BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 24 2012, pages 3195–3202
doi:10.1093/bioinformatics/bts601

Genome analysis

Advance Access publication October 9, 2012

De novo detection of copy number variation by co-assembly

Jurgen F. Nijkamp[1,2,3], Marcel A. van den Broek[2,3], Jan-Maarten A. Geertman[4], Marcel J. T. Reinders[1,3,5], Jean-Marc G. Daran[2,3] and Dick de Ridder[1,3,5,*]

[1]The Delft Bioinformatics Lab, Department of Intelligent Systems, Delft University of Technology, 2628 CD Delft, [2]Department of Biotechnology, Delft University of Technology, 2628 BC Delft, [3]Kluyver Centre for Genomics of Industrial Fermentation, 2600 GA Delft, [4]Heineken Supply Chain, Global Research & Development, 2382 PH Zoeterwoude and [5]Netherlands Bioinformatics Centre, 6500 HB Nijmegen, The Netherlands

Associate Editor: Michael Brudno

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 through a reference genome, within reach.

Results: Here, we develop and test an algorithm, named Magnolya, that uses 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 (CNV) 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 'coloured' 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 CNV probabilities are easily inferred. The copy number estimator and CNV 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 CNV in aneuploid brewing yeast genomes. Integer CNV was also accurately detected in a short-term laboratory-evolved yeast strain.

Availability: Magnolya is implemented in Python and available at: http://bioinformatics.tudelft.nl/

Contact: d.deridder@tudelft.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 2, 2012; revised on September 26, 2012; accepted on October 2, 2012

1 INTRODUCTION

Genomes can differ in many ways. Several types of variation are commonly distinguished, from small local differences such as single nucleotide polymorphisms and short insertions and deletions (indels), to variation involving DNA fragments >1 kbp, i.e. structural variation. Structural variations 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 analysing aberrantly mapped reads (Li *et al*., 2009), read-depth variation (Klambauer *et al*., 2012; Xie and Tammi, 2009), split-read mappings (Ye *et al*., 2009) or a combination of aberrantly mapped read pairs and read-depth variation (Medvedev *et al*., 2010). 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 coloured by the samples in which they are observed. By observing bifurcations in the graph that separate the colours (*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. 1). Because 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 (Myers, 2005). Medvedev and Brudno (2009) estimated the copy

[*]To whom correspondence should be addressed.

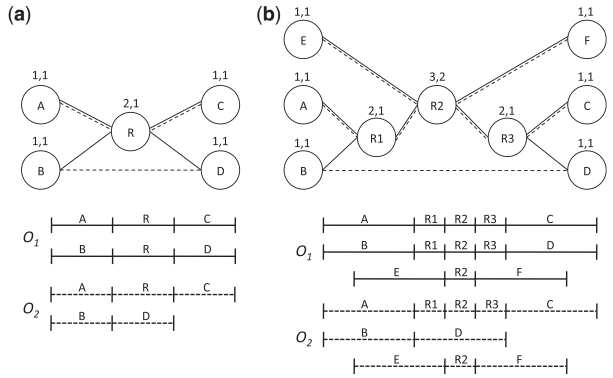


Fig. 1. Coloured string graphs with two samples, O_1 (solid) and O_2 (dashed). The pairs of numbers indicate the traversal count (left for O_1, right for O_2), which are estimated by a PMM. Solid and dashed edges represent reads spanning the connected contigs. Bars under the graphs represent the unknown 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

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 (CN) of a contig without making assumptions on genome size.

The use of a PMM for modelling read depths in segments across multiple samples that have been mapped to a reference genome was recently introduced (Klambauer *et al*., 2012). Here, we use 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 for segmentation. We show how this mixture model can be applied in coloured 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 use an overlap-layout-consensus 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 (Myers, 2005). 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, as for each pair of contigs connected by an edge the in- or out-degree (dependent on strand) is >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 String Graph Assembler, capable of efficiently assembling mammalian genomes (Simpson and Durbin, 2012). We extended the string graph by assigning each read a colour, corresponding to the sample from which it originated. In the resulting coloured contig string graph, we then detect variation by modelling read counts per colour with a mixture of Poissons and inferring the probability of a CNV.

To our knowledge, we here for the first time apply a PMM 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.

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 PMM (Section 2.1) and how these models are used in a coloured co-assembly string graph to detect CNV (Section 2.2). Detailed derivations of the formulas can be found in the Supplementary Material.

2.1 Contig copy number assignment with a Poisson mixture model

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 owing 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 and in- and out-degree of two. This is the case for sample O_1 (solid edges) in Figure 1a: 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 sample, contig R in Figure 1a 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.

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:

$$p(x_c) = \sum_{i=1}^{M} p(i)p(x_c|i) \tag{1}$$

The number of reads sampled from a certain nucleotide position in the genome can be modelled as a Poisson process (Myers, 2005), with rate parameter $\lambda = \frac{R}{G}$, 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 the following:

$$p(x_c) = \sum_{i=1}^{M} \pi_i \text{Pois}(x_c|\theta_{i,c}) \tag{2}$$

The parameters are the mixture coefficients the π_i, estimates of $p(i)$'s and the Poisson parameters $\theta_{i,c} = L_c i\lambda$, 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 λ directly because we do not know the genome size G; therefore, we estimate λ and the π_i from the data, which we do by Expectation Maximization (EM).

2.1.3 High–copy number repeats Specific repetitive regions in the DNA occur at high copy number, such as ribosomal DNA repeats and transposons. We are not interested in modelling 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

$$Z = \{c \in C : x_c \geq (M+1)L_c\lambda\} \tag{3}$$

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

$$p(x_c) = \begin{cases} 0 & \text{if } x_c \in Z \\ \text{Geom}(x_c - \theta_{M+1}|\lambda, \alpha, M) & \text{if } x_c \notin Z \end{cases} \tag{4}$$

where α is the rate parameter of the geometric distribution. Our model thus becomes

$$\begin{aligned} p(x_c|\pi, \lambda, \alpha) = & \sum_{i=1}^{M} \pi_i \text{Pois}(x_c|\theta_{i,c}) \\ & + \pi_{M+1} u(x_c) \text{Geom}(x_c - \theta_{M+1,c}|\lambda, \alpha, M) \end{aligned} \tag{5}$$

where $u(c)$ is an indicator function defined as

$$u(c) = \begin{cases} 1 & \text{if } c \in Z \\ 0 & \text{if } c \notin Z \end{cases} \tag{6}$$

2.1.4 Estimation of the model parameters by expectation maximization The mixture parameter π, the Poisson rate parameter λ and the geometric distribution rate parameter α are estimated from the data (N contigs) by optimizing the log likelihood,

$$-\mathcal{L}(C|\pi, \lambda, \alpha) = -\log \prod_{c=1}^{N} p(x_c|\pi, \lambda, \alpha)p(\pi) \tag{7}$$

In the E-step, current estimates of the parameters ($\theta_{i,c}^{old} = L_c i\lambda^{old}$ and π^{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 λ and α as follows:

$$\lambda^{new} = \frac{\sum_{c=1}^{N}\sum_{i=1}^{M} r_{i,c}x_c}{\sum_{c=1}^{N}\sum_{i=1}^{M} iL_c r_{i,c} + \sum_{c=1}^{N} u(x_c)(M+1)L_c r_{M+1,c} \log\left(1 - \frac{\alpha}{L_c}\right)} \tag{8}$$

$$\alpha^{new} = \frac{\sum_{c=1}^{N} u(x_c) r_{M+1,c}}{\sum_{c=1}^{N} u(x_c) r_{M+1,c}\left(\frac{x_c - \theta_{M+1,c}^{old}}{L_c} + 1\right)} \tag{9}$$

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*. (2012) to use a Dirichlet prior distribution with parameters γ for cases where we can incorporate prior knowledge on $p(i)$, where γ is an M-dimensional vector $(\gamma_1, \ldots, \gamma_M)$. The update rule for π_i then becomes as follows:

$$\pi_i^{new} = \frac{\sum_{c=1}^{N} r_{ic} + \frac{1}{N}(\gamma_i - 1)}{1 + \frac{1}{N}(\sum_{i=1}^{M+1} \gamma_i - (M+1))} \tag{10}$$

2.1.6 Model selection The model with the optimal number of Poisson distributions is selected among models with 3–20 Poissons with the lowest Bayesian information criterion (see Supplementary Material S1.2).

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

$$\hat{\imath}_{MAP} = \underset{i=1,\ldots,M+1}{\arg\max}\, p(i|x_c) = \underset{i=1,\ldots,M+1}{\arg\max}\, r_{i,c} \tag{11}$$

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

2.2 Copy number variation using co-assembly

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 CNV. We fit the proposed PMM for both samples; assuming independent, we can calculate the probability of a CNV for non-outlier contigs as follows:

$$p(CNV \text{ in } x_c) = 1 - \sum_{i=1}^{M} p_1(i|x_c)p_2(i|x_c) \quad (12)$$

3 METHODS

3.1 Simulated DNA sequencing data

The *Saccharomyces cerevisiae* strain S288C reference genome (Goffeau *et al*., 1996) was downloaded from the Saccharomyces Genome Database (Cherry *et al*., 2012) (www.yeastgenome.org, accessed March 03, 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 1 Kbp and 10 Kbp. 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 10 Kbp of another event. A total of 800 000 error-free sequencing reads with a length of 150 bp each were then simulated at 10× coverage for each sample using fragsim (Lysholm *et al*., 2011). Additionally, a 20× coverage read 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.

3.2 Real DNA sequencing data

In the experiments, we used a number of NGS data sets available for various yeast strains (Table 1). The sequencing of *S.cerevisiae* CEN.PK113-7D was described in Nijkamp *et al.* (2012), 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.

The genomic DNA of two brewing yeasts of the species *S.pastorianus* was purified as previously described (de Kok *et al*., 2012). The two *S.pastorianus* strains are indicated as SPA and SPB in this study. The fragments of ~180–200 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 for the *S.cerevisiae* strain Y12, part of the *S.cerevisiae* strain project (http://genome.wustl.edu/genomes/saccharomyces_cerevisiae_strain_project_genomes).

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 Newbler was coloured using the output file that describes the read layout on the contigs (the .ACE file). Alignments were performed with nucmer version 3.07, part of the Mummer 3 package (Delcher *et al*., 2002).

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 considered. 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. 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.

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 (Boeva *et al*., 2012), CN.MOPS v1.2.1 (Klambauer *et al*., 2012) and CNV-Seq (Xie and Tammi, 2009) 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 guanine-cytosine content (GC content) 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 (Li and Durbin,

Table 1. *Saccharomyces cerevisiae* strains used in this study

Accession	Species	Strain	Description	Ploidy	Dirichlet prior parameters
—	*S.pastorianus*	SPA	Lager brewing yeast	Aneuploid	None
—	*S.pastorianus*	SPB	Lager brewing yeast	Aneuploid	None
SRX129889	*S.cerevisiae*	CEN.PK113-7D	Laboratory strain	Haploid	$\gamma = (1 + G, 1, .., 1)$
SRX129995	*S.cerevisiae*	IMW004	Laboratory evolved strain	Aneuploid	$\gamma = (1 + G, 1, .., 1)$
SRX039438	*S.cerevisiae* *Subspecies uvarum*	Y12	Palm wine sample, single spore Derivative of NRRL Y-12633	Aneuploid	$\gamma = (1 + \frac{1}{2}G, 1 + \frac{1}{2}G, 1, .., 1)$

The 'Accession' column shows the Sequence Read Archive accessions numbers. The last column shows the γ vector used as parameters for the Dirichlet prior, reflecting prior expectations on copy number distribution. G is the hyper parameter that is set to the number of contigs.

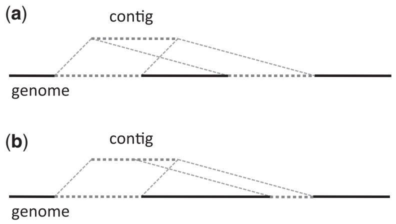


Fig. 2. A contig should be present an integer number of times in the genome. To assess whether this is true, contigs assembled from simulated reads were aligned to the genome from which the reads originate. (**a**) The contig occurs exactly two times in the genome and therefore is assigned an integrity value of 0. (**b**) The contig occurs 1.4 times in the genome and is assigned an integrity value of 0.4

2009). Sensitivities and specificities were calculated by counting true positive and false positive calls per base. In all, 250 bp around duplications were ignored, to prevent counting them as false positives owing to overhanging windows.

3.6 Timing

Sets of 1000, 2500, 5000, 7500, 10 000 and 12 500 contigs were randomly sampled without replacement from the Y12 dataset. Magnolya was run 25 times on each dataset, each with $M = 8$ and λ initialized on the 25th percentile. The algorithm was halted when an iteration resulted in a likelihood improvement $<10^{-5}$. The processor 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.

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.

4.1 Copy number estimation

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 2).

First, the assumption that contigs have an integer copy number was assessed by calculating copy number integrity (Section 3.4). The contigs were found to have low average integrity values (Table 2). Only 18 of the 3723 contigs had an integrity value >0.1, which proves the assembled contigs indeed mostly have an integer copy number. Second, the PMM copy number estimate $\hat{\imath}_{MAP}$ was compared with 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.

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 *Saccharomyces eubayanus* (Libkind *et al.*, 2011). The wine-making yeast VIN7 was shown to contain diploid *S.cerevisiae* and haploid *Saccharomyces kudriavzevii* genomic content (Borneman *et al.*, 2012). Furthermore, these hybrid genomes may contain many chromosomal rearrangements and aneuploidy (Nakao *et al.*, 2009). 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.

Table 2. Simulation results for copy number estimation

CN	CN_{align}	$\mathrm{CN}_{align} = \hat{\imath}_{MAP}$	Integrity
1	3638	3638 (100%)	0.0011 ± 0.014
2	67	66 (99%)	0.0111 ± 0.055
3	14	14 (100%)	0.0005 ± 0.001
4	4	4 (100%)	0.0014 ± 0.003

Only copy numbers 1–4 are listed, as no higher copy numbers were observed. The integrity column indicates averages and standard deviations of the integrity values (Fig. 2).

Saccharomyces 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 5893 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 Figure 3, i.e. the model explained the data as contigs having copy number 2 and 4. Visual inspection of the read count histogram (Fig. 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 (Borneman *et al.*, 2012). 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 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 $\sum_{k=1}^{N} L_k \hat{\imath}_k = 29.9$Mbp in size, 22% of which was represented one time (haploid) and 74% two times (diploid).

The contigs were mapped to the 12.1 Mbp *S.cerevisiae* S288C reference genome to investigate the possible *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 $\hat{\imath}_{ML} = 2$, indicating that the *S.cerevisiae* content in this hybrid genome is present in diploid form, as is the case for VIN7 (Borneman *et al.*, 2012).

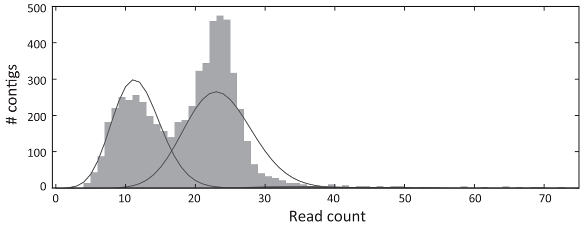


Fig. 3. PMM fitted on the assemblies of the wine yeast Y12. Grey bars: histogram of read counts on the contigs, normalized to a contig length of 300 bp. The fitted PMM is the plotted as a black line with $300\lambda = 11.7$.

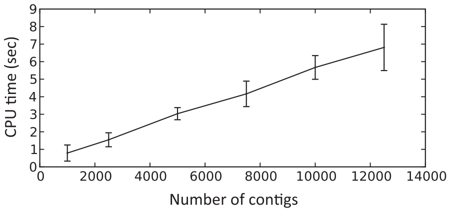


Fig. 4. CPU time consumed by the EM algorithm for different numbers of contigs, which were randomly sampled from the Y12 yeast dataset. On average the EM algorithm finished within 14 iterations

The Y12 assembly had the most contigs and was therefore used to time the algorithm (Fig. 4). Computational complexity is low, (compared with mapping and assembly of short read data) and scales linearly with the number of contigs.

4.2 Copy number variation detection

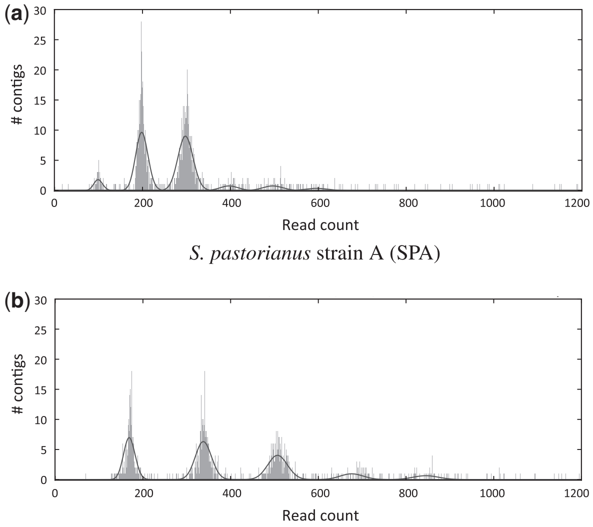
4.2.1 Benchmarking A simulation experiment was designed to benchmark the performance of Magnolya in a setting where a closely related reference genome is available versus the situation where this is not the case. We did not simulate biological noise or contamination, nor did we introduce uneven depth 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 *et al.* (2012). 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 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 because this rate is always high for all methods (as genomes are very large compared with the total length of the CNVs).

Table 3. Performance in simulation experiments expressed as the F-measure (the harmonic mean of precision and sensitivity)

	Reference genome	Shotgun assembly
CN.MOPS	0.57	0.34
Control-FREEC	0.86	0.12
CNV-seq global	0.91	0.44
CNV-seq contig	—	0.41
Magnolya	—	0.94

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.



S. pastorianus strain A (SPA)

S. pastorianus strain B (SPB)

Fig. 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)

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 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.

4.2.2 CNV in a laboratory evolution pair As another test case, a pre- and post-laboratory evolution pair was analysed. 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 close. We recently applied laboratory evolution to investigate lactate transport in the haploid yeast *S.cerevisiae* CEN.PK113-7D (de Kok *et al.*,

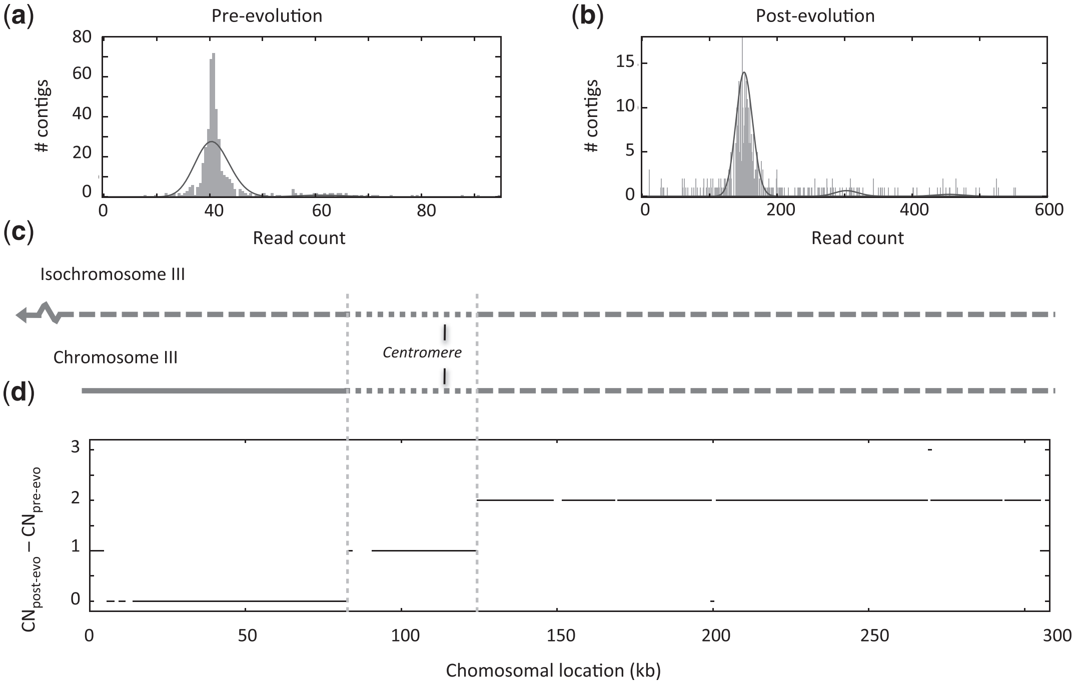


Fig. 6. *De novo* CNV detection using Magnolya. The yeast *S.cerevisiae* CEN.PK113-7D was evolved in the laboratory under strong selective pressure. The genomes of the evolved strain and its pre-evolution ancestor were both sequenced and assembled into contigs. The grey bars form a histogram of the read 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. $(x_c/L_c) * 1000$ 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 CNV plotted versus chromosomal location on the yeast reference genome

2012). This strain, used in systems biology research and in industry, deviates significantly from the yeast reference sequence S288C (Nijkamp *et al*., 2012) including multiple Kbp insertions and deletions and >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 using a more distant reference genome.

A mutant harbouring a *JEN1* deletion grew 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. 5c). 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. 5d). Compared with the mapping-based approach taken in de Kok *et al*. (2012), these results are easier to interpret.

4.2.3 CNV in aneuploid brewing strains CNV 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. 6), 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 (Nakao *et al*., 2009).

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. Figure 6 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 $p(i|x_c), i \in \{1, 2, \ldots, 6\}$, using equation (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 similar, this large scale CNV largely affects the gene dosage of thousands of genes, with an intriguing yet unexplored effect on the phenotypic characteristics.

5 CONCLUSIONS

Since the advent of NGS, variation detection has been performed by mapping short reads to a reference genome 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 through *de novo* assembly. Short sequencing reads and repetitive genomes have thus far hampered accurate reconstruction of individual genomes. However, with increasing read lengths and 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 PMM to *de novo* estimate copy numbers of contigs, and combined this with a co-assembly approach. This allows easy detection of CNV, 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 CNV. 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 with respect to a reference genome.

The Magnolya algorithm performs at higher precision and sensitivity than 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 CNV in matched experiments, such as tumor-normal pairs.

The co-assembled coloured string graphs that are used 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 coloured assembly graphs (Iqbal *et al*., 2012). We expect a combination of a bubble calling algorithm, our CNV detection and approaches exploiting read pair data in coloured assembly graphs to allow a move to fully unbiased detection of variation between individuals in the near future.

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

Conflict of Interest: none declared.

REFERENCES

Boeva,V. *et al*. (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, **28**, 423–425.

Borneman,A.R. *et al*. (2012) The genome sequence of the wine yeast VIN7 reveals an allotriploid hybrid genome with *Saccharomyces cerevisiae* and *Saccharomyces kudriavzevii* origins. *FEMS Yeast Res*., **12**, 88–96.

Cherry,J.M. *et al*. (2012) *Saccharomyces* genome database: the genomics resource of budding yeast. *Nucleic Acids Res*., **40**, D700–D705.

de Kok,S. *et al*. (2012) Laboratory evolution of new lactate transporter genes in a *jen1* mutant of *Saccharomyces cerevisiae* and their identification as *ADY2* alleles by whole-genome resequencing and transcriptome analysis. *FEMS Yeast Res*., **12**, 359–374.

Delcher,A.L. *et al*. (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res*., **30**, 2478–2483.

Goffeau,A. *et al*. (1996) Life with 6000 genes. *Science*, **274**, 546, 563–567.

Iqbal,Z. *et al*. (2012) *De novo* assembly and genotyping of variants using colored de bruijn graphs. *Nat. Genet*., **44**, 226–232.

Klambauer,G. *et al*. (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res*., **40**, e69.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al*. 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Libkind,D. *et al*. (2011) Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. *Proc. Natl Acad. Sci. USA*, **108**, 14539–14544.

Lysholm,F. *et al*. (2011) An efficient simulator of 454 data using configurable statistical models. *BMC Res. Notes*, **4**, 449.

Medvedev,P. and Brudno,M. (2009) Maximum likelihood genome assembly. *J. Comput. Biol*., **16**, 1101–1116.

Medvedev,P. *et al*. (2010) Detecting copy number variation with mated short reads. *Genome Res*., **20**, 1613–1622.

Myers,E.W. (2005) The fragment assembly string graph. *Bioinformatics*, **21** (**Suppl. 2**), ii79–ii85.

Nakao,Y. *et al*. (2009) Genome sequence of the lager brewing yeast, an interspecies hybrid. *DNA Res*., **16**, 115–129.

Nijkamp,J.F. *et al*. (2012) De novo sequencing, assembly and analysis of the genome of the laboratory strain *Saccharomyces cerevisiae* cen.pk113-7d, a model for modern industrial biotechnology. *Microb. Cell Fact*., **11**, 36.

Simpson,J.T. and Durbin,R. (2012) Efficient de novo assembly of large genomes using compressed data structures. *Genome Res*., **22**, 549–556.

Xie,C. and Tammi,M. (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, **10**, 80.

Ye,K. *et al*. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.