09578) and 25 mM MES buffer. The pH of the medium was set to 5.0 with 2 M KOH and 2 M H2SO4 prior to autoclaving (121 °C, 20 min). When growth was observed, a small aliquot (0.12 mL) was transferred to a new flask. After 10 transfers, the culture was plated on agar plates containing 20 g L$^{-1}$ glucose, 10 g L$^{-1}$ Bacto yeast extract and 20 g L$^{-1}$ Bacto peptone to obtain single colonies. Per evolution line, four colonies were tested for growth on lactate. One colony per evolution line was stocked and named as IMW004 and IMW005.
4. LABORATORY EVOLUTION OF LACTATE TRANSPORT

4.3.3 Cultivation and media

Shake-flask precultures for characterization experiments were inoculated with glycerol stocks and grown in synthetic medium \((233)\) with \(20 \text{ g L}^{-1}\) ethanol as carbon source.

Characterization of growth on lactate in shake flasks with an initial pH of 5.0 was performed in synthetic medium \((233)\) with \(5 \text{ g L}^{-1}\) lactate and 25 mM MES buffer, as described earlier. Cells growing exponentially on ethanol were used to inoculate 500-mL shake flasks containing 100 mL medium and incubated in an Innova incubator shaker (New Brunswick Scientific, Edison, NJ) at 200 rpm and 30 °C.

Aerobic batch cultures for transcriptome analysis were carried out at 30 °C in 2-L laboratory fermentors (Applikon, Schiedam, The Netherlands) with a working volume of 1 L. Synthetic medium (Verduyn et al., 1992) was supplemented with \(20 \text{ g L}^{-1}\) lactic acid (Fluka 09578) as sole carbon source. The pH of this medium was set to 5.0 before autoclaving (120 °C). Antifoam Emulsion C (Sigma) was autoclaved separately (120 °C) as a 20% (w/v) solution and added to a final concentration of 0.2 g L\(^{-1}\). The culture pH was maintained at 5.0 by automatic addition of 2 M KOH and 2 M H\(_2\)SO\(_4\). Cultures were stirred at 800 rpm and sparged with 500 mL air per min. The precultures were grown on ethanol as described earlier. Six hours before the inoculation of the batch cultures, 25 mM lactate was added to the preculture to induce lactate metabolism. Samples for RNA extraction were taken during the exponential growth phase at a cell dry weight concentration of 3-4 g L\(^{-1}\) and a lactate concentration of 57 g L\(^{-1}\). Sequential batch cultures for the characterization of growth rate and biomass yield were run in the same setup as described earlier, but with \(10 \text{ g L}^{-1}\) lactate to prevent possible nutrient limitations above 5 g L\(^{-1}\) cell dry weight. The first batch was inoculated with a preculture growing exponentially on ethanol, as described earlier. To increase reproducibility of the results, a sequential batch reactor (SBR) was used and the third cycle was sampled \((2)\).

4.3.4 Growth assays on agar plates

Growth on ethanol, lactate and pyruvate was tested by spotting 10 \(\mu\)L of serial dilutions of a culture growing exponentially on ethanol, on synthetic medium agarose (1% w/v) plates supplemented with 25 mM MES and either \(5 \text{ g L}^{-1}\) ethanol, lactate or pyruvate. Before spotting, the cultures were first washed 2 times and diluted to the appropriate cell concentration with synthetic medium \((233)\). Pictures were taken after 6 days incubation at 30 °C.

4.3.5 Analytical methods

Optical density at 660 nm was measured with a Libra S11 spectrophotometer (Biochrom, Cambridge, UK). Culture dry weights were determined via filtration of appropriate sample volumes (10-20 mL) over dry preweighed nitrocellulose filters (Gelman Laboratory, Ann Arbor, MI) with a pore size of 0.45 \(\mu\)m. After removal of the medium, the filters were washed twice with demineralized water, dried in a microwave oven for 20 min at 350 W and weighed. Culture supernatants were obtained after the centrifugation of the broth.
Supernatants and media were analysed via HPLC using an Aminex HPX-87H ion exchange column operated at 60 °C with 5 mM H2SO4 as mobile phase at a flow rate of 0.6 mL min⁻¹. Off-gas was first cooled in a condenser (2 °C) and dried with a Perma Pure Dryer (Permapure, Toms River, NJ). CO2 concentrations in the off-gas were measured with a NGA 2000 Rosemount gas analyzer (Rosemount Analytical, Orrville, OH).

4.3.6 Microarrays and analysis

Sampling of cells from the batch cultures and total RNA extraction was performed as described previously (1). Probe preparation and hybridization to Affymetrix Genechip microarrays were performed according to Affymetrix’s instructions. The one-cycle eukaryotic target labelling assay was used, starting with 15 µg of total RNA. The quality of total RNA, cRNA and fragmented cRNA was checked using the Agilent BioAnalyzer 2100 (Agilent Technologies, Amstelveen, the Netherlands). Results for each strain were obtained from two independent culture replicates. The Significance Analysis of Microarrays (SAM version 1.12; (225)) add-in to Microsoft Excel was used for comparison of replicate array experiments using a fold-change threshold of two and an expected false discovery rate of 5%. Transcript data have been deposited in the Genome Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE31634.

4.3.7 Whole-genome sequencing and analysis

DNA of the reference strain CEN.PK113-7D was prepared as described previously (25). A library of 200-bp fragments was created and sequenced paired-end using the Illumina Solexa system, generating c. 56 million 36-bp paired reads. These reads were mapped to the genome of S.cerevisiae strain S288C (SGD project) using the BurrowsWheeler Alignment tool (BWA) (136). Subsequently, a consensus sequence was generated using SAMtools’ pileup2fq (138) with default parameters, except for the minimum depth and root-mean-square mapping quality, which were set to 10 and 25, respectively. Consensus bases not passing the filtering thresholds, mostly in repetitive regions, were replaced by ‘N’. This consensus was used as CEN.PK113-7D reference genome.

Genomic DNA from IMW004 and IMW005 was isolated using the Qiagen 100/G kit (Qiagen, Hilden, Germany). A library of 200-bp genomic fragments was created and sequenced paired-end (50-bp reads) using an Illumina HiSeq 2000 sequencer by Baseclear BV (Baseclear). The individual reads were mapped onto the genome of CEN.PK113-7D, using BWA and further processed using SAMtools (138). Single-nucleotide variations and small insertions and deletions were extracted from the mapping using SAMtools’ varFilter. Default settings were used, except that the maximum read depth was set to 300× (-D300) and the minimum small insertions and deletions score for single-nucleotide variations filtering was set to 60 (-G60). To minimize false positive mutation calls, custom Perl scripts and Microsoft Excel were used for further mutation filtering. First, mutation calls containing ambiguous bases in either reference or mapping consensus were filtered out. Second, only the single-nucleotide variations with a quality of at least 20 and small insertions and deletions with a quality of at least 50 were kept. Variant quality is defined
4. LABORATORY EVOLUTION OF LACTATE TRANSPORT

as the Phred-scaled probability that the mutation call is incorrect (55). Third, mutations with a depth of coverage < 10× were discarded. Fourth, insertion and deletion mutation calls were only kept when at least 85% of the reads spanning the location confirmed the insertion or deletion. Fifth, small insertions and deletions that were close to an 'N' in the reference were removed, because this would complicate correct alignments and introduce false positive mutation calls. For small insertions and deletions, a window size of 40 bp was used, and for single-nucleotide variations, a window size of 5 bp. Copy number variation was analysed using CNV-seq (253).

4.3.8 Pulsed-field gel electrophoresis and southern blotting

Chromosomes were separated using the CHEF yeast genomic DNA Plug Kit (170-3593; Bio-Rad, Richmond, CA) according to manufacturer’s instructions and subsequently transferred onto Hybond-N+ nylon membranes (RPN303; Amersham Biosciences, Piscataway, NJ). Southern blotting, signal generation and signal detection were performed using the Gene Images AlkPhos Kit, CPD Star detection reagent and Hyperfilm ECL (RPN 3680, RPN3682 and 28-9068; Amersham Biosciences). The DNA probe used for southern blotting was amplified from CEN.PK113-7D genomic DNA with primers ADY2 inside Fw and ADY2 inside Rv (Table 4.3).

4.4 Results

4.4.1 Laboratory evolution of Jen1p-independent growth on lactate

To test whether laboratory evolution and subsequent analysis of the underlying mutations is a powerful strategy to identify mutants that can transport lactate in *S. cerevisiae*, a *jen1Δ* knockout strain was constructed and evolved for growth on lactate. Hitherto, Jen1p is the only efficient importer of lactate in *S. cerevisiae* described in literature (30). Consistent with earlier reports, *S. cerevisiae* IMK302 (*jen1*) did not show growth on plates with lactate after 6 days, whereas the reference strain CEN.PK113-7D (*JEN1*) grew normally (Figure 4.1). In shake-flask cultures with 5 g L\(^{-1}\) lactate at pH 5, the same results were obtained [\(\mu < 0.001\) h\(^{-1}\) for IMK302 (*jen1Δ*), but after prolonged incubation (515 days), growth was observed in two independent experiments. To select for faster growth, small aliquots of the shake-flask cultures were transferred to fresh medium for nine consecutive times. After c. 100 generations in 10 shake-flask cultures, single-colony isolates were obtained by plating on nonselective agar plates containing glucose, yeast extract and peptone. Per evolution line, four single-colony isolates were tested for growth on lactate and yielded essentially the same growth rates as the mixed population from which they were isolated. One single-colony isolate from each evolution experiment was chosen for further characterization. These resulting evolved strains were named IMW004 and IMW005 and grew at maximum specific growth rates of 0.14 and 0.18 h\(^{-1}\), respectively, on 20 g L\(^{-1}\) lactate in aerobic bioreactor batch cultures (pH 5.0).
**Figure 4.1**: Growth of *Saccharomyces cerevisiae* strains with different combinations of *JEN1* and *ADY2* alleles on ethanol, lactate and pyruvate. - 10 µL of a suspension with the indicated cell concentration was spotted onto synthetic medium agarose (1% w/v) plates at pH 5 with 5 g L⁻¹ ethanol (upper panel), 5 g L⁻¹ lactate (middle panel) or 5 g L⁻¹ pyruvate (lower panel) as the sole carbon source. Pictures were taken after 6 days of incubation at 30 °C.
4. LABORATORY EVOLUTION OF LACTATE TRANSPORT

4.4.2 Mutation analysis through whole-genome DNA sequencing and transcriptome analysis

Two techniques were used to study the molecular basis of Jen1p-independent growth on lactate in the evolved strains IMW004 and IMW005: transcriptome analysis with DNA microarrays and whole-genome DNA resequencing. In view of highly reproducible growth conditions and the ability to control the specific growth rate, chemostat cultures are normally the preferred cultivation type for comparative transcriptome analysis (10). However, chemostat cultures growing on lactate demonstrated persistent oscillations that precluded steady-state analysis (data not shown). Therefore, cell samples for transcriptome analysis were obtained from aerobic bioreactor batch cultures grown on 20 g L\(^{-1}\) lactate and at pH 5.0. To minimize differences in nutrient concentration and growth conditions, all cultures were sampled at a biomass dry weight concentration between 3 and 4 g L\(^{-1}\), corresponding to 5-7 g L\(^{-1}\) residual lactate. The average deviation of the mean of all genes with mRNA expression levels above background (in this case 12) was between 16% and 21%, which was only slightly higher than chemostat-based transcriptome analysis (10-15%; Basso et al., 2011). When comparing the transcriptome data of CEN.PK113-7D (\(JEN1\)) with IMW004 and IMW005 (both \(jen1\Delta\) evolved), only a very small number of genes (11 for IMW004 and 18 for IMW005) were expressed at different levels based on the statistical criteria applied in this study (absolute fold difference \(\leq 2\); false discovery rate 5%, see Materials and methods; Figure 4.2). In addition to \(JEN1\) (deleted in the evolved strain) and the introduced \(KANMX4\) dominant marker gene, the general amino acid permease \(GAP1\) was the only gene that was differentially expressed in both evolution lines IMW004 and IMW005. However, the transcript levels of \(GAP1\) were only 2.6- and 2.0-fold higher in IMW004 and IMW005, respectively, than in the reference strain CEN.PK113-7D.

In parallel, genomic DNA of IMW004 and IMW005 (both \(jen1\Delta\) evolved) was sequenced using Illumina technology, yielding c. 12.8 million 50-bp reads for IMW004 and c. 13.7 million reads for IMW005. The average depth of coverage was 51.5-fold for IMW004 and 56.6-fold for IMW005. The sequencing reads were mapped against the genome sequence of the reference strain CEN.PK113-7D, which was constructed by mapping CEN.PK113-7D sequencing reads to the S288C sequence. This resulted in 20 and 16 single-nucleotide variations for IMW004 and IMW005, respectively. No insertions or deletions were found. The set of relevant mutations was reduced to 7 for IMW004 and 10 for IMW005 (Table 4.4) by eliminating single-nucleotide variations outside coding regions, because no links were found with mRNA expression levels of adjacent genes. Nucleotide changes that only result in different codon usage for the same amino acid might affect translation efficiency and were therefore not discarded. Interestingly, several of the remaining single-nucleotide variations were either identical in both IMW004 and IMW005 or different mutations had occurred in the same genes (Table 4.4). Such mutations might (1) present genes that are relevant for the Jen1p-independent growth on lactate, (2) have accumulated during the construction of the \(jen1\Delta\) strain and have been present before the start of the evolution experiment or (3) represent errors in the reference sequence.
4.4 Results

**Figure 4.2: Differentially expressed genes** - Differentially expressed genes in IMW004 and IMW005 (both *jen1Δ* evolved) compared with the reference strain CEN.PK113-7D (*JEN1*) during exponential growth on 20 g L\(^{-1}\) lactate at pH 5.0. Expression data were compared using a fold-change threshold of two and an estimated false discovery rate of 5%.

**Gene Expression Scale:**
- **10** (Red) - Genes upregulated in both IMW004 and IMW005
- **1** (Orange) - Genes upregulated in IMW004 only
- **0.1** (Green) - Genes downregulated in IMW004 only
- **0.01** (Light Green) - Genes downregulated in both IMW004 and IMW005

**Gene List:**
- **KANMX4** - G418 resistance gene
- **GAP1** - General amino acid permease
- **JEN1** - Carboxylic acid transporter
- **AGT1** - α-glucoside transporter
- **YCR023C** - Membrane protein
- **YCR061W** - Hypothetical protein
- **ABP1** - Actin binding protein
- **YCR059C** - Hypothetical protein
- **YDL177C** - Hypothetical protein
- **MAL31** - Maltose transporter
- **YCR102C** - Hypothetical protein
- **SRD1** - Transcription regulator
- **HOR2** - Glycerol-3-phosphatase
- **YAR075W** - Hypothetical protein
- **CYC3** - Cystathionine γ-lyase
- **YNL300W** - Hypothetical protein
- **CYC3** - Cytochrome c heme lyase
- **DEP1** - Regulator of phospholipid metabolism
- **DAL2** - Allantoicase
- **HAC1** - Transcription factor
- **MBF1** - Transcriptional co-activator
- **YDR034W-B** - Hypothetical protein
- **DAP1** - Damage resistance protein
- **HXT2** - Hexose transporter
- **MUB1** - Involved in ubiquitination
- **SYM1** - Required for ethanol metabolism
- **RGI2** - Involved in respiratory energy metabolism
- **YBL049W** - Hypothetical protein
- **TAF11** - Involved in transcription (TFIID subunit)
- **IRA2** - GTPase activating protein
Table 4.4: Mutation calls inside genes in IMW004 and IMW005 identified via whole-genome sequencing

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Nucleotide change</th>
<th>Amino acid change</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genes mutated in both IMW004 and IMW005</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YAT1</td>
<td>Mitochondrial carnitine acetyltransferase</td>
<td>C75A</td>
<td>None</td>
</tr>
<tr>
<td>ADY2</td>
<td>Acetate transporter</td>
<td>C655G (IMW005)</td>
<td>Leu219Val (IMW005)</td>
</tr>
<tr>
<td>ADY2</td>
<td>Acetate transporter</td>
<td>C755G (IMW004)</td>
<td>Ala252Gly (IMW004)</td>
</tr>
<tr>
<td>BSC1</td>
<td>Hypothetical protein</td>
<td>T553A</td>
<td>Thr185Ser</td>
</tr>
<tr>
<td>FLO11</td>
<td>Glycoprotein (flocculin)</td>
<td>G1716A (IMW005)</td>
<td>None</td>
</tr>
<tr>
<td>FLO11</td>
<td>Glycoprotein (flocculin)</td>
<td>C1743T (IMW005)</td>
<td>None</td>
</tr>
<tr>
<td>FLO11</td>
<td>Glycoprotein (flocculin)</td>
<td>G1743T (IMW005)</td>
<td>None</td>
</tr>
<tr>
<td>DDC1</td>
<td>DNA damage checkpoint protein</td>
<td>G1513C</td>
<td>Ala505Pro</td>
</tr>
<tr>
<td><strong>Genes mutated only in IMW004</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YNL179C</td>
<td>Hypothetical protein</td>
<td>T156A</td>
<td>Leu52Phe</td>
</tr>
<tr>
<td>CAF120</td>
<td>Part of transcriptional regulatory complex</td>
<td>G1689A</td>
<td>None</td>
</tr>
<tr>
<td><strong>Genes mutated only in IMW005</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIT1</td>
<td>Cell wall mannoprotein</td>
<td>A1067G</td>
<td>Val356Ala</td>
</tr>
<tr>
<td>PAU2</td>
<td>Member of seripauperin multigene family</td>
<td>T318C</td>
<td>None</td>
</tr>
<tr>
<td>YHR0528W-A</td>
<td>Hypothetical protein</td>
<td>T217C</td>
<td>Ser73Phe</td>
</tr>
</tbody>
</table>

4.4.3 Rearrangement of chromosome III

Rearrangement of larger DNA fragments, including deletions and duplications, is a well-known phenomenon during (yeast) evolution. To analyse these larger structural variations, the genome-wide mRNA expression levels were plotted according to their chromosomal location. Interestingly, both the mRNA expression levels and depth of coverage of DNA sequencing showed 2log ratios between 1 and 2 for a group of genes on chromosome III, when comparing IMW004 and CEN.PK113-7D (Figure 4.3). Detailed investigation of the depth of coverage of DNA sequencing indicated a duplication of chromosome III between 83 and 126 kb, including the centromere, and a triplication between 126 kb to the end of the right arm (Figure 4.4a). These observations led to the hypothesis that, in addition to chromosome III, an additional pseudopalindromic isochromosome III was formed containing the centromere and two copies of the right arm of chromosome III. Karyotyping confirmed the formation of an additional chromosome with a size of c. 475 kb (Figure 4.5), which corresponds well to the proposed rearrangement (Figure 4.4).

To exactly determine the chromosomal rearrangement, the DNA sequences 5 kb upstream and downstream of the observed breakpoints (i.e. 7888 and 121131 kb) were aligned with each other. The highest homology was observed between the Ty1-type long terminal repeat retrotransposons YCLW\Delta15 and YCRC\Delta6, with 96% homology over a region of 176 bp. To verify whether crossover took place between YCLW\Delta15 and YCRC\Delta6 as proposed in Figure 4.4, primer combinations were designed (P5 and P6) that should give a 3041-bp PCR product if crossover had indeed taken place between YCLW\Delta15 and YCRC\Delta6. In line with the proposed mechanism of rearrangement, the expected 3-kb PCR product was obtained from IMW004, but not from CEN.PK113-7D genomic DNA. Sequencing of this DNA fragment showed that crossover took place between the last 98 homologous base pairs of YCLW\Delta15 and YCRC\Delta6. 
Figure 4.3: Comparison of depth of coverage of DNA sequencing (a, b) and mRNA expression levels (c, d) of CEN.PK113-7D (JEN1) and IMW004 (jen1Δ evolved).
- Values were plotted according to their chromosomal location. Chromosome numbers are indicated above the panels. Depth of coverage of DNA sequencing was averaged and plotted per 1500 bp. For mRNA expression levels, only genes with an expression level above 12 were considered.
Figure 4.4: Analysis of chromosomal rearrangements in IMW004.  - (a) Copy number estimates on chromosome III based on depth of coverage of DNA sequencing and mRNA expression levels. (b) Crossover between homologous regions of \textsc{YCLW}\textgreek{Δ15} and \textsc{YCRC}\textgreek{Δ6}, resulting in the formation of an additional pseudopalindromic isochromosome III, containing two extra copies of \textit{ADY}^{2}	extgreek{Δ7755G}. Indicated are the binding sites of primers P5 and P6, which were used to confirm that crossover took place as proposed. Duplicated parts of chromosome III are indicated in grey; triplicated regions in black. Relevant features are indicated, but not drawn to scale.
4.4 Results

Figure 4.5: Pulsed-field gel electrophoresis (PFGE) analysis of the reference strain CEN.PK113-7D and the evolved strains IMW004 and IMW005. Chromosome numbers and sizes are indicated on the left side. YNN295 (Bio-Rad #170-3605) was used as marker. The additional chromosome (c. 475 kb) present in IMW004 is indicated by the diagonal arrow.
4. LABORATORY EVOLUTION OF LACTATE TRANSPORT

4.4.4 Verification of the identified mutations in the acetate transporter \textit{ADY2}

The \textit{ADY2} gene, which encodes an acetate transporter (171), contained mutations that resulted in amino acid changes in both independently evolved strains (Table 4.4) and is located on the triplicated region of chromosome III of IMW004. To analyse whether the different copies of \textit{ADY2} in IMW004 all contained the same mutation, the individual sequencing reads mapping to \textit{ADY2} were analysed. At all positions, a constant signal was observed, showing that IMW004 carried three identical copies of \textit{ADY2} \textit{C755G}. \textit{ADY2} was not amongst the genes whose mRNA levels were significantly different from those in the reference strain (1.4-fold in IMW004 and 0.6-fold IMW005).

To investigate the physiological relevance of the mutation (C755G; Ala252Gly) and the apparent triplication of \textit{ADY2} in IMW004, knockout strains were constructed. Diagnostic PCR on genomic DNA of the single, double and triple knockout strains (IMW032, IMW040 and IMW041) only showed absence of \textit{ADY2} in the triple knockout strain (IMW041), thereby confirming that IMW004 contained three copies of \textit{ADY2} \textit{C755G}. Only the triple knockout strain was unable to grow on lactate (Figure 4.1), confirming that Ady2p \textit{Ala252Gly} was essential for lactate import in the evolved strain IMW004.

To analyse the effect of \textit{ADY2} \textit{C755G} gene dosage on the growth rate on lactate, aerobic shake-flask experiments were performed with 5 g L$^{-1}$ lactate at pH 5. IMW004 (3 copies of \textit{ADY2} \textit{C755G}) grew at a growth rate of 0.10 ± 0.01 h$^{-1}$; IMW032 (2 copies) at 0.09 ± 0.00 h$^{-1}$, IMW040 (1 copy) at 0.07 ± 0.00 h$^{-1}$ and IMW041 (0 copies) did not grow (\(\mu < 0.001\) h$^{-1}$).

To analyse whether the identified mutation (C655G; Leu219Val) in \textit{ADY2} could also fully explain the acquired phenotype in IMW005, a knockout of \textit{ADY2} \textit{C655G} was constructed. IMW033 (IMW005 \textit{ady}2\textit{Δ}) was not able to grow on lactate, thereby confirming the crucial role of Ady2p \textit{Leu219Val} in lactate import in IMW005 (Figure 4.1).

4.4.5 Reverse metabolic engineering of mutated \textit{ADY2} alleles

To verify whether mutation of \textit{ADY2} alone is sufficient to allow for growth on lactate, the mutated alleles were introduced into a nonevolved background (\textit{jen}1\textit{Δ ady}2\textit{Δ}). IMZ271 (\textit{jen}1\textit{Δ ady}2\textit{Δ}) and IMZ272 (\textit{jen}1\textit{Δ ady}2\textit{Δ} + \textit{ADY2}) did not show growth on agar plates with 5 g L$^{-1}$ lactate after 6 days of incubation, whereas IMZ273 (\textit{jen}1\textit{Δ ady}2\textit{Δ} + \textit{ADY2} \textit{C755G}) and IMZ276 (\textit{jen}1\textit{Δ ady}2\textit{Δ} + \textit{ADY2} \textit{C655G}) did grow on these plates (Figure 4.1). This confirmed that these alleles can confer the ability to transport lactate. IMZ271 (\textit{jen}1\textit{Δ ady}2\textit{Δ}) could still grow on acetate (data not shown), which prevented kinetic analysis of acetate transport mediated by the mutated Ady2p alleles. Interestingly, IMZ271 (\textit{jen}1\textit{Δ ady}2\textit{Δ}) could not grow on pyruvate, whereas IMZ272 (\textit{jen}1\textit{Δ ady}2\textit{Δ} + \textit{ADY2}), IMZ273 (\textit{jen}1\textit{Δ ady}2\textit{Δ} + \textit{ADY2} \textit{C755G}) and IMZ276 (\textit{jen}1\textit{Δ ady}2\textit{Δ} + \textit{ADY2} \textit{C655G}) could, showing that both the nonmutated and the mutated Ady2p alleles can function as efficient pyruvate importers in this strain background (Figure 4.1), in contrast to previous reports on other \textit{S.cerevisiae} strains (Akita et al., 2000).
Subsequently, two strains expressing the native \textit{ADY2} allele were compared. IMK302 (\textit{jen1}Δ \textit{ADY2}), which contains the native \textit{ADY2} gene on chromosome III, did not grow on lactate ($\mu < 0.001 \text{ h}^{-1}$) in shake-flask cultures with 5 g L$^{-1}$ lactate at pH 5. The fact that IMZ272 (\textit{jen1}Δ \textit{ady2}Δ + \textit{ADY2}), containing the native \textit{ADY2} under control of its own promoter on a centromeric vector, grew very slowly on lactate ($\mu_{\text{max}}$ around 0.02 h$^{-1}$) demonstrates that even nonmutated Ady2p can transport lactate, although at a very low rate. The observed differences in growth may be due to small differences in expression level between chromosomal (IMK302) and centromeric (IMZ272) expression. This might raise the lactate consumption rate in IMZ272 marginally above the flux required for maintenance metabolism and thus allow for very slow growth \cite{17, 177}.

To further characterize the mutated alleles of \textit{ADY2}, the reverse engineered strains IMZ273 (\textit{jen1}Δ \textit{ady2}Δ + \textit{ADY2}C$^{755}$G) and IMZ276 (\textit{jen1}Δ \textit{ady2}Δ + \textit{ADY2}C$^{655}$G) were tested in aerobic sequential batch fermentations with 10 g L$^{-1}$ lactate at pH 5.0. The reference strain CEN.PK113-7D (\textit{JEN1 ADY2}) was used for comparison, because in this strain, the large majority of lactate transport proceeds via Jen1p and not via Ady2p (Figure 4.1). Both IMZ273 (\textit{jen1}Δ \textit{ady2}Δ + \textit{ADY2}C$^{755}$G) and IMZ276 (\textit{jen1}Δ \textit{ady2}Δ + \textit{ADY2}C$^{655}$G) grew rapidly on lactate at maximum specific growth rates of 0.14 ± 0.00 and 0.12 ± 0.00 h$^{-1}$, respectively. In cultures of strain IMZ276, the Ady2p$^{\text{Ala252Gly}}$-dependent lactate consumption rate decreased towards the end of the fermentation, indicating a low affinity for lactate (Figure 4.6). The biomass yield on lactate of IMZ273 (0.37 ± 0.00 g g$^{-1}$; \textit{jen1}Δ \textit{ady2}Δ + \textit{ADY2}C$^{755}$G) was considerably higher than those of IMZ276 (0.30 ± 0.01 g g$^{-1}$; \textit{jen1}Δ \textit{ady2}Δ + \textit{ADY2}C$^{655}$G) and CEN.PK113-7D (0.26 ± 0.01 g g$^{-1}$; \textit{JEN1 ADY2}).

\textbf{4.5 Discussion}

\textbf{4.5.1 Ady2p as lactate transporter}

No protein structure of Ady2p or close homologues is available, and therefore, the effect of the mutations on the structure and function of Ady2p remains uncertain. However, the observation that single-nucleotide mutations turned the acetate transporter Ady2p into an efficient lactate transporter provides valuable information for future structure-function analysis. In both evolution lines, the identified mutations were located in putative transmembrane domains and resulted in amino acids that are one carbon atom smaller than the original amino acid (alanine to glycine in IMW004 and leucine to valine in IMW005). As lactate (C$_3$H$_6$O$_3$) is structurally similar, but one carbon atom larger than acetate (C$_2$H$_4$O$_2$), this suggests that extra space is created in the translocation site of Ady2p.

According to the literature, Jen1p catalyses an electroneutral lactate-proton symport, which is energetically equivalent to diffusion of the undissociated acid \cite{30, 33}. Complete aerobic dissimilation of 1 lactate then yields 7 ATP: 6 from oxidative phosphorylation at an in vivo P/O ratio of 1.0 \cite{12, 227} and 1 from substrate level phosphorylation in the tricarboxylic-acid cycle. To generate and maintain a proton-motive force, the plasma
4. LABORATORY EVOLUTION OF LACTATE TRANSPORT

Figure 4.6: Growth during aerobic sequential batch fermentation - Growth of CEN.PK 113-7D (jen1 ADY2, a), IMZ273 (jen1 ady2 + ADY2C655G, b) and IMZ276 (jen1 ady2 + ADY2C755G, c) during the third cycle of an aerobic sequential batch fermentation on lactate at pH 5.0. The results shown are from one representative experiment. Averages and mean deviations for the maximum specific growth rates ($\mu_{\text{max}}$) and biomass yields ($Y_{sx}$) were obtained from independent duplicate experiments.
membrane H\(^+\)-ATPase in yeast exports (only) a single proton for each ATP that it hydrolyses (228, 245). Lactate transport mechanisms that result in outward translocation of a proton and/or inward translocation of a negative charge can save ATP equivalents that would otherwise be used for the build-up or maintenance of the proton-motive force. These savings can therefore not exceed the in vivo activity of the plasma membrane H\(^+\)-ATPase. In this way, lactate import via a lactate anion uniport or via a lactate-proton antiport mechanism could increase the apparent ATP yield from lactate dissimilation to 8 or 9 mol mol\(^{-1}\), corresponding to increases in the biomass yield on lactate of 14\% and 29\%, respectively. The reverse engineered strain IMZ273 (Ady2p\textsubscript{Leu219Val}) displayed a 24 \pm 5\% higher biomass yield on lactate than IMZ276 (Ady2p\textsubscript{Ala252Gly}) and a 41 \pm 3\% higher biomass yield than the reference strain CEN.PK113-7D (Figure 4.6), in which lactate import primarily proceeds via Jen1p (Figure 4.1). Taking into account that an additional c. 5\% of the biomass yield increase can be explained owing to higher specific growth rates of the evolved strains, which causes a lower impact of maintenance-energy requirements (17, 177), these biomass yield values might be consistent with a different mode of energy coupling of lactate transport. Further analysis of the energy coupling of lactate transport by these Ady2p isoforms, involving in vitro studies, will provide valuable insights into the energetics of organic acid transport in yeast and its impact on growth energetics.

4.5.2 Chromosomal rearrangement

Chromosomal rearrangement is a well-known phenomenon during (yeast) evolution (55, 251) and occurs often by recombination of Ty1-type retrotransposons (reviewed by Garfinkel (74)). The occurrence of a c. 475-kb isocytochrome III in strain IMW004 shows how even a short (c. 100 generations) period of selective growth can select for a massive chromosomal rearrangement, in this case resulting in duplication of around 25 genes and triplication of about 100 genes. In addition to enhancing performance under the conditions that led to their selective advantage, such rearrangements may be ‘hopeful monsters’ that open up novel evolutionary paths via specialization of duplicated genes (220).

In-depth analysis of the sequencing reads showed that all three copies of ADY2 in IMW004 contained the same nucleotide change (C755G), which suggests that the mutation in ADY2 occurred before the triplication event. This indicates that first a gain of function mutation in ADY2 provided the cells with a selective advantage over the parental strain. The subsequent retrotransposon-mediated triplication of ADY2\textsuperscript{C755G} apparently gave sufficient evolutionary advantage to compensate for the burden of an additional chromosome. In line with this hypothesis, a strain derived from IMW004 containing only one copy of ADY2\textsuperscript{C755G} (IMW040) showed a 30\% lower growth rate on lactate. The observation that the second and third copy of ADY2\textsuperscript{C755G} did not lead to a two- and threefold higher growth rates can be explained by the saturation of transcription, possibly due to a limited availability of transcription factors that regulate ADY2 expression, such as Cat8p, Adr1p and Snf1p (31). Indeed, the ADY2 mRNA expression level of IMW004 (three ADY2\textsuperscript{C755G} copies) was only 40\% higher than that of CEN.PK113-7D (one ADY2 copy). Simulation of the evolution experiment, based on the observed cell densities and
4. LABORATORY EVOLUTION OF LACTATE TRANSPORT

the relatively small difference in growth rate between strains expressing one (0.07 h\(^{-1}\)) or three (0.10 h\(^{-1}\)) \(ADY2^{CT55G}\) copies, indicated that the formation of the isochromosome III occurred early in the evolution experiment before the first transfer. Accordingly, no further increase in the growth rate on lactate was observed in the remaining nine shake flasks of the evolution experiment.

4.5.3 Functional analysis through laboratory evolution and reverse engineering

Laboratory evolution combined with molecular analysis of the evolved strains enabled the successful identification in \(S.\)\textit{cerevisiae} of genes encoding transporter mutants that can import lactate. Although this strategy can only be used in cases where the gene function can be linked to a selective advantage, it can be very useful in situations where only a low residual enzyme activity is observed for which the genetic determinants are unknown, such as for glucose transport or disaccharide hydrolysis in mutant strains of \(S.\)\textit{cerevisiae} in which all known genes for these activities have been deleted \([49, 246]\). In addition, this approach can contribute genes and proteins with new, interesting properties for metabolic engineering. Elucidation of the genetic basis of the acquired phenotype is a crucial step in laboratory evolution-based functional analysis, which in this study was approached by a combination of transcriptome analysis and whole-genome resequencing. Subsequent verification via knockout studies and reverse metabolic engineering showed that the relevant changes could be identified via whole-genome resequencing, but not via transcriptome analysis. In yeast, only few studies have been performed to identify the relevant genetic changes and reverse engineer them into nonevolved strain background. When studies on both prokaryotic and eukaryotic laboratory evolved strains are considered, a clear picture emerges that, in many studies, transcriptome analysis did not allow rapid identification of relevant genetic changes that had been selected for in laboratory evolution experiments \([69, 71, 72, 97, 188, 229]\), even when combined with metabolome analysis \([235]\) or proteome analysis \([135]\). In contrast, studies in which the genetic basis for the acquired phenotype was elucidated and reconstructed via reverse metabolic engineering analysed the DNA level via whole-genome resequencing or hybridization of genomic DNA to microarrays \([9, 40, 84, 92, 95, 132, 223]\). These observations support the view that analysis of the genome sequence rather than the transcriptome is the most promising first-line analytical approach for elucidating relevant changes that occurred during evolution. Transcriptome analysis can then be used as a follow-up in cases where whole-genome resequencing does not generate obvious and productive leads. In this respect, mRNA sequencing is an interesting development that allows simultaneous measurement of mRNA levels and identification of mutations inside coding regions \([106, 238]\). Independent of the technique used, analysing multiple independently evolved strains strongly facilitates dissection of relevant and random genetic changes.
4.6 Acknowledgements

We thank Marinka Almering, Mark Bisschops, Lizanne Bosman, Daniel Gonzalez Ramos, Marit Hebly and Duygu Yilmaz for experimental assistance. This work was financially supported by Tate & Lyle Ingredients Americas Inc. The Kluyver Centre for Genomics of Industrial Fermentations is supported by the Netherlands Genomics Initiative.
4. LABORATORY EVOLUTION OF LACTATE TRANSPORT
De novo detection of copy number variation by co-assembly

Jurgen Nijkamp, Marcel van den Broek, Jan-Maarten Geertman, Marcel Reinders, Jean-Marc Daran and Dick de Ridder

Published in Bioinformatics
October 9, 2012, Volume 28, Issue 24, pages 3195-3202
Abstract

Motivation: Comparing genomes of individual organisms using next generation sequencing data is, until now, mostly performed using a reference genome. This is challenging when the reference is distant and introduces bias towards the exact sequence present in the reference. Recent improvements in both sequencing read length and efficiency of assembly algorithms have brought direct comparison of individual genomes by de novo assembly, rather than via a reference genome, within reach.

Results: Here, we develop and test an algorithm, named Magnolya, which employs a Poisson mixture model for copy number estimation of contigs assembled from sequencing data. We combine this with co-assembly to allow de novo detection of copy number variation between two individual genomes, without mapping reads to a reference genome. In co-assembly, multiple sequencing samples are combined, generating a single contig graph with different traversal counts for the nodes and edges between the samples. In the resulting “colored” graph the contigs have integer copy-numbers; this negates the need to segment genomic regions based on depth of coverage, as required for mapping-based detection methods. Magnolya is then used to assign integer copy numbers to contigs, after which copy number variation probabilities are easily inferred.

The copy number estimator and copy number variation detector perform well on simulated data. Application of the algorithms to hybrid yeast genomes showed allotriploid content from different origin in the wine yeast Y12, and extensive copy number variation in aneuploid brewing yeast genomes. Integer copy number variation was also accurately detected in a short-term laboratory evolved yeast strain.
5.1 Introduction

Genomes can differ in many ways. Several types of variation are commonly distinguished, from small local differences such as single nucleotide polymorphisms (SNPs) and short insertions and deletions (indels), to variation involving DNA fragments larger than 1kbp, i.e. structural variation (SV). SVs can be divided into balanced and unbalanced mutations. Balanced mutations, such as inversions and translocations, preserve the copy number of a given allele, whereas unbalanced mutations, such as indels and duplications, change the number of copies of the involved allele. Differences between genomes in the latter respect are referred to as copy number variation (CNV).

Algorithms for variant discovery have until recently started by mapping sequencing reads to a reference genome. Subsequently, variation between the sequenced genome and the reference is inferred by analyzing aberrantly mapped reads (138), read-depth variation (114, 253), split-read mappings (254) or a combination of aberrantly mapped read pairs and read-depth variation (153). Optionally, reads from the target sample in regions with many aberrantly mapped reads can then be locally assembled to infer genomic sequence not present in the reference genome.

Recently, the Cortex assembler was introduced, the first fully de novo variant detection algorithm, not reliant on a reference genome. Cortex is a de Bruijn graph assembler capable of co-assembling multiple sequencing samples. In the underlying data structure, the de Bruijn graph, the nodes and edges are colored by the samples in which they are observed. By observing bifurcations in the graph that separate the colors (bubbles), sequence variation is detected.

While bubble finding works well for detecting (simple) variation, it does not easily allow CNV detection. A duplication event introduces an (almost) identical sequence in the genome, i.e. a repeat; furthermore, larger CNV regions are likely to contain repetitive regions inherent to the genome, such as transposons and paralogous genes. The resulting repeats pose a problem for assemblers, which collapse them into single contigs (Fig. 5.1). Since it is unknown how such collapsed repeats should be traversed, i.e. which pairs of incoming and outgoing edges should be connected via the repeat, a bubble cannot be detected. Thus, in a de novo variation detection setting bubble calling is not suitable to detect CNVs.

The number of times a contig occurs in the genome can be inferred from read depths, exploiting the fact that assembly automatically segments the genome into contigs of integer copy number. Previously, the A-statistic has been proposed to determine whether a contig is unique or represents a collapsed repeat (158). Medvedev and Brudno (151) estimated the copy number of a contig using a maximum likelihood flow solving algorithm, assuming a known genome size. Here, we introduce a Poisson mixture model (PMM) approach to estimate the copy number of a contig without making assumptions on genome size.

The use of a PMM for modeling read depths in segments across multiple samples that have been mapped to a reference genome was recently introduced (114). Here, we employ a PMM in a de novo setting to obtain a genome wide model for one genome, rather than in a local segment across multiple samples. The algorithm relies on de novo assembly
DE NOVO DETECTION OF COPY NUMBER VARIATION

for segmentation. We show how this mixture model can be applied in colored assembly graphs to detect CNVs. The algorithm was named Magnolya (matched genomes de novo assembly graph analysis).

In contrast to Cortex’s De Bruijn graph approach, we employ an overlap-layout-consensus (OLC) assembler to generate a contig string graph. Contig string graphs are generated by first calculating pairwise overlaps between reads. A node in the resulting graph represents a read and an edge between two reads represents an overlap. The graph is subsequently simplified by transitive reduction, which removes redundant edges, followed by unitigging, a process of collapsing simple paths without branches. The result is a contig string graph, in which now the nodes represent collapsed reads called contigs and edges represent reads spanning two contigs. The contigs cannot be collapsed further, since for each pair of contigs connected by an edge the in- or out-degree (dependent on strand) is larger than 1.

String graph assemblers were first developed in the Sanger era, to assemble relatively long sequencing reads. The advent of next-generation sequencing (NGS), yielding short reads of length 36-100 bp, required development of assemblers not reliant on pairwise overlaps of reads, the widely used de Bruijn graph assemblers. However, increasing read lengths produced by NGS technology (including the third generation single-molecule sequencers) are renewing interest in overlap-layout-consensus assemblers, such as SGA (String Graph Assembler), capable of efficiently assembling mammalian genomes. We extended the string graph by assigning each read a color, corresponding to the sample from which it originated. In the resulting colored contig string graph we then detect variation by modelling read counts per color with a mixture of Poissons and inferring the probability of a CNV.

To our knowledge, we here for the first time apply a Poisson mixture model to detect CNVs fully de novo between samples from two individuals. This approach has two main advantages. First, no read count-based segmentation of the genome is required to distinguish regions with different copy numbers; this is handled in the co-assembly. Second, there is no bias to a, possibly distant, reference genome.

In experiments, copy number estimation using Magnolya is demonstrated both on simulated data and on data from the genome of an allotriploid wine yeast. Furthermore, CNV detection with Magnolya applied to co-assemblies is tested on simulated data and demonstrated on the aneuploid genomes of two Saccharomyces pastorianus brewing yeasts and a laboratory evolved yeast strain.

5.2 Algorithm

We propose an algorithm for CNV detection based on NGS data, not reliant on mapping to a reference genome. In this section we outline how contig copy number can be inferred using a Poisson mixture model (Sec. 5.2.1) and how these models are used in a colored co-assembly string graph to detect copy number variation (Sec. 5.2.2). Detailed derivations of the formulas can be found in the supplementary material.
5.2 Algorithm

Figure 5.1: Colored string graphs with two samples, \(O_1\) (solid) and \(O_2\) (dashed). The pairs of numbers indicate the traversal count (left for \(O_1\), right for \(O_2\)), which are estimated by a Poisson mixture model. Solid and dashed edges represent reads spanning the connected contigs. Bars under the graphs represent the unknown corresponding genome structures of the solid and the dashed sample. a) A “clean” duplication of contig R. b) A more complex duplication in which another repeat is enclosed. If the duplicated region is large, the centre node R is likely to contain repeats homologous to other sequence in the genome, complicating the graph structure.

5.2.1 Contig copy number assignment with a Poisson mixture model

5.2.1.1 Assembly segments the genome into integer copy number contigs

Assemblers would ideally assemble all sequencing reads into chromosome-sized contigs (contiguous sequences), but usually do not succeed in doing so because of repetitive sequences in the genome. Repeats are common in any genome due to transposons, rDNA repeats and paralogous genes and homozygous regions of two or more copies of a chromosome. Assemblers are unable to distinguish multiple copies of an (almost) identical sequence. As a result, reads originating from these identical sequences are merged into a single contig up until the position where the two sequences diverge. In the contig graph, a situation where a sequence occurs twice in the genome, flanked by four unique sequences, results in a node with in- and out-degree of two. This is the case for sample \(O_1\) (solid edges) in Fig. 5.1: it is unknown whether contig A and C or A and D should be connected via R; therefore, this repeat cannot be resolved without additional information. However, in the resulting contig graph, regions with different copy number are thus automatically segmented, and contigs will have integer copy numbers.

Co-assembling the two target samples is essential to obtain contigs with a single copy number for both samples. This negates the need for read-count based genome segmentation of read-mapping approaches to CNV detection. In our approach the genome is segmented by bifurcations in the contig graph, which is based on sequence information instead of read counts. For example, if a single sample assembly would have been performed on the dashed
5. DE NOVO DETECTION OF COPY NUMBER VARIATION

sample, contig R in Fig. 5.1 would not be repetitive, and contigs A-R-C and B-D would be formed. Mapping-based CNV detection between the solid and dashed sample would require segmenting contig B-R-D based on read depth variation in the solid sample, which introduces inaccuracies. By co-assembly we exploit the assemblers inability to resolve repeats to obtain precisely delineated contigs with a unique copy number difference.

5.2.1.2 Modelling read depths on contigs with a Poisson mixture model

The copy number of a contig can be inferred from the number of reads that start on a contig. We can model the observed number of reads \( x_c \) that start on a contig \( c \in C \) with a given copy number \( i \) as \( p(x_c|i) \). The data set contains contigs with different copy numbers, which together are modelled as a mixture model containing \( M \) components:

\[
p(x_c) = \sum_{i=1}^{M} p(i)p(x_c|i) \quad (5.1)
\]

The number of reads sampled from a certain nucleotide position in the genome can be modelled as a Poisson process \( \lambda = \frac{R}{G} \), where \( R \) is the total number of reads and \( G \) is the genome length. \( p(x_c|i) \) can thus be replaced by a Poisson distribution \( \text{Pois}(x_c|\theta_{i,c}) \), yielding:

\[
p(x_c) = \sum_{i=1}^{M} \pi_i \text{Pois}(x_c|\theta_{i,c}) \quad (5.2)
\]

The parameters are the mixture coefficients \( \pi_i \), estimates of \( p(i) \), and the Poisson parameters \( \theta_{i,c} = L_c \lambda \), with \( L_c \) the contig length and \( i \) the copy number (i.e. the mean number of reads in a contig with copy number \( i \) is modelled as directly proportional to \( i \)). We cannot compute \( \lambda \) directly because we do not know the genome size \( G \); therefore we estimate \( \lambda \) and the \( \pi_i \) from the data, which we do by Expectation Maximization (EM).

5.2.1.3 High copy number repeats

Specific repetitive regions in the DNA occur at very high copy number, such as ribosomal DNA repeats and transposons. We are not interested in modeling the copy number of these repeats. Therefore we view them as outliers and capture the contigs with a copy number higher than \( (M + 1)L_c\lambda \) in a shifted geometric distribution (model \( M + 1 \)). We define the set of high copy number repeat contigs \( Z \) as

\[
Z = \{c \in C : x_c \geq (M + 1)L_c\lambda\} \quad (5.3)
\]

The outlier distribution for the high copy number repeats is defined as

\[
p(x_c) = \begin{cases} 
0 & \text{if } x_c \in Z \\
\text{Geom}(x_c - \theta_{M+1}|\lambda, \alpha, M) & \text{if } x_c \notin Z
\end{cases} \quad (5.4)
\]
where $\alpha$ is the rate parameter of the geometric distribution. Our model thus becomes
\[
p(x_c | \pi, \lambda, \alpha) = \sum_{i=1}^{M} \pi_i \text{Pois}(x_c | \theta_{i,c}) \\
+ \pi_{M+1} u(c) \text{Geom}(x_c - \theta_{M+1,c} | \lambda, \alpha, M) \tag{5.5}
\]
where $u(c)$ is an indicator function defined as
\[
u(c) = \begin{cases} 
1 & \text{if } c \in Z \\
0 & \text{if } c \notin Z
\end{cases} \tag{5.6}
\]

5.2.1.4 Estimation of the model parameters by Expectation Maximization

The mixture parameters $\pi$, the Poisson rate parameter $\lambda$ and the geometric distribution rate parameter $\alpha$ are estimated from the data ($N$ contigs) by optimizing the log-likelihood,
\[
-\mathcal{L}(C | \pi, \lambda, \alpha) = - \log \prod_{c=1}^{N} p(x_c | \pi, \lambda, \alpha)p(\pi) \tag{5.7}
\]
In the E-step, current estimates of the parameters ($\theta_{i,c}^{old} = L_c i \lambda^{old}$ and $\pi^{old}$) are used to estimate the posterior probabilities, or responsibilities $r_{i,c} = \hat{p}(i|x_c)$ for each model and each contig. In the M-step, the newly obtained responsibilities are used to update $\lambda$ and $\alpha$ as:
\[
\lambda^{new} = 
\frac{\sum_{c=1}^{N} \sum_{i=1}^{M} r_{i,c} x_c}{\sum_{c=1}^{N} \sum_{i=1}^{M} i L_c r_{i,c} + \sum_{c=1}^{N} u(x_c) (M+1) L_c r_{M+1,c} \log \left(1 - \frac{\alpha}{L_c}\right)} \tag{5.8}
\]
\[
\alpha^{new} = 
\frac{\sum_{c=1}^{N} u(x_c) r_{M+1,c}}{\sum_{c=1}^{N} u(x_c) r_{M+1,c} \left(\frac{x_c - \theta_{M+1,c}^{old}}{L_c} + 1\right)} \tag{5.9}
\]

5.2.1.5 Incorporating prior knowledge on ploidy

In many biological experiments there is prior knowledge on the distribution $p(i)$. For example, in haploid yeast samples most contigs will correspond to mixture component $i = 1$; for diploid samples it is expected that $p(i = 2)$ will dominate. Note that ploidy is not the sole influence on the distribution of $p(i)$ in an unfinished assembly, but also the repeat content in the genome. We adopted the idea of Klambauer et al. [114] to use a Dirichlet prior distribution with parameters $\gamma$ for cases where we can incorporate prior
5. **DE NOVO DETECTION OF COPY NUMBER VARIATION**

knowledge on $p(i)$, where $\gamma$ is an $M$-dimensional vector $(\gamma_1, ..., \gamma_M)$. The update rule for $\pi_i$ then becomes:

$$
\pi_i^{new} = \frac{\sum_{c=1}^{N} r_{ic} + \frac{1}{N}(\gamma_i - 1)}{1 + \frac{1}{N}(\sum_{i=1}^{M+1} \gamma_i - (M + 1))}
$$

**(5.10)**

### 5.2.1.6 Model selection

The model with the optimal number of Poisson distributions is selected among models with three to twenty Poissons with the lowest Bayesian information criterion (BIC) (see supplementary material S1.2).

### 5.2.1.7 MAP estimation of integer copy numbers

We infer the integer copy number $\hat{i}_{MAP}$ for a given read count $x_c$ by maximum a posteriori estimation (MAP):

$$
\hat{i}_{MAP} = \arg \max_{i=1, ..., M+1} p(i|x_c) = \arg \max_{i=1, ..., M+1} r_{i,c}
$$

**(5.11)**

Contigs for which $\hat{i}_{MAP} = M + 1$ are outlier contigs.

### 5.2.2 Copy number variation using co-assembly

#### 5.2.2.1 Detecting copy number variation

Given two sequencing samples $O_1$ and $O_2$, we are interested in those contigs that display an aberrant copy number in the two samples, i.e. in copy number variation (CNV). We fit the proposed Poisson mixture model for both samples; assuming independent we can calculate the probability of a CNV for non-outlier contigs as:

$$
p(CNV \text{ in } x_c) = 1 - \sum_{i=1}^{M} p_1(i|x_c)p_2(i|x_c)
$$

**(5.12)**

### 5.3 Methods

#### 5.3.1 Simulated DNA sequencing data

The *Saccharomyces cerevisiae* strain S288C reference genome (S1) was downloaded from the Saccharomyces Genome Database (www.yeastgenome.org, accessed 3-3-2011). For testing CNV detection, a perturbed yeast genome containing 100 duplications (gains) was simulated from this yeast reference genome. Donor and insertion sites were randomly drawn. The duplication length was randomly drawn between 1Kbp and 10Kbp. Distance between the duplication events was guaranteed by rejecting a newly drawn event if either one of the edges of the donor or the insertion site was within 10Kbp of another event. 800,000 error-free sequencing reads with a length of 150 bp each were then simulated at 10X coverage for each sample using fragsim (145). Additionally, a 20X coverage read
5.3 Methods

dataset was simulated from the reference genome to generate a shotgun reference assembly, which was used to gauge performance of mapping based CNV detection methods on an unfinished genome.

5.3.2 Real DNA sequencing data

In the experiments, we used a number of NGS data sets available for various yeast strains (see Table 5.1). The sequencing of \textit{S. cerevisiae} CEN.PK113-7D was described in Nijkamp et al. (163), yielding a Illumina library (Illumina, San Diego, CA) with a 50 bp read length and a 454 GS FLX library (454 Life Sciences, Branford, CT) with an average read length of 350 bp.

Table 5.1: \textit{S. cerevisiae} strains used in this study. The ”Accession” column shows the Sequence Read Archive accessions numbers. The last column shows the $\gamma$ vector used as parameters for the Dirichlet prior, reflecting prior expectations on copy number distribution. $G$ is the hyper parameter which is set to the number of contigs.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Species</th>
<th>Strain</th>
<th>Description</th>
<th>Ploidy</th>
<th>Dirichlet prior parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>\textit{S. pastorianus} SPA</td>
<td>lager brewing yeast</td>
<td>aneuploid</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td></td>
<td>\textit{S. pastorianus} SPB</td>
<td>lager brewing yeast</td>
<td>aneuploid</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>SRX129889</td>
<td>\textit{S. cerevisiae} CEN PK113-7D</td>
<td>laboratory strain</td>
<td>haploid</td>
<td>$\gamma = (1 + G, 1, \ldots, 1)$</td>
<td></td>
</tr>
<tr>
<td>SRX129995</td>
<td>\textit{S. cerevisiae} IMW004</td>
<td>laboratory evolved strain</td>
<td>aneuploid</td>
<td>$\gamma = (1 + G, 1, \ldots, 1)$</td>
<td></td>
</tr>
<tr>
<td>SRX039438</td>
<td>\textit{S. cerevisiae} Y12</td>
<td>palm wine sample, single spore</td>
<td>aneuploid</td>
<td>$\gamma = (1 + \frac{1}{2}G, 1 + \frac{1}{2}G, 1, \ldots, 1)$</td>
<td></td>
</tr>
<tr>
<td>subsp. uvarum</td>
<td>derivative of NRRL Y-12633</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The genomic DNA of two brewing yeasts of the species \textit{S. pastorianus} was purified as previously described (50). The two \textit{S. pastorianus} strains are indicated as SPA and SPB in this study. The fragments of approximately 180 to 200 bp were sequenced paired-end on a Genome Analyzer IIx (Illumina) with a read length of 100 bp at Baseclear (Leiden, The Netherlands). Afterwards the overlapping read pairs were merged into single longer reads.

DNA sequencing reads (454/Roche) were downloaded from the short-read archive (SRA) for the \textit{S. cerevisiae} strain Y12, part of the \textit{S. cerevisiae} strain project [http://genome.wustl.edu/genomes/saccharomyces_cerevisiae_strain_project_genomes].

5.3.3 Assembly and alignment

Genome assembly was performed using the GSAssembler 2.6, aka Newbler (454/Roche) using default settings. The contig string graph that results from an assembly with the Newbler was colored using the .ACE file that describes the read layout on the contigs. Alignments were performed with nucmer version 3.07, part of the Mummer 3 package (53).

5.3.4 Copy number integrity

A contig has to be present an integer number of times in its genome if the assembler correctly bifurcated the contig graph. This was assessed by aligning contigs that were assembled using simulated reads to the genome. Only contigs longer than 500 bp were con-
5. DE NOVO DETECTION OF COPY NUMBER VARIATION

Figure 5.2: A contig should be present an integer number of times in the genome. To assess whether this is true, contigs assembled from simulated reads were aligned to the genome from which the reads originate. a) The contig occurs exactly two times in the genome and therefore is assigned an integrity value of 0. b) The contig occurs 1.4 times in the genome and is assigned an integrity value of 0.4.

sidered. Alignments with an identity lower than 95% were discarded. For each position in the reference only the best query hit was kept, allowing for query overlaps (delta-filter -i95 -r in the Mummer package). The number of times the contig is covered in the remaining alignments is then summed. The closest integer to this sum is denoted as CN_{align}. We define copy number integrity (Fig. 5.2) as the absolute deviation from CN_{align}, optimal at 0 and with a maximum of 0.5. For example, a contig that is covered 1.6 times in the genome has a integrity value of 0.4, which implies the contig is far from having an integer copy number.

5.3.5 Validation

Validation was performed on the genome with the simulated duplication events (gains) described above. Magnolya was run on the co-assembly with haploid settings. FREEC v5.6 [19], CN.MOPS v1.2.1 [114] and CNV-Seq [253] were used to benchmark the performance of Magnolya. The read datasets used to perform the co-assembly were independently mapped to the reference genome and the reference shotgun assembly. CN.MOPS was provided three times the reference sample and one time the perturbed sample, to enable it to model ”normal read count”. The haploid CN.MOPS version was used with minWidth=1 and priorImpact=0.5. The minimum and maximum expected GC content required for FREEC were set to 0.3 and 0.5, respectively (the S. cerevisiae GC content is 0.38). For both FREEC and CN.MOPS the window size was set to 500 bp, CNV-seq inferred a window size itself. CNV-seq was run with global normalization on the reference genome and the shotgun assembly. Additionally, contig normalization was performed on the shotgun assembly to account for possible collapsed repeats. Reads were mapped to the reference genome and the reference assembly using the Burrows-Wheeler Aligner (BWA) version 0.5.9-r16 [136]. Sensitivities and specificities were calculated by counting true positive and false positive calls per base. 250 bp around duplications were ignored, to prevent counting them as false positives due to overhanging windows.
5.4 Results and Discussion

Table 5.2: Simulation results for copy number estimation. Only copy numbers 1 to 4 are listed, since no higher copy numbers were observed. The integrity column indicates averages and standard deviations of the integrity values (Fig. 5.2).

<table>
<thead>
<tr>
<th>CN</th>
<th>CN_{align}</th>
<th>CN_{align} = \hat{i}_{MAP}</th>
<th>integrity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3638</td>
<td>3638 (100%)</td>
<td>0.0011 ± 0.014</td>
</tr>
<tr>
<td>2</td>
<td>67</td>
<td>66 (99%)</td>
<td>0.0111 ± 0.055</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>14 (100%)</td>
<td>0.0005 ± 0.001</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4 (100%)</td>
<td>0.0014 ± 0.003</td>
</tr>
</tbody>
</table>

5.3.6 Timing

Sets of 1000, 2500, 5000, 7500, 10000 and 12500 contigs were randomly sampled without replacement from the Y12 dataset. Magnolya was run 25 times on each dataset, each with $M = 8$ and $\lambda$ initialized on the 25th percentile. The algorithm was halted when an iteration resulted in a likelihood improvement less than $10^{-5}$. CPU time for the EM procedure was measured using python’s `time.clock()` on a single core of a Dell T7400 Workstation with Intel Xeon X5272 dual core processor.

5.4 Results and Discussion

Magnolya can be used in estimating the copy number of the contigs in the genome. This was first verified on simulated data and then applied on data obtained from an allotriploid hybrid genome of a palm wine yeast. In a multi-sample setting the PMM can then be used as a statistical approach to CNV detection. This has been tested on simulated data and on the complex aneuploid genomes of beer brewing yeasts.

5.4.1 Copy number estimation

5.4.1.1 Simulations

The performance of the PMM to estimate the number of times a contig is present in a genome was assessed using an assembly of simulated reads from the S288C yeast reference genome. Two individual features of our method were tested using this simulation (Table 5.2).

First, the assumption that contigs have an integer copy number was assessed by calculating copy number integrity (Sec. 5.3.4). The contigs were found to have very low average integrity values (Table 5.2). Only 18 of the 3723 contigs had an integrity value over 0.1, which proves the assembled contigs indeed mostly have an integer copy number. Second, the PMM copy number estimate $\hat{i}_{MAP}$ was compared to the number of times a contig appears in the genome $CN_{align}$. Of the 3723 tested contigs, the PMM incorrectly estimated the copy number for just a single contig.
5. **DE NOVO DETECTION OF COPY NUMBER VARIATION**

Figure 5.3: PMM fitted on the assemblies of the wine yeast Y12. Grey bars: histogram of read counts on the contigs, normalized to a contig length of 300 bp. The fitted Poisson mixture model is plotted as a black line with $300\lambda = 11.7$.

5.4.1.2 **Copy number and genome size estimation of an allotriploid hybrid genome.**

*Saccharomyces* yeast species used in industrial fermentation processes, such as beer and wine making, are often hybrid species, containing DNA of several *Saccharomyces* origins. For example, the lager beer brewing yeast *S. pastorianus* was shown to be a hybrid between *S. cerevisiae* and *S. eubayanus* ([140]). The wine making yeast VIN7 was shown to contain diploid *S. cerevisiae* and haploid *S. kudriavzevii* genomic content ([24]). Furthermore, these hybrid genomes may contain many chromosomal rearrangements and aneuploidy ([159]). Aneuploid genomes have an irregular number of chromosomes. Magnolya can be used to estimate the copy number per contig in aneuploid genomes and thereby the total genome size.

*S. cerevisiae* Y12 is such a hybrid yeast. Genomic sequencing reads from its genome were assembled into 14,551 contigs containing 16.1 Mbp of total sequence, of which 5,893 large contigs ($\geq 500$ bp) contained 13.9 Mbp. Fitting the mixture model on the read counts without incorporation of prior knowledge on the ploidy resulted in $i = 2$ and $i = 4$, to fit the two peaks in Fig. 5.3, i.e., the model explained the data as contigs having copy number 2 and 4. Visual inspection of the read count histogram (Fig. 5.3) led us to the belief that these peaks stem from haploid and diploid genomic content instead, which is in line with previous results describing the allotriploid strain VIN7 ([24]). The uneven coverage of the 454 sequencing data resulted in a better explanation of the data using more Poissons.

The parameters for the Dirichlet prior distribution were then set to favour single and double copy numbers for the contigs (Table 5.1), representing our belief that the Y12 genome has haploid and diploid content, rather than diploid and tetraploid. The mixture model was fit on the contigs ($\geq 500$ bp) using the model selection procedure. A total of $M = 8$ Poissons was found to be optimal. The *S. cerevisiae* Y12 genome was estimated to be $\sum_{k=1}^{N} L_k \hat{i}_k = 29.9$Mbp in size, 22% of which was represented one time (haploid) and 74% two times (diploid).
The contigs were mapped to the the 12.1 Mbp \textit{S. cerevisiae} S288C reference genome to investigate the possible \textit{S. cerevisiae} origin of parts of the genome. Only alignments of contigs that aligned reliably to the reference genome were kept (> 95% contig coverage, > 99% identity). In total 9.2 Mbp could thus be aligned. These aligned bases originated for 97% from contigs with $i_{ML} = 2$, indicating that the \textit{S. cerevisiae} content in this hybrid genome is present in diploid form, as is the case for VIN7 (21).

The Y12 assembly had the most contigs and was therefore used to time the algorithm (Fig. 5.4). Computational complexity is low, (compared to mapping and assembly of short read data) and scales linearly with the number of contigs.

5.4.2 Copy number variation detection

5.4.2.1 Benchmarking

A simulation experiment was designed to benchmark the performance of Magnolya in a setting where a closely related reference genome is available versus the situation where this is not the case. We did not simulate biological noise or contamination, nor did we introduce uneven depth of coverage. Therefore, this benchmark does not gauge how well the methods deal with such biological effects, but merely illustrates the potential performance drops when no reference is available.

Magnolya was benchmarked against the three top performing methods in the study performed by Klambauer \textit{et al.} (114). The reference and a perturbed yeast genome containing 100 duplications between 1 Kbp and 10 Kbp were used. The sequencing reads were mapped to the reference genome and to a shotgun assembly to infer CNVs. Sensitivity and precision are the two measures relevant to CNV detection. Table 5.3 reports the F-measure, i.e. the harmonic mean of sensitivity and precision, for both situations. The F-measure does not take the true negative rate into account, but this is not an issue since this rate is always high for all methods (since genomes are very large compared to the total length of the CNVs).

For all methods two samples were used, a reference and a perturbed sample, except for CN.MOPS. CN.MOPS is designed to be used with multiple samples, so it can model read depth across multiple samples. It cannot be used on only two samples, because it cannot distinguish what is normal read count and what is not. We therefore provided it with three times the reference sample and one time the perturbed sample. CNV-seq performs best on the reference genome (Table 5.3). Possibly this is because it is model free, only looks at read depth ratios and does not account for noise, which is absent in this simulation. For all methods performance dropped when applied to the shotgun assembly.

Magnolya has the highest F-score, with a sensitivity of 0.93 and a precision of 0.94. Its performance is close to the other methods on a finished reference genome, but surpasses the other methods with at least a double F-measure on the shotgun assembled reference genome.
5. **DE NOVO DETECTION OF COPY NUMBER VARIATION**

Figure 5.4: CPU time consumed by the EM algorithm for different numbers of contigs, which were randomly sampled from the Y12 yeast dataset. On average the EM algorithm finished within 14 iterations.

5.4.2.2 **CNV in a laboratory evolution pair**

As another test case, a pre- and post-laboratory evolution pair was analyzed. Short-term laboratory evolution (up to a few months) generally leads to only few mutations. An evolved strain and its pre-evolution ancestor are therefore expected to be genetically very close. We recently applied laboratory evolution to investigate lactate transport in the haploid yeast *S. cerevisiae* CEN.PK113-7D (50). This strain, used in systems biology research and in industry, deviates significantly from the yeast reference sequence S288C including multiple Kbp insertions and deletions and over 20,000 single nucleotide variations. Our *de novo* CNV detection algorithm enables us to directly compare the two CEN.PK113-7D individuals, without having to resort to comparison via a more distant reference genome.

A mutant harboring a *JEN1* deletion grew very poorly in liquid culture with lactate as sole carbon source. Laboratory evolutionary engineering was then used to evolve a *JEN1* independent fast growing strain on lactate. A mutation in *ADY2*, an acetate transporter, was identified as responsible; but additionally a copy number increase was suspected to give the evolved strain a competitive advantage. Read mapping analysis confirmed that the gene dosage of *ADY2* was indeed increased through the formation of a novel isochromosome carrying two additional copies of *ADY2* (Fig. 5.6c). Here, we reanalyzed the sequencing data of the evolved strain. The obtained integer copy numbers indeed indicate the chromosomal regions corresponding to the novel isochromosome (Fig. 5.6d). Compared to the mapping-based approach taken in de Kok *et al.* (50), these results are easier to interpret.

5.4.2.3 **CNV in aneuploid brewing strains**

Copy number variation was detected between two aneuploid lager brewing yeast strains of the species *S. pastorianus*. Co-assembly of the Illumina reads showed aneuploidy for both strains (Fig. 5.5), with a large percentage of contigs present one, two or three times in the genomes. This is in agreement with previous observations in a different *S. pastorianus* strain (159).
5.5 Conclusions

Since the advent of next-generation sequencing, variation detection has been performed by mapping short reads to a reference genome in order to detect aberrant read mappings. Comparing two individuals by mapping them to a (perhaps distant) reference inevitably introduces bias. The only unbiased approach to comparisons of individuals is via de novo assembly. Short sequencing reads and repetitive genomes have thus far hampered accurate reconstruction of individual genomes. However, with increasing read lengths and

---

Figure 5.5: Poisson mixture distribution fitted on the assemblies of two brewing yeasts. The grey bars are the read counts $x_c$ on the contigs, normalized to a contig length of 1000 bp. The plotted distributions indicate the mixture components ($M = 6$, $1000\lambda = 99$ for SPA and $1000\lambda = 169$ for SPB).

The mixture model was trained on both datasets using the model selection procedure. The Bayesian information criterion indicated a total of $M = 6$ Poisson distribution was optimal to use on these data. Fig. 5.5 shows that the two *S. pastorianus* strains have different karyotypes. For example, strain SPB appears to have more contigs appearing only once in the genome. For each contig, the probability of a copy number difference between the strains was calculated using the posterior probabilities $p(i|x_c), i \in \{1, 2, ..., 6\}$ using equation (5.12). More than 13 Mbp ($p(CNV) > 0.95$) were found to be present at a different copy number. Although the nucleotide composition of these two strains may be very similar, this large scale CNV largely affects the gene dosage of thousands of genes, with an intriguing yet unexplored effect on the phenotypic characteristics.

Table 5.3: Performance in simulation experiments expressed as the F-measure (the harmonic mean of precision and sensitivity). The columns show a scenario with and without the availability of a finished reference genome. CNV-seq was run with global normalization and normalization per contig.

<table>
<thead>
<tr>
<th>Reference genome</th>
<th>Shotgun assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN-MOPS</td>
<td>0.57</td>
</tr>
<tr>
<td>Control-FREEC</td>
<td>0.86</td>
</tr>
<tr>
<td>CNV-seq global</td>
<td>0.91</td>
</tr>
<tr>
<td>CNV-seq contig</td>
<td>-</td>
</tr>
<tr>
<td>Magnolya</td>
<td>-</td>
</tr>
</tbody>
</table>

5.5 Conclusions
Figure 5.6: *De novo* CNV detection using Magnolya. The yeast *S. cerevisiae* CEN.PK113-7D was evolved in the laboratory under strong selective pressure. The genomes of the evolved strain and its pre-evolution ancestor were both sequenced and assembled into contigs. The grey bars form a histogram of the read counts on the contigs for a) the pre-evolution ancestor and b) the evolved yeast strain. The read counts have been normalized for visualization purposes to a contig length of 1000 bp, i.e. \( \frac{x_c}{L_c} \times 1000 \) for each contig c. The mixture model consisting of 4 Poissons has been plotted as a black line with a) \( 1000\lambda = 41 \) and b) \( 1000\lambda = 152 \). Note that because of the large difference in number of reads in the two datasets the axes are unequally scaled. c) Schematic representation of the original chromosome III and the newly formed isochromosome III during evolution. The solid line represents \( i = 1 \), the dotted line \( i = 2 \) and the dashed line \( i = 3 \). d) Integer copy number variation plotted versus chromosomal location on the yeast reference genome.
recently developed de novo assemblers that efficiently assemble mammalian sized genomes, reconstruction and comparison of individual genomes is coming within reach.

We here proposed a Poisson mixture model to de novo estimate copy numbers of contigs, and combined this with a co-assembly approach. This allows easy detection of copy number variation, one of the most abundant types of genomic variation, with severe phenotypic effects. The mixture model estimates copy numbers per sample at high specificity, exploiting the fact that assemblers automatically segment a genome into regions of integer copy number, and allows inference of copy number variation. The resulting de novo CNV detection algorithm has two main advantages over mapping-based approaches: foregoing the need for read count-based segmentation and the lack of bias w.r.t. a reference genome.

The Magnolya algorithm performs at higher precision and sensitivity then other methods when no finished reference is available. The method was shown to perform well on yeast genomes, a simple eukaryote for which good assemblies can be obtained using current sequencing technology. When long enough reads can be obtained to generate human assemblies a de novo approach might be preferred over reference based approaches to detect copy number variation in matched experiments, such as tumor-normal pairs.

The co-assembled colored string graphs that are employed in this study enclose all genomic variation between the two assembled individuals, including inversions, insertions, deletions and translocations. While in this work we focused on the detection of CNVs, an algorithm named bubble calling was recently proposed to mine other classes of variation from colored assembly graphs (103). We expect a combination of a bubble calling algorithm, our CNV detection and approaches exploiting read pair data in colored assembly graphs to allow a move to fully unbiased detection of variation between individuals in the near future.

5.6 Acknowledgements

Funding: This work was supported by the Kluyver Centre for Genomics of Industrial Fermentation, supported by the Netherlands Genomics Initiative.

5.7 Supplementary data

Supplementary data is available at http://bioinformatics.oxfordjournals.org/content/28/24/3195.long
5. *DE NOVO* DETECTION OF COPY NUMBER VARIATION
Genome duplication and mutations in \textit{ACE2} cause multicellular, fast-sedimenting phenotypes in evolved \textit{Saccharomyces cerevisiae}


\textit{In press, Proceedings of the National Academy of Sciences (PNAS)}
Abstract

Laboratory evolution of the yeast *Saccharomyces cerevisiae* in bioreactor batch cultures has, under different selection pressures, yielded variants that grow as multicellular, fast-sedimenting clusters. Knowledge on the molecular basis of this phenomenon may contribute to the understanding of the natural evolution of multicellularity and to manipulating cell sedimentation in laboratory and industrial applications of *S. cerevisiae*. Multicellular, fast-sedimenting lineages obtained from two independent evolution experiments with a haploid *S. cerevisiae* strain were analyzed by whole-genome resequencing. The two independent cell lines showed different frameshift mutations in a stretch of eight adenosines in *ACE2*, which encodes a transcriptional regulator implicated in cell cycle control. Introduction of the two *ace2* mutant alleles into the haploid parental strain led to slow-sedimenting cell clusters that consisted of just a few cells, thus representing only a partial reconstruction of the evolved phenotype. In addition to single-nucleotide mutations, a whole-genome duplication event had occurred in both evolved multicellular strains. Construction of a diploid reference strain with two mutant *ace2* alleles led to complete reconstruction of the multicellular-fast sedimenting phenotype. This study shows that whole-genome duplication and a frameshift mutation in *ACE2* are sufficient to generate a fast-sedimenting, multicellular phenotype in *S. cerevisiae*. The nature of the *ace2* mutations and their occurrence in two independent evolution experiments encompassing fewer than 500 generations of selective growth suggest that switching between unicellular and multicellular phenotypes may be relevant for competitiveness of *S. cerevisiae* in natural environments.
6.1 Introduction

Ease of cultivation and genome analysis, short generation times and large population sizes have contributed to the popularity of micro-organisms as model systems in experimental evolution. In addition to providing insights into evolutionary adaptation mechanisms and strategies, laboratory evolution of microorganisms provides a powerful tool to improve characteristics that are relevant to microbial biotechnology. This application of laboratory evolution, known as evolutionary engineering (197) has, for example, contributed to expanding substrate range (57, 118, 128, 248, 249), functional implementation of alternative product pathways (50, 257) and increased tolerance to inhibitors (3, 57, 252) in various production organisms (reviewed in (181)). Recent advances in DNA sequencing and genetic modification facilitate characterization and reconstruction of the molecular basis of evolved phenotypes obtained in laboratory evolution, thus enabling experimental testing of hypotheses on evolutionary strategies and underlying molecular mechanisms (237). This approach has generated new insights into mutation rates (13, 62, 143), genetic drift (62, 212), epistasis (129), clonal interference (110) and other important aspects of evolution by natural selection (reviewed in (41)). In microbial biotechnology, reverse engineering of evolved phenotypes, known as inverse metabolic engineering (168), has similarly benefited from the availability of these genomic methodologies (219). In this applied research context, knowledge on the genetic basis of an industrially relevant phenotype not only increases understanding, but also enables its reconstruction and improvement in other microbial strains and species (27, 79, 133, 168, 219).

In unicellular organisms such as the yeast Saccharomyces cerevisiae, laboratory evolution is facilitated by the ease with which single-cell lines can be isolated from evolving cultures. Recently, however, Ratcliff et al. (184) described evolution of multicellularity in S. cerevisiae cells within a single long-term cultivation experiment. The multicellular variant, in which daughter cells did not separate from the mother cell upon cell division, dominated the population within a few generations when fast sedimentation was selected for in test tubes. Evolution of these multicellular clusters of S. cerevisiae, which even showed signs of cellular differentiation, was proposed to be a laboratory model for the origin of multicellularity in eukaryotes (184).

At least 25 occurrences of the shift from unicellular to multicellular life forms have been recognized in the evolution of life on Earth (20, 86, 116, 190). However, knowledge on the evolutionary pressures resulting in the selection of multicellular life forms and the underlying molecular mechanisms is far from complete. It has been proposed that multicellularity can contribute to phenotypes as diverse as stress tolerance (141, 242), affinity for substrates (121) and relief of predatory pressure (21).

Knowledge on the mutations that cause the switch from unicellular to multicellular growth in yeast may contribute to understanding of the events leading to the transition to multicellular life forms. Moreover, such knowledge can contribute to a better modulation of biomass sedimentation in laboratory research and industrial application of S. cerevisiae. In our research on evolutionary engineering of S. cerevisiae, we frequently
observed multicellular, fast-sedimenting clusters that, upon microscopic examination, resemble the phenotype described by Ratcliff et al. (184). The goal of the present study was to elucidate mutations that are responsible for the generation of multicellular variants. To this end, we monitored the formation of multicellular variants in two independent laboratory evolution experiments with a haploid laboratory strain of *S. cerevisiae*. Subsequently, representative mutants from the two experiments evolutions were characterized. Genetic changes identified by whole-genome resequencing were reverse engineered in the unicellular parental strain, enabling the identification of two changes that, together, were sufficient to reproduce the multicellular, fast-sedimenting phenotype.

6.2 Results

6.2.1 Selection of multicellular clusters in sequential bioreactor batch cultures

Where previous reports studied evolution of *S. cerevisiae* in serial shake ask cultures (95, 105, 106, 193), we reproducibly observed the occurrence of large multicellular clusters during prolonged anaerobic cultivation of the haploid *S. cerevisiae* strain CEN.PK113-7D (163) in sequential bioreactor batch cultures. The phenotype of these clusters was similar to the “snowake yeast” previously described by Ratcliff and coworkers (184, 185). The design of the “fill and draw” system used in our bioreactors provided an unintended selective advantage to fast-sedimenting cell lines. The vertical pipe used to empty the bioreactor after each cultivation cycle did not reach the bottom of the vessel. Consequently, fast-sedimenting cells were enriched in the small remaining volume used as inoculum for the next batch cultivation cycle. To facilitate identification of mutations contributing to the multicellular phenotype (95, 168), two identical independent anaerobic evolution experiments were started on a mixture of 20 g·L$^{-1}$ glucose and 20 g·L$^{-1}$ galactose. Although the specific growth rate on galactose doubled during both evolution experiments (from 0.11 to 0.22 and 0.20 h$^{-1}$; Fig. 6.1A and Fig. S1A) and the length of the batch cultivation cycles decreased by at least 35% (Fig. S1 H and I), the morphology of *S. cerevisiae* changed dramatically as large, multicellular clusters became dominant in both evolution experiments (Fig. 6.1 B-F and Fig. S1 B-G). The sedimentation index, calculated from the time-dependent decrease of the optical density of statically incubated cell suspensions, strongly increased, in parallel with the increasing abundance of multicellular clusters (Fig. 6.1 B-F and Fig. S1 B-G). Culture samples taken at the end of the two evolution runs [after 4,200 (~900 generations) or 2,880 h (~500 generations)] showed almost complete sedimentation after 5 min of static incubation (Fig. 6.1I). In *S. cerevisiae*, reversible aggregation of individual cells into fast sedimenting clusters can occur via flocculation, which involves a Ca$^{2+}$-dependent interaction of yeast cell wall proteins and carbohydrates (210). However, the multicellular clusters observed in the evolved cultures could not be reverted to a single-cell morphology by incubation with well-known antiflocculent agents such as EDTA (0.5 M) (37), mannose (211), or protease (trypsin 1,500 units·mL$^{-1}$) (215).
This indicated that the phenotype did not result from interaction of unicellular yeasts, but rather from an incomplete cell division.

6.2.2 Whole-genome sequence analysis of two evolved multicellular isolates

To investigate the molecular basis of the evolved multicellular phenotype, a fast-sedimenting mutant was isolated from each of the two evolution experiments. Strains IMS0267 and IMS0386 originated from evolution 1 and from evolution 2, respectively. To verify the genetic stability of the mutations responsible for the multicellular phenotype, strains IMS0267 and IMS0386 were grown for at least 10 generations on glucose in shake flask cultures. This did not result in observable changes in multicellularity or sedimentation behavior. Genomic DNA of strains IMS0267 and IMS0386 was sequenced at high genome coverage (81.6-fold and 38.5-fold coverage for IMS0267 and IMS0386, respectively) and compared to the reference genome of the parental strain CEN.PK113-7D [163]. The high coverage enabled accurate analysis of genome-wide copy number variation (CNV) by co-assembly of the evolved and the reference strains [162] as well as identification of single nucleotide variations (SNV) and indels.

To estimate the ploidy of the evolved strains we de novo co-assembled sequence reads of the evolved and the CEN.PK113-7D reference strains. Copy numbers of the assembled contigs were estimated using the Poisson mixture model-based algorithm Magnolya [162]. Surprisingly, this analysis revealed that both evolved mutants had undergone a whole-genome duplication event relative to the haploid MATa ancestor CEN.PK113-7D (Figure 6.2A and 6.2B). Both IMS0267 and IMS0386 were for the most part diploid with triplicated genome islands. IMS0267 exhibited triplication of parts of CHRII, XIII and XVI while IMS0386, besides triplication of parts of CHRIII, VIII and quadruplication of XIII, had a complete trisomy of CHRII and XI (Figure 6.2A). However, IMS0267 and IMS0386 kept haploid characteristics as the strains were not able to form tetrads, but were able to mate with a MATa strain (IMI081) and sporulate at a low rate.

Table 6.1: Insertions (INS) and deletions (DEL) detected in the genomes of two evolved multicellular mutant isolates (IMS0267 and IMS0386) relative to the sequence of the parental haploid reference strain CEN.PK113-7D.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Nucleotide change</th>
<th>Type mutation</th>
<th>Amino acid change</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genes mutated in IMS0267:</strong></td>
<td>[162]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACE2</td>
<td>Transcription factor that activates expression of early G1-specific genes.</td>
<td>*1112A</td>
<td>INS</td>
<td>frame-shift</td>
</tr>
<tr>
<td><strong>Genes mutated in IMS0386:</strong></td>
<td>[162]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACE2</td>
<td>Transcription factor that activates expression of early G1-specific genes.</td>
<td>A1112*</td>
<td>DEL</td>
<td>frame-shift</td>
</tr>
</tbody>
</table>
6. LABORATORY EVOLUTION OF MULTICELLULAR YEAST

Mapping of sequence reads of the evolved strains onto the genome sequence of CEN.PK113-7D revealed no single nucleotide variations (SNVs) and only two high-probability indels (Table 6.1). Only a single gene, *ACE2*, was affected by mutation in both evolved strains (Table 6.1). *ACE2* encodes a transcriptional regulator of, amongst others, *CTS1*, a gene involved in the latest phase of the cell cycle and more specifically required for septum destruction after cytokinesis \((58, 196, 236)\). Interestingly, two differently mutated *ACE2* alleles were identified in the evolved isolates. These mutations were found in the same region of *ACE2*: in IMS0267 an adenosine was introduced at position 1112 while in IMS0386 an adenosine was deleted at the same position. The resulting alleles were named *ace2-1* and *ace2-2*. Both mutations caused the introduction of a premature stop codon, at position 1165 or position 1114 in IMS0267 and IMS0386 respectively (Fig. S2). Based on its occurrence in both evolved strains and its role in the yeast cell cycle, we hypothesized that the mutations in *ACE2* contributed to the evolved multicellular phenotype.

6.2.3  *Ace2-1 and Ace2-2 exhibit reduced transcriptional activation of Ace2 targets*

The predicted proteins encoded by *ace2-1* and *ace2-2* alleles were 388 and 371 amino acids long instead of 770 amino acids for the original protein (Fig. S2). As a result, the three C2H2-type zinc finger domains and the nuclear localization signal sequence (NLS) located at the C terminus of the Ace2 protein sequence were lost. Conversely, the truncated Ace2 versions retained the nuclear export signal sequence and the interaction domain with Cbk1, a protein kinase involved in the regulation and localization of Ace2. To study the impact of the *ace2* mutations in the evolved multicellular strains, transcription of the previously characterized Ace2 targets *DSE1/YER124C*, *DSE2/YHR143W*, *CTS1*, and *SCW11* \((58, 198)\) was analyzed in the *ace2-1* and *ace2-2* strains by real-time RT-PCR. Expression of all these four Ace2 targets was at least 90% lower in the evolved strains than in the parental strain CEN.PK113-7D (Fig. 6.3A). Among the targets of Ace2, *CTS1* is of special interest, because it encodes an endo-chitinase required for degradation of the mother-daughter septum \((126)\). Cell wall staining with Calcofluor White, which specifically stains chitin \((93)\), confirmed that within the multicellular clusters, the cells remained attached at the chitin bud neck site (Fig. 6.3B). Consistent with a key role of reduced chitinase expression in the multicellular phenotype, treatment with chitinase led to dispersal of the multicellular clusters into single cells (Fig. 6.3C and D). To test whether reduced expression of *CTS1* is sufficient to cause a multicellular phenotype, we analyzed the phenotype of *cts1Δ* mutants. A homozygous *cts1Δ/cts1Δ* strain showed large cell aggregates relative to an isogenic unicellular reference strain (Fig. S3). Sedimentation of the *cts1Δ/cts1Δ* strain was not as fast as in the evolved strains (Fig. S3), which may either reflect differences in strain background or indicate that, in addition to a key role of reduced *CTS1* expression, other factors contribute to the fast-sedimenting phenotype.
6.2 Results

6.2.4 Reverse engineering of different ace2 alleles in unicellular strains

To further investigate the importance of the ace2-1 and ace2-2 mutations in evolution of multicellular, fast-sedimenting *S. cerevisiae* strains, the wild type ACE2 allele in the haploid ancestor strain CEN.PK113-7D was replaced by either of the two mutant versions. Neither reverse engineering of these mutant ace2 alleles nor complete deletion of ACE2 in CEN.PK113-7D (strain IMK395) resulted in complete reconstruction of the multicellular phenotype of the evolved strains (Figure 6.4). The clusters formed by strains IMK395, IMI197 (ace2-1-HphNT1) and IMK245 (ace2-2-KanMX) were much smaller and their sedimentation indices, although significantly higher than that of CEN.PK113-7D, were 10-fold lower than those of the evolved isolates IMS0267 and IMS0386. Conversely, replacement of one of the ace2-1 or ace2-2 copies in IMS0267 and IMS0386, respectively, by the wild type ACE2 allele led to a complete reversion of the phenotype to single cells (Figure 6.4). This observation confirmed the recessive character of the ace2 mutations (IMI220 and IMI221) that was expected based on the loss of transcriptional activation activity (Figure 6.3A).

Estimation of ploidy by flow cytometry analysis of DNA content was not possible with the multicellular evolved strains IMS0267 and IMS0386. We therefore performed the analysis with strains IMI220 (ace2-1/ACE2), IMI221 (ace2-2/ACE2) and the strains IMS0267 and IMS0386 pretreated with chitinase. Cytometry values confirmed the prediction from sequence coassembly that the evolved strains had undergone a whole genome duplication (Fig. 6.2 A and B). IMI220 and IMI221 exhibited a 1.9- and a 2.1-fold increase in DNA content, whereas chitinase treated IMS0267 and IMS0386 showed 2.0- and 2.2-fold increased DNA contents, respectively, relative to the haploid reference CEN.PK113-7D (Fig. 6.2C). To exclude the possibility of transformation associated selection of unicellular mutants, we confirmed that re-exchanging the ACE2 WT allele introduced in IMI220 and IMI221 by ace2-1 [IMW064 (ace2-1/ace2-1) and IMW066 (ace2-1/ace2-2)] restored formation of large clusters (Fig. S4). Because the introduction of the ace2-1 or ace2-2 alleles in a haploid strain was not sufficient to reconstruct the multicellular phenotype observed in the evolved strains, we investigated the impact of the change in ploidy of the evolved strains on the multicellular phenotype. To this end, the MATα strain IMI246 (ace2-2-KanMX) was constructed by replacing ACE2 in CEN.PK113-13D and crossed with the MATa strain IMI197 (ace2-2-HphNT1). The resulting diploid strain IMD014 (ace2-2-KanMX/ace2-2-HphNT1) formed large multicellular clusters (Fig. 6.4) and exhibited a sedimentation index similar to that of the evolved strains IMS0267 and IMS0386 (Fig. 6.4). Similarly, the homozygous diploid strains IMD015 (ace2-1-KanMX/ace2-1-HphNT1) and IMD017 (ace2::loxP-HphNT1-loxP/ace2::loxP-KanMX-loxP), as well as the heterozygous diploid strain IMD016 (ace2-2-KanMX/ace2-1-HphNT1) exhibited a multicellular, fast-sedimenting phenotype comparable to that of the two evolved strains IMS0267 and IMS0386 (Fig. S5). These results demonstrate complete reverse engineering of an evolved multicellular, fast sedimenting phenotype by introduction, in diploid *S. cerevisiae*, of specific recessive mutations in ACE2 that drastically reduce or eliminate transcriptional activation of Ace2 target genes. Consistent with the ploidy-dependent phenotype of ace2 null
mutants, deletion of the Ace2 target gene **CTS1** in a haploid strain background did not result in the multicellular phenotype observed in diploid **cts1Δ/cts1Δ** strain (Fig. S3).

Figure 6.1: Sequential batch cultivation in bioreactors on glucose-galactose mixtures resulted in evolution of multicellular *S. cerevisiae* - A- Maximum specific growth rate (**µ**\text{max}) estimated from CO\textsubscript{2} production during glucose consumption in the glucose-galactose batch cultures (●) and the **µ**\text{max} on galactose estimated from galactose batch cultures (○) in evolution experiment 1. Culture samples were taken at different stages of the evolution experiment, grown to stationary phase in shake flasks containing YP medium with 20 g l\textsuperscript{-1} glucose and were left to settle for 30 min in a 1 ml cuvette. Sedimentation indices (■) represent the difference in OD\textsubscript{660} within 30 min. The data represent the average and the mean deviation of duplicate experiments. B- Microscopic pictures of evolution line 1 after 0, C-1196 h, D-2105 h, E-3209 h and F-4200 h of evolution. G- Sedimentation of the reference strain CEN.PK113-7D and a culture sample of evolution line 1 and 2 at 4200 and 2877 hours respectively, photographed after 5 min of static incubation.

6.3 Discussion

This study provides the first complete identification of a molecular mechanism by which the unicellular eukaryote *S. cerevisiae* can evolve into a multicellular, fast-sedimenting phenotype. Considering the impact of multicellularity in evolution, the molecular events underlying the transformation of unicellular yeast to multicellular clusters were surprisingly simple, requiring only a mutation in a single gene and a whole genome duplication. The recessive characteristic of the **ace2-1** and **ace2-2** mutations strongly suggests that they preceded or even facilitated the origin of the genome duplication event that occurred during laboratory evolution of strains IMS0267 and IMS0386. Although generation of multicellular clusters is easily observable, numerous shake flask-based laboratory evolution studies with *S. cerevisiae* strains, including the strain used in our study, do not report this phenotype [95, 105, 193]. The fast and reproducible selection of multicellular mutants in the present study was, in all likelihood, a consequence of the design of the effluent-removal system in our bioreactor setups. We thereby inadvertently mimicked the experimental design of Ratcliff and coworkers [184] who intentionally selected for a fast-sedimenting snowflake phenotype by including a biomass settling phase in their serial-batch laboratory evolution experiments.

The accelerated diauxic consumption of glucose-galactose mixtures (Fig. 6.1A and Fig. S1) by the evolved cultures cannot be completely attributed to the mutations that
6.3 Discussion

Figure 6.2: Ploidy of the evolved mutants IMS0267 and IMS0386. - Ploidy of the evolved mutants IMS0267 and IMS0386. A Prediction of DNA content in the evolved strains S. cerevisiae IMS0267 (Upper) and IMS0386 (Lower), using the Magnolya algorithm (162). The numbers indicate chromosome position. + (red) indicates the ploidy of the ancestral genome (strain CEN.PK113-7D) and x (blue) indicates the ploidy of the evolved genome. B Determination of cell size (white bar) and DNA content measurements (black bar) of strains CEN.PK113-7D (MATa), CEN.PK122 (MATa/MATa), IMS0386, IMS0267, IMI220 (ACE2/ace2-1-HphNT1), and IMI221 (ACE2/ace2-2-HphNT1) by ow cytometry. Strains IMI220 and IMI221 are unicellular strains derived from IMS0267 and IMS0386 by reintroduction of a WT ACE2 allele. *For IMS0386 and IMS0267 the analysis was preceded by treatment with Trichoderma viride chitinase. Data are presented as average ± mean deviation of duplicate biological replicates.

caused multicellularity (Fig. S6), suggesting that additional mutations contributed to this characteristic. Analysis of several of these mutations, which is outside the scope of this study, was complicated by their heterozygous nature.

The observed ploidy dependency of the phenotype caused by the ace2 alleles identified in the evolved strains is probably at least partly due to the different bud-site selection preferences of haploid and diploid S. cerevisiae strains (32, 35). Haploid cells exhibit axial budding, during which a new bud is formed directly adjacent to the bud scar. Conversely, diploid cells exhibit a polar budding pattern, in which daughter cells bud distally (32). Different bud-site selection strategies will inevitably affect the morphology of multicellular aggregates in mutants with compromised cell division. For example, polar budding should result in less steric hindrance, thereby facilitating generation of larger structures, consistent with the larger size of multicellular clusters in diploid ace2/ace2 strains. Additionally, ploidy may affect separation of mother and daughter cells even in unicellular strains. Of a set of only 17 S. cerevisiae genes whose expression is affected by ploidy (73), two (CTS1 and DSE4, of which only the endo-chitinaseencoding CTS1 gene is a known Ace2 target) are associated with mother-daughter cell separation. The strong positive correlation of ploidy and CTS1 gene expression suggests that, in diploid cells, separation of mother and daughter cells requires more endo-chitinase than in haploids. This assumption would be consistent with the observed stronger phenotype of reduced CTS1 expression in diploids.

The strong ploidy dependence of ace2 phenotypes underlines the importance of analyzing whole or partial genome duplication in the analysis of evolved strains (36, 122, 255). In addition to facilitating the identification of key mutations, research on genome duplication and subsequent further evolution in laboratory experiments may lead to further insight.
Figure 6.3: Effect of mutations in *ACE2* on gene expression and multicellularity

- (A) Quantification of the expression of characterized Ace2 regulated genes (*CTS1*, *SCW11*, *DSE1* and *DSE2*) in the strains CEN.PK113-7D (black bar; *ACE2*) IMS0267 (white bar; *ace2-1/ace2-1*) and IMS0386 (grey bar; *ace2-2/ace2-2*). Samples were taken in midexponential phase from a shake flask culture grown on YPD medium. Relative gene expression data represent the expression of *CTS1*, *SCW11*, *DSE1* and *DSE2* normalized to *ACT1*. The expression ratios were further normalized relative to CEN.PK113-7D. The data represented are average ± mean deviation of duplicate biological replicates. (B) Calcofluor White staining of an IMS0267 multicellular cluster. This picture is representative for the entire culture as well as for the two other singlecolony isolates obtained from evolved hyper-sedimenting cultures. Microscopic observations of a multicellular cluster of IMS0386 resuspended in 100 mM potassium phosphate buffer (C) before and (D) after 7h incubation with 60 units of chitinase at 25 °C.
6.3 Discussion

Figure 6.4: Reverse engineering of the multicellular phenotype - (A) cellular morphology of strains A- CEN.PK113-7D (Mata ACE2), B- IMK395 (Mata ace2Δ::loxP-HphNT1-loxP), C- IMI197 (Mata ace2-1 loxP-HphNT1-loxP), D- IMI246 (ace2-2), E- CEN.PK122 (Mata/Mata ACE2/ACE2), F- IMD014 (Mata/Mata ace2-2- loxP-HphNT1-loxP/ace2-2-loxP-KanMX-loxP), G- IMS0267 (ace2-2/ace2-1), H- IMI220* (ACE2/ace2-1- loxP-HphNT1-loxP), I- IMS0386 (ace2-2/ace2-2), J- IMI221# (ACE2/ace2-2- loxP-HphNT1-loxP). K- Sedimentation indices of the reference haploid strain CEN.PK113-7D, of the diploid reference CEN.PK122 (Mata/Mata), the evolved multicellular fast-sedimenting strains IMS0267 and IMS0386 and the reverse engineered mutants IMK395, IMI197, IMI246, IMD014, IMI320* and IMI221#. The sedimentation index indicates the fraction of cells that sediment in a cuvette after 30 min. The data represented are average ± mean deviation of duplicate biological replicates. * denotes strains constructed in the IMS0267 background, # denotes strains constructed in the IMS0386 strain background.
in the evolutionary past of *S. cerevisiae*, in which a whole genome duplication played a pivotal role \(251\).

Lack of degradation of the chitin septum between the mother and the daughter cells appears to be the predominant mechanism underlying the formation of the multicellular clusters observed in the present study. This mechanism may have played a role in the transition from unicellular fungi to dimorphic and filamentous organisms, because these organisms share a conserved role for chitin in cell wall architecture. Inactivation of the ACE2 ortholog in the pathogenic yeast *Candida glabrata* led to cell clusters and hypervirulence in a murine model \(55, 56\). Similarly, *C. albicans* strains with an *ace2Δ/Δ* genotype showed altered separation and morphology and, moreover, resistance to azole antifungal drugs \(147\). However, outbreaks of hypervirulent and/or antibiotic-resistant mutants of these pathogens have hitherto not been reported.

Although mutations in the endo-chitinase-encoding *CTS1* gene and/or in other components of the regulation of Ace2 and morphogenesis (RAM) pathway can be expected to have similar impacts on sedimentation characteristics, only mutations in *ACE2* were found in two independent evolution experiments. Moreover, the *ace2-1* and *ace2-2* mutations occurred in the same homopolymer of eight adenosine residues (Fig. S2). Poly-(dA:dT) tracts occur frequently in *S. cerevisiae* genome \(146\)(57, 58), and these regions may participate in the yeast genome evolution by creating mutagenesis hot-spots \(52\). Poly-(dA:dT) tracts are, however, less abundant in coding regions than in intergenic regions (Table S2), presumably because a resulting evolvability confers a selective disadvantage in most protein-encoding DNA. In contrast, acquisition of a fast-sedimentation phenotype may offer selective advantages in nutrient-rich environments where single cells are easily washed away, such as flowers or fruits subjected to frequent bursts of intensive rainfall. Close inspection of the nucleotide sequences of *Candida ACE2* orthologs, and *S. cerevisiae* genes of the RAM pathway did not reveal homopolymers longer than five residues. In pathogenic *Candida* strains, this might limit the frequency with which hypervirulence occurs as a consequence of loss of function mutations in *ACE2*.

Knowledge of the mutations responsible for a multicellular, fast-sedimenting phenotype in *S. cerevisiae* allows modulation of this property by genetic engineering. The results presented in this study indicate that stable, fast-sedimenting yeast strains for use in cell retention systems can be constructed by inactivation of both copies of *ACE2* in diploid strains. Formation of multicellular clusters, as observed in the evolved strains investigated in this study, does not hinder cell growth. In fact, the evolved strains IMS0267 and IMS0386 showed higher growth rates than their ancestor CEN.PK113-7D in chemically defined medium with glucose and galactose (Fig. 6.1 and Fig. S1). Additionally, it may be possible to prevent or delay occurrence of multicellular phenotypes in adaptive evolution experiments, where it is not always a desirable feature, by ectopic integration of multiple *ACE2* genes.
6.4 Material and Methods

Strain Maintenance  S. cerevisiae strains used in this study (Table 6.2) were derived from the CEN.PK family (64) and from the BY lineage (26). Strains were maintained on YP medium [demineralized water; 10 g·L⁻¹ Yeast extract (BD Difco); 20 g·L⁻¹ Peptone (BD Difco)] with 20 g·L⁻¹ glucose (Dextrose) (YPD). Culture stocks were prepared from shake flask cultures, which were incubated at 30 °C and shaken at 200 rpm, by the addition of 20% (vol/vol) glycerol and were stored at -80 °C.

Laboratory Evolution of CEN.PK113-7D and Batch Cultivations  Long-term cultivation in sequential batch reactors was the method used to improve the anaerobic growth characteristics of CEN.PK113-7D in a mixture of 20 g·L⁻¹ glucose and 20 g·L⁻¹ galactose. Bioreactors were inoculated by adding a shake flask culture that had been grown overnight on synthetic medium (SM) [5 g·L⁻¹ (NH₄)₂SO₄, 3 g·L⁻¹ KH₂PO₄, 0.5 g·L⁻¹ MgSO₄·0.7H₂O, trace elements, and vitamins as described in ref. 61], and 20 g·L⁻¹ glucose at 30 °C. An alternating batch regime was conducted with every first batch containing 20 g·L⁻¹ glucose and 20 g·L⁻¹ galactose medium and every second batch containing 20 g·L⁻¹ galactose as the sole carbon source in the medium. The cycles on galactose-only medium were included to balance the number of generations of growth on the two sugars.

The strains CEN.PK113-7D, CEN.PK122, IMS0267, IMS0386, and IMD014 were compared with respect to fermentation time by batch cultivation in bioreactors. Bioreactors containing SM with 20 g·L⁻¹ glucose and 20 g·L⁻¹ galactose were inoculated by adding a shake flask culture that had been incubated overnight in synthetic medium and 20 g·L⁻¹ galactose at 30 °C. Cultivation was carried out in 2 L laboratory bioreactors (Applikon) with a working volume of 1 L. SM supplemented with 0.01 g·L⁻¹ ergosterol and 0.42 g·L⁻¹ Tween 80 dissolved in ethanol and trace elements was used as the medium to which either 20 g·L⁻¹ glucose and 20 g·L⁻¹ galactose or only 20 g·L⁻¹ galactose was added. Antifoam Emulsion C (Sigma-Aldrich) was autoclaved separately (120 °C) as a 20% (wt/vol) solution and added to a final concentration of 0.2 g·L⁻¹. Cultures were stirred at 800 rpm, cultures were kept anaerobic by sparging 0.5 L·min⁻¹ nitrogen gas (<10 ppm oxygen), and culture pH was kept at 5 by automatically adding 2 M KOH. The bioreactor was equipped with Norprene tubing (Cole Palmer Instrument Company) to minimize oxygen diffusion. The bioreactor was automatically drained when off-gas CO₂ levels dropped below 0.05% after the CO₂ production peak, leaving 25 (evolution 1) or 5 mL (evolution 2) as inoculum for the next batch. The bioreactor was filled to 1 L using a feed pump controlled by an electric level sensor. For each cycle, the specific growth rate on either glucose or galactose was estimated from the off-gas CO₂ production in the exponential phase by fitting an exponential function through the data points. The number of generations was estimated to range from 3.5 to 5 per batch culture based on dry-weight measurements. The culture was regularly checked for purity by plating on lithium-containing agar plates (62) and by microscopy. Culture samples were stored by the addition of 20% (vol/vol) glycerol and kept at -80 °C.
Single Colony Isolation. Representative single colony isolates from the end of both evolution experiments were obtained by biomass samples on YP medium with 20 g·L⁻¹ galactose. Single colonies were restreaked twice before inoculating a 15-mL plastic tube containing 1 mL synthetic medium supplemented with vitamins, trace elements, and 20 g·L⁻¹ galactose. After incubation for 1 d at 30 °C, these cultures were used to inoculate shake flasks containing 100 mL of the same medium. Fully grown cultures of these shake flasks were stocked. The mutant with the highest sedimentation indices from evolution experiments 1 and 2 were named IMS0267 and IMS0386, respectively.

Calcofluor White Staining. Two hundred microliters of a fully grown shake flask culture on YPD medium was washed thrice in PBS buffer (3.3 mM NaH₂PO₄, 6.7 mM Na₂HPO₄, 0.2 mM EDTA, and 130 mM NaCl) and resuspended in 500 µL PBS buffer. One hundred microliters of the resulting suspension was incubated with 10 µL Calcofluor White stain (Calcofluor White M2R 1 g·L⁻¹ and Evans Blue 0.5 g·L⁻¹; Fluka). After 15 min, the cell suspension was washed once more. Directly thereafter, phase-contrast and fluorescence microscopy was performed with a Zeiss Imager.D1 microscope equipped with a 40× Plan Neofluor lens and Filter Set 01 (excitation bandpass filter width from 353 to 377 nm, emission long-pass filter from 397 nm, 395-nm beam splitter filter; Carl Zeiss). Images were taken with a Zeiss Axiocam MRC using Axiovision 4.5 software.

Chitinase Assay. One hundred microliters of an overnight shake flask culture on YPD medium was centrifuged and resuspended in either 100 µL 100 mM potassium phosphate buffer (KPB) (pH 6.0) (13.2 mM KH₂PO₄, 86.8 mM K₂HPO₄) or 100 µL 100 mM KPB buffer (pH 6.0) with 1 mg ml⁻¹ chitinase [chitinase from Trichoderma viride, >600 units mg⁻¹ (Sigma-Aldrich)].

Sedimentation Assay. To visualize sedimentation in test tubes, yeast cells were harvested from fully grown shake flask cultures on YPD medium, washed twice, and resuspended in SM to a biomass concentration of 2 g dry weight·L⁻¹. After vortexing thoroughly to ensure a homogeneous suspension, samples were rapidly placed in test tubes, and the clock was immediately started. Photographs were taken after 5 min. To quantify the rate of sedimentation, shake flask cultures were grown to stationary phase in YPD medium, washed twice, and resuspended in SM to a biomass concentration of 0.42 g dry weight·L⁻¹. The cell suspension was left to settle in a 1-mL cuvette for 30 min while OD660 was continuously recorded using a Hitachi U-3010 spectrophotometer (Hitachi High-Technologies Europe). The sedimentation index was defined as the ratio of the decrease of OD660 during the 30-min incubation period and the initial OD660 value (ΔOD660/OD660, initial).

Whole Genome Sequencing. Genomic DNA from the two evolved strains and CEN.PK113-7D was isolated using the Qiagen 100/G kit (Qiagen). A library of 200-bp genomic fragments was created and paired-end (50-bp reads) sequencing was performed with an Illumina HiSeq 2000 sequencer at Baseclear BV. The individual reads were mapped
onto the reference genome of CEN.PK113-7D (163). Single-nucleotide variations, small
insertions, and deletions were extracted from the mapping under the assumption that the
analyzed genome was diploid. Default settings were used, except that the minimum and
maximum read depths were set to 10× and 400×, respectively. To minimize false-positive
mutation calls, custom scripts and manual curation were used for further mutation filtering.
First, mutation calls that contained ambiguous bases in either reference or mapping
consensus were filtered out. Second, only single nucleotide variations with a quality of at
least 20 and small insertions and deletions with a quality of at least 60 were kept. Variant
quality was defined as the Phred-scaled probability that the mutation call is incorrect.
Third, mutations with a depth of coverage smaller than 10× were discarded. All varia-
tions were manually verified by comparing with raw sequencing data of CEN.PK113-7D.
The Magnolya algorithm (162) was used to analyze copy number variation, using Newbler
(454 Life Sciences) for the coassembly. Haploid settings were used for CEN.PK113-7D
diploid settings for the evolved strains to determine their ploidy levels. The raw se-
quencing data were deposited at the NCBI Sequence Read Archive under BIORproject ID
PRJNA193417.

**Flow Cytometric Analysis.** Cell volumes and the DNA contents of the evolved iso-
lates and a haploid and a diploid reference strain (CEN.PK113-7D and CEN.PK122,
respectively) were analyzed by flow cytometry. A culture volume corresponding to 1 ×
107 cells·mL⁻¹, determined with a Z2 Coulter Particle Count and Size Analyzer (Beck-
man Coulter), was centrifuged (5 min, 3,425 × g). The pellet was washed once with
phosphate buffer (NaH₂PO₄ 3.3 mM, Na₂HPO₄ 6.7 mM, NaCl 130 mM, and EDTA 0.2
mM) and resuspended in phosphate buffer. Cells were briefly sonicated (~3 s) in an MSE
Soniprep 150 sonicator (150-W output, 7-µm peak-to-peak amplitude; MSE) to prevent
cell aggregation. For analysis of evolved strains IMS0267 and IMS0386, cell suspensions
were centrifuged and resuspended in 50 mM potassium phosphate buffer (pH 6.0) with
1 mg·mL⁻¹ Trichoderma viride chitinase (Sigma-Aldrich) and incubated at 30 °C for at
least 60 min to disperse cell clusters. After centrifuging (15 min, 1,700 × g), the pellet
was washed once in 100 mM potassium phosphate buffer and finally culture samples were
resuspended in diluted in IsotonII diluent (Beckman Coulter) to a cell density of ~107
mL⁻¹. Cellular DNA was then stained with the Vybrant DyeCycle Orange Stain Kit (In-
vitrogen) and incubated in the dark for 30 min at 37 °C. Stained and unstained samples
were analyzed on a Cell Lab Quanta SC MPL flow cytometer equipped with a 488-nm
laser (Beckman Coulter). Quantification of the fluorescence intensity (DNA content) and
electronic volume (EV, as a measure for cell volume) was performed by using the free
CyFlogic software (version 1.2.1; CyFlo Ltd.).

**Quantitative PCR.** Transcript levels of Ace2 targets in CEN.PK113-7D, IMS0267,
and IMS0386 were determined in duplicate shake flask cultures grown on YPD medium
to midexponential phase, when the culture was cooled on ice, and 20 mL of broth was
harvested by centrifugation. Total RNA extraction was based on a method described
previously (46). Cells were centrifuged and resuspended in one pellet volume of TAE buffer, two pellet volumes of acid phenol-chloroform (5:1, pH 4.5), and 0.1 pellet volume 10% (wt/vol) SDS. The tubes were placed in a water bath at 65 °C for 5 min before being aliquoted in three 1-mL tubes and stored at -80 °C. RNA was extraction as described by Schmitt et al. (204). cDNA was synthesized using the QuantiTect Reverse Transcription Kit (Qiagen). The QuantiTect SYBR Green PCR Kit (Qiagen) was used for quantitative PCR, performed in triplicate and at two dilutions in the Rotor-Gene Q (Qiagen). A primer concentration of 0.5 µM in a total reaction volume of 20 µL was used. All quantitative PCR (qPCR) primers are listed in Table S3. Expression of each transcript relative to the expression in CEN.PK113-7D and normalized to the transcript level of ACT1 was calculated using the program REST (Qiagen) by entering take-off and amplification values. A 100% efficient reaction would give an amplification value of 2 for every sample, meaning that the amplicon doubled in every cycle. The actual amplification of the reactions was similar with that obtained using primers for actin ACT1 (1.65-1.9). Outliers (<1.65) were manually removed. The takeoff represents the cycle at which the second derivative is at 20% of the maximum level, indicating the end of the noise and the transition to the exponential phase. The take-off value was calculated for each gene of interest by the Rotor-Gene Q Series Software (Qiagen). Average relative transcript levels were determined from two to four technical replicates. Results presented are averages of at least two biological replicates.

**Strain Construction.** The protocol described by Gietz and Woods (78) was used to transformation linear DNA fragments into *S. cerevisiae* strains. Transformants were selected on YPD agar plates containing 200 mg·L⁻¹ hygromycin B or 200 mg·L⁻¹ G418. Transformants were restreaked once before they were confirmed to have the correct integration by PCR (Table S3) on colony material suspended in 0.02 M NaOH and boiled for 10 min. To confirm the presence of the correct allele(s), single read (Sanger) sequencing was performed on selected PCR products by Baseclear BV on an ABI3730XL sequencer (Life Technologies Ltd.).

Disruption of *ACE2* in CEN.PK113-7D was done by integrating the ACE2KO construct, which was amplified by PCR from the plasmid pUGhphNT1 (50) with primers ACE2KOf and ACE2KOr. Correct replacement of the ACE2 gene by the hygromycin B resistance gene was confirmed by PCR with primers sets ACE2fw-Hph NT1 fw, ACE2rv-Hph NT1, and ACE2fw-ACE2rv. The resulting strain was named IMK395 (*ace2Δ::loxP-HphNT1-loxP*). Introduction of the WT *ACE2* allele (resulting in IMI196), the *ace2-1* allele (resulting in IMK245), and the *ace2-2* allele (resulting in IMI197) into CEN.PK113-7D or introduction of the *ace2-2* allele (resulting in IMI246) and of the *ace2-1* allele (resulting in IMK484) in CEN.PK113-13D was done by cotransformation of two overlapping DNA fragments that recombine with each other and integrate side-by-side into the same chromosomal locus (Fig. S7A). The first fragment contained either the WT *ACE2* allele or an *ace2*, flanked by a unique overlapping sequence with the second fragment. This first construct was obtained by PCR on genomic DNA of CEN.PK113-7D or on genomic DNA
of IMS0386 using primers ACE2idf and ACE2tagA. For IMS0267, the first construct was amplified from genomic DNA of IMS0267 using primers ACE2idf and ACE2tagB. The second fragment also contained the unique sequence, together with the hygromycin B or kanamycin resistance gene and a sequence homologous to a sequence 204 bp downstream of \textit{ACE2} (Fig. S7). This second construct was obtained by PCR on the plasmid pUG-hphNT1 using primers tagApUG and pUGACE2r or by a PCR on pUG6 using primers tagBpUG and pUGACE2r. After integration of the two constructs in the CEN.PK113-7D genome, correct insertion of the constructs was confirmed by PCR using primers pairs ACE2seqf-Hph NT1 rv or ACE2seqf-KanA, ACE2hygidrv-Hph NT1 fw or ACE2hygidrv-KanB, and ACE2seqf-ACE2hygidrv. By sequencing the PCR product obtained from the primer pair ACE2seqf-Hph NT1 rv or ACE2seqf-KanA, the insertion of the correct allele was confirmed using the primer ACE2seqf.

Because the introduction of two genetic elements into the multicellular mutants proved more difficult than in the unicellular ancestor, allele switching in these mutants was done by integrating one complete construct into the \textit{ACE2} locus (Fig. S7B). The construct was obtained by amplifying the complete ACE2-tagA-HphNT1-ACE2 construct from genomic DNA of the appropriate mutants constructed in CEN.PK113-7D by PCR with primers ACE2seqf and ACE2hygidrv. After integration of those constructs in IMS0267 (resulting in IMI220) and IMS0386 (resulting in IMI221), correct insertion of the construct was confirmed by PCR using primer pairs ACE2f-Hph NT1 rv, ACE2TARcheck-Hph NT1 fw, and ACE2f-ACE2TARcheck. By sequencing the PCR product obtained from the primer pair ACE2f-Hph NT1 rv and by sequencing the smaller PCR product from the primer pair ACE2f-ACE2TARcheck using the primer ACE2seqf, presence of the expected alleles was confirmed.

Construction of a diploid \textit{ace2-2/ace2-2} mutant (IMD014) was done by crossing strain IMI197 and strain IMI246 on YPD agar plates. The resulting diploid strain was selected on synthetic agar medium with 200 mg-L$^{-1}$ G418 and hygromycin by restreaking twice on this medium. Correct insertion of the correct alleles was confirmed by sequencing the PCR product obtained from the primer pair ACE2f-Hph NT1 rv and by sequencing the PCR product obtained from the primer pair ACE2f-KanA. Similarly, the strain IMD015 was constructed by crossing IMI246 and IMK245. Reintroduction of relevant \textit{ace2} alleles into IMI220 (\textit{ace2/ace2-1}) and IMI221 (\textit{ace2/ace2-2}), resulting in strains IMW064 (\textit{ace2-1/ace2-1}) and IMW066 (\textit{ace2-2/ace2-2}), respectively, was done by integrating two overlapping constructs into the \textit{ACE2} locus, thereby replacing the \textit{ACE2}-tagA-HphNT1-ACE2 construct (Fig. S7A). The first construct contained an \textit{ace2-1} or \textit{ace2-2} allele, flanked by a unique overlapping sequence with the second construct. The first construct was obtained by PCR on genomic DNA of IMS0267 or IMS0386, using primers ACE2idf and ACE2tagB. The second construct also contained the unique sequence, together with the kanamycin resistance gene and a sequence homologous to a sequence 204 bp downstream of \textit{ACE2}. This second construct was obtained by PCR on the plasmid pUG6 using primers tagBpUG and pUGACE2r. After transformation of the two constructs to the
appropriate strain, correct insertion was confirmed by PCR using primer pairs ACE2f-KanA, ACE2TARcheck-KanB, and ACE2f-ACE2TARcheck, as well as by demonstrating resistance to G418 plates but not to hygromycin. Presence of the desired alleles was confirmed by sequencing the PCR product obtained from the primer pair ACE2f-Hph NT1 rv and by sequencing the smaller PCR product from the primer pair ACE2f-ACE2TARcheck using the primer ACE2seqf. Introduction of a hygromycin resistance gene into the MAT? CEN.PK113-16B strain was done by transforming a genetic construct obtained by PCR from the plasmid pUG-hphNT1 [50] using primers MTH1markfw and MTH1markrv. The resulting strain was named IMI081 (ACE2 loxP-HphNT1-loxP). Constructs were made by PCR amplification on genomic DNA by using Expand high fidelity Polymerase (Roche) according to the manufacturers instructions in a Biometra TGradient Thermocycler (Biometra). Isolation of fragments from gel was done with the Zymoclean Gel DNA Recovery kit (Zymo Research). PCR amplification on colony material was done using FastStart Taq DNA Polymerase (Roche) according to the manufacturers instructions on colony material suspended in 0.02 M NaOH and heated for 10 min at 100 °C.

**Mating and Sporulation.** Strains IMS0267 and IMS0386 were mated with IMI081 by streaking both strains on YPD plates. After overnight incubation at 30 °C, the strains were streaked over each other. After another 4 h of incubation at 30 °C, diploids were selected by streaking on selective medium (SM medium with 20 g·L\(^{-1}\) glucose and 200 mg·L\(^{-1}\) hygromycin). Resulting single colonies were restreaked twice on the same medium. Sporulation was performed by incubating a culture in YP medium supplemented with 10 g·L\(^{-1}\) potassium acetate for 2 d at 23 °C. Subsequently, the entire culture was washed twice, resuspended in 20 g·L\(^{-1}\) potassium acetate, and incubated for 3-4 d at 23 °C. Spores were segregated on YPD plates using a micromanipulator (Singer Instruments) and incubated at 30 °C.

**Homopolymer Distribution.** The *S. cerevisiae* reference genome and its annotation (release 64-1-1, February 3, 2011) were downloaded from the Saccharomyces Genome Database (www.yeastgenome.org/) [63]. A file ”domains.tab”, containing domains predicted using InterProScan [157], was downloaded from the same site (March 10, 2013). The number of occurrences of dA:dT homopolymers of eight or more residues was counted in the overall genome, in genes (i.e., sequences annotated as gene in the reference genome), in coding sequences within genes, in introns, and in domains. Homopolymers were considered present when all bases fell inside the genomic feature. For each of these features, a Fisher exact test (two-tailed) was then performed under the null hypothesis that the occurrence of homopolymeric stretches is independent of the underlying genomic feature (genes, coding sequences, introns, and domains).

### 6.5 Supplemental data

Supplemental data and figures are available at http://www.pnas.org/.
### Table 6.2: Strains used in this study.

* denotes strains constructed in the IMS0267 and # strains constructed in the IMS0386 strain backgrounds.

<table>
<thead>
<tr>
<th>Strain Description and Genotype Source</th>
<th>Description and Genotype Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEN.PK113-7D</td>
<td>Mata ACE2</td>
</tr>
<tr>
<td>CEN.PK113-13D</td>
<td>Mata ade2-52</td>
</tr>
<tr>
<td>CEN.PK113-16D</td>
<td>Mata ACE2 leu2-3-112</td>
</tr>
<tr>
<td>CEN.PK122</td>
<td>Mata/Mata ACE2/ACE2</td>
</tr>
<tr>
<td>IMS0267</td>
<td>ace2-1 / ace2-1</td>
</tr>
<tr>
<td>IMK095</td>
<td>ace2-2-loxP-HphNT1-loxP</td>
</tr>
<tr>
<td>IMK196</td>
<td>Mata/ace2-loxP-HphNT1-loxP</td>
</tr>
<tr>
<td>IMK485</td>
<td>Mata ace2-loxP-KanMX-loxP</td>
</tr>
<tr>
<td>IMK197</td>
<td>Mata ace2-2-loxP-HphNT1-loxP</td>
</tr>
<tr>
<td>IMK484</td>
<td>Mata ace2-loxP-KanMX-loxP</td>
</tr>
<tr>
<td>IMK1246</td>
<td>Mata waa3-52 ace2-2-loxP-KanMX-loxP</td>
</tr>
<tr>
<td>IMD614</td>
<td>Mata/Mata waa3-52/URA3</td>
</tr>
<tr>
<td>IMI220</td>
<td>ace2-2-loxP-HphNT1-loxP/ace2-2-loxP-KanMX-loxP</td>
</tr>
<tr>
<td>IMW064</td>
<td>ace2-1/ace2-2-loxP-KanMX-loxP</td>
</tr>
<tr>
<td>IMI221</td>
<td>ace2-2-loxP-HphNT1-loxP</td>
</tr>
<tr>
<td>IMW066</td>
<td>ace2-1/ace2-2-loxP-KanMX-loxP</td>
</tr>
<tr>
<td>IMI081</td>
<td>Mata ace2 leu2-3-112 loxP-HphNT1-loxP</td>
</tr>
</tbody>
</table>
Exploring variation aware contig graphs for (comparative) metagenomics using Marygold

Jurgen Nijkamp, Mihai Pop, Marcel Reinders and Dick de Ridder

Published in Bioinformatics
September 20, 2013, Volume 29, Issue 22, pages 2826-2834
Abstract

Motivation: While many tools are available to study variation and its impact in single genomes, there is a lack of algorithms for finding such variation in metagenomes. This hampers the interpretation of metagenomics sequencing datasets, which are increasingly acquired in research on the (human) microbiome, in environmental studies and in the study of processes in the production of foods and beverages. Existing algorithms often depend on the use of reference genomes, which poses a problem when a metagenome of a priori unknown strain composition is studied. In this paper, we develop a method to perform reference-free detection and visual exploration of genomic variation, both within a single metagenome and between metagenomes.

Results: We present the MARYGOLD algorithm and its implementation, which efficiently detects bubble structures in contig graphs using graph decomposition. These bubbles represent variable genomic regions in closely related strains in metagenomic samples. The variation found is presented in a condensed Circos-based visualization which allows for easy exploration and interpretation of the found variation.

We validated the algorithm on two simulated datasets containing three resp. seven Escherichia coli genomes and showed that finding allelic variation in these genomes improves assemblies. Additionally, we applied MARYGOLD to public real metagenomic datasets, enabling us to find within-sample genomic variation in the metagenomes of a Kimchi fermentation process, the microbiome of a pre-mature infant and in microbial communities living on acid mine drainage. Moreover, we used MARYGOLD for between-sample variation detection and exploration, by comparing sequencing data sampled at different time points for both of these datasets.
7.1 Introduction

Recently, large-scale microbiome analyses, such as metaHIT (182) and the NIH Human Microbiome Project (99) have demonstrated the importance and the richness of our second genome: the microbes inhabiting many different sites of our bodies. To get the complete picture of all DNA present, its biological functions and the genetic variation, increasingly whole-metagenome sequencing is performed (94, 182). Large-scale analysis of variation within and between the resulting metagenomes – single nucleotide variants (SNVs), short insertions and deletions (indels) and structural variation – is essential to fully understand their genetic makeup. However, relatively few studies have thus far ventured into such analyses. Schloissnig et al. (203) combined the data obtained from stool samples in the metaHIT and HMP projects and a study comparing leanness and obesity (224). They took a read mapping approach to detect millions of SNPs and other genetic variation in bacterial metagenomes. In other recent work, Morowitz et al. (155) studied the metagenome of stool samples of a pre-mature infant using an assembly approach. By manual analysis of the assembly output they found genetic variation in Citrobacter strains.

A major reason for this lack of metagenomic variation analysis is a lack of available tools. Variation detection algorithms have mainly been developed for the isolated genomes of single individuals and can be roughly divided in two categories: reference-based and reference-free. Several reference-based methods that map reads to a pre-constructed reference genome are available to detect single nucleotide variants and structural variants, as reviewed in (160) and (152), respectively. However, in metagenomics reference genomes are usually not available for all species. Additionally, many bacterial species are not cultivable in the lab, which complicates construction of reference genomes.

Reference-free methods in contrast detect variation between sequencing samples directly, without resorting to reference genomes. Fasulo et al. (66) showed that genomic regions with polymorphisms give rise to bubbles in the fragment assembly graph and presented an algorithm to find and smooth the bubbles, which was implemented in the Celera assembler. The Cortex assembler (103) was designed to co-assemble multiple genomes and search for such bubbles in the de Bruijn graph that represent variation between the assembled genomes. Decomposition of a de Bruijn graph into biconnected components followed by bubble detection has been shown to be useful in detecting splicing variants in RNA sequencing data (191).

Similar to co-assembly of individual genomes, metagenome assembly involves simultaneous assembly of multiple genomes. The presence of two closely related genomes in a sample leads to similar bubbles. The metagenome scaffolder Bambus 2 was designed to detect these bubbles and collapse them, thereby elongating scaffolds while preserving divergence in the data (119). Bambus 2 finds bubbles by looking for subgraphs where multiple paths begin at a source node and collapse to one sink node. To keep computation tractable the number of hops considered is limited to two. Meta-IDBA similarly searches bubbles in a de Bruijn graph, but the search is limited to a maximum of 300 bp (176).

Here we set out to detect complex bubbles such as presented in Fig. 7.2a, i.e. longer than can currently be detected by Bambus 2 and Cortex, typically found when more than
two homologous sequences (closely related genomes) are present in a sample. We overcome the computational limitations in the above mentioned methods by developing an efficient way to find genome variation in assembly graphs, by introducing a graph decomposition into bi- and tri-connected components that allows us to select a subset of the nodes as source nodes for bubble finding. Identifying bubbles representing multi alleles has three applications. First, it can be used to detect sequence variation, i.e. multiple alleles, between strains within a single-sample metagenome. Second, co-assembling multiple metagenomes allows to detect variation between samples. This latter application is particularly novel and is set to become more prominent as more metagenomics datasets will become available over the coming years. Third, although not the main focus in this paper, it allows to simplify assemblies, generating longer scaffolds and contigs by collapsing bubbles into a single linear sequence instead of breaking into multiple contigs. Co-assembly has previously been employed by (61) to calculate global distances between metagenomes, but here we aim to find sequence level differences.

We demonstrate MaryGold, our algorithm to find bubbles and compare them across samples using co-assembly, by comparing metagenome samples obtained using 454 technology of different time points obtained from the stool sample of a pre-mature infant (155) and a fermentation process of the traditional Korean food Kimchi (109). We also apply the algorithm on sequencing data of microbial populations found on acid mine drainage (54), obtained using both 454 and Illumina technologies, and compare MaryGold to Bambus 2. MaryGold exports the variation detected in a format that can be visualised by Circos, allowing for a rich visual exploration and interpretation of the genomic variation in metagenomes.

7.2 Algorithm

7.2.1 Generating a contig graph

De novo assembly of metagenomes relies on the same principles as that of single genomes. It can be done either by assemblers based on de Bruijn graphs or overlap-layout-consensus (OLC) assemblers as reviewed by (139). Both of these types of assembler use graphs to merge the millions of short reads into longer contiguous sequences, with the ultimate goal of reconstructing full genomes. Although de Bruijn graph and OLC assemblers have different ways of building their graphs, i.e. by means of k-mer hashing and pairwise read alignment respectively, both paradigms produce graphs that contain similar information and have similar limitations. In both, repeats longer than the read length (or k-mer length) will cause branching in the graph and unique regions correspond to unambiguously reconstructable sequences.

The algorithm we propose uses the graphical output of assemblers to reconstruct metagenomes. The assembly graphs are assumed to have nodes representing contigs and edges representing reads spanning the connected of contigs. The graph is bidirected since it has both contigs coming from the forward strand and from the reverse strand.
7.2 Algorithm

Figure 7.1: Two genomes, their contig graph and collapsed contig graph, where the two simple bubbles have been compressed into multi-allelic contigs. Before compression the assembly would contain six contigs. After compression the contigs are connected by a simple path without branching and can be merged into a single contig.

of the DNA molecule (125). From this point onwards we refer to this graph as the contig graph \( G = (V, E) \) that has a set of nodes \( V = \{v_1, v_2, \ldots, v_n\} \) and a set of edges \( E = \{(v_i, v_j, o_1, o_2) \mid v_i, v_j \in V \land o_1, o_2 \in \{\triangleright, \triangleleft\}\} \). The set of nodes is created by assigning each contig \( c \in C \) to a unique node. A path from \( v_i \) to \( v_j \) through the bidirected contig graph is a sequence of nodes and edges \( v_i, e_{i1}, v_{i1}, e_{i2}, \ldots, v_{im}, e_{im+1}, v_j \) such that each intermediate node \( v_{il} \) in subsequence \( e_{il}, v_{il}, e_{il+1} \) has matching orientations of its incoming edge \( (e_i = (v_{il-1}, v_{il}, o_1, o_2)) \) and outgoing edge \( (e_{il+1} = (v_{il}, v_{il+1}, o'_1, o'_2)) \), i.e. \( o_2 = o'_1 \) (125).

The DNA sequence corresponding to a node in a path depends on its traversal. The reverse complement DNA sequence of a node \( v_{il} \) is used when either \( o_2 = \triangleleft \) or \( o'_1 = \triangleleft \) (Fig. 7.1 node GGTG).

7.2.2 Metagenome graph compression

In the contig graph, two contigs can be reliably concatenated, with or without a gap, if they are connected by a single edge and no other edge contradicts their concatenation. Ambiguity arises when multiple edges of a contigs link to other contigs. In metagenome assemblies where we are dealing with closely related strains, such branches in the graph can be caused by sequence divergence between two strains of the same (or closely related) species in the sample, forming bubbles (Fig. 7.1). In this work we try to elongate scaffolds and contigs by identifying these regions and collapsing them to multi-allelic contigs.

The main challenge tackled here is the detection of complex bubble structures. We define a bubble as a subgraph where all maximally extended paths within the bubble start in a single source node and end in a single sink node or vice versa (the source and sink node are interchangeable because of the two-stranded nature of DNA, i.e. they can be swapped by taking the reverse complement of all nodes in the subgraph).
7. EXPLORING VARIATION AWARE CONTIG GRAPHS

7.2.2.1 Finding separation pairs using graph decomposition

In Bambus 2 source-sink pairs are iteratively found by traversing all paths leaving a node and checking if these paths all collapse to a sink node within a certain number of hops. The number of hops is limited to two to keep this search tractable. Here we set out to find source-sink pairs efficiently so that we can detect source and sink nodes that are more than two hops apart from each other in one or more of the connecting paths. This allows us to detect more complex multi-allelic sites. We achieve this by finding separation pairs, i.e. pairs of nodes that increase the number of connected components in a graph when they are removed. All source and sink pairs are a separation pair (Theorem 1) though the converse does not hold. We can therefore first find all separation pairs, and then select the subset of separation pairs that are also source-sink pairs.

Theorem 1. A separation pair is a source-sink pair of nodes forming a bubble if and only if all paths in the corresponding subgraph starting in the source node end in the sink node when extended, and all paths starting in the sink node end in the source node when extended.

Separation pairs are detected in an undirected version of the contig graph using graph decomposition methods. An initial set of separation pairs is found by decomposing each connected component into its biconnected components, which can be done in $O(|V| + |E|)$. A biconnected component is a maximal biconnected subgraph. The cut vertices of a graph are the nodes whose removal increase the number of connected components. If a biconnected component has exactly two such cut vertices, then these two are stored as separation pair. The set of separation pairs is extended by decomposing each biconnected component into its triconnected components. The triconnected components of a biconnected component describe the 2-vertex cuts in a graph. A 2-vertex cut is a pair if nodes that, if removed, increases the number of connected components, i.e. a separation pair.

Battista and Tamassia (14) developed an algorithm to decompose a biconnected undirected graph into its triconnected components by building an SPQR tree, which was later improved and implemented in time $O(|V| + |E|)$ by Gutwenger and Mutzel (89). The SPQR tree represents all triconnected components of a biconnected graph from which all 2-vertex cuts can be extracted. Fig. 7.2 shows such an SPQR tree for a toy graph example; an elaborate description of how to construct the SPQR tree can be found in Weiskircher (241). The pairs of 1-vertex cuts obtained from the decomposition in biconnected components and the 2-vertex cuts obtained from the triconnected components form the set of separation pairs. These separation pairs are next validated to be source-sink pairs using the bubble search algorithm.

7.2.2.2 Validating separation pairs as source-sink pairs

The separation pairs are only source and sink pairs when all maximally extended paths inside the bubble originate from the source node and end in the sink node or vice versa. For example, the pair {1,6} in the biconnected component in Fig. 7.2a is a separation pair, because when it is removed it increases the number of connected components, but
Figure 7.2: (a) Toy example of a biconnected contig graph. Separation pairs are hatched, but only valid source-sink pairs are hatched vertically, otherwise horizontally. (b) The corresponding SPQR tree describing the triconnected components. Virtual edges are indicated by dashed lines; branches in the SPQR tree link shared virtual edges. By merging these edges the original graph can be reconstructed. A detailed description on how to construct an SPQR tree can be found in Weiskircher (241).

It does not form a valid source-sink pair of a bubble. The only valid source-sink pairs in Fig. 7.2a are \{1,5\}, \{6,9\} and \{10,11\}.

**Theorem 2.** Node \(v_{i_1}\) is part of a bubble if and only if all paths starting at \(v_{i_1}\) with a certain orientation \(o_1\) end in the source node when extended and all paths starting with the opposite orientation end in the sink node when extended.

Valid bubbles are found with a search based on breadth-first search (BFS) in the bidirected contig graph, starting from a random node in the list of nodes that form separation pairs. Since a bubble can be present both upstream and downstream of the contig, the search is performed in two directions independently. In a normal BFS all outgoing edges of a node \(v\) that has been visited are queued for visiting. In our bubble search, edges of a node \(v_{i_1}\) with orientation \(o_1'\) are only queued when all its edges for which \(o_2 = o_1'\) have been visited, since theorem 2 states that if there is no path from the source leading to one of these not-visited edges of \(v_{i_1}\) a valid bubble will never be present and there is no need for further exploration (Fig. 7.3).

**Theorem 3.** A contig can be the source of at most one bubble and the sink of at most one other bubble.

To limit computation time, we try to minimise the number of bubble searches that we have to do by reducing the set of separation pairs. When a bubble is successfully found, all separation pairs that have at least one node in this bubble are removed from the set. Only the sink node can still be used to start the bubble search algorithm, if it forms a separation pair with another node outside the established bubble (theorem 3). If we failed to find a bubble from a source node, then all separation pairs that contain the source node are removed from the set.
7. EXPLORING VARIATION AWARE CONTIG GRAPHS

Figure 7.3: The bubble search algorithm. Solid edges indicate visited edges. (a) The search is started from node 1, whereby the queue $Q$ holding the edges to-be-visited is initialised with all edges with a given orientation; $o'_1 = \triangleright$ in this case. (b) Edge $(1, 2, \triangleright, \triangleleft)$ is visited and it is observed that all edges incident to node 2 with $o_2 = \triangleleft$ have been visited. Therefore, all edges of node 2 with $o'_2 = \triangleleft$ are queued for visiting. (c) Edge $(1, 3, \triangleright, \triangleright)$ is visited, but in this case not all edges incident to node 3 with $o_2 = \triangleright$ have been visited yet. Therefore, the edges of node 3 are not queued. (d) The next edge $(2, 4, \triangleleft, \triangleleft)$ in the queue is traversed to target node 4. Again, not all edges incident to node 4 with $o_2 = \triangleleft$ have been visited, therefore no edges are added to $Q$. The queue is now empty and the algorithm terminates: node 1 in forward orientation is not a valid source node.

7.2.3 Assembly simplification by collapsing multi-allele bubbles

Scaffold and assembly graphs can be simplified by collapsing the bubbles into supercontigs. The inner contigs of the bubble are replaced in the contig graph by a single super node (Fig 7.1). All edges to and from the inner nodes are removed and replaced by two new edges: from source to supernode and from supernode to sink. Subsequently a so-called unitigging step is performed, which is a search for simple paths without branches. The contigs in the simple paths are then combined into a single supercontig.

These supercontigs represent different alleles between the source and sink node present in the genomes in sample. Any path from source to sink through this supercontig spells a valid assembly (158) and is a potential allele in one of the genomes. The native output of MARYGOLD is a graph describing the multi-allele. Nonetheless, downstream processing is facilitated by sequence-based tools, as there is also a sequence output, which is the highest depth path in the multi-allele (with ties broken arbitrarily).

7.2.4 Comparative metagenomics: Genomic variation detection between metagenomes by co-assembly

The MARYGOLD algorithm enables variation detection between metagenome sequencing samples through co-assembly. By keeping track of the sample origin of the reads during the assembly phase, a read depth per contig per sample can be calculated. Bubbles are then detected and the potential alleles are enumerated by finding all maximal paths form source to sink. The number of alleles corresponds to the number of paths from source to sink, although in reality there might be more paths possible than actual alleles present in genomes.

The alleles (and their carrying organisms) are present in unknown relative abundances. The sequencing depth for a single allele in a complex bubble, reflecting its relative abundance, is not trivial to find, since some of the nodes will be shared by two or more paths, e.g. node 5 in Fig 7.2a is shared by three paths between nodes 10 and 11. We infer the allele depth from the read depth of the contigs in its path by minimising the difference between the sum of the depths of the paths $j$ that cross a node $i$ and its read depth $d_i$ for...
7.3 Methods

all node $i = \{1, \ldots, n\}$ using a nonnegative linear least squares approach. The residual of node $i$ is defined for paths $j = \{1, \ldots, m\}$ in a multi-allele $a$ as

$$r_i = d_i - \sum_j^m x_{ij} \beta_j$$  

(7.1)

where $X$ is the design matrix with binary regressors:

$$X_{i,j} = \begin{cases} 
1 & \text{path } j \text{ crosses node } i \\
0 & \text{otherwise}
\end{cases}$$  

(7.2)

and $\beta_a$ is an $m \times 1$ vector with the unknown path weights. The least square method then finds path weights $\hat{\beta}_a$ that minimise the sum of the squared residuals

$$S = \sum_{i=1}^n r_i^2$$  

(7.3)

under the constraint that the allele readdepths in $\hat{\beta}$ are positive.

7.2.4.1 Measure of variation between alleles

To get an impression of the degree of the genomic difference in a bubble, we use a normalised version of the Levenshtein distance (or edit distance), which summarises the sequence distances of the alleles in a bubble. The average distance $D_M$ is obtained by calculating the Levenshtein distance $lev(s,t)$ for each pair of alleles $s$ and $t$ in a multi-allele $M$ with $N$ alleles and normalising for the length of the longest allele of the pair.

$$D_M = \left( \frac{N}{2} \right)^{-1} \sum_{s.t,s\leq t} \frac{lev(s,t)}{\max(|s|,|t|)}$$  

(7.4)

7.3 Methods

Most of the MARYGOLD algorithm has been implemented in C++, with some additional Python scripts. The algorithm interfaces with the AMOS bank, where Bambus 2 and metAMOS store their assembly and scaffold information, using the AMOS library (221). The graph decomposition in bi- and tri-connected components was implemented using the Open Graph Drawing Framework (OGDF) (38). The multi-allele search algorithm has been parallelized using OpenMP (42). The non-negative least squares solver from the SciPy optimisation package was used to obtain sequencing depths for the individual alleles, which is a wrapper around Fortran code (131).

Two shotgun datasets were simulated with the 454sim v1.04 package (145) using default settings from the genomes of three reps. seven E. coli strains, to illustrate the effect on metagenome assembly of multiple closely related genomes. Four real sequencing datasets from the public domain were used to demonstrate the comparative metagenomics functionality of MARYGOLD. The first was obtained from the microbiome of a pre-mature infant.
7. EXPLORING VARIATION AWARE CONTIG GRAPHS

Table 7.1: Assembly statistics for the two simulated *E. coli* and two real metagenomics datasets.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Number of contigs</th>
<th>Number of multi-alleles</th>
<th>Number of contigs after simplification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infant microbiome</td>
<td>5686</td>
<td>154</td>
<td>5251</td>
</tr>
<tr>
<td>Kimchi fermentation</td>
<td>11466</td>
<td>133</td>
<td>11325</td>
</tr>
<tr>
<td>3 <em>E. coli</em> strains</td>
<td>7291</td>
<td>644</td>
<td>5259</td>
</tr>
<tr>
<td>7 <em>E. coli</em> strains</td>
<td>12378</td>
<td>1582</td>
<td>6156</td>
</tr>
</tbody>
</table>

Morowitz *et al.* (155) sequenced the genomic DNA of the fecal microbiome, sampled on day 10, 16, 18 and 21 of the infants life. The second dataset was obtained at 10 different time points during the 29-day fermentation process of Kimchi, the traditional Korean food (109). Both datasets are well suited for comparative metagenomics, since we can compare the metagenomes of the different time points. Finally, two datasets from the C75 location biofilms within the Richmond mine (54) were downloaded from the Sequence Read Archive to compare performance on low-coverage, but long 454 reads (SRA:SRR358990, 454 Titanium FLX, 133M bases) versus high coverage, but shorter Illumina reads (SRR359000, Illumina HiSeq2000 3.6G bases, 100 bp per read). These datasets were also used to compare the performance of Bambus 2 to MARYGOLD. Bambus’ OrientContigs program was run followed by OutputMotifs to output all multi-allelic regions found by Bambus 2.

Assemblies were performed within the metAMOS v1.1 package (222) using the Newbler v2.6 assembler (Roche, Branford, CT). Newbler was run with parameters '-a 0 -m -ace' and additionally '-large' for the Illumina data assembly to speed-up the processing of the fragment assembly graph. Assigning taxonomic attributions was performed using the Fragment Classification Package v1.0 (FCP) (174). Circular plots were generated with Circos v0.63 (124) and graph layouts were performed with Cytoscape v2.8.3 (205). The DAVID functional annotation tool (98) was applied on the genes in the Kimchi sample that overlapped with the multi-alleles to find enrichment of functions and sequence features. DAVID was run with default settings except that all predicted genes in the metagenome were used as background.

7.4 Results

7.4.1 MARYGOLD finds multi-allelic regions by detecting bubbles in the contig graph

The bubbles MARYGOLD detects are multi-allelic sites in the metagenome, meaning that for the genomic regions represented in the bubble multiple alleles are present in the sample. Morowitz *et al.* (155) resolved the bubbles in the assembly graph by manual curation. This was possible since only two *Citrobacter* strains were present, and therefore only bi-allelic bubbles had to be resolved. Our simulation experiments confirm that for a low number of strains the contig graph is rather simple and bubbles can be resolved manually (Fig.
7.4a). If more strains are present, such as shown in our simulation experiment with seven *E. coli* genomes, the graph becomes too complex for manual curation (Fig. 7.4b).

MARYGOLD detected 644 and 1582 bubbles in the simulated sequencing samples with three resp. seven strains (Table 7.1). The number of bubbles does not increase linearly with the number of strains, which is expected. If two genomes A and B differ at a locus, creating a bubble, then a third genome C that is again different at that locus will not result in an additional bubble, but just an additional path from source to sink within the bubble. Only at already present loci where A and B are identical, but C is different, the addition of genome C will result in an additional bubble.

In the pre-mature infant datasets 154 bubbles were found (Table 7.1), predominantly in *Citrobacter* spp. (Fig. 7.5) as also described in the original article (155). In the original study of the Kimchi fermentation no allele variation study was performed. Here we show that allele variation is present by detecting 133 bubbles in *Lactobacillus* spp, *Leuconostoc* spp and *Weisella* spp (Fig. 7.6), which are thought to be the key players in this vegetable fermentation process (109).

![Figure 7.4: Largest connected components of the contig graphs for two simulated sequencing datasets for a three and b seven *E. coli* genomes. Red nodes are source and sink nodes. Blue nodes are nodes within a bubble.](image)

### 7.4.2 Bubble collapsing allows studying genetic variation in genomic context

In the contig graph, contigs can be reliably concatenated into longer contiguous sequences if they are connected by a simple path without branching. The contig graph branches at multi-allelic loci in the metagenome, leading to highly fragmented assemblies. By collapsing the bubbles that were detected by MARYGOLD (Fig. 7.1) the number of contigs...
in the assembly with three *E. coli* strains could be reduced by 28% and in the assembly of seven *E. coli* strains even by over 50% (Table 7.1).

The reduction in number of contigs in the real sequencing datasets was less apparent. In true metagenomic samples also genomes that have no close homologues are present. For such genomes bubble compression does not reduce the number of contigs, since it only improves of the assembly of multiple closely related strains. Also, low sequencing depth causes low abundant strains to stay under the detection limit.

Bubbles are collapsed by choosing a single path from source to sink in the contig graph, with the highest average sequencing depth, to represent the multiple alleles, which allows the output of linear instead of fragmented sequences as generated by the assembler. The linear sequences aid further analysis, such as gene prediction and taxonomic assignments. Fig. 7.5 shows a zoomed in version of one such linear contig in the pre-mature infant sample with gene annotations for each gene that overlaps with a multi-allelic locus. The genes that differed in the *Citrobacter* strains were found to be involved in iron and copper uptake as well as immunoglobulin (Ig) binding regulators, which bind to human Ig in an non-immune manner (195).

Bubble collapsing enabled gene prediction across multi-allelic loci. The features ’Retro-tranposable element Tf2’ and ‘*E. coli* ABC transporter YbhF’ were found to be significantly overrepresented ($p = 2.2 \times 10^{-9}$ resp. $p = 6.0 \times 10^{-3}$, Bonferroni corrected) in genes in the Kimchi sample that overlapped with the multi-alleles. Retrotransposons are expected to be found in multi-alleles, since they amplify themselves in a genome and create an additional allele after each insertion. The ABC transporter protein superfamily is involved in many cellular processes, including transport of a wide range of substrates and has been linked to development of resistance to multiple drugs, including antibiotics (47).

7.4.3 Co-assembly allows variation detection between metagenomes

Co-assembly of multiple metagenomics samples followed by bubble detection using MARYGOLD allows for reference-free sequence variation detection. Subsequent visualisation of the bubbles enables visual analysis and interpretation, such as in Fig. 7.5 that shows the sequencing depths at the four days for which the microbiome of the infant was sampled. The alleles and hence their carrying strains vary in abundance over time. The bubble of the *Pantoea* spp. with the longest sequence (red in Fig. 7.5) also has a high average distance between the alleles. The sequencing depths show that the ’green’ allele is the most abundant at the second time point (day 16), whereas for the two subsequent time points this allele diminishes and the red allele becomes the most abundant one.

7.4.4 Comparing MARYGOLD to Bambus 2 on 454 and Illumina data

Both the microbiome and Kimchi data were obtained using 454/Roche sequencing technology. To study the influence of the technology used and accompanying difference in sequencing depth, we finally applied MARYGOLD on DNA sequenced using both Illumina and 454 technology which originated from biofilms at the same location within the Richmond Mine, but sampled on different dates (51). More sequence could be assembled using
7.4 Results

Figure 7.5: Overview of the multi-allele bubbles of the pre-mature infant’s microbiome. Tracks in the circular plot from outside to inside are: the multi-allele bubbles coloured by species; their average length-weighted edit-distance $D_M$ indicating variation; the inferred sequencing depths $\hat{\beta}_a$ for day 21, 18, 16 and 10 with each allele $a$ indicated by a different colour. Sequencing depths were scaled by $\log_2(\hat{\beta}_a + 1)$ to allow variable sequencing depths between genomes and their grey axis are at 0.5. Below the circle, a zoomed version of one of the supercontigs, containing eight bubbles. Top: supercontig with the variable locations marked in purple. Middle: a gene track in which genes overlapping with a multi-allele site are functionally annotated. Bottom: the contig graph.
Figure 7.6: Overview of the multi-allele bubbles of the Kimchi fermentation. Tracks in the circular plot from outside to inside are: the multi-allele bubbles coloured by species; their average length-weighted edit-distance $D_M$ indicating variation; the inferred sequencing depths $\hat{\beta}_a$ for the time points chronologically sorted from inside to outside, with each allele $a$ indicated by a different colour. Sequencing depths were scaled by $\log_2(\hat{\beta}_a + 1)$ to allow variable sequencing depths between genomes.
7.4 Results

![Figure 7.7: Comparison of the number of bubbles found by MaryGold and Bambus 2 for two acid mine drainage datasets: (a) the 454 dataset (SRR358990) and (b) the Illumina dataset (SRR359000). Indicated are the total number of contigs within the bubbles, with the number of bubbles between brackets. The overlapping section indicates the number of Bambus 2 bubbles that shared at least three contigs with a MaryGold bubble/vice versa.](image)

the Illumina data since the higher number of sequenced bases allowed lower abundance bacterial strains to rise above the detection limit. In total 16.8 Mbp (N50=5850) were assembled with the Illumina data, whereas only 6.1 Mbp (N50=1669) were assembled with the 454 data.

We also used this data to compare the variation found by MaryGold to that found by Bambus 2. Although Bambus 2 is optimised for contig graphs generated with paired-end data, i.e. scaffold graphs rather than assembly graphs, the methods could be compared since they both output their bubbles. Other assemblers also contain bubble detection algorithms, such as Velvet (258), the Celera assembler (66) and meta-IDBA (176), but the bubble finding algorithms are embedded in the assembler and it is unclear how to access these methods. Moreover, the goal of these assemblers is to smooth the assemblies by collapsing bubbles, rather than using them for variation detection.

Both methods found more bubbles using the Illumina data than using the 454 data (Fig. 7.7). Although the samples were taken at different time points, we assume that they have similar microbial complexity since the sample location in the Richmond mine was identical. We therefore attribute the increased amount of variation found to the higher sequencing depth of the Illumina dataset. MaryGold found more bubbles than Bambus 2 in both datasets (Fig. 7.7). For the Illumina and 454 datasets, 69 (73%) resp. 3 (27%) Bambus 2 bubbles overlapped by at least three contigs with a MaryGold bubble. The difference in the number of bubbles found by MaryGold and Bambus 2 can be explained by looking at the algorithmic differences. Bambus 2 finds only bubbles detectable within two hops of the source node, which are then iteratively composed into larger bubbles. Additionally, the definition of a bubble slightly differs: unlike MaryGold, Bambus 2 allows for missing edges within the bubble due to lack of coverage and requires the paths through a bubble to be roughly the same length.

Finally, to give an idea of the computational complexity of MaryGold we recorded its runtime on the contig graph with 31,191 nodes and 9,586 edges of the Illumina acid mine drainage dataset, using an Intel Xeon CPU with four cores and four threads per core. Finding the set of separation pairs using graph decomposition took \( \sim 37 \) seconds (CPU time). The subsequent bubble search algorithm took \( \sim 12.5 \) minutes on a single
7. EXPLORING VARIATION AWARE CONTIG GRAPHS

core. Using two threads the wall clock time spent by the bubble search algorithm dropped by 42%. The full analysis, from raw reads to MARYGOLD output, is mostly limited by the assembly step.

7.5 Conclusion

We presented MARYGOLD, a tool to detect and explore genomic variation within and between metagenomic sequencing samples, not reliant on reference genomes. MARYGOLD finds multi-allelic regions that reflect sequence variation and improves the contiguity of assemblies by collapsing these. As input it expects a contig graph, which makes it generic and applicable to any type of sequencing data; we demonstrated the algorithm both on 454 and Illumina data. In this work we generated the contig graph using contig links from fragment assembly string graphs, but the contig graph can be extended by adding edges derived from other linking sources such as paired-end sequencing data, de Bruijn graphs or even reference genomes. The multi-allelic regions are efficiently detected by searching for bubble structures in a contig graph, enabled by decomposition of the graph into bi- and triconnected components. Note that while MARYGOLD can detect local sequence variation, i.e. multiple alleles of a genomic locus, it ignores variation that does not result in a bubble, such as that induced by horizontal gene transfer events.

Sequence variation in metagenomes and in particular between metagenomes, i.e. comparative metagenomics, is still a largely unexplored field. Our algorithm not only finds sequence variation within a single metagenomic sample, but also between samples, and allows for convenient visualisation and interpretation of variation of even multiple metagenomic sequencing samples. The importance of sequence variation detection in metagenomes was demonstrated by finding variability in ABC transporters that are involved in resistance to multiple drugs in the Kimchi sample and variability in bacterial cell surface proteins that bind to human Ig in the infant’s microbiome. Comparing data sampled at different time points in the infant’s microbiome showed varying relative abundances during the early colonisation of its gut, suggesting competition between closely related strains until a stable composition is reached.

As sequencing gets cheaper, more assembly-grade metagenomics sequencing datasets become available. In the future it will be possible to compare cohorts of samples and detect sequence level difference in microbiomes or other environmental samples explaining specific phenotypes and link the meta-genotypes to disease. The development of tools, such as MARYGOLD, that find this sequence variation, but also allow further analysis and interpretation, is essential for effectively mining the metagenomics sequencing data.

Funding This work was funded by Kluyver Centre for Genomics of Industrial Fermentation, supported by the Netherlands Genomics Initiative. The project was initiated during a visit of JFN to the lab of MP, supported by an EMBO Short-term fellowship.
Discussion

Over the past decade the price of DNA sequencing dropped remarkably. A steep price decrease was initiated by the introduction of second generation sequencing in the beginning of 2008, which made it feasible for individual labs to sequence genomes, rather than just large international consortia. Reverse engineering genotypes of microbial strains with desirable phenotypes using this cost effective genome sequencing has great potential to establish unknown genotype-phenotype links. Such links enable rational metabolic engineering and synthetic biology, which are major opportunities for modern industrial biotechnology.

This thesis presents several algorithms for the analysis of next-generation sequencing data. Algorithms have been developed to help reconstruct genomic sequences, detect copy number variation and find sequence differences in metagenomes. These methods have been inspired by industrial microbiology research and have been applied in (meta-) genome sequencing of yeast and bacterial populations and in reverse engineering of laboratory evolved yeast.

8.1 The sequencing process and its challenges

The short sequencing reads produced by current sequencers cannot be reconstructed to chromosome length. Genome sequences with high homology to sequences elsewhere in the genome pose a problem when sequencing reads are not long enough to span over these repetitive sequences. In such cases the true locations of the reads falling within a repeat remain unknown, since we cannot infer from which copy of the repeat the reads originate.

The problem of reads being shorter than repeats in the genome has impact on both reconstructing novel genomes (de novo assembly) and comparing a newly sequenced individual to an existent reference genome (resequencing). In resequencing applications, where a finished reference genome is available with resolved repeats, mate-pair libraries help when one of the mates can be uniquely mapped. In de novo assembly, repetitive DNA can be partly resolved with paired-end and mate-pair sequences, but only when the pairs tightly span the repeat (244). Wetzel et al. (244) suggest that the insert sizes of mate-pair libraries should be tuned to the repeat structure of the genome, which is in current practice
8. DISCUSSION

hardly ever considered. Ideally, several libraries with different mate-pair distances would have to be used, but unfortunately such libraries are laborious and therefore expensive to generate.

Repeats in the genome have many different lengths, e.g. in yeast there are transposon sites of $\sim 6$ Kbp, remnants of spliced-out transposons, homologous genes, conserved functional domains and duplications of functional blocks of multiple genes (e.g. the MAL locus in chapter 3). The gaps in the assembly of the CEN.PK113-7D genome in chapter 3 were shown to be mostly caused by such repetitive elements.

In chapter 2 we showed that the longer 454 (454 Life Sciences, Branford, CT) reads ($\sim 450$ bp) provide a more contiguous assembly than the shorter Solexa/Illumina (Illumina, San Diego, CA) reads (50 bp). Although the 454 technology has lower throughput and higher error rates than Illumina, in particular caused by the homopolymer problem, the 454 data are still valuable for de novo assembly. The SMRT technology (Pacific Biosciences, Menlo Park, CA) provides even longer reads, but suffers from low throughput and high error rates.

Reads long enough to span repetitive regions in the DNA of an organism will bring relieve most de novo assembly problems. Accurate chromosome-length assemblies of yeast genomes are possible with at least 10 kbp reads, which may be supplied in the future by Moleculo (Illumina, San Diego, CA), which has such technology under development.

8.2 De novo assembly

8.2.1 Assemblers

De novo assembly in the Sanger era was dominated by overlap-layout-consensus (OLC) assemblers. OLC assemblers are based on pairwise read overlap computation and the algorithmic complexity is therefore in the order of reads. The rise of next-generation sequencing made OLC assemblers impractical because the short reads required very high sequencing depth, and therefore huge numbers of reads. To deal with this amount of data, de Bruijn graph (DBG) assemblers were developed which have a complexity in the order of the size of the genome.

DBG assemblers were ideal for the short reads generated by Solid (Life technologies, Carlsbad, CA) and Illumina sequencers, but less suitable for longer but more error prone reads such as 454 and PacBio reads, which poses a challenge if a combinations of these sequencing types is available for assembly as in chapter 3. OLC assemblers do not require exact overlaps as is the case for DBG assemblers and are therefore more suitable for mixing error prone long reads and high quality short reads. Increasing read lengths and scalable OLC assemblers (208) are bringing OLC assembly with mixed datatypes within reach.

The short Illumina reads and longer 454 reads available for the CEN.PK113-7D genome could not be efficiently assembled by a single assembler. Therefore, we assembled the two datasets separately with different assemblers and developed MAIA (chapter 2) to integrate these separate assemblies. Comparative assemblies can be used as input to MAIA in addition to de novo assemblies. Although the MAIA approach was effective for our yeast
8.2 De novo assembly

genome, a drawback of the method is that a closely related reference genome is required, which is only available for a few species. Currently, MAIA only uses contig sequences since the raw sequencing information is lost in the assembly phase. Future work on the improvement of MAIA would be to include read depths and base qualities in integration phase.

Assembly quality  Assessing the quality of a de novo assembly is intrinsically challenging, since the assembly is usually performed to gain knowledge of the true genome sequence, and cannot be used for validation. N50, total sequence length and number of contigs have been the popular metrics since it requires no a priori knowledge. A major drawback of these metrics is that they only assess contiguity, but do not detect erroneously stitched reads and contigs. Nevertheless, quality assessment in typical assembly projects, including the CEN.PK113-7D project, has been done for NGS assemblies with above metrics.

The lack of quality oriented (rather than contiguity oriented) assessment tools popularised metrics such as the N50. The community recognises that metrics less biased towards contiguity are required, but only recently such a quality assessment tool was published that which estimates the likelihood of an assembly given the set of reads (183). Future assembly projects could benefit from such tools when choosing the optimal assembler for their project and testing parameter ranges.

Assembly representation  The most widely used output of assemblers is the linear FASTA file, useful for downstream processing of the sequences, such as ORF finding. A conservative assembler would break contigs on unresolvable branches in the assembly graph, which is fine for most applications, but information is lost on how the contigs are linked. It would be better to output the internal graph structure, like for example Allpaths 2 does (148). A standard format should be developed for assembly graphs, containing information on links between contigs, read depths, base qualities and possibly read layouts, although the latter requires much space. Such a graph can be used as input for the downstream processing tools. For example, an ORF finder could use the graph information to predict genes across linked contigs.

Modular assemblers  Many assemblers have been developed in the past decade. The majority of these are based on only a few paradigms, i.e. greedy assembly (SSAKE, VCAKE and SHARCGRS), de Bruijn graphs (Soapdenovo, Abyss, Velvet) and OLC (Newbler, Minimus, Celera). Under the hood the software packages perform many steps, such as read processing, read hashing, overlap computation, overlap filtering, assembly graph building, pruning, bubble popping and consensus sequence generation.

Most assemblers are built monolithically, i.e. all these steps reside in a single package. Only few, such as Minimus, have modular designs. It is arguable that assemblers should be built modularly instead of monolithic for several reasons. First, this would make
them robust to yet-to-come sequencing data types with their own typical properties, challenges and algorithmic requirements. Second, it would allow the scientific community to contribute to specific steps in the assembly process, such as the studies on machine learning based overlap correction (172), consistency-based consensus calling (186) and bubble finding (chapter 7). Finally, it could help reduce the time lost implementing the same routines.

8.3 Reference-free variant detection

The majority of the variant detection algorithms finds sequence variants by mapping sequencing reads of newly sequenced genomes to a reference genome. Subsequently, variation between target genomes and the reference is inferred by analysing aberrantly mapped reads.

Unbiased reference-free variant detection through \textit{de novo} coassembly has come within reach (103) as a result of increasing read lengths and efficient assemblers. Comparing sequencing samples directly instead of via a reference genome has several advantages over reference-based methods. First, many species do not have a reference genome. Constructing a draft genome is possible with NGS shotgun-sequencing data, but mapping-based variant finding algorithms have in general been developed for finished (human, diploid) reference genomes and perform poorer on draft genomes (chapter 5). Second, often the goal is not to find difference to the reference genome, but between pairs of newly sequenced genomes, such as the pre- and post-laboratory evolution strains (chapter 4 and 6) and tumour-normal pairs, which are more closely related to each other than to the reference genome. Comparing them via a more distant genome biases the results towards this reference and such analyses might call mutations that are not present and miss true variation. Third, it is unclear what a reference genome should be for a metagenome. For several cultivable bacterial species finished genomes exist which can be combined into a reference metagenome, but generating high quality genomes for uncultivable species is not trivial with current sequencing technologies.

8.3.1 Annotating reference-free variants

Genome annotation is essential for interpretation of genetic variation. The reference genome for a species, next to the genome sequence itself, also serves as a database with several types of annotation related to these sequences, such as genes and transcription factor binding sites. The number of strains having their genomes sequenced, e.g. for yeast (63), is rapidly increasing, uncovering sequences not present in the reference genome (chapter 3). Currently, these genomes are individually stored in genome databases such as SGD and Genbank. A need for a species wide pan-genomes emerges to allow annotation of newly sequenced individuals using a single non-redundant annotation reference database.

Relating variation that has been detected reference-free is still an outstanding challenge. An approach for bubbles (chapter 7) is to align the source and sink contig to the
8.4 Challenges in sequence-based reverse engineering

reference and link the sequence and its annotation between the aligned contigs to the variation. If the reference is closely related to the tested samples it can also be co-assembled to obtain reference locations for the contigs, although this will create additional complexity in the assembly graph. Instead of relating the graph to the reference an unbiased approach would be to annotate the graph itself, for example by performing gene prediction across connected nodes, although this might require to adapt ORF finders to accept graphs. Annotating the graph is in particular suitable for metagenomes and species without a high quality reference genome.

There is no standard format for delivering reference-free detected genome variation to end users. Current formats (VCF, GFF, ...) have been developed for if variation related to a reference genome. Reference-free variation delivery is challenging, since bubbles or other subgraphs are not informative for most users. Marygold (chapter [7]) delivers the variation in a Circos plot, showing the degree of sequence variation within a bubble and variation over time. A standard for graph-based delivery of variation would help the community in building downstream tools for interpretation and visualisation. Developing a quality score for graph-based variants, similar to SNP quality [137], would help prioritisation and interpretation.

8.3.2 Visualising reference-free variants

Current visualisation software such as IGV and the UCSC Genome Browser have been developed for linear genomes, presenting only a single contig at a time. Development of a graph-based genome browser would greatly help users to interpret the reference-free found variation by visualising relationships between contigs. Ideally such a browser visualises an assembly graph differently at different zoom levels, which can be achieved by collapsing bubbles as found in chapter [7] into single nodes when zooming out. For example, a genome would be the single root node of the SPQR tree when maximally zoomed out. Zooming in results in descending the SPQR tree and expanding the bubbles at each traversed edge.

8.4 Challenges in sequence-based reverse engineering

This thesis describes studies that use evolutionary engineering to enhance *S. cerevisiae* strains in chapter 4 and chapter 6. We applied next generation sequencing successfully in several cases to reverse engineer the relevant genotype of these strains.

Discovering the mutations responsible for the phenotypic change is a challenge. We found false positive mutations in particular, which could often be related to draft status of the CEN.PK113-7D genome that we assembled and annotated in chapter 3. Reads map to wrong locations because their true location is missing from the template and map poorly around gaps. Upgrading this genome to a finished status [150], with errors corrected, repeats resolved and gaps closed, was not feasible with the available resources. Nevertheless, thorough filtering of the mutations led us in both chapter 4 and chapter 6 to the causal genetic changes in the DNA. Additionally, parallel evolution lines proved
In several evolutionary engineering experiments, under different selection pressures, clumping and fast-sedimenting clusters of yeast cells were observed. In chapter 6 we found this clumping phenotype to be caused by a frameshift mutation in \textit{ACE2} in combination with polyploidy. The evolved strains in both chapter 4 and chapter 6 exhibited aneuploid genomes. In chapter 6 this aneuploidy was even essential for the phenotype, although reverse engineering a diploid strain in combination with the \textit{ACE2} mutation also sufficed. Aneuploidy is often observed in short term evolution experiments, and can be detected with Magnolya. However, detecting perfect polyploidy, where all chromosomes have the exact same copy number, remains a challenge for both Magnolya and other copy number detection methods, since the number of cells is unknown and fixed amount of DNA is used for the sequencing process. Future developments on single-cell sequencing protocols could solve this, since then signals can be compared between experiments based on intensity.

The superior phenotypes that were reverse engineered in this thesis were constructed using evolutionary engineering, a non-targeted approach. Evolutionary engineering is a valuable complement when used in combination with rational metabolic engineering approaches, but the short time-frame of this technique limits the evolution to only small genetic changes. Additionally, only a limited number of parallel evolutions can be created in the shake flask set-up as used in chapter 4 and chapter 6.

The time limitation of laboratory evolution can be overcome by harnessing the endless source of diversity in nature, which offers a wealth of genotypes and phenotypes. Robotising random crossing, screening and selection in microwells would enable massively parallel creation of new yeast variants with optimal traits. Now already and even more in the future, automated and miniaturised screening systems provide fast strain construction platforms. An \textit{in vivo} genetic algorithm approach would speed-up evolution by selecting strong evolutionary lines, discarding weak ones and randomly recombining the strong with each other.

This thesis shows sequence-based reverse engineering can efficiently resolve the genetic factors influencing an observed phenotype. Determining genotypes by next-generation sequencing is highly automated and still evolving rapidly. Future developments on strain generation and phenotype screening will put sequence-based reverse engineering in an increasingly prominent place in industrial biotechnology research.
References


REFERENCES
for discovering structural variation with next-generation sequencing. Nat Methods, 6(11 Suppl), S13–20.


REFERENCES

965–74. 60


Summary

Microorganisms are employed on a large scale in a variety of industrial production processes, such as production of food and ingredients, beverages, pharmaceutical components and fine- and bulk chemicals. Improving the behaviour of these microorganisms is of crucial importance in the design of efficient production processes.

In this thesis we focus use directed evolution, which is a method to improve the phenotype of a microorganism without knowing what to change in the DNA to obtain the desired effect. Improved strains are obtained through the natural evolution processes of variation and selection by enforcing a strong selective pressure in the laboratory. The mutations that occur during such short evolution can be determined by sequencing the DNA of the improved strains and comparing it to the starting strain. Finally, relating the relevant mutations to the enforced phenotype is called reverse engineering. We describe the reverse engineering of mutations in ADY2 enabling lactate transport across the cell wall and mutations in ACE2, which in combination with aneuploidy, could be related to the origin of multi-cellular growth in yeast.

Next-generation sequencing was employed to reconstruct genomes and infer mutations. This technology has made tremendous progress since its introduction in 2008 and has become affordable for individual laboratories. Sequencing poses many challenges to the field of bioinformatics, since genomes are long and repetitive (e.g. yeasts have millions of base pairs and humans billions), but current sequencing technology can only read up to 100 to 1000 base pairs in sequence.

The computational methods developed in this thesis enable the use of sequencing technology in industrial microbiology research. A starting strain in this field is the yeast Saccharomyces cerevisiae CEN.PK113-7D. To determine its genome sequence we developed an algorithm that uses a Tabu-search that exploits previously sequenced genomes and heterogeneous sequencing data types. Additionally, two methods were developed to infer mutations between genomes in absence of a high-quality closely related reference genome. First, we describe an algorithm that estimates the number of times a DNA subsequence exists in the full genome using a Poisson mixture model. We used this method to infer copy number differences between aneuploid lager brewing yeasts. Second, we describe an algorithm that uses graph decomposition to find variation in metagenomes, which are a combination of many bacterial genomes within one sample and play an important role in industrial fermentation, such as in production of cheese and yoghurt.
Samenvatting

Micro-organismen worden op grote schaal gebruikt in allerlei industriële productieprocessen, zoals in de productie van voedsel en ingrediënten, dranken, farmaceutische componenten en fijn- en bulkchemicaliën. Het verbeteren van het gedrag van deze micro-organismen is van cruciaal belang om efficiënte productieprocessen te kunnen ontwerpen.

In dit proefschrift focussen we op een methode, gerichte evolutie, om het fenotype van een micro-organisme te verbeteren zonder te weten wat er daarvoor in het DNA aangepast moet worden. Door in het laboratorium een sterke selectiedruk op te leggen worden verbeterde stammen verkregen door de natuurlijke evolutionaire principes van variatie en selectie. De mutaties die optreden in de korte periode van enkele weken tot maanden evolutie kunnen bepaald worden door de DNA volgorde van de verbeterde stammen af te lezen en te vergelijken met die van de uitgangsstam. Het relateren van deze mutaties aan het geselecteerde fenotype wordt reverse engineering genoemd. Met deze techniek hebben we mutaties in ADY2 kunnen relateren aan melkzuurtransport over de celwand en mutaties in de transcriptiefactor ACE2, in combinatie met aneuploïdie, kunnen relateren aan het onstaan van multi-cellulaire groei van gist.

We hebben next-generation sequencing gebruikt om genomen te reconstrueren en mutaties te ontdekken. Deze technologie heeft sinds haar introductie in 2008 een enorme vlucht genomen en is betaalbaar geworden voor individuele laboratoria. DNA sequencing creëert enorme uitdagingen binnen de bioinformatica, doordat volledige genomen miljoenen (b.v. gist) tot miljarden (b.v. mens) base paren tellen en repetitief zijn, terwijl er maximaal ongeveer 100 tot 1000 base paren afgelezen kunnen worden.

De computationele methodes die in dit proefschrift ontwikkeld zijn maken het gebruik van sequencing mogelijk binnen het industrieel microbiologisch onderzoek. Een uitgangsstam in dit vakgebied is de gist CEN.PK113-7D. Om het DNA volgorde van het genoom van CEN.PK113-7D te bepalen, in aanwezigheid van al bekende genomen en heterogene sequensing datatypes, ontwikkelde we een algoritme dat gebruikt maakt van Tabu-search. Daarnaast hebben we twee methodes ontwikkeld om mutaties te vinden in genomen zonder gebruik te maken van een hoge kwaliteit referentiegenoom. Ten eerste beschrijven we een algoritme dat het aantal keer dat een stuk DNA in een genoom voorkomt bepaalt met behulp van een Poisson mixture model. We hebben dit algoritme gebruikt om verschillen te vinden in de genomen van aneuploïde bierbrouwgisten, welke het DNA van twee soorten in hun celkern hebben. Ten tweede beschrijven we een algoritme om variatie te vinden in zogenaamde meta-genomen, welke bestaan uit een combinatie van vele bacteriële soorten die o.a. een belangrijke rol spelen in industriële fermentatie, zoals bij de productie van kaas en yoghurt.
Curriculum Vitae

Jurgen Nijkamp was born in Houten, the Netherlands on the 30th of October, 1982 and lived there the first 18 years of his life. He attended the pre-university education at the Stedelijk Gymnasium in Utrecht where he graduated in 2001. Afterwards, he moved to Delft and enrolled into Life Science & Technology, which was taught as a collaboration between the Delft University of Technology and the Leiden University. During his time in Delft he was very active in the Students’ Association LIFE and spent one year as its president. He was co-owner and for over two years financial manager of the company Dimax v.o.f., which provided students as personal drivers. As part of a summer internship in 2007, he worked at the corporate finance division of the merchant bank Kempen & Co.

After finishing his bachelor’s degree in 2006 Jurgen proceeded with the Bioinformatics master track in Delft, which was part of the MSc Media & Knowledge Engineering at the faculty Electrical Engineering, Mathematics and Computer Science. In 2007, Jurgen spent three months at the ETH Zürich, where he performed a research project in the computational biochemistry group of Gaston Gonnet. Being interested in the industrial applications of Life Sciences, he finished his master degree in 2008 with a thesis project at the bioIT group of the DSM Biotechnology Center.

The research theme of bioinformatics within industrial microbiology, which was initiated during his Master phase, was continued in 2009 with a PhD project within the Kluyver Centre for Genomics of Industrial Fermentation under the supervision of Dick de Ridder. In between Master and PhD, Jurgen worked as consultant at Accenture, Amsterdam. The performed research during his PhD studies was in close collaboration with the Industrial Microbiology group. In 2012 Jurgen was awarded an EMBO short-term fellowship to be a visiting researcher at the Center for Bioinformatics and Computational Biology at the University of Maryland, USA. Currently, Jurgen works at the Marker Technology and Genomics lab at Bejo Seeds, located in Warmenhuizen, the Netherlands.
Dankwoord - A word of thanks

Er zijn onnoemelijk veel mensen die in enige mate hebben bijgedragen aan dit proefschrift. Hieronder wil ik graag een aantal mensen persoonlijk bedanken voor hun bijdrage.

Allereerst Dick de Ridder en Marcel Reinders, die mij vanuit het Patroonherkenning & Bioinformatica (PRB) lab in Delft hebben begeleid. Tijdens mijn afstudeeronderzoek in 2008 bij DSM had ik al ervaren dat jullie zeer betrokken begeleiders zijn. Toen ik vanuit mijn korte loopbaan als consultant op zoek ging naar een promotieplaats twijfelde ik dan ook geen moment toen dat bij jullie mogelijk bleek. Veel dank voor de vrijheid die ik heb mogen genieten tijdens mijn onderzoek. Ik heb hierdoor mijn eigen pad kunnen bepalen en me kunnen ontwikkelen tot een onafhankelijke onderzoeker. Daarnaast ook bedankt voor de mogelijkheden die jullie geboden hebben om in Maryland een deel van mijn onderzoek uit te voeren en (inter)nationaal congressen te bezoeken om mijn werk te presenteren en inspiratie op te doen.

Dick, je hebt een onuitputtelijk vermogen om me net weer dat stapje meer te laten doen. Als ik vast zat wist je het probleem snel weer te doorgronden en een oplossing aan te dragen. De deadlines voor conferenties koos je altijd strak om mij op tijd een rond verhaal af te laten leveren, wat zijn vruchten heeft afgeworpen. Op een wonderlijke manier weet jij te allen tijde op je mails te reageren. Het maakt niet uit of je met vakantie bent, het nacht is, je in de auto zit of in vergadering bent. Tijdens de congressen in Lunteren en op de Wageningseberg toonde je jezelf een wandelende encyclopedie: met Dick in je team win je de pubquiz gegarandeerd!

Ik prijs mij zeer gelukkig dat ik tijdens mijn promotie omringt ben geweest door zeer getalenteerde mensen in de PRB groep. Voor mij onbeantwoordbare vragen konden altijd wel door iemand op een whiteboard worden opgelost. Alexey, Bastiaan, Marc en Wynand jullie inzichten in de wiskunde en informatica zijn cruciaal geweest voor de tostandkoming van dit boekje.

Samenwerking met experimentele groepen is een essentieel onderdeel van het bioinformatica-vakgebied. Ik heb het grote voorrecht gehad om vier jaar lang te mogen samenwerken met de Industriële Microbiologie (IMB) groep. Bart, onze projecten waren vanaf het begin op elkaar afgestemd. We hebben heel veel samengewerkt en dat heeft, zowel in de informatica als de biologie, tot prachtige resultaten geleid. Stefan, met jou kwam het eerste succesverhaal door gebruik van sequentiedata tot stand en realiseerde we ons voor het eerst wat voor grootschalige veranderingen er op DNA niveau plaatsvonden. Marcel, jij was al in de next-generation sequencing gedoken voordat ik begon en we zijn er samen verder ingegroeid. Jean-Marc en Jack, jullie kennis van gist en haar genetica beantwoordt vele vragen die je met een DNA sequentie alleen niet boven water krijgt. Marit, wie had gedacht in ons Life jaar dat we samen in de gist groep zouden promoveren? Door jou heb ik veel maandagavonden doorgewerkt, wanneer je weer Grey’s Anatomie met Nicole kwam kijken. Ik hoop dat je nog vaak bij ons thuis komt. Niels, heel leuk dat jij na LST ook in de gistkelder bent beland en dat we daar drie jaar de kamer hebben gedeeld.
Verder hebben natuurlijk nog veel meer mensen van zowel PRB als de IMB groep een belangrijke rol gespeeld. Van de TU: Ahmed, Chris, Emrah, Erdogan, Erik, Lodewyk, Robbert, Saskia, Sepideh, Thies, Wouter en Yuching. Ik wil Umesh, Hanka, Jayne, Jeroen, Inken en Miranda bedanken voor de samenwerking binnen de RSG en Karin, Celia en Femke voor het faciliteren van deze groep vanuit NBIC.

I would like to thank Mihai Pop for hosting me as a visiting PhD student in Maryland. Mihai, your work has played an important role in many chapters of this thesis. Thank you for the fruitful stay in MD in spring 2012. Also many thanks to the guys letting me feel at home there. Hisham, Joe, Joyce, Katie, Kwame, Lee, Matt, Mike, Rich and the others: thank you for the D.C. experience.

Mijn dank gaat ook uit naar het ‘Kluyver Centre for Genomics of Industrial Fermentation’ en iedereen die het Kluyver mogelijk maakt, waaronder Jenny, Maita, Elly en Jack. Deze publiek-private samenwerking heeft niet alleen mijn promotie gefinancierd, maar ook gezorgd voor inspiratie en talloze contacten tijdens de congressen en programmadagen.

Met veel plezier heb ik tijdens mijn promotie met verschillende bedrijven binnen het Kluyver Centre samengewerkt. Dank aan Jan-Maarten (Heineken) voor de nauwe samenwerking aan biergisten. Dank aan Hans, Paul en Wilbert (DSM) voor de samenwerking aan het CEN.PK genoom.

Veel dank gaat ook uit naar mijn familie, die er altijd voor me zijn geweest en ontzettend belangrijk zijn. Lieve pap en mam, jullie steun is onvoorwaardelijk en nog altijd onmisbaar. Nicole, dank je voor je liefde en oneindige geduld als er weer eens weekenden opgeofferd werden om artikelen en een proefschrift te schrijven. We hebben een hele mooie tijd voor ons met onze prachtige dochter Carlijn.
Publications


