BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 3 2009, pages 309–314
doi:10.1093/bioinformatics/btn632

Phylogenetics

PanCGH: a genotype-calling algorithm for pangenome CGH data

Jumamurat R. Bayjanov[1,*], Michiel Wels[1,2,3], Marjo Starrenburg[2,4], Johan E. T. van Hylckama Vlieg[2,3,4], Roland J. Siezen[1,2,3,4] and Douwe Molenaar[2,3,4]

[1]Center for Molecular and Biomolecular Informatics, Nijmegen Center for Molecular Life Sciences, Radboud University Medical Centre, P.O. Box 9101, 6500 HB Nijmegen, [2]NIZO food research, P.O. Box 20, 6710 BA Ede, [3]TI Food and Nutrition, P.O. Box 557, 6700 AN Wageningen and [4]Kluyver Centre for Genomics of Industrial Fermentation, Delft, The Netherlands

Received on September 9, 2008; revised on November 20, 2008; accepted on December 4, 2008

Advance Access publication January 7, 2009

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Pangenome arrays contain DNA oligomers targeting several sequenced reference genomes from the same species. In microbiology, these can be employed to investigate the often high genetic variability within a species by comparative genome hybridization (CGH). The biological interpretation of pangenome CGH data depends on the ability to compare strains at a functional level, particularly by comparing the presence or absence of orthologous genes. Due to the high genetic variability, available genotype-calling algorithms can not be applied to pangenome CGH data.

Results: We have developed the algorithm PanCGH that incorporates orthology information about genes to predict the presence or absence of orthologous genes in a query organism using CGH arrays that target the genomes of sequenced representatives of a group of microorganisms. PanCGH was tested and applied in the analysis of genetic diversity among 39 *Lactococcus lactis* strains from three different subspecies (*lactis*, *cremoris*, *hordniae*) and isolated from two different niches (dairy and plant). Clustering of these strains using the presence/absence data of gene orthologs revealed a clear separation between different subspecies and reflected the niche of the strains.

Contact: J.Bayjanov@cmbi.ru.nl

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

1 INTRODUCTION

Detection of genomic variation between related organisms can elucidate relations between genotypic and phenotypic traits of organisms, for example, those related to diseases with a genetic origin (Inazawa *et al*., 2004; Kallioniemi *et al*., 1992) or to functional traits of microorganisms (Pretzer *et al*., 2005). Comparative Genomic Hybridization (CGH) microarrays allow the detection of variation between a reference genome, whose sequences are targeted by the probes, and query genomes. The type of genetic variations that can be detected depends on the array design and the sequence similarity of reference and query genomes. Using short oligonucleotides, single nucleotide polymorphisms (SNPs) may be detected between highly similar genomes, like those of different human individuals. However, bacterial strains belonging to the same species often display extensive sequence variations (Lan and Reeves, 2000; Medini *et al*., 2005). In these cases, CGH microarrays generally only allow the detection of deletions, insertions and amplifications of relatively large pieces of DNA, like entire genes. Nevertheless, even this coarse-grained information can be very helpful in understanding the genetic basis of functional differences between strains of the same bacterial species. CGH data were used to show that highly variant parts of genomes of 20 *Lactobacillus plantarum* strains encode proteins that have a major role in the adaptation of these strains to different environments (Molenaar *et al*., 2005). CGH arrays can also be used to provide insight into evolutionary processes by analyzing the diversity among strains of the same species (Earl *et al*., 2007; Rasmussen *et al*., 2008) or different species (Fukiya *et al*., 2004).

Current microarray chips can contain several hundreds of thousands of probes, and make it possible to design an array from genomes of several reference strains of the same species at high probe density. These microbial species-level 'pangenome' arrays overcome the limited variability that is detectable with arrays based on a single reference genome. Several genotype-calling algorithms (Hua *et al*., 2007; Plagnol *et al*., 2007; Teo *et al*., 2007; Xiao *et al*., 2007) have been proposed for the interpretation of these data. However, these algorithms are mainly suited for detecting SNPs or other genomic variations between closely related organisms. The biological interpretation of pangenome microarrays in terms of the presence and absence of genetic functionalities in strains with unknown sequences poses a problem, because the probes target different homologous genes with various degrees of sequence similarity. To solve this problem, we have devised the genotype-calling algorithm PanCGH that combines orthology (Fitch, 1970) information about genes with species-level pangenome array data to determine the presence or absence of orthologous genes in bacterial strains. In this study, we test and apply PanCGH to CGH data of 39 *Lactococcus lactis* strains to investigate their genotypic variation. To our knowledge, PanCGH is the first algorithm addressing the problem of deducing gene content from data obtained with CGH microarrays that target the pangenome of a group of relatively diverse microorganisms.

*To whom correspondence should be addressed.

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

2 METHODS

2.1 DNA preparation

DNA was prepared from *L. lactis* strains (Supplementary Table 1) using the QiaAmp DNA Mini Kit (Qiagen GmbH, Hilden, Germany) according to the manufacturer's protocol for the isolation of genomic DNA from Gram-positive bacteria.

2.2 Microarray design and hybridization data acquisition

All genomic, plasmid and single gene or operon DNA sequences (1988 sequences in July 2005, constituting 10.7 Mb) of *L. lactis* were collected from the NCBI CoreNucleotide database and were deposited in a local database. This included complete genome sequences of *L. lactis* strain IL1403 (2.35 Mb, accession number AE005176) and fragments of the genome of strain SK11 (2.43 Mb, Genbank record GI:62464763). Additionally, draft genome sequences consisting of 547 contigs (2.3 Mb) of *L. lactis ssp. lactis* strain KF147 (NIZOB2230) and 961 contigs (2.6 Mb) of *L. lactis ssp. lactis* KF282 (NIZOB2244W) were added to this database. Redundant stretches of DNA were removed from the database, where a stretch of DNA was defined as redundant if it differed from another piece of DNA by at most 2 nt over a window of 100 nt. For the remaining non-redundant 7 Mb of DNA, a 32-mer tiling design was defined by starting an oligomer approximately every 19 nt, resulting in a total of 386 298 probes. We also designed 3181 random probes with their sequence absent in the non-redundant 7 Mb of DNA and they were randomly located on the array. Description of the platform with probe information has been deposited in the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) with an accession number GPL7231.

Array production and DNA hybridization, using fragmented DNA, were performed by NimbleGen Systems Inc. (Madison, WI, USA). The raw hybridization data, as well as annotations of the sequences, were stored in a custom relational database. Additionally, raw and normalized hybridization data of 39 *L. lactis* strains have been deposited in the GEO database with an accession number GSE12638. The annotations (gene definitions and putative protein function descriptions) were, in case of publicly available sequences, extracted from the GenBank files. For the draft sequences of *L. lactis* strains KF147 and KF282, GLIMMER (Salzberg *et al*., 1998) was used to define the genes and InterProScan (Zdobnov and Apweiler, 2001) was used to generate protein function descriptions.

2.3 Normalization of CGH microarray data

Many of the available normalization techniques do not take positional information of probes into account, yet spatial artifacts do contaminate array data. Such artifacts can be minimized by incorporating positional information of probes into normalization (Khojasteh *et al*., 2005; Neuvial *et al*., 2006; Yuan and Irizarry, 2006). Since a multiplicative noise model works better to minimize spatial artifacts than the additive noise model (Sasik *et al*., 2002), the normalization process is carried out on a logarithmic scale. We tested both the loess (Cleveland *et al*., 1992) and the fields (Fields Development Team, 2006) algorithms to normalize array data in two dimensions (R Development Core Team, 2007). Both methods fit a smooth 3D surface to the data. The height of this surface at a specific position represents the local average signal. For each individual spot, the height of the surface at the position of that spot is then subtracted from its raw signal intensity value. In order to avoid negative values the overall mean of the smooth fit is added to all signal intensities. We compared normalized data of both methods and concluded that the fields algorithm was faster and yielded better results. Therefore, we used the fields algorithm with its default Nadaraya–Watson kernel for spatial normalization of the array data.

Although this normalization minimizes within-array spatial biases, there is still a difference in overall signal intensity between arrays, which makes it difficult to compare them. Therefore, after spatial normalization, signal intensities in each array were divided by the median of their distribution.

2.4 The genotype-calling algorithm—PanCGH

The purpose of this genotype-calling algorithm is to facilitate the biological interpretation of pangenome CGH data by inferring the presence of a gene in a query strain using signal intensities of probes matching an orthologous gene of a reference strain. Since in a pangenome array several orthologs from different reference strains are represented on the array, the question is generalized to whether the query strain contains a member from a group of orthologous genes. Therefore, our algorithm also requires ortholog groups as input. Each ortholog group g_i contains the gene identifiers of a single gene or of several orthologous genes from the reference strains (reference orthologs). The set of all ortholog groups is represented as $G = \{g_1, g_2, \ldots, g_k\}$. To predict the presence or absence of a member gene from ortholog group g_i in the query strain, one cannot generally simply use the average signal from the set P_i of all probes targeting all genes from g_i as an indicator. Since short 32-mers are used, only probes that almost perfectly match the query gene will display a high fluorescence. Generally, these are a subset of P_i targeting the most similar reference ortholog. Therefore, the PanCGH algorithm uses these subsets of probes, and calculates the presence score from that subset of which the largest majority of probes has a high fluorescence (see Fig. 1 for an illustration of this principle). The output of the algorithm is a prediction of the presence or absence in the query strain of a member gene for each of the ortholog groups from the set G. In addition, if it predicts a gene to be

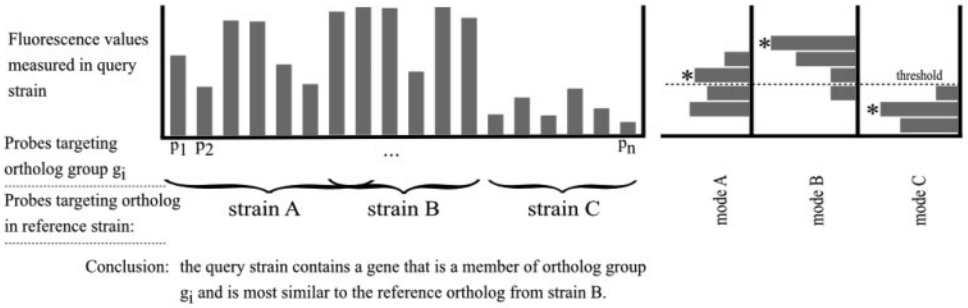

Fig. 1. Schematic representation of the PanCGH algorithm for a CGH experiment. The left panel shows the fluorescence of a query strain to a set of probes (p_1 to p_n) targeting different reference orthologs (homologous genes from reference strains A, B and C) of an ortholog group g_i. Some probes target several reference orthologs, as shown by the overlap between the probe sets targeting the reference orthologs from strains A and B. In the right panel, a schematic representation of the calculation of the presence score is shown. For each reference ortholog, the mode (indicated with a star) is calculated from the distribution of (log) signals of the corresponding probes. The presence score is the highest of these mode values. In this case, the presence score is above the threshold and equals the mode of the signals targeting the reference ortholog from strain B.

present it also predicts which of the reference orthologs is most similar to the gene in the query strain.

The algorithm proceeds as follows.

For each group of orthologs g_i in the set G perform Steps 1 to 4.

Step 1: For the set of reference strains $\{A, \ldots, X\}$, get the sets of probes $P_A, \ldots, P_X$ that match a sequence of a gene in the ortholog group g_i. Construct the union set $\Pi_i = P_A \cup P_B \cup \ldots \cup P_X$.

Step 2: Construct the set S of ordered pairs (p_k, s_k), where $p_k \in \Pi_i$ and s_k is the normalized fluorescence intensity value of probe p_k from the CGH array of the query strain.

Step 3: Calculate the 'presence score' S_i and the reference strain strain_Y with the closest homolog for a group g_i in the query genome as follows.

- For each reference strain Y in $\{A, \ldots, X\}$ compute the mode value m_Y over signals s_k in the sets $\{(p_k, s_k) | \, p_k \in P_Y\}$ (see below how the mode is calculated).
- Define S_i as the maximum of the modes, $S_i = \max(m_A, \ldots, m_X)$, or if all modes are undefined then S_i = NA.
- If there is only one strain Y which has a mode $m_Y = S_i$, then this is the strain with the closest homolog.
- Else, if there is more than one strain and only one of them has a mode S_i then $m_Y = S_i$ and the strain with the closest homolog is $\text{strain}_Y = \{Y \mid m_y = S_i\}$.
- Else, if there is more than one strain and more than one of them has a mode S_i then $m_Y = S_i$ and the strain with the closest homolog is $\text{strain}_Y = \{Y \mid m_Y = S_i, \, n_Y = \max(n_A, \ldots, n_X)\}$, where $n_Y = |P_Y|$ is number of probes in a set P_Y.

Step 4: Assign presence or absence of an ortholog in a query strain for the gene with closest similarity to that of strain_Y in group g_i using the following criteria.

- If S_i = NA (undefined) then the presence or absence of a member of g_i in the query strain cannot be decided from the data, hence presence = NA.
- Else, if $S_i >$ threshold, the query strain has a gene in an ortholog group g_i, hence presence = 1. The most similar reference ortholog is found in strain_Y. (See the results section for a determination of the optimal threshold presence score.)
- Else, if $S_i <$ threshold, the query strain possesses no gene in ortholog group g_i, hence presence = 0.

The mode over the signals s_k of a set of ordered pairs $\{(p_k, s_k) | \, p_k \in P_Y\}$ is calculated as follows.

(1) Define n_Y as the number of probes in the set P_Y.

(2) If $n_Y < 10$ then the mode is undefined: mode = NA.

(3) Else, bin the signals $\log(s_k)$ into $B = \text{round}(\log_2(n_Y + 1))$ equal sized intervals on the logarithmic scale. Calculate the frequencies $\{f_j | j \in \{1, \ldots, B\}\}$ as the number of signals $\log(s_k)$ in each bin and define mode as the mean of lower and upper limits of the bin with the highest associated frequency.

In the case study of *L. lactis* CGH arrays, the majority of genes in the false-positive group had <10 matching probes. Therefore, a minimum of 10 probes was required. This is not a strict requirement, and it might differ, depending on the probe size and the size of genes. The binning procedure in Step 3 of the calculation of the mode is recommended by Sturges (1926).

2.5 Defining orthologous groups of *L. lactis* genes

In order to predict orthology among genes, the genome sequence of three fully sequenced public *L. lactis* strains (*ssp. lactis* IL1403, *ssp. cremoris* SK11 and *ssp. cremoris* MG1363, accession numbers AE005176, CP000425 and AM406671, respectively) and incomplete genome sequences of two *L. lactis* strains (*ssp. lactis* KF147, *ssp. lactis* KF282) isolated from plants were used (Siezen *et al*., 2008). The orthology prediction program InParanoid (Remm *et al*., 2001) was run with default settings to find orthologous genes among the three completely sequenced genomes. All possible pairwise comparisons between the three genomes were performed. In cases where inconsistencies regarding bidirectionality of the ortholog relation were found between the pairwise InParanoid predictions, genes were regarded as not being orthologous and each treated as single genes in an orthologous group of size 1. As incomplete genomes are not suited for bidirectional best-BLAST analyses like InParanoid, the genes of the two incomplete genomes were added by performing a pairwise BLAST analysis of the genes from the incomplete genomes against the three complete genomes. If a gene in the incomplete genome had a best-BLAST hit with a member of one of the ortholog groups derived from the completely sequenced genomes, this gene was added to that ortholog group. In cases where best-BLAST hits referred to different ortholog groups, the gene was assigned to a new ortholog group, unless the difference in *E*-value of the BLAST searches was larger than 10^{-10}. In those cases, the gene was added to the ortholog group of the gene with the hit having the lowest *E*-value. We found a total of 4571 ortholog groups of which 1389 groups had a gene in all five *L. lactis* reference strains.

3 RESULTS

3.1 Microarray design and data normalization

Species-level pangenome CGH arrays containing oligonucleotides that target, among others, sequences of four reference strains *L. lactis ssp. lactis* IL1403, *L. lactis ssp. cremoris* SK11, *L. lactis ssp. lactis* KF147 and *L. lactis ssp. lactis* KF282 were designed. During the course of our work, the complete sequences of *L. lactis ssp. cremoris* strains SK11 and MG1363 were published (Makarova *et al*., 2006; Wegmann *et al*., 2007), and we remapped the probe targets of the existing design on these genomes. The availability of the complete MG1363 genome sequence also allowed us to use this strain as a test case (query strain) for the PanCGH algorithm. We analyzed genomic DNA isolated from 39 different *L. lactis* strains, including the reference strains.

The raw data from the hybridization experiments was biased. In particular, spatial artifacts on the microarrays were apparent. Hence, we applied a spatial normalization method to improve the data set. Visual inspection of the corrected data indicated that the spatial bias was minimized. To confirm the correctness of this procedure, a hierarchical clustering of strains using either raw or normalized signal intensities of all probes was carried out. Using the normalized signals, all except one *ssp. cremoris* strain clustered together and all *ssp. lactis* strains made another cluster, whereas strains from different subspecies clustered together when raw signals were used for clustering. This shows that, normalized microarray data correspond better with independent experimental criteria, namely those used for subspecies determination.

3.2 Determination of a presence score threshold for the genotype-calling algorithm

The pangenome microarray for *L. lactis* used in this work contains probes for several representatives of orthologous genes in different reference strains (reference orthologs). To predict whether a representative gene from a group of orthologous genes is present in a query strain with unknown sequence, a presence score for that group is calculated from the normalized fluorescence signals of probes that target the different reference orthologs (Fig. 1). A target sequence is predicted to be present when the presence score lies above a

threshold value. To define this threshold value, we used CGH data from the reference strains SK11 and IL1403 and calculated presence scores for sets of ortholog groups known to be either present or absent in SK11. An ideal threshold score value should separate all present from all absent genes. Supplementary Figure 1 shows that there is a clear separation between present and absent genes, although there is some overlap of the distributions. The PanCGH algorithm was also applied to strain MG1363. This is an ideal test strain for the procedure, because its gene content is known from the genome sequence, but just like any of the other query strains, its genome was not used for the design of the array. The distribution of presence scores was also bimodal for this strain, clearly separating present and absent genes. To determine the best threshold value, we tested all possible threshold values between the minimum and maximum presence score. As the best possible threshold, we defined the value at which the total error rate (false-positive + false-negative) was minimal. Supplementary Figure 2 shows that the position of the best possible threshold is 5.5 in an ROC curve (Hanley and McNeil, 1982) for SK11, IL1403 and MG1363. We estimated the accuracy (Table 1) of the algorithm using the gene annotation of the genome of strain MG1363 at the same threshold. Ortholog groups predicted as absent in MG1363 separated clearly from the groups predicted to be present.

Table 1. True-positive rate (sensitivity) and true-negative rate (specificity) of the PanCGH genotype-calling algorithm for three *L. lactis* strains

Strain	True-positive rate (%)	True-negative rate (%)
SK11	97.6	90.5
IL1403	97.9	86.2
MG1363	95.4	96.4

3.3 Applying the PanCGH algorithm

The PanCGH algorithm was applied to hybridization data from 39 *L. lactis* strains to assign corresponding genotypes to each strain. Strains were hierarchically clustered based on the presence or absence of genes of ortholog groups in these strains (Fig. 2). The observed clustering is in agreement with a number of independent genotypic and phenotypic observations on the strains (Rademaker *et al*., 2007; see Supplementary Table 1) supporting the robustness of the method developed in this article. Most strains group in either of the two large subclusters representing the two different subspecies: *L. lactis ssp. lactis* genotype (bottom subcluster) and *L. lactis ssp. cremoris* genotype (top subcluster). In the dendrogram, strains

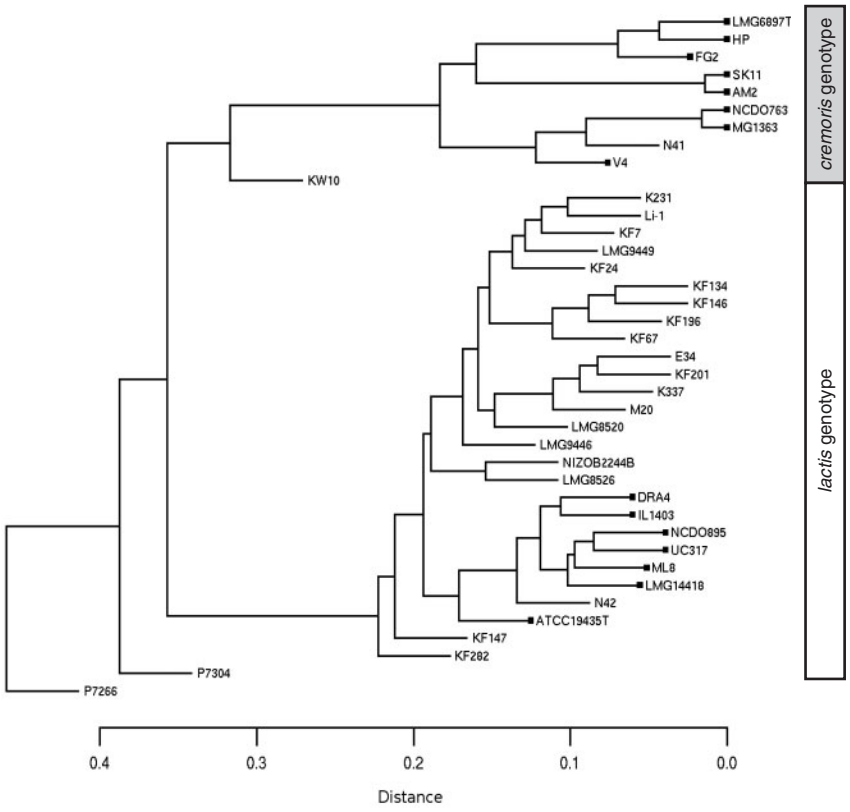

Fig. 2. Hierarchical clustering of *L. lactis* strains based on presence/absence predictions of representatives of 4571 ortholog groups of *L. lactis*. The pairwise binary distance was used as a distance metric and clustering was performed using the average linkage agglomeration method (Hastie *et al*., 2001). The cluster of strains at the top represents the subspecies *cremoris* genotype, while the large cluster at the bottom, excluding strains P7266 and P7304, contains strains of subspecies *lactis* genotype and one strain (LMG8520) of subspecies *hordniae* phenotype. In these two clusters 1341 groups from the total of 4571 ortholog groups are present in all strains. Though strains P7266 and P7304 have subspecies *lactis* phenotype, they are far apart from other subspecies *lactis* strains (see explanation in text). Branches with a solid rectangle are dairy isolates and other strains are isolated from plants.

P7266 and P7304 formed two distinct branches. Although these two strains have a *L. lactis ssp. lactis* phenotype, they have been shown to be highly different in genotype compared to the *L. lactis ssp. lactis* and the *L. lactis ssp. cremoris* genotypes (Rademaker *et al*., 2007). Further divisions within these two subclusters also reflect functional differences among strains. For instance, the top subcluster (*cremoris* genotype) is divided into three branches with 1, 4 and 5 strains; the latter branch contains five strains with both *cremoris* genotype and phenotype, whereas the other two branches contain strains having a *cremoris* genotype but displaying a *lactis* phenotype (Supplementary Table 1). The large subcluster at the bottom (*lactis* genotype) is divided into different branches, of which the largest contains 17 strains isolated from plants, while the next largest branch contains mostly strains of dairy origin.

4 DISCUSSION

The predictions of the PanCGH algorithm on *L. lactis* strains show a high true-positive rate (sensitivity) and low to moderate false-positive rate, as shown by tests of the algorithm with CGH data from sequenced strains (Table 1). Two types of sources that increase total error rate (false-positive + false-negative) can be distinguished: those that are inherent to the CGH method, like noise and limitations of the array platform, and those that are due to external factors. To the first type belong, for example, errors due to low sequence similarity (leading to poor hybridization) or due to the small size of some genes, as it is difficult to determine the presence or absence of small genes with low numbers of targeting probes. Errors due to low sequence similarity can be avoided by basing the array design on reference genomes from strains in different branches of the phylogenetic tree of a species. Errors due to external factors mainly originate from inconsistencies in the ORF calling and annotation of the reference strains or the InParanoid orthology prediction. A large part of the false-positive and false-negative predictions are due to the latter type of errors. For example, analysis of the genomes and genome annotations of strains MG1363 and SK11 showed that ORF-calling criteria differ between the two annotations. Many of the small ORFs defined only in strain SK11 were found by us to be also present in MG1363, but they were not identified as such in the original annotation. This caused positive gene calls by PanCGH in strain MG1363 for those ORFs that are not identified in the original annotation, but whose sequences are nonetheless present in this strain. These appear as false calls in the test of PanCGH, but are in fact correct. Imperfections in the orthology prediction also caused errors. In particular, for genes with many paralogs, it is difficult to correctly assign orthology relations using automated prediction methods that rely only on gene sequence information (Koonin, 2005; Notebaart *et al*., 2005). For example, in strain MG1363, we found that almost half of the apparent false-positive calls concerned hypothetical proteins. The remaining false-positive calls concerned mainly transporters and transposases, which often have many paralogs (Table 2).

Despite these sources of errors, the PanCGH algorithm has a high accuracy, which shows the robustness of the method. In order to avoid the errors originating from inconsistencies in ORF-calling and annotation, the same ORF-calling algorithm and definitions should be applied to all reference genomes. The orthology grouping can also be improved by including additional sources of information from e.g. phylogenetic trees and 3D structures (Francke *et al*., 2008; Golding and Dean, 1998).

Table 2. Functional categories in ortholog groups with frequent false calls in test strain *L. lactis* MG1363

Functional category	False-positive[a] (%)	False-negative[a] (%)
Hypothetical genes	49.9	60
Transposases	29.2	0
Related to transporters	5.3	7.2

[a]As a percentage of the total number of false cells.

In summary, we have developed a novel genotype-calling algorithm—PanCGH—for the biological interpretation of species-level pangenome CGH arrays. In contrast to conventional CGH arrays, these pangenome arrays allow the comparison of strains that are relatively diverse in terms of genome sequence. Information obtained from sequenced reference strains was incorporated to compare strains not only by signal intensities of individual probes, but also at the level of the inferred genotype, or more specifically, the presence and absence of members of ortholog groups. The results show that our genotype-calling algorithm predicts a genotype with high accuracy from a species-level pangenome CGH array data, which enables the extraction of relevant biological information for unsequenced strains. Since the threshold is determined from training data, the PanCGH algorithm can be applied to arrays that target the pangenome of any microorganism. Currently we are working on biological interpretation of the PanCGH analysis of *L. lactis* diversity (G.Felis *et al*., unpublished data).

ACKNOWLEDGEMENTS

We thank Giovanna Felis for useful discussions.

Funding: BSIK grant [through the Netherlands Genomics Initiative (NGI)]; BioRange programme [as part of, the Netherlands Bioinformatics Centre (NBIC)]; NGI (as part of the Kluyver Centre for Genomics of Industrial Fermentation).

Conflict of Interest: none declared.

REFERENCES

Cleveland,W.S. *et al*. (1992) Local regression models. In Chambers,J.M. and Hastie,T.J. (eds) Chapter 8 of *Statistical Models in S*. Wadsworth & Brooks, Cole, pp. 312–316.

Earl,A.M. *et al*. (2007) Bacillus subtilis genome diversity. *J. Bacteriol.*, **189**, 1163–1170.

Fields Development Team. (2006) *Fields: Tools for Spatial Data*. National Center for Atmospheric Research, Boulder, CO. Available at http://www.image.ucar.edu/Software/Fields/ (last accessed August, 2008).

Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.

Francke,C. *et al*. (2008) A generic approach to identify transcription factor-specific operator motifs; inferences for LacI-family mediated regulation in Lactobacillus plantarum WCFS1. *BMC Genomics*, **9**, 145.

Fukiya,S. *et al*. (2004) Extensive genomic diversity in pathogenic Escherichia coli and Shigella Strains revealed by comparative genomic hybridization microarray. *J. Bacteriol.*, **186**, 3911–3921.

Golding,G.B. and Dean,A.M. (1998) The structural basis of molecular adaptation. *Mol. Biol. Evol.*, **15**, 355–369.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Hastie,T. *et al*. (2001) *The Elements of Statistical Learning*. Springer, New York.

Hua,J. *et al*. (2007) SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. *Bioinformatics*, **23**, 57–63.

Inazawa,J. *et al*. (2004) Comparative genomic hybridization (CGH)-arrays pave the way for identification of novel cancer-related genes. *Cancer Sci.*, **95**, 559–563.

Kallioniemi,A. *et al*. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Khojasteh,M. *et al*. (2005) A stepwise framework for the normalization of array CGH data. *BMC Bioinformatics*, **6**, 274.

Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics, *Annu. Rev. Genet.*, **39**, 309–338.

Lan,R. and Reeves,P.R. (2000) Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends Microbiol*., **8**, 396–401.

Makarova,K. *et al*. (2006) Comparative genomics of the lactic acid bacteria. *Proc. Natl Acad. Sci. USA*, **103**, 15611–15616.

Medini,D. *et al*. (2005) The microbial pan-genome. *Curr. Opin. Genet. Dev.*, **15**, 589–594.

Molenaar,D. *et al*. (2005) Exploring Lactobacillus plantarum genome diversity by using microarrays. *J. Bacteriol.*, **187**, 6119–6127.

Neuvial,P. *et al*. (2006) Spatial normalization of array-CGH data. *BMC Bioinformatics*, **7**, 264.

Notebaart,R.A. *et al*. (2005) Correlation between sequence conservation and the genomic context after gene duplication, Nucleic Acids Res., **33**, 6164–6171.

Plagnol,V. *et al*. (2007) A method to address differential bias in genotyping in large-scale association studies. *PLoS Genet.*, **3**, e74.

Pretzer,G. *et al*. (2005) Biodiversity-based identification and functional characterization of the mannose-specific adhesin of Lactobacillus plantarum. *J. Bacteriol.*, **187**, 6128–6136.

R Development Core Team (2007) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rademaker,J.L. *et al*. (2007) Diversity analysis of dairy and nondairy Lactococcus lactis isolates, using a novel multilocus sequence analysis scheme and (GTG)5-PCR fingerprinting. *Appl. Environ. Microbiol.*, **73**, 7128–7137.

Rasmussen,T.B. *et al*. (2008) Streptococcus thermophilus core genome: comparative genome hybridization study of 47 strains, *Appl. Environ. Microbiol.*, **74**, 4703–4710.

Remm,M. *et al*. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.

Salzberg,S.L. *et al*. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **26**, 544–548.

Sasik,R. *et al*. (2002) Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. *Bioinformatics*, **18**, 1633–1640.

Siezen,R.J. *et al*. (2008) Genome-scale genotype-phenotype matching of two Lactococcus lactis isolates from plants identifies mechanisms of adaptation to the plant niche. *Appl. Environ. Microbiol.*, **74**, 424–436.

Sturges,H.A. (1926) The Choice of a Class Interval. *J. Am. Stat. Assoc.*, **21**, 65–66.

Teo,Y.Y. *et al*. (2007) A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*, **23**, 2741–2746.

Wegmann,U. *et al*. (2007) Complete genome sequence of the prototype lactic acid bacterium Lactococcus lactis subsp. cremoris MG1363. *J. Bacteriol.*, **189**, 3256–3270.

Xiao,Y. *et al*. (2007) A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics*, **23**, 1459–1467.

Yuan,D.S. and Irizarry,R.A. (2006) High-resolution spatial normalization for microarrays containing embedded technical replicates. *Bioinformatics*, **22**, 3054–3060.

Zdobnov,E.M. and Apweiler,R. (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.