A peer-reviewed version of this preprint was published in PeerJ on 29 May 2018.

Seitz A, Hanssen F, Nieselt K. 2018. DACCOR–Detection, characterization, and reconstruction of repetitive regions in bacterial genomes. PeerJ 6:e4742 https://doi.org/10.7717/peerj.4742

DACCOR - Detection, charACterization, and reconstruction of Repetitive regions in bacterial genomes

Alexander Seitz[1], **Friederike Hanssen**[1], **and Kay Nieselt**[1]

[1]**Center for Bioinformatics (ZBIT), Integrative Transcriptomics, Eberhard-Karls-Universität Tübingen**

Corresponding author:
Alexander Seitz[1]

Email address: alexander.seitz@uni-tuebingen.de

ABSTRACT

The reconstruction of genomes using mapping based approaches with short reads experiences difficulties when resolving repetitive regions. These repetitive regions in genomes result in low mapping qualities of the respective reads, which in turn lead to many unresolved bases of the genotypers. Currently, the reconstruction of these regions is often based on modified references in which the repetitive regions are masked. However, for many references such masked genomes are not available or are based on repetitive regions of other genomes. Our idea is to identify repetitive regions in the reference genome *de novo*. These regions can then be used to reconstruct them separately using short read sequencing data. Afterwards the reconstructed repetitive sequence can be inserted into the reconstructed genome. We present the program `DACCOR`, which performs these steps automatically. Our results show an increased base pair resolution of the repetitive regions in the reconstruction of *Treponema pallidum* samples, resulting in fewer unresolved bases.

INTRODUCTION

Modern genome reconstruction often relies on mapping programs such as BWA (Li and Durbin, 2009) to align short reads generated by Next-Generation-Sequencing (NGS) technologies to a known reference genome (Veeramah and Hammer, 2014). The consensus sequence of the aligned reads can then be used to generate the genomic sequence of the newly sequenced sample, assuming that the sample was sequenced with a sufficent coverage depth. This allows for the fast identification of short insertions, deletions, and single-nucleotide polymorphisms (SNPs). The mapping programs typically calculate a score for each aligned read that corresponds to the quality of the alignment (Li et al., 2008). The score quantifies the probability that a read is placed at the correct genomic position. Reads with a low mapping quality can be filtered out to remove reads that might stem from contaminations or were sequenced with low sequencing quality (Smith et al., 2008). Besides bad quality reads, also reads mapping to repetitive regions could yield low quality scores if they cannot be mapped to a unique position. Filtering of reads with low mapping qualities would also include these reads. This filtering is often conducted in the context of ancient DNA (aDNA) (Bos et al., 2016), so that for such samples the repetitive regions of the respective reconstructed genomes are generally affected.

However, repetitive regions play an important role in the genome (Shapiro and von Sternberg, 2005). Hundreds to thousands of such regions are present in prokaryotic and eukaryotic chromosomes (Treangen et al., 2009). The human genome, for example, consists of approximately 50% repetitive regions (Lander et al., 2001). Tandem repeat regions appear to encode outer membrane proteins, which suggests that they help pathogens to adapt to their hosts (Denoeud and Vergnaud, 2004). In the case of the bacterium *Treponema pallidum*, repetitive sequences in the *arp* gene are used to distinguish between the subspecies that cause veneral syphilis (*Treponema pallidum pallidum*), nonveneral yaws (*Treponema pallidum pertenue*), and bejel (*Treponema pallidum endemicum*), which is not possible using serological tests (Harper et al., 2008).

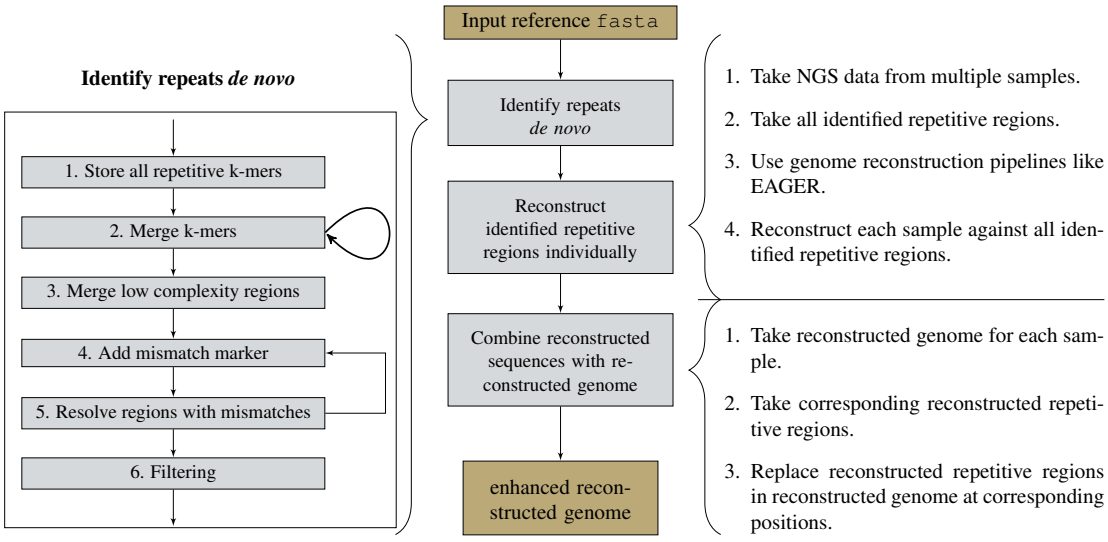

Figure 1. Workflow of the identification of repetitive sequences, is separated into six steps.

New sequencing technologies like the Illumina SLR platform, PacBio, or Oxford Nanopore are able to create long reads that can span most repetitive regions in order to resolve them (Huddleston et al., 2014). However, it is not always possible to apply these technologies to DNA samples. For example, in aDNA projects the average extracted fragment length is approximately 44 to 72 base pairs (Sawyer et al., 2012). Sequencing these fragments with long read technology would not result in any information gain. Additionally, the sequencing of hard to cultivate pathogens, like *Treponema pallidum*, also result in short DNA fragments similar to aDNA (Arora et al., 2016).

In order to better resolve the repetitive regions using short reads, researchers often first mask duplicated and low-complexity regions prior to the read mapping (Frith et al., 2010). For the human reference genome, for example, a masked reference is already available (UCSC, 2014). If no masked genome is available, programs like *RepeatMasker* (Smitt et al., 1996) can identify these regions and create a masked reference. *RepeatMasker* uses libraries of known repetitive regions and compares them to the input sequence. While this allows masking of the genome, *de novo* identification of repetitive regions is not possible.

A program that can identify repetitive regions *de novo* is `VMatch` (Kurtz, 2003). It uses suffix-arrays (Weiner, 1973) to identify the repetitive regions. `VMatch` has been applied in multiple genome projects for annotation of repetitive regions (e.g. by Lindow and Krogh (2005)), as well as masking tasks (Assuncao et al., 2010).

The general idea of our approach first starts with a *de novo* identification of all repetitive regions in a given reference genome. These identified regions, together with the full reference genome, are then used as separate references to reconstruct them for each individual sample, again using a mapping approach. The reconstructed repetitive regions can then be combined with the reconstruction of the full genome to increase the base-pair resolution of the reconstructed genome for each sample.

METHODS

The methodology of `DACCOR` (see Figure 1) first identifies all repetitive regions in a genome *de novo*. Each of these regions are then reconstructed individually, before they are combined with the rest of the reconstructed genome.

To create an integrated version to identify repetitive regions, our *de novo* approach uses a k-mer based approach, similar to WindowMasker (Morgulis et al., 2005). The workflow to identify these regions can be split up into six steps (see left part of Figure 1). In the first step, the reference genome is divided into its distinct k-mers and all k-mers that occur more than once are stored. In the second step, matching k-mer pairs overlapping at $(k-1)$ positions are combined into $(k+1)$-mers and stored again. This second step is repeated until all maximal unique repetitive regions are identified. Afterwards low complexity regions, e.g. long regions consisting only of one base, are identified. This step is necessary, because they are identified as two identical repetitive regions directly next to each other and can be combined

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3480v1 | CC BY 4.0 Open Access | rec: 20 Dec 2017, publ: 20 Dec 2017

into one region. In the following two steps, repetitive regions with mismatches are identified. Here, a mismatch-marker representing any unknown base is added to the end of all currently identified maximally exact repetitive regions. This representation allows us to combine two previously identified repetitive regions that are separated by these mismatches. Matching pairs of separated regions can then be combined into repetitive regions with mismatches. In the last step, leading and trailing mismatch markers are removed. All remaining mismatch markers are replaced with the character *N*.

DACCOR is not limited to this repeat finding approach. It fully supports the output format of Vmatch, meaning that the integrated repeat finding approach can be substituted with the result of a previous run of *Vmatch*. The regions identified by *Vmatch* are then extracted and saved in separate fasta files to be used in the next step.

The repetitive regions that are identified after the six steps are saved to a multi-fasta file. In order to be able to use each repetitive region directly as reference for a mapping pipeline, it is also possible to write each identified sequence to a separate file. Additionally a summary of all identified repetitive regions, as well as a bed file to view the location of the regions in a genome viewer are created. Users can specify a minimal length for all reported repeats.

The repeat regions can be used as separate references for the reconstruction of the individual repeat regions in multiple NGS sequencing samples in addition to the reconstruction of the full genome. To correctly reconstruct both ends of each respective repetitive region, a user-defined length for the flanking region on both the 3' and 5' end of each region is added, so that reads overlapping only part of the respective region can be mapped correctly.

In order to reconstruct the repetitive region from sequenced fastq files, we use the EAGER pipeline developed by Peltzer et al. (2016). It can preprocess the raw reads, including adapter clipping and quality trimming, map them against a given reference, and generate a consensus sequence in the fasta file format. For the consensus reconstruction, it uses the results of the genotyping results, following GATK BEST Practice's guidelines (Van der Auwera et al., 2013).

The *de novo* identification of repeats in a genome, the automatic separate mapping of NGS data against all repetitive regions, as well as the subsequent enhanced reconstruction of the genome of NGS samples has been implemented in DACCOR (Detection charACterization and reConstruction of Repetitive regions in genomes), a stand-alone program written in Java.

The repeat identification methodology, as described above, is implemented in the identify subprogram of our program and can be used to identify repetitive regions in a given reference genome *de novo*.

The reconstruct subprogram of DACCOR automatically generates EAGER configuration files for a given reference, its identified repetitive regions, and multiple sequencing samples.

Additionally, the combine subprogram can use the EAGER output of the reconstructed regions, as well as the EAGER output of the whole genome and combine these reconstructions. Because the origin of the reconstructed subsequences are known, they can replace the bases in the original reconstruction generated without specific repeat resolution. For each repetitive region, the respective positions in the genome are replaced if the original reconstruction resulted in an unknown (*N*) character.

To be able to automatically assemble a genome with all its repetitive regions, the subprogram pipeline first identifies all repetitive regions in a given reference. Afterwards the NGS samples are automatically reconstructed against both the complete reference and each identified repetitive region individually using the EAGER pipeline. Finally, these individual regions are combined with the reconstructed genome sequence to increase the resolution in repetitive regions.

The identify subprogram can identify repetitive regions within as well as between different chromosomes or a bacterial genome and its plasmids. This is done by combining the identified *k*-mers of all sequences in a given multi-fasta reference file. To be able to match identified repetitive regions to the corresponding sequence, a unique offset is added to the indices of the start location of each region.

To evaluate our method, we applied DACCOR to several bacterial genomes of various lengths and repetitiveness. We first compared the step of repeat identification with VMatch (Kurtz, 2003), allowing for up to five mismatches in repetitive regions of a minimum length of 101 base-pairs (the length typical Illumina HiSeq reads). We then applied our proposed reconstruction method to the syphilis samples published by Arora et al. (2016), Pinto et al. (2016), and Sun et al. (2016) to reconstruct the sequences of the 16S and 23S rRNA, which are duplicated in the bacterium *Treponema pallidum*. For this, we first used DACCOR to identify all repetitive regions with at most five mismatches and a minimal length

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3480v1 | CC BY 4.0 Open Access | rec: 20 Dec 2017, publ: 20 Dec 2017

Table 1. Comparison of the identified repetitive positions in different bacterial genomes of DACCOR and VMatch, which was seen as the golden truth to evaluate against. Both programs were run allowing for one mismatch and reporting only regions of at least 101 bp. The *k*-mer size for DACCOR was set to 17.

	T. pallidum	*S. flexneri*	*E. coli*	*M. leprae*
true positives	22,382	376,669	107,456	74,406
true negatives	1,116,478	4,211,322	4,535,106	3,192,194
false positives	0	354	463	131
false negatives	773	18 857	3 307	1 472
accuracy (%)	99.93	99.58	99.92	99.95

of 101 base pairs in the *Nichols* strain. We also reconstructed the full genomes of the samples using the standard EAGER pipeline. This allowed us to compare our reconstruction of the two genes to the reconstruction generated by the standard method using the full genome as a reference. In order to identify specific variations in either copy of the genes, we searched for heterozygous positions in the individual reconstructions of the extracted sequences that show an allele frequency between 25 and 75%.

Finally we applied the full DACCOR pipeline to reconstruct the whole genomes, including the identified repetitive regions, of two syphilis samples. For this we chose two samples from Arora et al. (2016), one with a moderate coverage (*AR1*, 7X) and one with a very high coverage (*AR2*, 157X).

RESULTS

We first evaluated the identification of repetitive regions by comparing these to the repeats identified by VMatch. This comparison (see Table 1) shows that the results of both programs are almost identical. We considered VMatch as the "golden truth" and could therefore compute an accuracy for the repetitive regions reported by DACCOR. We achieved a very good accuracy with over 99% in all tested cases with only very few false negatives as well as false positives.

Next we reran DACCOR with a higher sensitivity allowing for up to five mismatches in repetitive regions of lengths at least 101 base pairs. These results of the different bacterial genomes (see Table 2) show that the *Shigella* genome contains by far the most repetitive regions (1,249 compared to 29 in *Nichols*, 242 in *E. coli*, and 190 in *M. leprae*). It also contains the longest repetitive regions of the four bacteria (5,383 compared to 3,283, 3,141, and 2,578). The average lengths of the repetitive regions are quite similar in all four bacteria (between 570 and 823 base-pairs). The number of repetitive regions that can be identified when allowing for up to five mismatches also varies between the different bacteria. There are 127 of regions containing mismatches in the *Shigella* genome, whereas there are only 47 in the genome of *E. coli*, 9 in *M. leprae*, and 1 in *Nichols*. Overall 8.3% of the *Shigella* genome is comprised of repetitive regions. The genomes of *E. coli* and *M. lepra* are comprised of 2.4% and 2.3% repetitive regions respectively, and only 2.0% of the *Nichols* genome is repetitive.

The runtime linearly correlates with the number of identified repetitive bases (see Figure 2). After an initial preprocessing step, DACCOR identifies about 12,000 repetitive bases per minute. The *k*-mer size

Table 2. Statistics of identified repetitive regions using several bacterial genomes for a *k*-mer size of 17, at most five mismatches, and a minimum length of 101 base pairs.

	T. pallidum	*S. flexneri*	*E. coli*	*M. leprae*
genome size [bp]	1,139,633	4,607,202	4,646,332	3,268,203
# repetitive regions	29	1,249	242	190
different repetitive regions	14	482	104	76
max length of repetitive regions	3,283	5,383	3.141	2,578
average length	823	570	706	643
repetitive regions with mismatches	1	127	47	9
sum of repetitive bases	23,892	660,261	160,897	119,871
% of genome repetitive (non overlapping)	2.0	8.3	2.4	2.3

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3480v1 | CC BY 4.0 Open Access | rec: 20 Dec 2017, publ: 20 Dec 2017

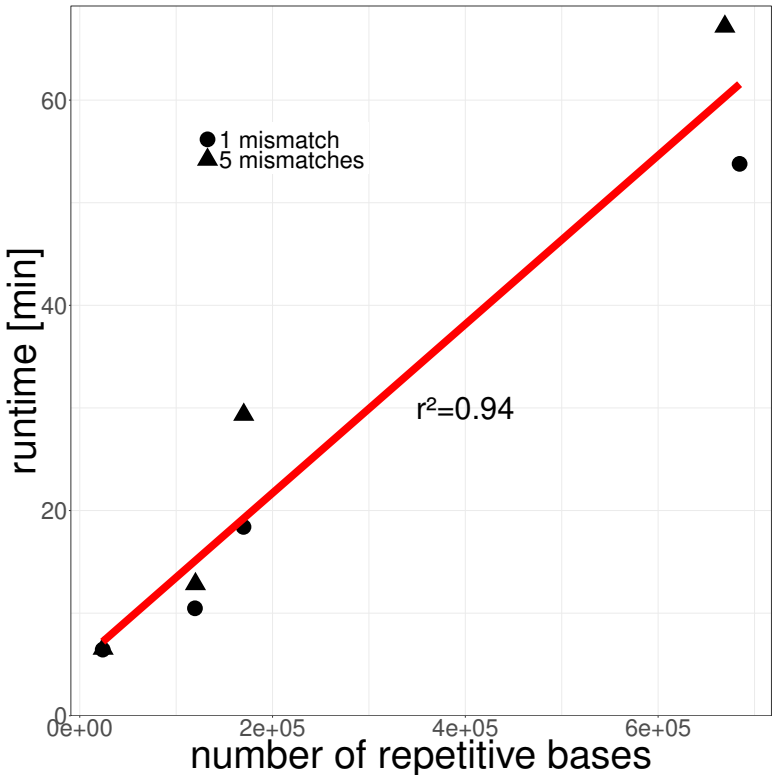

Figure 2. Runtime of DACCOR in relation to the number of identified repetitive bases.

does not influence the runtime (see supplementary Figure 2).

The two longest identified repetitive regions in the *Nichols* genome correspond to the 16S and 23S rRNAs, respectively. Both operons contain two copies of the gene. We extracted these regions and used

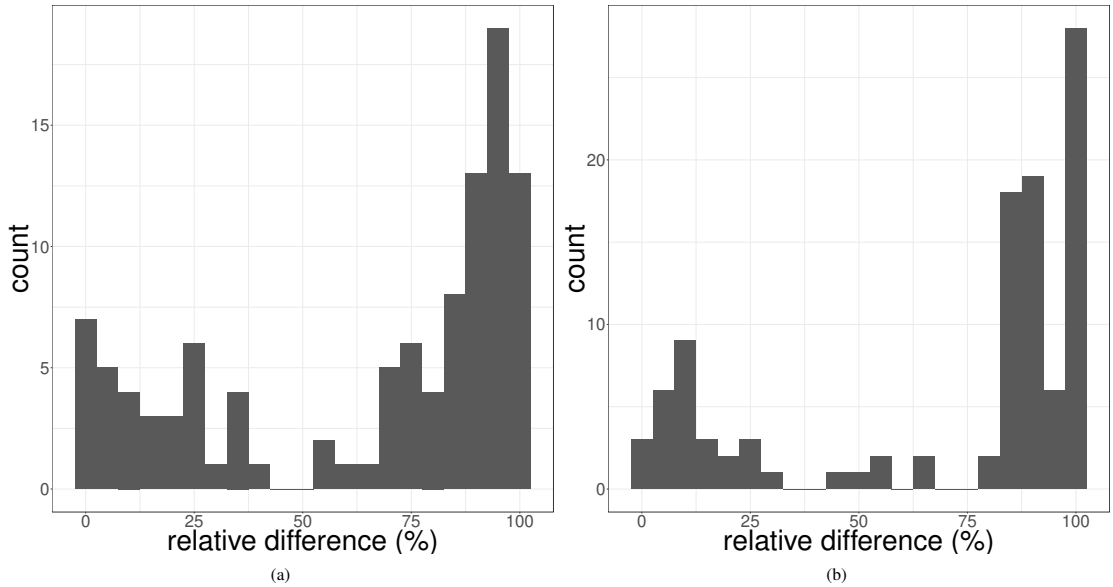

Figure 3. Histogram of resolved bases (as fraction relative to the length of the respective gene) of the 16S rRNA (a) and 23S rRNA (b) in 106 clinical syphilis samples (Arora et al., 2016; Pinto et al., 2016; Sun et al., 2016). The number of resolved base-pairs when mapping against each copy of the gene individually in comparison to the standard mapping approach using both copies has been computed.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3480v1 | CC BY 4.0 Open Access | rec: 20 Dec 2017, publ: 20 Dec 2017

Table 3. SNPs in the 16S and 23S rRNAs identified after extracting the repetitive region. *Site-specific* positions refers to positions that appear to be different in the two copies of the respective gene (allele frequency between 25 and 75%).

	16S rRNA	23S rRNA
length of gene	1495	2900
number of variant positions	36	46
number of site-specific positions	30	43

them as two independent references for their reconstruction off all samples published by Arora et al. (2016), Pinto et al. (2016), and Sun et al. (2016). We mapped against each copy of the gene individually and count the number of resolved bases in each copy. This number was compared to the number of resolved bases in the two copies of the respective gene when mapping against the whole genome without repeat masking. We then computed the difference of these two numbers and divided by the length of the gene. We could improve the base-pair resolution (see Figure 3) by a median value of 82.7% for the 16S and 87.4% for the 23S rRNA. It shows that the percentage of the resolved base-pairs was at least as high when mapping only against the extracted sequences, compared to the mapping against the whole genome for all analyzed samples. This means that we do not lose resolution when mapping only against these sequences, but in almost all cases gain information up to an improvement of 100% of the sequence of the respective gene.

Using the individual reconstructions of the 16S rRNA and the 23S rRNA gene sequences, we tried to identify SNPs that are specific for only one copy of the respective gene. The results of this analysis identified 36 positions that have a SNP call in at least one of the 106 samples for the 16S rRNA (see Table 3). Of those 36 positions, 30 show evidence for a site-specific SNP in at least one of all analyzed clinical samples. There are two positions (884 and 888 relative to the start of the 16S rRNA), which show site-specific variance in about 20% of the samples (see supplementary material). In the 23S rRNA there are 46 positions where at least one sample has a SNP. Of these 46, 43 show evidence for site-specific SNPs in at least one sample. Here, one position (2003) shows evidence for site-specificity in 37% of the samples. Additionally, there are nine positions that show site-specificity in at least 15% of the samples.

Finally, we compared the total number of unresolved positions in the samples *AR1* and *AR2*, published by Arora et al. (2016), between DACCOR and EAGER with the full genome as reference (see Table 4). On the sample *AR1*, the approach using the standard mapping approach without repeat resolution resulted in 23,348 unresolved bases, compared to the 4,473 unresolved bases using our enhanced repeat resolution approach. This means that using DACCOR, 82.81% of the repetitive regions are resolved, compared to the 10.27% without DACCOR. For the high coverage sample *AR2*, the number of unresolved repetitive bases could be decreased by 16,585 from 17,549 to 964 bases. As a result, 96.3% of the repetitive bases could be resolved, compared to the 82.8% using the standard mapping approach.

DISCUSSION

We have developed DACCOR, an approach to increase the base-pair resolution of repetitive regions in the reconstruction of full genomes using short reads. For this we first identify the repetitive regions *de novo*. These regions are then used as individual references for mapping short reads of NGS samples. Finally, a

Table 4. Enhanced genome resolution of two clinical syphilis samples (*AR1* and *AR2* from Arora et al. (2016)). EAGER indicate the results using only the full genome as a reference without the extra repeat resolution of DACCOR. The values refer to the repetitive regions only, including the margin regions (in total 30330 bp).

	AR1		AR2	
	EAGER	DACCOR	EAGER	DACCOR
#N	23,348	4,473	17,549	964
%N	76.98	14.75	57.86	3.18

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3480v1 | CC BY 4.0 Open Access | rec: 20 Dec 2017, publ: 20 Dec 2017

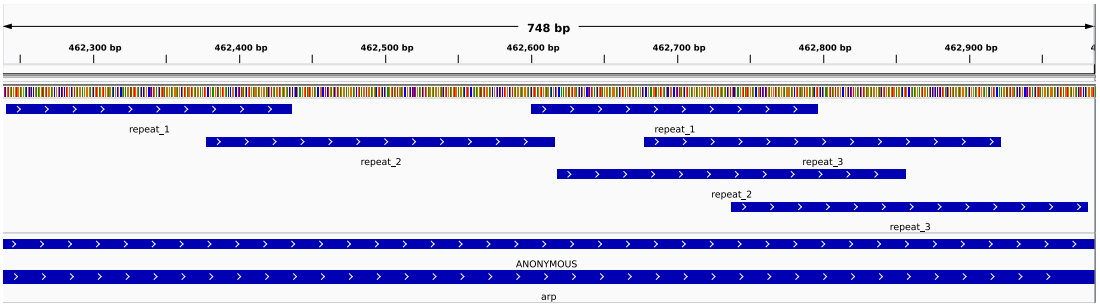

Figure 4. Repetitive regions contained in the *arp* gene in *Treponema pallidum*

new draft genome is created by combining the reconstructed repeat regions with the rest of the genome that has been reconstructed using a standard mapping approach.

Our de novo identification of repeats uses a *k*-mer approach. The choice of *k* is important for the identification of repeat regions. For *de novo* assembly based on de Bruijn graphs, the optimum depends on the genome length, the coverage, the quality, and the length of the reads (Zerbino and Birney, 2008). In the case of different read lengths, as often observed in aDNA projects, it has been shown that using multiple different *k*-mers improves the assembly (Seitz and Nieselt, 2017). Other approaches choose an optimal *k*-mer size so that the uniqueness is maximized (Gardner and Hall, 2013). However, as we want to identify repetitive regions, we do not want to use a *k*-mer size in order to maximize uniqueness. In our approach the *k*-mer size defines the minimum length of the repetitive regions that can be identified in the first step. Thus, a *k*-mer size as small as possible should be used to be able to identify all putative repeat regions. However, very small *k*-mer sizes lead to an exponential increase in the runtime (see supplementary Figure 2), due to the increasing number of random occurrences of these small *k*-mers that have to be accounted for. Therefore, we propose a minimum size for *k* of 17. On the other hand, the *k*-mer size should also not be longer than half the length of the input reads.

The runtime of `DACCOR` correlates mainly with the repetitiveness of the genome, i.e. with the number of repetitive bases that are identified in the respective genome. When allowing mismatches, a slight increase in runtime is also noted.

To identify all repetitive *k*-mers, all possible *k*-mers are stored in the first step of the `identify` subprogram of `DACCOR`. Those that are not repetitive are removed after the screening. Nonetheless, this results in high memory usage if the genome contains many repetitive regions. For the example of *Shigella flexneri*, we observed a memory footprint of 8 GB with a *k*-mer size of 17.

The comparison between `DACCOR` and `VMatch` showed that `VMatch` is slightly more sensitive, probably due to its suffix array approach. A possibility to increase the sensitivity of `DACCOR` would be to elongate all identified repetitive regions based on a local alignment.

When mapping short reads from typical NGS data, a number of approaches recommend to use genomes as references whose repetitive regions are masked (Tarailo-Graovac and Chen, 2009). However, this may be problematic, because repetitive regions often overlap and can be quite complex. An example for this is the *arp* gene of *Treponema pallidum* (see Figure 4). It contains several overlapping repetitive regions. It can be seen that the region labeled *repeat_3* is partly repetitive with itself. Thus a masking of one of the repetitive regions would mask most of itself. Additionally, the masking of the second occurrence of *repeat_1* would also mask most of both occurrences of *repeat_3*. When masking the first occurrence of *repeat_1*, the masking of either occurrence of *repeat_2* would also mask part of the unmasked region of *repeat_1*. Thus, masking of repetitive regions could result in either loosing genome information or leaving repetitive regions unmasked. We therefore propose to use the identified repetitive regions as separate references for the mapping, and merge all individually reconstructed regions into a common draft genome.

Since the *de novo* identification of repeats in genomes, using these then as individual references for the mapping and merging all reconstructed genomic regions into a final draft genome, require many different steps, our main goal of `DACCOR` was to present a fully automatic procedure encompassing all these steps. `DACCOR` makes use of `EAGER`, a pipeline for the automatic reconstruction of genomic data sets. For highly identical repeat regions, each copy is stored as individual sequences together with a margin at the 5' and 3' end of the region. We propose to set the margin to twice the maximum read length to cover all

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3480v1 | CC BY 4.0 Open Access | rec: 20 Dec 2017, publ: 20 Dec 2017

reads spanning over the ends of the repeats.

Using DACCOR, we have shown that a higher base-pair resolution, compared to the reconstruction using the standard mapping approach in repeat regions, can be achieved. In the standard approach, reads that can be mapped to different locations result in a mapping quality of zero, which in turn decreases the genotyping quality (McKenna et al., 2010). By mapping to the repeat region only, these reads have a higher mapping quality, a higher genotype quality, and thus result in more resolved positions. However, one has to acknowledge that other unresolved bases, that may be due to lack of coverage or low sequencing quality, can not be resolved with our approach. In the case of the two syphilis samples, we could show that the majority of unresolved bases in repeat regions could be resolved, and that the remaining unresolved bases lie outside of the identified repetitive regions. Furthermore, we have shown that using DACCOR, the identification of SNPs in repeat regions can be improved. This is especially useful for the 23S rRNA, as it is known to play a role in the antibiotics resistance of bacteria (Arora et al., 2016).

In conclusion, we have developed a fully automatic pipeline that first conducts a *de novo* repeat identification in bacterial genomes and then uses the repeat regions for an enhanced mapping of short read NGS data. Increasing the resolution of a draft genome has an effect on many downstream analyses, such as population genetics or phylogenetic analyses. For future improvements we plan to reduce the runtime and memory usage by adjusting our data structure and by adding more parallelization to some of the compute steps. With this we hope to eventually be able to identify repetitive regions also in large eukaryotic genomes, like the human genome.

SOFTWARE AVAILABILITY

We have developed an automated software pipeline, written in Java, which allows other researchers to use our methodology. This pipeline is available on github:
`https://github.com/Integrative-Transcriptomics/Daccor`

ACKNOWLEDGMENTS

We received support from Deutsche Forschungsgemeinschaft and the Open Access Publishing Fund of University of Tübingen. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Arora, N., Schuenemann, V. J., Jäger, G., Peltzer, A., Seitz, A., Herbig, A., Strouhal, M., Grillová, L., Sánchez-Busó, L., Kühnert, D., Bos, K. I., Davis, L. R., Mikalová, L., Bruisten, S., Komericki, P., French, P., Grant, P. R., Pando, M. A., Vaulet, L. G., Fermepin, M. R., Martinez, A., Centurion Lara, A., Giacani, L., Norris, S. J., Šmajs, D., Bosshard, P. P., González-Candelas, F., Nieselt, K., Krause, J., and Bagheri, H. C. (2016). Origin of modern syphilis and emergence of a pandemic Treponema pallidum cluster. *Nature Microbiology*, 2(December):16245.

Assuncao, A. G. L., Herrero, E., Lin, Y.-F., Huettel, B., Talukdar, S., Smaczniak, C., Immink, R. G. H., van Eldik, M., Fiers, M., Schat, H., and Aarts, M. G. M. (2010). Arabidopsis thaliana transcription factors bZIP19 and bZIP23 regulate the adaptation to zinc deficiency. *Proceedings of the National Academy of Sciences*, 107(22):10296–10301.

Bos, K. I., Herbig, A., Sahl, J., Waglechner, N., Fourment, M., Forrest, S. A., Klunk, J., Schuenemann, V. J., Poinar, D., Kuch, M., Golding, G. B., Dutour, O., Keim, P., Wagner, D. M., Holmes, E. C., Krause, J., and Poinar, H. N. (2016). Eighteenth century Yersinia pestis genomes reveal the long-term persistence of an historical plague focus. *eLife*, 5(JANUARY2016):1–11.

Denoeud, F. and Vergnaud, G. (2004). Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a web-based resource. *BMC bioinformatics*, 5:4.

Frith, M. C., Hamada, M., and Horton, P. (2010). Parameters for accurate genome alignment. *BMC bioinformatics*, 11(1):80.

Gardner, S. N. and Hall, B. G. (2013). When whole-genome alignments just won't work: KSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. *PLoS ONE*, 8(12).

Harper, K. N., Liu, H., Ocampo, P. S., Steiner, B. M., Martin, A., Levert, K., Wang, D., Sutton, M., and Armelagos, G. J. (2008). The sequence of the acidic repeat protein (arp) gene differentiates

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3480v1 | CC BY 4.0 Open Access | rec: 20 Dec 2017, publ: 20 Dec 2017

venereal from nonvenereal Treponema pallidum subspecies, and the gene has evolved under strong positive selection in the subspecies that causes syphilis. *FEMS Immunology and Medical Microbiology*, 53(3):322–332.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P. H., Graves, T. a., Alkan, C., Dennis, M. Y., Wilson, R. K., Turner, S. W., Korlach, J., and Eichler, E. E. (2014). Reconstructing complex regions of genomes using long-read sequencing technology Reconstructing complex regions of genomes using long-read sequencing technology. *Genome research*, 7(11):688–696.

Kurtz, S. (2003). The Vmatch large scale sequence analysis software. *Ref Type: Computer Program*, 412.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

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

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858.

Lindow, M. and Krogh, A. (2005). Computational evidence for hundreds of non-conserved plant microRNAs. *BMC genomics*, 6:119.

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

Morgulis, A., Gertz, E. M., Sch{\"a}ffer, A. A., and Agarwala, R. (2005). WindowMasker : window-based masker for sequenced genomes. *Bioinformatics*, 22(2):134–141.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3480v1 | CC BY 4.0 Open Access | rec: 20 Dec 2017, publ: 20 Dec 2017

Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., and Nieselt, K. (2016). EAGER: Efficient Ancient Genome Reconstruction. *Genome Biology*, 17(1):60.

Pinto, M., Borges, V., Antelo, M., Pinheiro, M., Nunes, A., Azevedo, J., Borrego, M. J., Mendonça, J., Carpinteiro, D., Vieira, L., and Gomes, J. P. (2016). Genome-scale analysis of the non-cultivable Treponema pallidum reveals extensive within-patient genetic variation. *Nature microbiology*, 2(October):16190.

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE*, 7(3).

Seitz, A. and Nieselt, K. (2017). Improving ancient DNA genome assembly. *PeerJ*, 5:e3126.

Shapiro, J. A. and von Sternberg, R. (2005). Why repetitive DNA is essential to genome function. *Biological reviews of the Cambridge Philosophical Society*, 80(2):227–50.

Smith, A. D., Xuan, Z., and Zhang, M. Q. (2008). Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics*, 9(1):128.

Smitt, A., Hubley, R., and Green, P. (1996). RepeatMasker Open-3.0. http://www.repeatmasker.org.

Sun, J., Meng, Z., Wu, K., Liu, B., Zhang, S., Liu, Y., Wang, Y., Zheng, H., Huang, J., and Zhou, P. (2016). Tracing the origin of Treponema pallidum in China using next-generation sequencing. *Oncotarget*, 7(28).

Tarailo-Graovac, M. and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, (SUPPL. 25):1–14.

Treangen, T. J., Abraham, A. L., Touchon, M., and Rocha, E. P. (2009). Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiology Reviews*, 33(3):539–571.

UCSC (2014). HGDownload. *http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/*, (accessed: 2017-12-19).

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and DePristo, M. A. (2013). *From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline*. Number SUPL.43.

Veeramah, K. R. and Hammer, M. F. (2014). The impact of whole-genome sequencing on the reconstruction of human population history. *Nat Rev Genet*, 15(3):149–162.

Weiner, P. (1973). Linear pattern matching algorithms. *Switching and Automata Theory, 1973. SWAT '08. IEEE Conference Record of 14th Annual Symposium on*, pages 1–11.

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3480v1 | CC BY 4.0 Open Access | rec: 20 Dec 2017, publ: 20 Dec 2017