Taxoblast – a simple pipeline for homology-based identification of contaminating and hybrid sequences in assembled genomes

Simon M Dittami Corresp., [1] , **Erwan Corre** [2]

[1] UMR8227 - Sorbonne Universités CNRS UPMC, Station Biologique de Roscoff, Roscoff, Brittany, France

[2] FR2424 - Sorbonne Universités CNRS UPMC, Station Biologique de Roscoff, Roscoff, Brittany, France

Corresponding Author: Simon M Dittami
Email address: simon.dittami@gmail.com

Modern genome sequencing strategies are highly sensitive to contaminations making the detection of foreign DNA sequences an important part of sequencing pipelines. Here we present Taxoblast, a simple pipeline with a graphical user interface, for the post assembly detection of contaminating sequences. Analyses are based on multiple blastn searches with short sequence fragments, therefore also enabling the identification of hybrid scaffolds that contain both target and potential contaminant sequences or horizontal gene transfers. The pipeline was applied to the published genome of the kelp *Saccharina japonica*, revealing a number of probable bacterial contaminations as well as hybrid scaffolds that contain both bacterial and algal sequences. This or similar types of analysis, in combination with manual curation, may be a useful complement to standard bioinformatics analyses prior to submission of genomic data to public repositories. Taxoblast is open-source and designed to be usable also by researchers with little background in informatics. It is freely available at http://sdittami.altervista.org/taxoblast and via sourceforge.

TAXOBLAST – A SIMPLE PIPELINE FOR HOMOLOGY-BASED IDENTIFICATION OF CONTAMINATING AND HYBRID SEQUENCES IN ASSEMBLED GENOMES

Simon M. Dittami[1*] Erwan Corre[2]

[1]Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR8227, Integrative Biology of Marine Models , Station Biologique, CS 90074, 29688 Roscoff cedex, France. Email: simon.dittami@sb-roscoff.fr

[2]ABiMS platform, FR 2424 CNRS UPMC, Station Biologique, CS 90074, 29688 Roscoff cedex, France. Email: erwan.corre@sb-roscoff.fr

*Corresponding author: phone +33 - 29 82 92 362, fax +33 - 29 82 92 324

Abstract

Modern genome sequencing strategies are highly sensitive to contaminations making the detection of foreign DNA sequences an important part of sequencing pipelines. Here we present Taxoblast, a simple pipeline with a graphical user interface, for the post assembly detection of contaminating sequences. Analyses are based on multiple blastn searches with short sequence fragments, therefore also enabling the identification of hybrid scaffolds that contain both target and potential contaminant sequences or horizontal gene transfers. The pipeline was applied to the published genome of the kelp *Saccharina japonica*, revealing a number of probable bacterial contaminations as well as hybrid scaffolds that contain both bacterial and algal sequences. This or similar types of analysis, in combination with manual curation, may be a useful complement to standard bioinformatics analyses prior to submission of genomic data to public repositories. Taxoblast is open-source and designed to be usable also by researchers with little background in informatics. It is freely available at http://sdittami.altervista.org/taxoblast and via sourceforge.

Key words: Genome assembly; BLAST (Basic Local Alignment Search Tool); contaminating sequences; horizontal gene transfer.

Introduction

Modern genome sequencing strategies rely strongly on the amplification of low quantities of Deoxyribonucleic acid (DNA), making them highly sensitive to even small contaminations in the samples. A recent study by Longo et al. (Longo, O'Neill & O'Neill, 2011), for example, has shown almost 1/4th of non-primate genomes available in the NCBI databases to be contaminated by repeated elements frequently found in human cells. Samples may furthermore be contaminated by airborne bacteria or other eukaryotes, ingested food, or symbionts living within or attached to the target organism. The detection of contaminants in genome datasets may be accomplished pre-assembly, post-assembly, or using a combination of both approaches.

Pre-assembly removal of potential contaminants has the advantage of reducing the complexity of the assembly process by producing smaller and more homogenous data sets. A first step may be filtering according to kmer-coverage or according to per-read guanine-cytosine (GC) contents (Schmieder & Edwards, 2011) or applying more advanced binning techniques based on oligonucleotide composition (e.g. [3,4]). This bears the risk of also removing genomic sequences from the target organism e.g. repeated elements or regions resulting from recent horizontal gene transfers, and of missing contaminants with similar properties as the target genome**.** A complementary approach is to search for potential contaminants based on sequence similarity. Here the main limitation lies in the quality and completeness of the reference databases. Furthermore, sequence similarity searches may be time-consuming due to the amount of raw data to treat, and are classically limited to smaller sets of contaminants such as vectors or individual species (e.g. to screen for human DNA contaminants), or ribosomal sequences. However, the recent development of fast search and classification algorithms that do not rely on marker genes dedicated to the analysis of (mainly microbial) metagenomes, such as RITA

(MacDonald, Parks & Beiko, 2012), KRAKEN (Wood & Salzberg, 2014), or Kaiju (Menzel, Ng & Krogh, 2016) (see [8] for a recent review), could also be applied to standard genome datasets with the aim of identifying contaminations prior to assembly. Just as for kmer- and GC contents-based approaches the limitation remains that foreign sequences recently integrated into the host genome, may be falsely removed.

Post-assembly cleaning of genomic sequences is frequently performed once manual annotation reveals the presence of contaminants (e.g. [9,10]), and can be carried out without impacting non-contaminated scaffolds of the target species. As for pre-assembly approaches, criteria to remove sequences are properties such as GC contents and read coverage (e.g. [11]) or sequence similarity (e.g. [9,10,12]). As mentioned above, GC and coverage based approaches will fail to detect contaminants that resemble the target organism with respect to these characteristics, while the performance of sequence-similarity based approaches is strongly dependent on the quality of the reference database. A key advantage of post-assembly approaches, however, is that they allow for the detection of recent horizontal gene transfer events as long as these elements are embedded into scaffolds containing also host DNA.

Here we present Taxoblast, a simple analysis pipeline with a graphical user interface to aid non-expert users in the post-assembly detection of potential contaminating sequences from different taxa as well as horizontal gene transfers by sequence similarity searches.

Methods

Two approaches are frequently used for the sequence-similarity based detection of potential contaminants: One is to perform searches with nucleotide sequences (either as blastn against a nucleotide database or blastx against a protein database), but if such searches are carried out with the entire scaffolds, they may be biased by highly conserved regions, which frequently have very little discriminatory power (transposons, virus insertions etc.). Alternatively, protein-based searches may be performed with all predicted proteins of a genome against a reference database (frequently NCBI nr or uniref90). Based on these results, scaffolds for which most proteins have best matches with expected contaminants are removed. The advantage of this approach is that, if there are several predicted proteins on a scaffold, then several independent results can be combined to reach a conclusion. Furthermore, protein sequences are more conserved than nucleotide sequences thus making it easier to find matches. The main disadvantage, however, is that this approach depends on the available protein predictions. Not all DNA encodes proteins, and especially bacterial contaminations in eukaryote genomes frequently lack protein predictions, because the latter are based on the availability of RNAseq data, which, in turn usually includes PolyA-selection during library preparation, removing bacterial messenger ribonucleic acid (mRNA) from the final dataset.

Taxoblast chooses a compromise between both approaches. It works with nucleotides to eliminate the impact of protein predictions. However, in a first step (Figure 1) each genomic scaffold is split into small fragments of e.g. 500 base pairs (bp) or 1 kbp. Then, each of these fragments is compared individually to a reference database. This way each sequence fragment has the same weight in the analysis, i.e. analyses are not biased towards highly conserved elements as would be the case for blast searches with

a single long query sequence. Internal tests aiming to identify bacterial sequences in algal genomes have resulted in very little differences between blastx searches against the National Center for Biotechnology Information (NCBI) nr database and blastn searches against the the NCBI nt database, except that blastx searches were significantly slower. MegaBLAST searches are designed specifically for highly similar sequences with > 95% sequence identity (McGinnis & Madden, 2004) and were not sufficiently sensitive to detect potential bacterial contaminants in the genome of brown algae; the algorithm, may, however, be preferable when attempting to detect well-known contaminants such as human DNA. In a third step, the taxa corresponding to the best blast hits are summarized for the entire scaffold with respect to two target taxa (the target group and a group of expected contaminants), where the taxonomic resolution can range from specific species to entire domains. The output is a table summarizing for each scaffold what percentage of fragments had best matches with potential contaminants or target sequences. As a rule of thumb, Taxoblast considers that sequences with >90% of hits matching potential contaminants are likely to be contaminants, whereas scaffolds with >90% of hits belonging to the target group are most likely not; Taxoblast automatically generates lists with all scaffolds matching these criteria.

Results and Discussion

To demonstrate the usefulness of Taxoblast, we have applied the pipeline to the recently published genome of the kelp *Saccharina latissima*. This genome has already been treated post assembly to remove potential bacterial contaminations using blastx searches of entire scaffolds against the NCBI nr database in combination with ORF density and GC contents (Ye et al., 2015). As there is little use in running taxoblast with very short sequences, only scaffolds >2kb were considered for our analyses. A fragment size of 500 bp was chosen, and blastn searches were carried out against the NCBI nr database (version July 19th, 2016) with an e-value cutoff of 0.01 (please note that decreasing this cutoff to 1e-5 had virtually no impact on the results obtained). The total calculation time for the blast analyses on the cluster of the “Analysis Bioinformatics for Marine Science” (ABIMS) platform (http://abims.sb-roscoff.fr) was 5 days. Taxonomic summaries were calculated first with the aim of distinguishing eukaryote (taxon 2759) from bacterial (taxon 2) sequences, and in a second round to distinguish diatom (taxon 2836) from brown algal (taxon 2870) sequences. Both bacteria and diatoms are common contaminants in cultures of marine algae. Despite previous cleaning efforts undertaken prior to the publication of the *S. japonica* genome, several bacterial or bacteria-like sequences were found in the published assembly (see Supplemental Information File 1), but Taxoblast did not identify any potential diatom sequences.

In total, 894 scaffolds >2 kbp (of 6,985) corresponding to 7.96 Mbp of sequence information were identified as bacterial based on the 90% bacterial hits threshold defined above (Figure 2A). Please note that only 148 of these 894 scaffolds were predicted to comprise protein coding sequences (201 proteins in total), underlining the usefulness of blastn searches rather than blastp or blastx searches, at least in some cases. Nine of the scaffolds classified as bacterial contained 16S fragments and could be assigned to different bacterial taxa using RDP classifier (Wang et al., 2007). These were *Haliea* (scaffold3770), *Methylophaga* (scaffold4403), an unclassified *Gammaproteobacterium* (scaffold6634), *Lewinella* (scaffold4968), an unclassified *Proteobacterium* (scaffold4223), *Marinoscillum* (scaffold3608), *Pseudonocardia* (scaffold7323), an unclassified *Alphaproteobacterium* (scaffold5114), *Flexibacter* (scaffold4987). We also manually examined the five longest potential bacterial sequences without 16S

and confirmed the automatic classification: For scaffold2350 (50 kbp) long fragments matched with *Rhodobacteraceae*; scaffold2282 (54 kbp) contained large sequence portions throughout the scaffold conserved with *Actinobacteria*; scaffold2573 (40 kbp) contained several conserved regions with sequences from *Planctomycetes*; scaffold2615 (39 kbp) was clearly part of a *Methylophaga* genome (also detected via a 16S sequence in a different scaffold, see above); and scaffold2647 (37 kbp) contained large regions that were homologous to genomes of different *Alphaproteobacteria*. Finally, we examined the distribution of GC contents for all 894 scaffolds classified as bacterial, and compared it to that of the remaining algal and unclassified scaffolds using the prinseq server version 0.20.4 (Schmieder & Edwards, 2011) (Figure 2B). While the algal reads exhibited a unimodal distribution with a narrow peak in GC contents at ca. 49%, reads classified as bacterial had a wide multimodal distribution with peaks at 30, 38, 53, and 64% GC. This clearly supports the hypothesis of diverse phylogenetic origins of the scaffolds classified as bacterial.

Taxoblast also highlighted 1,060 scaffolds (corresponding to a total of 82 Mbp of sequence information) that could not be attributed to either bacteria or eukaryotes with the 90% threshold. These scaffolds may comprise assembly artifacts the have merged contaminating sequences with sequences of the target species as well as recent horizontal gene transfers. It is not the aim of this publication to re-analyze the *S. japonica* genome, but for the purpose of illustration we have selected the two scaffolds with the highest numbers of blast hits from this category, and manually examined them.

The first, scaffold159, is approximately 500 kbp long and has 557 blast hits for different 500 bp-segments, 26% of which (towards the beginning of the sequence) are with bacteria. Consequently, we uploaded scaffold159 to the GC profile server (Gao & Zhang, 2006), which searches for variations in the GC contents within a sequence, revealing one segmentation point in the scaffold: bases 1 to 79743 exhibited an average GC content of 40.7 % while for bases 79,744 to 504,453 the average GC content was 49.9 %. Manual examination of both parts of the scaffold separately confirmed that the first part was highly similar to a cyanobacterial genome of the genus *Stanieria* (79% sequence identity in matching regions), whereas the second segment best aligned with genomic sequences of the brown alga *Ectocarpus siliculosus*. For the second examined scaffold, scaffold248 with a total length of 422 kbp, the situation was similar, except that the segmentation point was found at position 151,859 and that the first part of the sequence was highly similar to published genomes of bacteria belonging to the *Flammeovirgaceae*. The corresponding GC contents were 51.6% for the first (bacterial) part of the sequence and 50.3% for the rest.

Interestingly, neither of the two examined scaffolds contained undefined bases (Ns) between the bacteria-like parts and those conserved with other brown algae, but more information would be required to distinguish between recent horizontal gene transfers and assembly errors. In particular, differences in sequencing coverage between different parts of the scaffold could provide an indication for the latter, but unfortunately, no genome browser is available for *S. latissima* yet. In the case of similar coverage, ultimately polymerase chain reactions (PCRs) would be required to confirm horizontal transfers. Although not all of the results obtained by the Taxoblast pipeline are as clear as these examples, the presented data demonstrates that the approach yielded results that would have been very useful to point out scaffolds requiring complementary analyses and manual curation in *S. japonica*.

Conclusions

Taxoblast is a simple analysis pipeline to detect potential contaminations of different phylogenetic origins and horizontal gene transfers based on multiple blastn searches with small sequence fragments. Results are summarized across several independent searches making the tool more robust and allowing for the detection of hybrid scaffolds that contain sequences of different phylogenetic origin. Using the published genome of the kelp *S. latissima*, we show that our procedure was able to highlight several contaminating bacterial sequences as well as hybrid scaffolds. A key advantage of taxoblast is that it is accessible also to non-specialist users: it comes with a simple graphical user interface and the output format is a tab-separated text file compatible with most spreadsheet programs. This makes it easy to combine results with other sources of information (e.g. GC content or coverage). Moreover, sequences requiring further attention such as the aforementioned hybrid scaffolds can be easily identified and further investigated either manually or using semi-automatic pipelines such as PhyloGena (Hanekamp et al., 2007). Currently an important limitation, however, is that for larger (i.e. eukaryote) genomes, blast searches still need to be run on a dedicated blast server. One possibility here will be to replace BLAST by recent and accelerated algorithms such as PLAST (Nguyen & Lavenier, 2009), DIAMOND (Buchfink, Xie & Huson, 2014), or NSimScan (Novichkov et al., 2016) that are able to output the standard tabular BLAST format. Ultimately, a solution may also be to integrate taxoblast or a taxoblast-like pipeline in the galaxy environment, thus combining the resources of a computational cluster with a simple interface. Despite the fact the taxoblast, like all similarity-based tools, is limited by the quality of the reference database, we believe that the example of *S. japonica* has demonstrated that the approach was useful and complementary to pre-assembly screening. Our hope is that this pipeline will also prove useful for other genomes.

Acknowledgements:

We would like to thank the ABIMS platform (http://abims.sb-roscoff.fr) for access to their computing facilities and C. Boyen for critical reading of the manuscript. This work benefited from the support of the French Government via the National Research Agency investment expenditure program IDEALG (ANR-10-BTBR-04).

Additional Information

Data Availability: TaxoBlast 1.0 (including the source code) is freely available at http://sdittami.altervista.org/taxoblast and via sourceforge (https://sourceforge.net/projects/taxoblast/). The *S. japonica* genome analyzed is available in Genbank under accession number JXRI00000000.1 (Ye et al., 2015).

References

Buchfink B., Xie C., Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12:59–60. DOI: 10.1038/nmeth.3176.

Cock JM., Sterck L., Rouzé P., Scornet D., Allen AE., Amoutzias G., Anthouard V., Artiguenave F., Aury J-M., Badger JH., Beszteri B., Billiau K., Bonnet E., Bothwell JH., Bowler C., Boyen C., Brownlee C.,

Carrano CJ., Charrier B., Cho GY., Coelho SM., Collén J., Corre E., Da Silva C., Delage L., Delaroque N., Dittami SM., Doulbeau S., Elias M., Farnham G., Gachon CMM., Gschloessl B., Heesch S., Jabbari K., Jubin C., Kawai H., Kimura K., Kloareg B., Küpper FC., Lang D., Le Bail A., Leblanc C., Lerouge P., Lohr M., Lopez PJ., Martens C., Maumus F., Michel G., Miranda-Saavedra D., Morales J., Moreau H., Motomura T., Nagasato C., Napoli CA., Nelson DR., Nyvall-Collén P., Peters AF., Pommier C., Potin P., Poulain J., Quesneville H., Read B., Rensing SA., Ritter A., Rousvoal S., Samanta M., Samson G., Schroeder DC., Ségurens B., Strittmatter M., Tonon T., Tregear JW., Valentin K., von Dassow P., Yamagishi T., Van de Peer Y., Wincker P. 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* 465:617–21. DOI: 10.1038/nature09016.

Collén J., Porcel B., Carré W., Ball SG., Chaparro C., Tonon T., Barbeyron T., Michel G., Noel B., Valentin K., Elias M., Artiguenave F., Arun A., Aury J-M., Barbosa-Neto JF., Bothwell JH., Bouget F-Y., Brillet L., Cabello-Hurtado F., Capella-Gutiérrez S., Charrier B., Cladière L., Cock JM., Coelho SM., Colleoni C., Czjzek M., Da Silva C., Delage L., Denoeud F., Deschamps P., Dittami SM., Gabaldón T., Gachon CMM., Groisillier A., Hervé C., Jabbari K., Katinka M., Kloareg B., Kowalczyk N., Labadie K., Leblanc C., Lopez PJ., McLachlan DH., Meslet-Cladiere L., Moustafa A., Nehr Z., Nyvall Collén P., Panaud O., Partensky F., Poulain J., Rensing SA., Rousvoal S., Samson G., Symeonidi A., Weissenbach J., Zambounis A., Wincker P., Boyen C. 2013. Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the Archaeplastida. *Proceedings of the National Academy of Sciences of the United States of America* 110:5247–52. DOI: 10.1073/pnas.1221259110.

Gao F., Zhang C-T. 2006. GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. *Nucleic acids research* 34:W686-91. DOI: 10.1093/nar/gkl040.

Hanekamp K., Bohnebeck U., Beszteri B., Valentin K. 2007. PhyloGena--a user-friendly system for automated phylogenetic annotation of unknown sequences. *Bioinformatics* 23:793–801. DOI: 10.1093/bioinformatics/btm016.

Kumar S., Jones M., Koutsovoulos G., Clarke M., Blaxter M. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Frontiers in Genetics* 4:237. DOI: 10.3389/fgene.2013.00237.

Longo MS., O'Neill MJ., O'Neill RJ. 2011. Abundant human DNA contamination identified in non-primate genome databases. *PloS One* 6:e16410. DOI: 10.1371/journal.pone.0016410.

MacDonald NJ., Parks DH., Beiko RG. 2012. Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Research* 40:e111–e111. DOI: 10.1093/nar/gks335.

McGinnis S., Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research* 32:W20-5. DOI: 10.1093/nar/gkh435.

McHardy AC., Martín HG., Tsirigos A., Hugenholtz P., Rigoutsos I. 2007. Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods* 4:63–72. DOI: 10.1038/nmeth976.

Menzel P., Ng KL., Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications* 7:11257. DOI: 10.1038/ncomms11257.

Nguyen VH., Lavenier D. 2009. PLAST: parallel local alignment search tool for database comparison. *BMC Bioinformatics* 10:329. DOI: 10.1186/1471-2105-10-329.

Novichkov V., Kaznadzey A., Alexandrova N., Kaznadzey D. 2016. NSimScan: DNA comparison tool with increased speed, sensitivity and accuracy. *Bioinformatics* 32:2380–2381. DOI: 10.1093/bioinformatics/btw126.

Olsen JL., Rouzé P., Verhelst B., Lin Y-C., Bayer T., Collen J., Dattolo E., De Paoli E., Dittami S., Maumus F., Michel G., Kersting A., Lauritano C., Lohaus R., Töpel M., Tonon T., Vanneste K., Amirebrahimi M., Brakel J., Boström C., Chovatia M., Grimwood J., Jenkins JW., Jueterbock A., Mraz A., Stam WT., Tice H., Bornberg-Bauer E., Green PJ., Pearson GA., Procaccini G., Duarte CM., Schmutz J., Reusch TBH., Van de Peer Y. 2016. The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea. *Nature* 530:331–335. DOI: 10.1038/nature16548.

Sangwan N., Xia F., Gilbert JA. 2016. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* 4:8. DOI: 10.1186/s40168-016-0154-5.

Schmieder R., Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)* 27:863–4. DOI: 10.1093/bioinformatics/btr026.

Teeling H., Waldmann J., Lombardot T., Bauer M., Glöckner F. 2004. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* 5:163. DOI: 10.1186/1471-2105-5-163.

Wang Q., Garrity GM., Tiedje JM., Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology* 73:5261–7. DOI: 10.1128/AEM.00062-07.

Wood DE., Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* 15:R46. DOI: 10.1186/gb-2014-15-3-r46.

Ye N., Zhang X., Miao M., Fan X., Zheng Y., Xu D., Wang J., Zhou L., Wang D., Gao Y., Wang Y., Shi W., Ji P., Li D., Guan Z., Shao C., Zhuang Z., Gao Z., Qi J., Zhao F. 2015. Saccharina genomes provide novel insight into kelp biology. *Nature communications* 6:6986. DOI: 10.1038/ncomms7986.

Figure legends

Figure 1: Overview of the Taxoblast pipeline (A), the corresponding graphical user interface (B) and the generated output (C).

Figure 2: Application of the Taxoblast pipeline to identify potential bacterial sequences in the published *S. japonica* genome (Ye et al., 2015). (A) shows the percentage of bacterial/eukaryote blast hits over the 6731 scaffolds >2 kbp with blast hits (254 scaffolds > 2kbp had no hits). Dotted lines show the 90% cutoff proposed to consider a sequence as “contaminant”. (B) and (C) illustrate the different distribution of GC contents in the sequences considered bacterial, and those considered eukaryotic or unclassified.

Supplemental Information

Supplemental Information File 1: Taxoblast output when run with the *S. japonica* genome (scaffolds >2 kbp) to distinguish between eukaryote and bacterial sequences.

Figure 1

Overview of the Taxoblast pipeline

Figure 1 - Overview of the Taxoblast pipeline (A), the corresponding graphical user interface (B) and the generated output (C).

C

List of scaffolds with more than 90% of hits belonging to taxon 1

List of scaffolds with more than 90% of hits belonging to taxon 2

List of scaffolds accociated with neither taxon

Table of results per taxon for manual fitlering

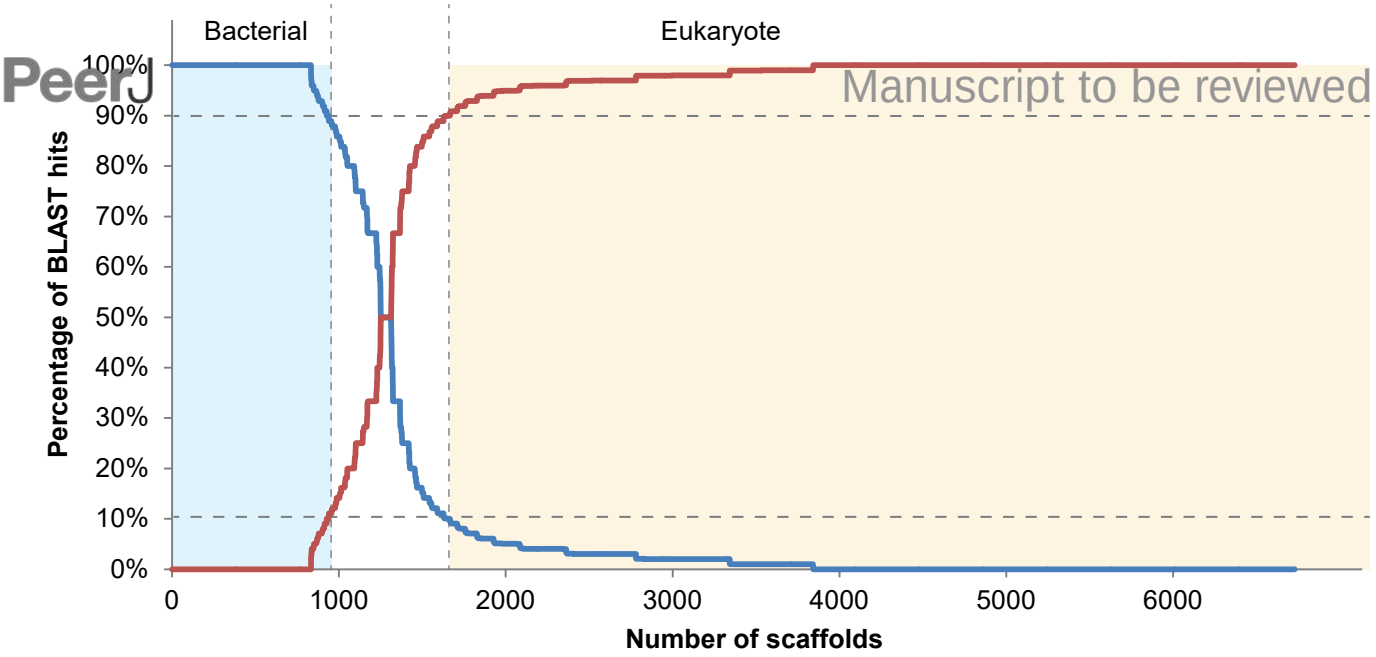
Scaffold	#Hits	#Queries	%Tax:2	%Tax:2759
Saccharina_japonica_cultivar_Ja_scaffold2065	45	133	2	97
Saccharina_japonica_cultivar_Ja_scaffold5519	2	13	100	0
Saccharina_japonica_cultivar_Ja_scaffold6435	5	9	100	0
Saccharina_japonica_cultivar_Ja_scaffold1407	112	237	0	97
Saccharina_japonica_cultivar_Ja_scaffold6825	4	7	100	0
Saccharina_japonica_cultivar_Ja_scaffold2815	11	64	0	100

Figure 2(on next page)

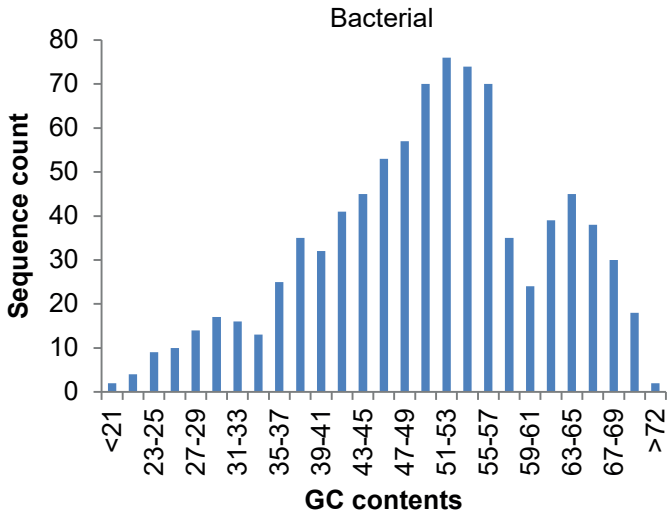
Taxoblast analysis of the *Saccharina japonica* genome

Figure 2 - Application of the Taxoblast pipeline to identify potential bacterial sequences in the published *S. japonica* genome (Ye et al., 2015). (A) shows the percentage of bacterial/eukaryote blast hits over the 6731 scaffolds >2 kbp with blast hits (254 scaffolds > 2kbp had no hits). Dotted lines show the 90% cutoff proposed to consider a sequence as "contaminant". (B) and (C) illustrate the different distribution of GC contents in the sequences considered bacterial, and those considered eukaryotic or unclassified.

A



B



C