W560–W566 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki356

Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies

Sacha A. F. T. van Hijum*, Aldert L. Zomer, Oscar P. Kuipers and Jan Kok

Department of Molecular Genetics, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, PO Box 14, 9750 AA Haren, The Netherlands

Received December 14, 2004; Revised and Accepted January 27, 2005

ABSTRACT

With genome sequencing efforts increasing exponentially, valuable information accumulates on genomic content of the various organisms sequenced. Projector 2 uses (un)finished genomic sequences of an organism as a template to infer linkage information for a genome sequence assembly of a related organism being sequenced. The remaining gaps between contigs for which no linkage information is present can subsequently be closed with direct PCR strategies. Compared with other implementations, Projector 2 has several distinctive features: a user-friendly web interface, automatic removal of repetitive elements (repeat-masking) and automated primer design for gap-closure purposes. Moreover, when using multiple fragments of a template genome, primers for multiplex PCR strategies can also be designed. Primer design takes into account that, in many cases, contig ends contain unreliable DNA sequences and repetitive sequences. Closing the remaining gaps in prokaryotic genome sequence assemblies is thereby made very efficient and virtually effortless. We demonstrate that the use of single or multiple fragments of a template genome (i.e. unfinished genome sequences) in combination with repeat-masking results in mapping success rates close to 100%. The web interface is freely accessible at http://molgen.biol.rug.nl/websoftware/projector2.

INTRODUCTION

Sequencing of a genome often starts with a random shotgun sequencing strategy (1) or, as shown more recently, with direct sequencing on genomic DNA (http://www.fidelitysystems.com/). The DNA sequences of the clones or sequenced genome fragments often overlap, yielding enlarged DNA sequences (contigs). These contigs can subsequently be positioned on larger genomic fragments in large-insert genomic libraries, for instance from phage and cosmid banks or bacterial artificial chromosomes, using various techniques (2–4). Thus, gaps between these linked contigs can be closed by PCR or cloning strategies. Often the final closing of physical gaps between contigs is a time- and money-consuming phase in any genome sequencing effort (5).

Linkage information for contigs can be derived from the genomic sequences of related organisms. As new genome sequences are released on a weekly basis, the chance increases for the matching of an unfinished genome with a related genome. Software packages, such as Projector (6), MGview (7) and MUMmer (8), have been developed to order contigs using a template genome. MGview requires a hardware implementation of BLAST (http://www.timelogic.com/) rather than the freely available NCBI BLAST software (9). The algorithms used by MUMmer and MGview result in positioning of target contigs at multiple places on the template genome, requiring contig mappings to be inspected manually. Since these software implementations do not allow automatic repeat-masking or automatic ordering of contigs, subsequent primer design for gap-closing purposes has to be performed manually.

In contrast to MUMmer and MGview, the Projector mapping algorithm can handle up to at least 40% of size difference between the target and template contigs (6). This feature allows taking genomic insertions and deletions into consideration and makes it possible to position contigs of the genome being sequenced on less related genomes. Projector 2 will select the best positions for the contigs, based on a set of experimentally validated rules (6). An added benefit of using Projector 2 is the automated selection of, and primer design on, the sequences from the ends of the contigs. This task can become very tedious when it has to be performed manually, especially when dealing with a large number of contigs. As the contig ends used for primer design often

*To whom correspondence should be addressed. Tel: +31 50 3632092; Fax: +31 50 3632348; Email: S.A.F.T.van.Hijum@rug.nl

contain unreliable sequences, gap-closing primer design is not a trivial task.

The Projector algorithm proved to be very efficient: over 90% success rate was obtained in gap-closing PCRs of *Lactococcus lactis* MG1363 contigs (target) mapped onto the genome of *L.lactis* IL1403 (template) (6). Projector 2 uses the same mapping algorithm as Projector and has now various additional features: (i) a web interface, (ii) automatic filtering of repetitive sequences (repeat-masking) to minimize incorrect mapping events, (iii) ready-to-use publicly available prokaryotic genome sequences, (iv) automated primer design for gap closure by direct PCR and multiplex PCR strategies [see (v)], and (v) the possibility of positioning target contigs on user-supplied (multiple) template chromosomes or contigs. The latter possibility involves mapping the contigs onto the fragments created from the template sequences. The use of this feature allows mapping onto genomes that are less coding-dense. Chromosomes and large contigs from unfinished genome sequences can thus be used as templates, further increasing the chance to match the genome being sequenced with a closely related genome of which the sequence is (being) finished.

The power of these methods is demonstrated by using the mapping of a genome sequence assembly on its finished counterpart and on a genome sequence of a closely related organism. We show that mapping onto the repeat-masked single or multiple fragments of a template genome of various organisms results in estimated success rates of gap-closing PCRs close to 100%.

IMPLEMENTATION

System requirements and the web interface

Projector 2 runs on a UNIX platform. Linux Fedora Core 1 (http://fedora.redhat.com) is used as the operating system. It requires a locally installed copy of the BLAST software (2.2.9) (http://genome.nhgri.nih.gov/blastall/) and an Apache web server (2.0.48) with PHP (4.3.4). Projector 2 consists of 14 sub-programs written in Pascal and compiled by Free-Pascal, version 1.0.10 (http://www.freepascal.org/). These programs are linked by a shell script. The web interface consists of three parts, such as (i) upload DNA sequence(s), (ii) select desired settings and (iii) the status page, which presents the mapping results after a successful run. Each run is assigned a session id, which allows the user to inspect the run status or the results at any given time.

Input

Target contigs need to be uploaded in the FASTA format. As template, the user can either (i) use preformatted and, optionally, repeat-masked (i.e. repetitive elements removed, see below) finished prokaryotic genome sequences (Figure 1, left-hand side) or (ii) upload FASTA formatted contig(s) or chromosome(s) (Figure 1, right-hand side).

The mapping procedure

Figure 1 presents a flow scheme of Projector 2. For multiple fragments of a template genome, each of the sequences meeting the minimum size criterion (*n* in Figure 1; in this study: 100 000 bp) is selected. The selected templates are (i) reduced to fragments (1000 bp fragments were used), which most likely, contain parts of open reading frames (ORFs), and (ii) used in a separate mapping procedure with the target contigs (*n* times; Figure 1). The contigs of the organism being sequenced that meet the minimum size criterion (1200 bp was used in this study) are selected. From these contigs, fragments (300 bp in this study) were taken from either side (5′ are designated as L fragments and 3′ end as R fragments) going from the outside to the inside of the contig DNA sequences. These L and R fragments were aligned to the template DNA sequence using BLAST (9), resulting in aligned positions (mapped positions). This combination of L and R fragments was selected whose mapped positions resemble their actual positions on the contig as closely as possible, based on three criteria, i.e. order, orientation and spacing (Figure 1) (6).

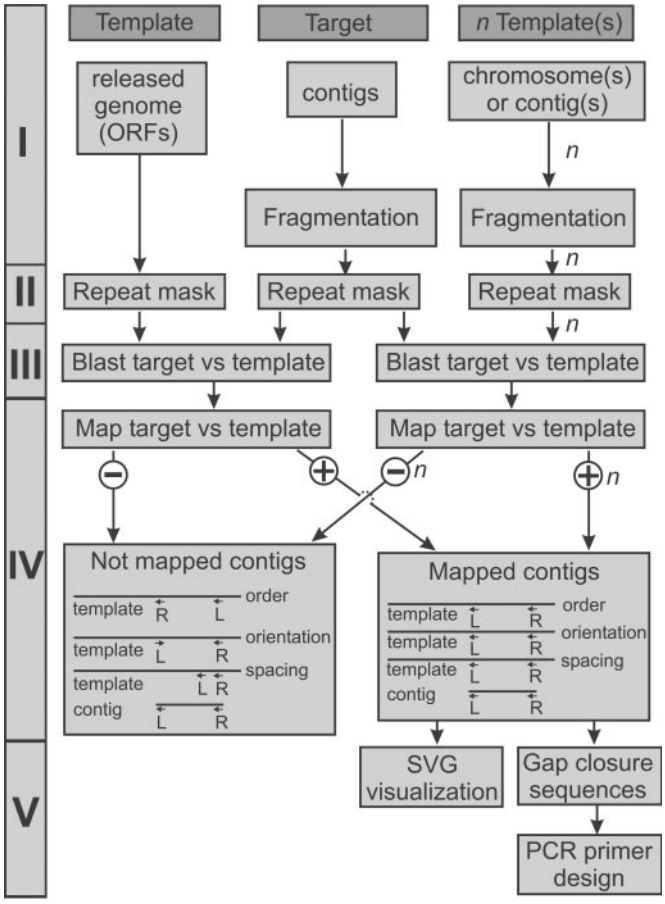

Figure 1. The Projector 2 procedure. From top to bottom: **I**, single (left) or fragmented *n*-multiple templates (right) are optionally repeat masked; **II**, contigs (middle) are fragmented and also optionally repeat masked; **III**, the (unique) contig fragments are compared against the (unique) template fragments using BLAST; yielding **IV**, mapped contigs. Arrows with a plus sign (+) signify contigs that were successfully mapped; and those with a minus (−) sign could not be mapped. **V**, For the mapped contigs, gap-closing sequences and PCR primers are designed and a visual SVG output is generated.

Repeat-masking

The mapping procedure has of an optional step in which repeat-masking is performed. In this step, repetitive DNA

sequences are removed: (i) the target or template DNA sequences are fragmented (in this study, target sequences of 300 bp and template sequences of 1000 bp); (ii) these fragments are aligned to all the target or template sequences, respectively, using BLAST; and (iii) repetitive sequences are removed, i.e. fragments with a significant BLAST hit (an *E*-value of 1×10^{-20} was used) to another fragment. We recommend performing this step for the target as well as the template sequences, as both may contain repetitive sequences (see below). The repetitive DNA sequences identified by Projector 2 can be retrieved and be used to correctly re-assemble the target genome assembly.

Primer design

The automated mapping and subsequent primer design is the most important feature that distinguishes Projector 2 from other software implementations. In the mapping procedure, the relative position and orientation of the contigs is determined, allowing Projector 2 to design PCR primers on the contig ends for gap-closure purposes. For a circular genome sequence, primers are also designed to close the gap between contigs that are on either side of the origin of replication. The contigs ends of in a genome sequence assembly often contain repetitive DNA sequences and such ambiguous DNA sequences (e.g. phage DNA, IS elements or gene duplications) cannot be used for assembly. The number of contigs with repetitive elements on their ends depends on the organism, but can be quite considerable (Table 1). Projector 2 removes these repetitive elements prior to primer design, thereby greatly reducing mis-priming events of the primers in subsequent gap-closing PCRs. In addition, sequence redundancy in contig ends is generally lower and in some cases these ends contain a high percentage of G and C residues (Table 1). Projector 2 also skips regions with a G+C-content >75% or <25%, allowing the design of gap-closing PCR primer pairs that are more balanced in G+C-content.

Primer design for multiplex PCR purposes

In the case that multiple DNA sequences are used as template, multiplex PCR primers are designed to close gaps between contigs that are mapped onto the edges of different template sequences. Obtaining the desired gap-closing PCR products in multiplex PCRs will only be feasible when a limited number of template DNA sequences are used (this study; 10). In the case of a large number of template sequences, one could (i) perform multiple multiplex PCRs with a limited number of multiplex PCR primers or (ii) perform a separate mapping procedure with the template sequences on another related genome sequence to obtain additional linkage information for the template DNA sequences.

Output

After a successful mapping run, a web page is constructed that gives an overview of the mapping results. For each of the template sequences (*n* or 1 for a finished genome sequence; Figure 1), a scalable vector graphics map (SVG) (http://www.w3.org/Graphics/SVG/Overview.html) is produced. The use of SVG has several advantages: the graphics are scalable, can be viewed on any operating system platform and can be embedded into a web page. The user determines the size of the SVG map, the font size, and whether or not to show contig names. The SVG map provides valuable additional information, for example, it shows the L and R fragments used for the positioning of the contigs and their BLAST hits. Tables with two types of gap-closing primers designed by Primer3 (10) are provided: (i) primers that allow gaps to be closed by direct PCR strategies and (ii) primers that allow gaps between multiple fragments of a template genome to be closed by multiplex PCR strategies. The user is notified if no primer could be designed on a contig end. Selection of an alternative primer pair is facilitated by inspecting the gap-flanking sequences table, which contains the DNA sequences of the contig ends that can be used for primer design by software other than Primer3, e.g. in our case GenomePrimer (11). All files generated in the mapping procedure can be retrieved for future reference.

Sources of genome sequence data

The genome sequence of *L.lactis* MG1363 has been sequenced and is being annotated in a consortium consisting of the Microbiology Department (University College, Cork, Ireland), the Institute of Food Research (Norwich, UK) and the Molecular Genetics Department (University of Groningen, The Netherlands). Unfinished genomic sequences for *Nitrosomonas europaea* ATCC 19718 (sequence assembly of August 24, 2000) and *Rickettsia typhi* str. wilmington (sequence assembly of July 15, 2002) were obtained from the US DOE Joint Genome Institute (http://www.jgi.doe.gov). Preliminary sequence data for *Mycobacterium tuberculosis* CDC1551 (sequence assembly of September 7, 2002) was obtained

Table 1. The number of contigs containing repetitive elements and a high G+C-content on either end was determined for the four sequence assemblies shown (see Table 2 for details)

Genome sequence assembly origin	Total number of contigs in sequence assembly	Total number of contigs used for mapping[a]	Number of contig ends containing repeats[b]			Number of contigs ends with high G+C-content[b]		
			One	Both	%[c]	One	Both	%[c]
L.lactis (A)	210	131	34	12	35	6	1	5
M.tuberculosum (A)	516	491	176	25	41	0	30	6
N.europaea	477	409	31	12	11	2	4	1
R.typhi (A)	154	108	49	50	92	1	27	26

The genome sequence assemblies used were *L.lactis* MG1363 (A), *M.tuberculosis* CDC1551 (A), *N.europaea* ATCC 19718 and *R.typhi* str. wilmington (A).
[a]The contigs used for mapping were >1200 bp.
[b]On one or both ends of a contig.
[c]The number of contigs with at least one end that contains repeats, or with a high G+C-content, divided by the total number of contigs used for mapping × 100%.

from The Institute for Genomic Research (http://www.tigr.org). Finished genome sequences were obtained from the NCBI genome database (http://www.ncbi.nlm.nih.gov/).

RESULTS AND DISCUSSION

Reference mappings are used to validate mapping results

Four different prokaryotic genome sequence assemblies were mapped onto their isogenic genome sequences as a control test of the performance of Projector 2. For this purpose, similar template genome sequences were available for three of the four genome sequence assemblies (Table 2). The success of ordering contigs of an unfinished genome is highly dependent on the similarity of the template genome. As we will demonstrate below, the best mapping results are obtained when a repeat mask is applied to both target and template sequences and when mapping is performed on a single template genome. For this reason, the mapping result using a single isogenic repeat-masked template sequence was taken as a reference. These reference mappings were verified by using BLAST and MUMmer (8) (data not shown).

Contig ends often contain unreliable sequences

There are large differences in the completion stage of the four genome sequence assemblies used in this study. This is indicated by differences in the total number of contigs as well as the number of small contigs present in these assemblies (Table 1). Approximately 25% of the *Rickettsia typhi* contigs have a high G+C-content on one or both ends (Table 1). In addition, the *R.typhi* genome sequence contains many repetitive elements, which is reflected by the very high number of contigs with repetitive sequences on one or both ends (Table 1). These examples clearly illustrate the necessity to perform gap-closing primer design on high-quality and non-repetitive DNA sequences.

The prediction of successful gap-closure PCRs is accurate

To estimate the number of successful gap-closing PCRs, we determined the number of gaps <15 kb (Table 2). Such a gap size should be straightforward to close by direct PCR. The percentage of gaps that can be closed by direct PCR methods never reaches 100% because some gaps in the genome sequence assembly are >15 kb. The number of these large

Table 2. Mapping results for unfinished genome assemblies on their isogenic counterparts and related genomes

Target genome	Template genome	Number of templates	Repeat mask performed	Contigs correct mapped	Total mapped contigs	Gaps <15 kb	% Gaps PCR[a]
L.lactis A	***L.lactis A***	**1**	+	**106**	**106**	**102**	**96**
L.lactis A	*L.lactis A*	1	–	106	125	83	66
L.lactis A	*L.lactis A*	10	+	106	109	99	91
L.lactis A	*L.lactis A*	10	–	106	163	69	42
L.lactis A	*L.lactis B*[b]	1	+	82	85	79	93
L.lactis A	*L.lactis B*[b]	1	–	82	90	74	82
L.lactis A	*L.lactis B*[b]	10	+	80	86	74	86
L.lactis A	*L.lactis B*[b]	10	–	81	116	61	53
M.tuberculosis A	***M.tuberculosis A***	**1**	+	**436**	**436**	**428**	**98**
M.tuberculosis A	*M.tuberculosis A*	1	–	426	488	389	80
M.tuberculosis A	*M.tuberculosis A*	10	+	430	440	417	95
M.tuberculosis A	*M.tuberculosis A*	10	–	434	545	351	65
M.tuberculosis A	*M.tuberculosis B*	1	+	426	432	419	97
M.tuberculosis A	*M.tuberculosis B*	1	–	429	487	386	80
M.tuberculosis A	*M.tuberculosis B*	10	+	423	434	418	97
M.tuberculosis A	*M.tuberculosis B*	10	–	428	548	349	64
M.tuberculosis A	*M.leprae*[c]	1	+	195	231	125	54
N.europaea	***N.europaea***	**1**	+	**63**	**63**	**62**	**98**
N.europaea	*N.europaea*	1	–	61	65	55	85
N.europaea	*N.europaea*	10	+	62	63	60	95
N.europaea	*N.europaea*	10	–	61	157	32	21
R.typhi	***R.typhi***	**1**	+	**79**	**79**	**75**	**95**
R.typhi	*R.typhi*	1	–	79	97	60	62
R.typhi	*R.typhi*	10	+	78	79	74	94
R.typhi	*R.typhi*	10	–	78	97	60	62
R.typhi	*R.prowazekii*[b]	1	+	73	79	74	94
R.typhi	*R.prowazekii*[b]	1	–	74	96	64	67
R.typhi	*R.prowazekii*[b]	10	+	77	79	75	95
R.typhi	*R.prowazekii*[b]	10	–	79	96	61	64

For each mapping procedure, the results were compared with the reference results obtained by using an isogenic template and repeat-masking (represented in boldface). Genome origins: *L.lactis* MG1363 (A), *L.lactis* IL1403 (B), *M.tuberculosis* CDC1551 (A), *M.tuberculosis* H37Rv (B), *M.leprae* TN, *N.europaea* ATCC 19718, *R.typhi* str. wilmington (A) and *R.prowazekii* str. Madrid E (B).

[a]The number of gaps that could be closed with direct PCR is defined as the number of gaps with sizes <15 kb divided by the total number of PCRs (= total number of mapped contigs) × 100%.

[b]The target genomes *R.typhi* (13) (Supplementary Figure S1) and *L.lactis* MG1363 (14) contain an inversion compared with their respective templates resulting in two incorrectly mapped contigs.

[c]This mapping was performed to demonstrate the mapping success when using a template genome with very limited colinearity under optimal conditions (one template with repeat-masking).

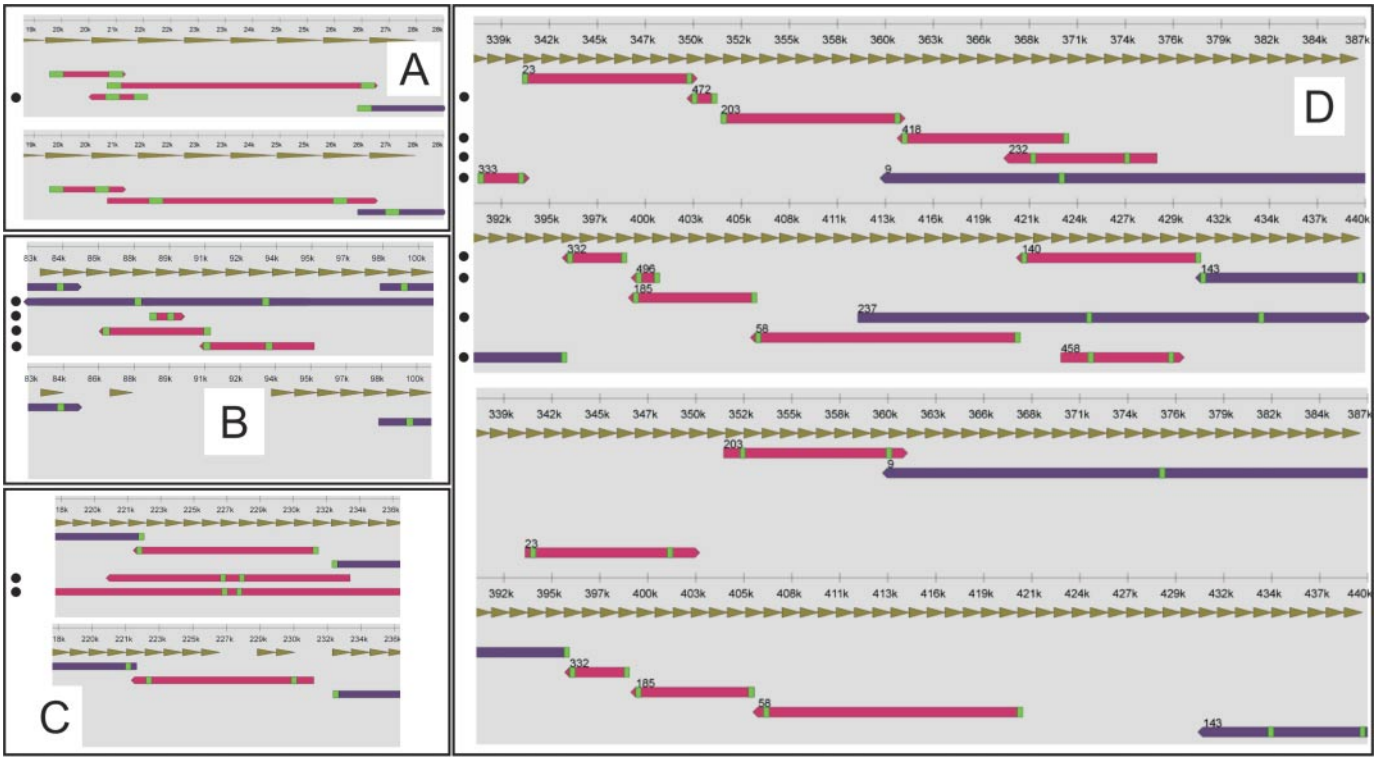

Figure 2. SVG output of Projector 2 runs performed with and without repeat-masking. (A–D) Details of the results for mapping of: (**A**) *R.typhi* contigs on its isogenic template; (**B**) *N.europaea* contigs on its isogenic template; (**C**) *L.lactis* MG1363 contigs on *L.lactis* IL1403; and (**D**) *M.tuberculosis* contigs on *M.tuberculosis*. For each inset, the mapping results with (lower panel) and without (upper panel) repeat-masking of the target and template sequences are shown. A ruler in base pairs (kb or Mb) is shown above each mapping. The template fragments are indicated in dark green triangles below this scale. The mapped contigs, one or at most two on each line, are shown below the template fragments. Within each mapped contig, the L and R fragments used to map the contig are indicated with green boxes. Contig numbers are shown for mapping in (D). A dot (•) indicates an incorrectly mapped contig on that line.

gaps depends on the coverage of the target genome sequence assembly.

The anticipated percentage of successful gap-closing PCRs for the *L.lactis* MG1363 contigs mapped onto the *L.lactis* IL1403 template is calculated at 93% (Table 2), which is in good agreement with the reported 94% of gap-closing PCR products obtained in our previous study (6). This result indicates that the method for estimating the number of successful gap-closure PCRs is accurate. For *M.tuberculosis*, primer design failed for four contigs due to the fact that insufficient sequence was left after repeat-masking and to a high G+C-content of the remaining DNA sequence. Primer pairs for these gaps were designed on flanking contig ends.

Repeat-masking provides robust mapping results

The estimated success percentages of the gap-closing PCRs of the isogenic reference mappings using repeat-masked sequences, are close to 100% for the four different assemblies, which indicates that the number of large gaps is relatively small in the four genome sequence assemblies (Table 2 and Figures 2 and 3). Mappings with repeat-masking on the non-isogenic templates yield slightly lower numbers of estimated successful gap-closing PCRs than the reference mappings (Table 2 and Figure 3). Without repeat-masking, a significant drop in the number of correctly mapped contigs and, thus, in successful gap-closing PCRs is observed for all assemblies (Table 2 and Figure 3). For the non-isogenic mappings, the

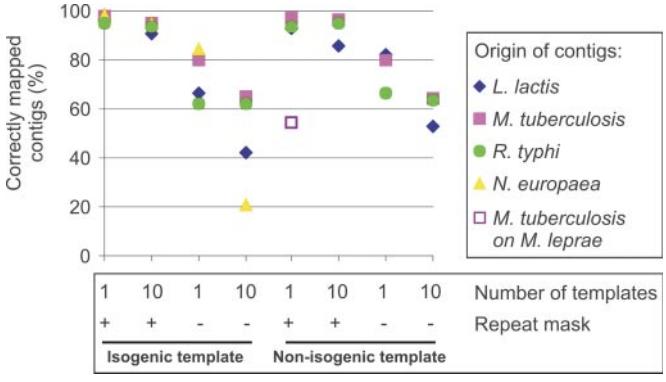

Figure 3. Results of the mapping procedures described in Table 2. The percentage of mapped contigs is plotted for four bacterial genome assemblies mapped onto (non-) isogenic templates (for details see Table 2).

number of correctly mapped contigs is approximately the same for *M.tuberculosis* and *R.typhi*. Only when the genome sequence of *L.lactis* IL1403 is used as template, the number of correctly mapped contigs drops from 106 to 85, indicating that, of the three template genomes, *L.lactis* IL1403 is least similar to the corresponding target genome.

Figure 2 shows the effect of repeat-masking on mapping results. In the case of *N.europaea* (Figure 2B), four incorrect mapping events were avoided in this way. The same occurred for *L.lactis* MG1363 mapped onto *L.lactis* IL1403, where two incorrect mapping events were thus avoided (Figure 2C).

The necessity to perform repeat-masking for the template sequences becomes clear from that a number of template fragments were removed after repeat-masking, thereby preventing a number of incorrect mapping events (lower panels in Figure 2A–D). The need to repeat-mask the contig sequences becomes evident from the fact that the L and R fragments used for contig mapping are, in most cases, shifted inwards of contigs after repeat-masking (Figure 2A–D). For instance, the *R.typhi* genome sequence assembly contained very large numbers of contigs with repetitive sequences on their ends. The fragments used for mapping of the *R.typhi* contigs (green boxes) overlap (Figure 2A; upper panel), while the fragments used for contig mapping are positioned more inwards in the contigs after repeat-masking (Figure 2A; lower panel).

The use of multiple fragments of a template genome does not affect the mapping results

To evaluate the performance of Projector 2 when using multiple fragments of a template genome, i.e. unfinished genomic sequences, mapping was performed for the template genomes listed in Table 2 after they had been divided into 10 equally sized template fragments. Mapping is performed on each of these 10 template sequences separately, yielding 10 separate maps. Multiplex PCR has to be used to obtain linkage information for these separate maps. As relatively low numbers of template fragments were used, multiplex PCR is expected to be successful in closing the gaps between contigs bridging multiple fragments of a template genome. In case repeat-masked target sequences were mapped onto multiple repeat-masked template fragments, almost all of the contigs were identically ordered when compared with the reference mapping (Table 2 and Figure 3). The number of correctly mapped contigs decreases considerably when the multiple templates were not repeat-masked. In the case of *N.europaea*, the percentage of correctly mapped contigs onto multiple templates dropped from 95% with repeat-masking on both template and target sequences to 21% without repeat-masking. The latter extreme case illustrates that there are many repetitive elements on the template genome, resulting in a large number of incorrect mapping events. For *M.tuberculosis*, more contigs were mapped (545 and 548, repectively; Table 2) than the total number of contigs used for mapping (491; Table 1), because some contigs were mapped onto multiple fragments of the template genome.

Mapping success depends on sequence similarity and colinearity between target and template genomes

In a previous study, we showed that a correlation exists between sequence similarity and mapping success. Although *L.lactis* MG1363 is distantly related to *L.lactis* IL1403 (sequence similarity of 85%; A. L. Zomer, U. Wegmann, unpublished data), the percentage of successful gap-closing PCRs of *L.lactis* MG1363 mapped onto *L.lactis* IL1403 is only moderately lower than that of, for instance, the closely related *M.tuberculosis* CDC 1551 mapped onto *M.tuberculosis* H37Rv (Table 2). *Mycobacterium leprae* is a well-documented case of an organism containing extensive genomic rearrangements compared with *M.tuberculosis* CDC1551 (12). Supplementary Figure S2 clearly demonstrates this case of limited colinearity in a dot-plot. The estimated percentage of successful gap-closing PCRs for *M.tuberculosis* CDC1551 mapped onto *M.leprae* is at an acceptable 54% (Table 2), demonstrating that by using template genome with limited colinearity, over half of the gaps are closed with a single run of the Projector 2 software.

Conclusions

Projector 2 accurately positions contigs of an unfinished genome sequence onto a genome of a related organism. It was not possible to compare the performance of MGview (7) and MUMmer (8) to Projector 2, as the former software implementations do not allow automatic generation of a list of mapped contigs. Manual inspection of the results from MUMmer did yield similar mapping results as those obtained with Projector 2. For MGview, similar results could not be generated because it requires a specialized hardware implementation of BLAST. In addition, both MGview and MUMmer do not perform primer design. As we have demonstrated, the ends of contigs often contain repetitive and unreliable DNA sequences, making the primer design a time-consuming task, even for experts.

Projector 2 allows closing the majority of gaps in a genome sequence assembly using minimum resources. The user is provided with pre-processed genomes, automated primer design on reliable and non-repetitive sequences, and, finally, the possibility of using multiple fragments of a template genome. Additionally, primers are designed that can be used in multiplex PCR strategies for closing gaps between different template sequences. As Projector 2 allows obtaining linkage information for template DNA sequences for which no ORF information is known, it can also use genomic DNA sequences that are less coding-dense.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENT

The Open Access publication charges for this article were paid by the Department of Molecular Genetics.

Conflict of interest statement. None declared.

REFERENCES

1. Fraser,C.M. and Fleischmann,R.D. (1997) Strategies for whole microbial genome sequencing and analysis. *Electrophoresis*, **18**, 1207–1216.
2. Stjepandic,D., Weinel,C., Hilbert,H., Koo,H.L., Diehl,F., Nelson,K.E., Tummler,B. and Hoheisel,J.D. (2002) The genome structure of *Pseudomonas putida*: high-resolution mapping and microarray analysis. *Environ. Microbiol.*, **4**, 819–823.
3. Crowe,M.L., Rana,D., Fraser,F., Bancroft,I. and Trick,M. (2002) BACFinder: genomic localisation of large insert genomic clones based on restriction fingerprinting. *Nucleic Acids Res.*, **30**, e118.
4. Zabarovska,V.I., Gizatullin,R.Z., Al Amin,A.N., Podowski,R., Protopopov,A.I., Lofdahl,S., Wahlestedt,C., Winberg,G., Kashuba,V.I., Ernberg,I. *et al.* (2002) A new approach to genome mapping and sequencing: slalom libraries. *Nucleic Acids Res.*, **30**, e6.
5. Tettelin,H., Radune,D., Kasif,S., Khouri,H. and Salzberg,S.L. (1999) Optimized multiplex PCR: efficiently closing a whole-genome shotgun sequencing project. *Genomics*, **62**, 500–507.
6. Van Hijum,S.A.F.T., Zomer,A.L., Kuipers,O.P. and Kok,J. (2003) Projector: automatic contig mapping for gap closure purposes. *Nucleic Acids Res.*, **31**, e144.

7. Herron-Olson,L., Freeman,J., Zhang,Q., Retzel,E.F. and Kapur,V. (2003) MGView: an alignment and visualization tool to enhance gap closure of microbial genomes. *Nucleic Acids Res.*, **31**, e106.
8. Kurtz,S., Phillippy,A., Delcher,A.L., Smoot,M., Shumway,M., Antonescu,C. and Salzberg,S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.
9. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
10. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
11. Van Hijum,S.A.F.T., de Jong,A., Buist,G., Kok,J. and Kuipers,O.P. (2003) UniFrag and GenomePrimer: selection of primers for genome-wide production of unique amplicons. *Bioinformatics*, **19**, 1580–1582.
12. Cole,S.T., Eiglmeier,K., Parkhill,J., James,K.D., Thomson,N.R., Wheeler,P.R., Honore,N., Garnier,T., Churcher,C., Harris,D. *et al.* (2001) Massive gene decay in the leprosy bacillus. *Nature*, **409**, 1007–1011.
13. McLeod,M.P., Qin,X., Karpathy,S.E., Gioia,J., Highlander,S.K., Fox,G.E., McNeill,T.Z., Jiang,H., Muzny,D., Jacob,L.S. *et al.* (2004) Complete genome sequence of *Rickettsia typhi* and comparison with sequences of other rickettsiae. *J. Bacteriol.*, **186**, 5842–5855.
14. Le Bourgeois,P., Lautier,M., van den,B.L., Gasson,M.J. and Ritzenthaler,P. (1995) Physical and genetic map of the *Lactococcus lactis* subsp. *cremoris* MG1363 chromosome: comparison with that of *Lactococcus lactis* subsp. *lactis* IL 1403 reveals a large genome inversion. *J. Bacteriol.*, **177**, 2840–2850.