MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and explicit phylogenetic tree approaches

Weizhi Song[1,2], Bernd Wemheuer[1,3], Shan Zhang[1,2], Kerrin Steensen[1,4] and Torsten Thomas[1,3]

[1]Centre for Marine Bio-Innovation, University of New South Wales, Sydney NSW 2052, Australia

[2]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney NSW 2052, Australia

[3]School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney NSW 2052, Australia

[4]Department of Genomic and Applied Microbiology, Georg-August-University Göttingen, Grisebachstr. 8, D-37077 Göttingen, Germany.

Corresponding Author: Torsten Thomas

Email addresses:

Weizhi Song: weizhi.song@student.unsw.edu.au

Bernd Wemheuer: b.wemheuer@unsw.edu.au

Shan Zhang: shan.zhang@student.unsw.edu.au

Kerrin Steensen: kerrin.steensen@stud.uni-goettingen.de

Torsten Thomas: t.thomas@unsw.edu.au

Abstract

Background: Metagenomic datasets provide an opportunity to study horizontal gene transfer (HGT) on the level of the microbial community. However, current HGT detection methods cannot be applied to community-level datasets or require reference genomes. Here, we present MetaCHIP, a pipeline for reference-independent HGT identification at the community-level.

Results: Assessment of MetaCHIP's performance on simulated datasets revealed that it can predict HGTs having various degrees of genetic variation from metagenomic datasets. The results also indicated that the detection of very recent gene transfers, having a low level of genetic mutations, from metagenomics datasets is largely affected by the reads assembly step. Assessment of MetaCHIP's performance on real datasets confirmed the role of HGT in the spread of genes related to antibiotic resistance in the human gut microbiome.

Conclusion: MetaCHIP provides an opportunity to study HGTs among members of a microbial community and therefore has several applications in the field of microbial ecology and evolution. It is implemented in Python and freely available at: https://github.com/songweizhi/MetaCHIP.

Keywords: Metagenomics, horizontal gene transfer, HGT identification, taxonomy classification, phylogenetic clustering, bioinformatics

Background

Genome reconstruction (binning) of uncultured microorganisms has recently become feasible due to the comprehensive sequencing of microbial community DNA (metagenomic DNA) and novel computational approaches [1-3]. These reconstructed genome bins have provided new insights into the biochemistry, physiology and adaptation of previously uncharacterized microbial groups [4-8]. Moreover they offer the opportunity to study HGTs within communities of uncultured microorganisms.

Horizontal gene transfer (HGT), the transmission of genetic information between organisms, is thought to be an important driver of microbial evolution and adaptation, including the development of antibiotics resistance and virulence [9, 10]. Several bioinformatics tools have been developed using a range of algorithms and features to identify HGTs. For example, GIST [11] and IslandViewer [12] utilize compositional features of genome sequences to predict HGT events, while DarkHorse [13] and HGTector [14] use sequence similarities (best-matches). Explicit phylogenetic approaches are employed by Ranger-DTL [15] and AnGST [16], which predict HGTs through the reconciliation of gene tree with corresponding species tree.

However, current HGT detection methods cannot be applied to entire communities or require reference genomes. For example, HGTector [14] can only detect HGTs from members in a defined distal group to defined self-group members, which limits its application to predict HGTs among all members of a microbial community, while DarkHorse [13] requires suitable references genomes to predict HGTs, which are often not available for metagenomic datasets.

Hence, we developed here MetaCHIP ("Meta" for "metagenomics", "CHIP" for "Community-level HGT Identification Pipeline"), a pipeline for the reference-independent and community-level identification of HGTs. Our analysis of simulated and real data showed that MetaCHIP can detect HGTs from communities with high degree of confidence and to give new biological and ecological insights.

Methods

The workflow of MetaCHIP is presented in **Figure 1**. MetaCHIP uses both best-match and phylogenetic approaches (see above). MetaCHIP's input is the sequence file of a set of genome bins derived from metagenomic data. Gene prediction is performed with Prodigal [17] and genomes are clustered according to their organismal phylogeny. As the 16S rRNA gene, which is the most commonly used phylogenetic and taxonomic marker, is often missing in genome bins [18-20], we build a phylogenetic tree for all input genomes based on the protein sequences of 43 universal single-copy genes (SCG) used by CheckM [21]. Predicted protein sequences for input genomes are searched for the PFAM [22] and TIGRFAM [23] hmm profiles (version 31.0 and 14.0, respectively) of these SCG proteins using HMMER 3.1b2 [24]. Protein sequences for each hmm profile are aligned using MAFFT [25], then concatenated and a phylogenetic tree is built using FastTree 2.1.9 [26]. The produced SCG protein tree is then converted into a distance matrix and clustered using the Nearest Point Algorithm implement in SciPy [27]. As no "best" algorithm exists to cluster phylogenetic sequences [28], clustering profiles generated in this step should be manually curated by comparing them with taxonomic classifications of the input genomes (if available) prior to the HGT identification step. Taxonomy classification of the input bins can be obtained either with published pipelines, like PhyloSift [29], or with the recently developed GTDB-Tk [30], which is based on the Genome Taxonomy Database (GTDB) [31].

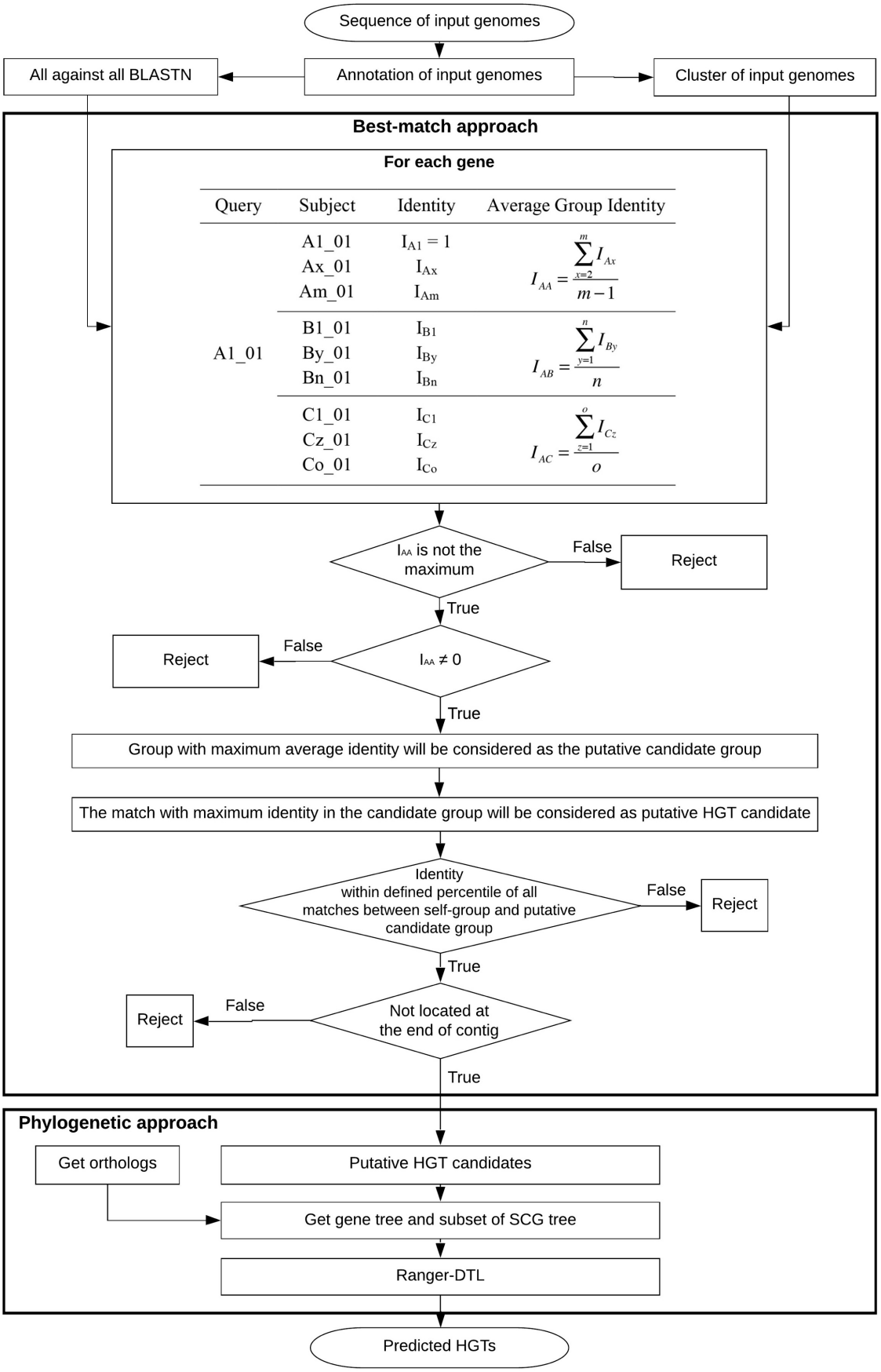


Figure 1. Workflow of MetaCHIP.

Best-match approach

An all-against-all BLASTN search is performed among all predicted open reading frames (ORFs) from the input genomes. The BLASTN results are first filtered with user-defined alignment length (e.g. 200 bp) and coverage cut-offs (e.g. 70%). The filtered matches are then compared between groups of genomes using the following steps. Here, we suppose all input genomes are divided into three groups (A, B and C), with individual genomes referred to as Ax, By and Cz, respectively **(Figure 1)**. Genes from each genome are represented as Ax_N, By_N, Cz_N. Take gene A1_01 as an example, the number of its BLASTN matches from group A, B and C are m, n and o, respectively, with their corresponding identities being I_{Ax}, I_{By} and I_{Cz}. The average identities of the matches from each group are I_{AA}, I_{AB} and I_{AC}, respectively **(Figure 1)**. The following analysis are then performed for each gene (here as an example with A1_01):

1. If I_{AA} is the maximum, which means all its best matches are coming from the self-group, then gene A1_01 is not a candidate for HGT.
2. If $I_{AA} = 0$ (that is only the self-match was found from group A), then all BLASTN matches from other groups will be ignored. This is because, if the non-self-group subject with maximum identity was considered a HGT candidate, then it is very likely to be a false positive due to the lack of self-group matches.
3. If $I_{AA} \neq 0$ and I_{AA} is not the maximum, then the non-self-group with maximum average identity (e.g. I_{AB} or I_{AC}) will be considered as a putative candidate group for HGT.
4. The BLASTN match with maximum identity in the candidate group will be considered the putative HGT candidate.
5. Identity distribution of all genes between the self-group and the putative candidate group are summarized. The identity cut-off corresponding to pre-defined percentile

(e.g. the highest 10%) is calculated. Only putative HGT candidates which have identities higher than this cut-off, will be further considered.

Analysis of regions flanking putative HGTs

Assembly algorithm based on DeBruijn graphs (e.g. SOAP, Velvet, SPAdes, IDBA) will produce "bubbles" for sequence regions with sequencing error, but high similarity [32]. The resolution of such bubbles may produce two contigs with overlapping sequences of the same region. This duplication could be falsely considered in the HGT analysis and to avoid this, putative HGT candidates located at the end of contigs are disregarded.

To further corroborate the predicted HGT candidates, their flanking sequences within user-defined length cut-off (e.g. 3 Kbp) are extracted from the annotation files. A pairwise BLASTN is performed between each pair of flanking regions. The genomic regions are plotted with GenomeDiagram [33] for visual inspection (**Figure 2**).

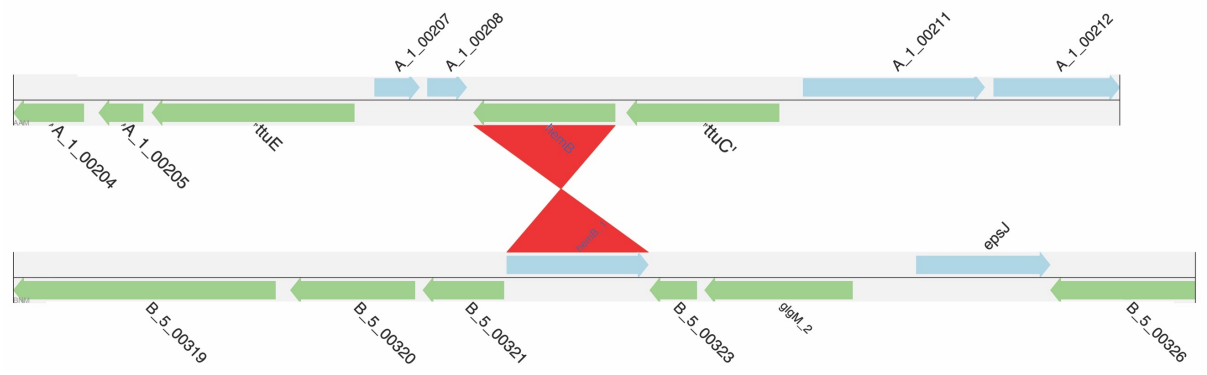


Figure 2. Example output for the flanking regions of an identified HGTs. Genes coded on the forward strand are displayed in light blue, and genes coded on the reverse strand are displayed in light green. The names of matched genes are displayed in blue.

Phylogenetic approach

An explicit phylogenetic approach is used to further corroborate the results given by the best-match approach and to provide information on the direction of the gene transfer. All protein

orthologs within the set of input genomes are obtained using Get_Homologues [34] with 70% sequence identity and 70% coverage as minimum BLASTN cut-off, the minimal number of proteins for each cluster is set to three and their differences in length are no less than 70%. The protein tree for the orthologous groups, which includes the HGT candidates predicted by the best-match approach, are constructed as follows: First, amino acid sequences are aligned with MAFFT and a phylogenetic tree is constructed using FastTree with default parameters. A subset of the SCG protein tree that includes all members in the corresponding ortholog is constructed as described above. The reconciliation between each pair of ortholog tree and SCG protein tree is performed using Ranger-DTL 1.0. Briefly, Ranger-DTL predicts HGTs by performing a duplication-transfer-loss (DTL) reconciliation between a protein family phylogeny and its corresponding organismal phylogeny [15].

Performance on simulated dataset

To assess the performance of MetaCHIP, ten complete chromosomes each from the class Alphaproteobacteria and Betaproteobacteria were random;y selected from the NCBI database (**Table S1**). To assess how reliable SCG protein trees are to reconstruct organismal phylogenies from partial genome bins, the selected 20 genomes were each divided into 100 contigs with equal length and 20, 40, 60 and 80 contigs were randomly selected to represent genome bins with 20, 40, 60 and 80% completeness, respectively. The similarities between the SCG protein trees with these different levels of completeness and the tree based on 16S rRNA gene sequences were then assessed by performing Mantel tests [35].

To simulate HGTs, ten genes from each of the 10 Alphaproteobacteria genomes were selected and randomly transferred into the 10 Betaproteobacteria genomes (**Table S2**) with different levels of genetic variation (0, 5, 10, 15, 20, 25 and 30%) using HgtSIM [36]. The

six-frame stop codon sequence “TAGATGAGTGATTAGTTAGTTA” was added to the two ends of transferred genes to facilitate correct gene prediction. The donor genomes and mutated recipient genomes were either used directly as inputs into MetaCHIP or sequencing reads were simulated. For the latter, 10 donor genomes were combined with the 10 recipient genomes for each level of genetic variation separately, and sequencing reads for each group were simulated three times with different abundance profiles (**Table S3**) using GemSIM [37].

As the reconstruction of genes involved in HGT are highly affected by sequencing depth or the assembler used [36], 3, 6, 9 and 12 million reads, corresponding to average coverage of approximately 6, 11, 17 and 23x, were simulated for each level of genetic variation. The paired-end reads were quality filtered using Trimmomatic [38] with a quality cut-off of 20 and a sliding window of 6 bp. Reads from three replicates were combined and then assembled with IDBA_UD 1.1.1 [39] or metaSPAdes 3.9.0 [40] and contigs were filtered with a length cut-off of 2500 bp. A gene transfer was considered to be reconstructed during the assembly process, if at least one of the gene’s two flanking regions was >1 Kbp and the flanking region matched the recipient genome [36]. The existence of gene transfers in the filtered contigs was analysed by performing a pairwise BLASTN between transferred genes and the contigs for each level of genetic variation. The BLASTN results were then filtered with an identity cut-off of 98% and a coverage cut-off of 98% for the transferred genes.

Genome binning was performed with MetaBAT [1] and MyCC [2] and the results were refined with Binning_refiner [41]. The overall precision (defined as how pure a bin is) and recall (defined as how complete a bin is) of the generated bins were assessed with Evaluate.py from MyCC. Bin completeness and contamination were also assessed with CheckM. The correlations between the genome bins and the reference genomes were

obtained by running pairwise BLASTN searches. The correlations between MetaCHIP's predicted HGTs and the known/simulated gene transfers were determined by running pairwise BLASTN searches with identity and coverage cut-off of 98%. We also investigated how the recovery of gene transfers is influenced by setting a cut-off for the distance to the end of the contig (1, 2, 5, 10 and 50 Kbp).

Performance on real dataset

Genome bins derived from metagenomic datasets for microbiomes from human guts [1, 42] and seawater samples taken in the North Sea [43] were used to test the performance of MetaCHIP on real datasets. For the human gut dataset, genome bins previously produced by MetaBAT [1] were used directly here. For the North Sea dataset, all sequencing reads were quality filtered with Trimmomatic as previously described [43] and assembled using metaSPAdes (version 3.10.1). Binning was performed with MetaBAT (0.32.5) and MyCC (v2017) and bins were refined using Binning_refiner. CheckM was subsequently used to assess the quality (contamination and completeness) of the genome bins with the non-lineage specific marker gene mode. The SCG protein tree of these bins was built with the algorithm described above. Annotation of predicted HGTs was performed by running RPS-BLAST [44] against the COG database [45]. COGs related to antibiotic resistance were retrieved from the antibiotic resistance genes database (ARDB; April 2018) [46].

Results and discussion

Performance on simulated dataset

MetaCHIP initially clusters the input genomes based on phylogenetic information to identify clusters between which HGT should be analyzed. We therefore first assessed how reliable the reconstruction of a SCG-based phylogeny is for incomplete genomes. The results showed a

high degree of congruence between the SCG protein trees and the tree based on 16S rRNA gene sequences for genome bins with completeness higher than 40% (**Figure 3**). This value is thus suggested for the completeness cut-off for genomes bins used as input for MetaCHIP.

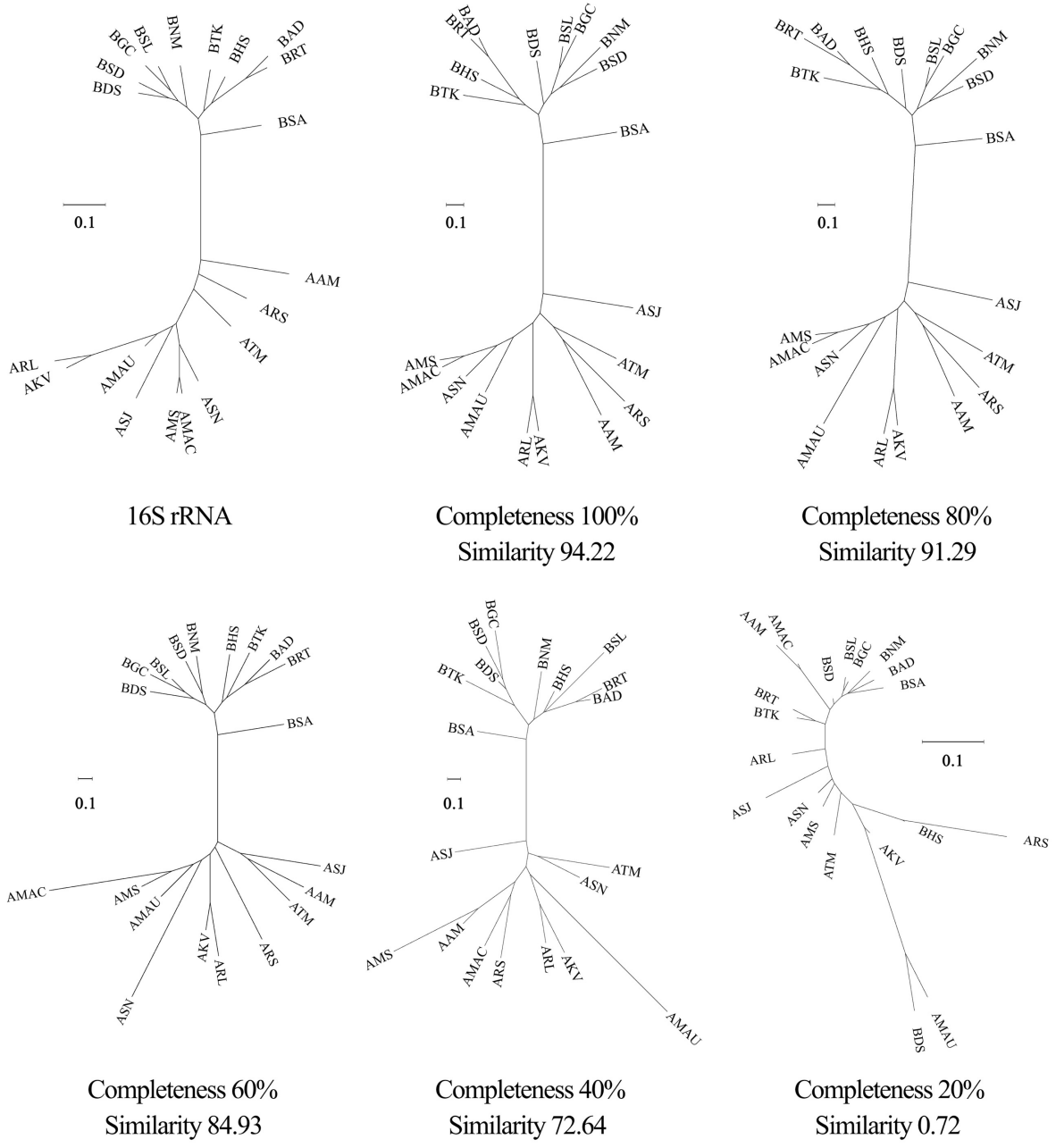


Figure 3. The similarity between the tree based on 16S rRNA gene sequences and the SCG protein trees with different level of genome completeness. Similarities were assessed by Mantel tests.

We next tested how effective MetaCHIP is in recovering HGTs from completed genomes (i.e. the genomes used here without any read simulation, assembly or binning). No less than 98% of introduced gene transfers were recovered by the best-match approach when genetic variations were 10% or below, of which no less than 66% were also identified by the phylogenetic approach. With higher levels of genetic variation, a steady decline in recovery was observed with only 12% of HGT events being found at a variation level of 30%. The phylogenetic analysis predicted the correct directions of gene flow in more than 80% of cases **(Figure 4)**.

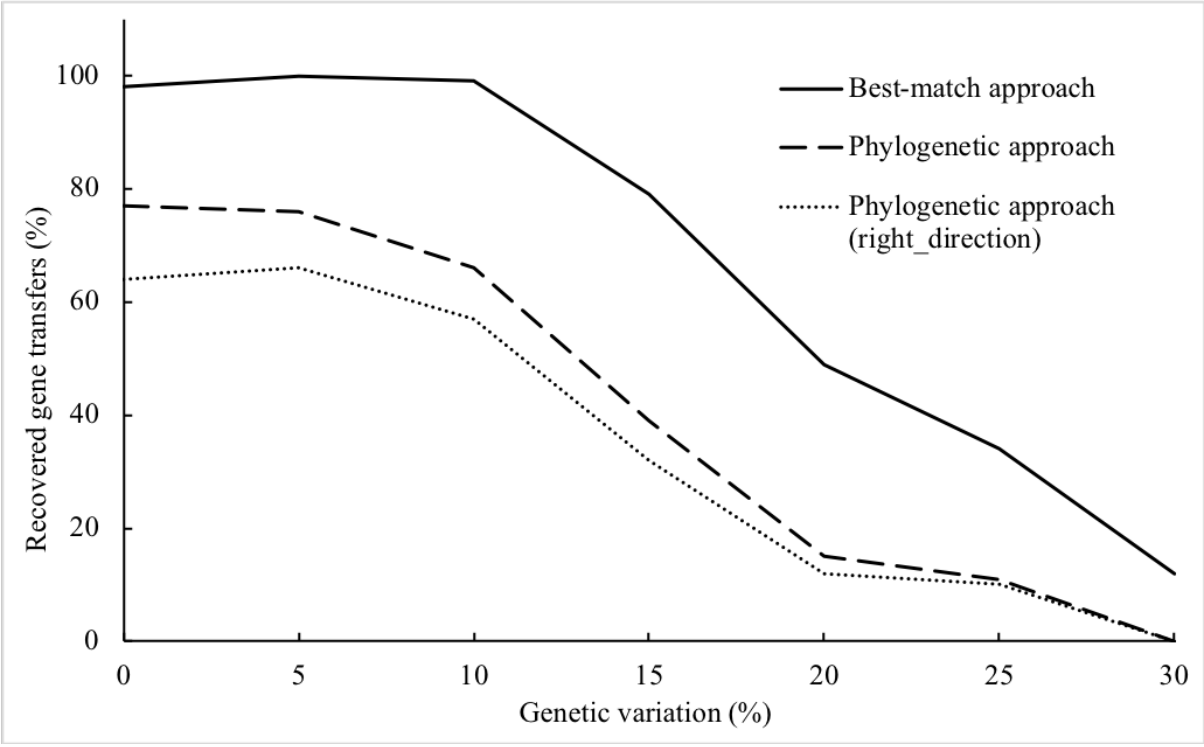


Figure 4. The number of recovered gene transfers as a function of genetic variation for genomes without read simulation.

We next evaluated how different assemblers and sequencing depths influence the recovery of HGTs with different genetic variations. When no variations were introduced to the

transferred genes, more transferred genes were recovered by metaSPAdes than with IDBA_UD. For 5% genetic variation both assembler performed overall quite poorly in terms of the recovery rate of introduced gene transfers, but IDBA-UD had generally a better recovery rate than metaSPAdes. IDBA_UD showed also better recovery for HGTs with variation levels of 10-30% **(Figure 5)**. MetaSPAdes was therefore used for the assembly of metagenomic reads with no variations, while IDBA_UD was selected for the other levels of genetic variation. For gene transfers with no variations, the recovery rate for metaSPAdes assemblies was highest with a sequencing depth of 11.33x, beyond which it declined. For the 5% genetic variation, the best recovery from the IDBA_UD assemblies was at sequencing depths of 11.3x or greater **(Figure 5)**. As a compromise for the non-linear behaviour of recovery rates, a sequencing depth of 17x (9 million reads) was selected for all subsequent simulations.

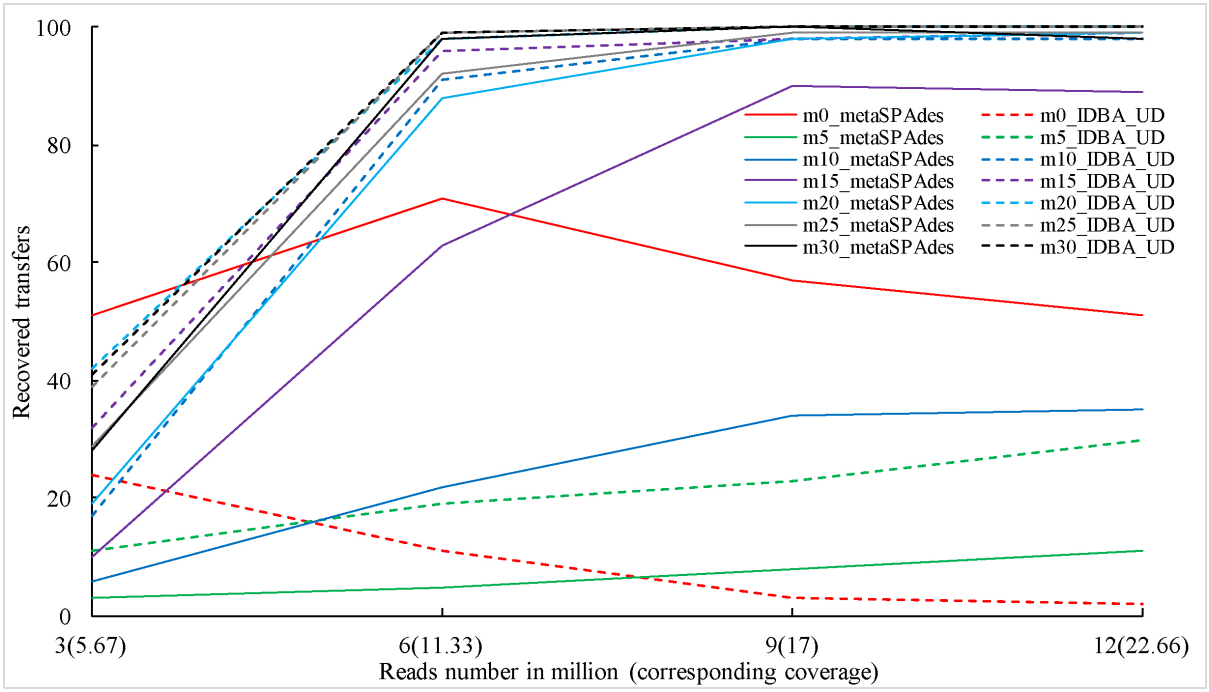


Figure 5. The effect of sequencing depth on the recovery of introduced gene transfers with different assemblers (metaSpades and IDBA_UD) and different levels of genetic variation.

Based on these choices of coverage and assembler we next binned genomes from the simulated datasets. The overall precision and recall of the genome bins we generated were for all variation groups not lower than 99.73% and 89.49%, respectively (**Table 1**).

Table 1. Precision and recall rate of refined genome bins at different level of genetic variation of HGTs

Genetic variation (%)	0	5	10	15	20	25	30
Precision (%)	99.73	99.96	99.95	99.97	99.93	99.97	100.00
Recall (%)	89.49	93.20	95.92	96.46	95.41	96.45	96.35

We next investigated the presence of introduced gene transfers in these genome bins. For 0% genetic variation, 30% of introduced gene transfers were identified in the genome bins and all of them were found in the recipient genomes. For the levels of genetic variation greater than 5%, no less than 73.7% of transferred gene copies were found in both the donor and recipient genome bins (**Figure 6**).

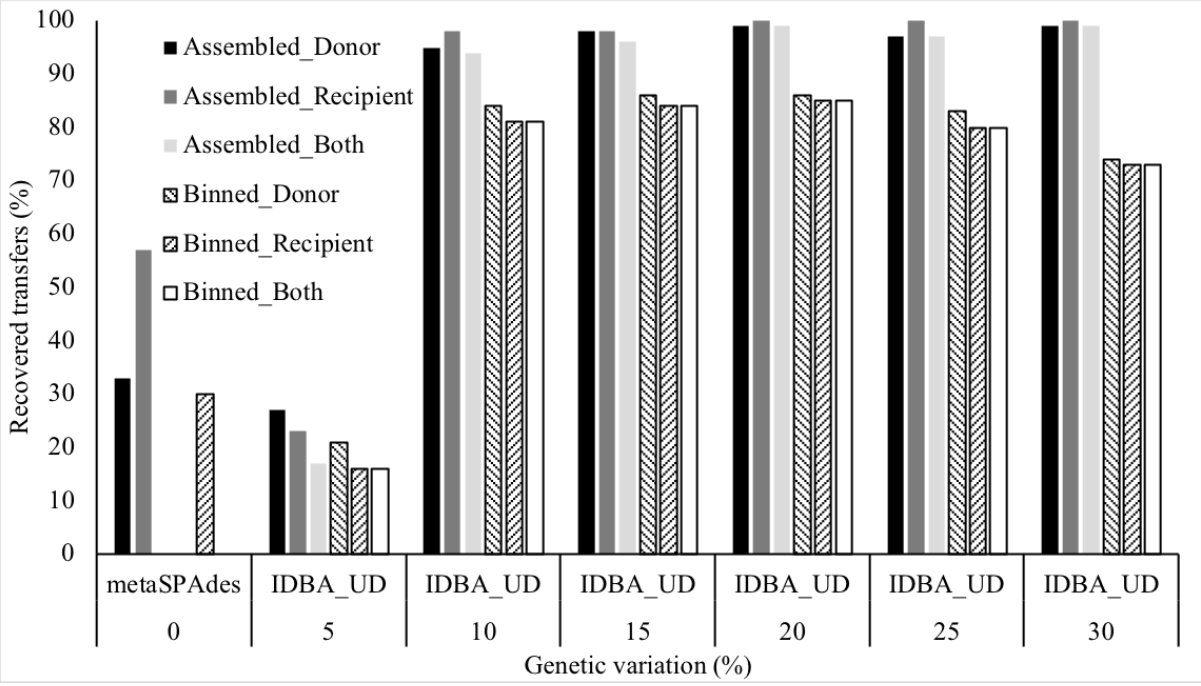


Figure 6. The percentage of gene transfers recovered in genome bins during assemble and binning.

The influence of setting a cut-off in MetaCHIP for the distance of the predicted gene transfer to the end of the assembled contigs was also investigated. This showed that for distance cut-off of 1-10Kb there were minimal changes in the recovery rate of transfers for any given levels of genetic variation (**Figure 7**). By default, MetaCHIP is using a cut-off of 1 Kbp.

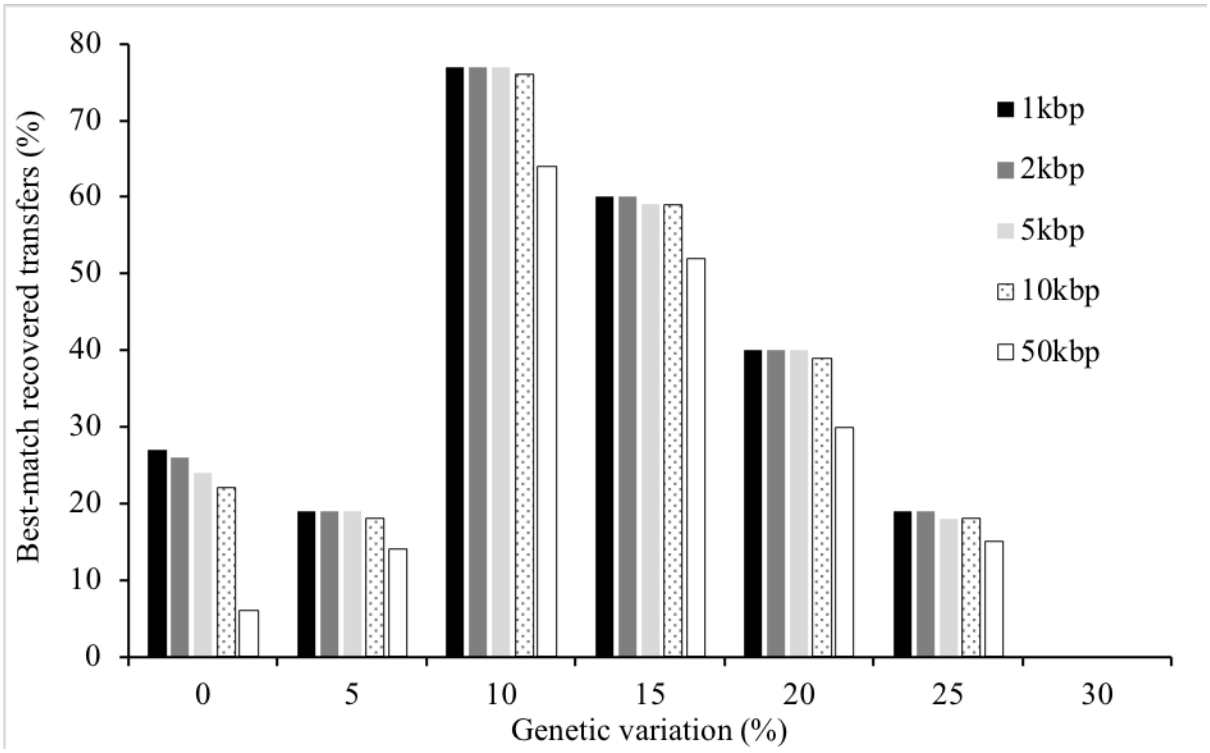


Figure 7. The effect of end-sequence length cut-offs on the recovery of gene transfers by the best-match approach.

By applying MetaCHIP to the genome bins, 27% of the 100 introduced gene transfers were recovered by the best-match approach for the 0% genetic variation and 6 of them were validated by the phylogenetic approach (**Figure 8**), which accounts for 90% and 20%, respectively, of the gene transfers that actually exist in the genome bins. For a 5% genetic variation, all introduced gene transfers that were found in the bins were also identified by the

best-match approach and 63% of them were validated by the phylogenetic approach (**Figure 8**). The best recovery rates were obtained when the genetic variation is 10%, where at least 77% of introduced gene transfers were recovered by the best-match approach and 55 of them were validated by the phylogenetic approach, which accounted for 92% and 67% of all the binned gene transfers, respectively. A steady decline in the ability of MetaCHIP to detect HGT was also observed with higher genetic variations (**Figure 8**), similar to what was seen for the original genomes (**Figure 4**).

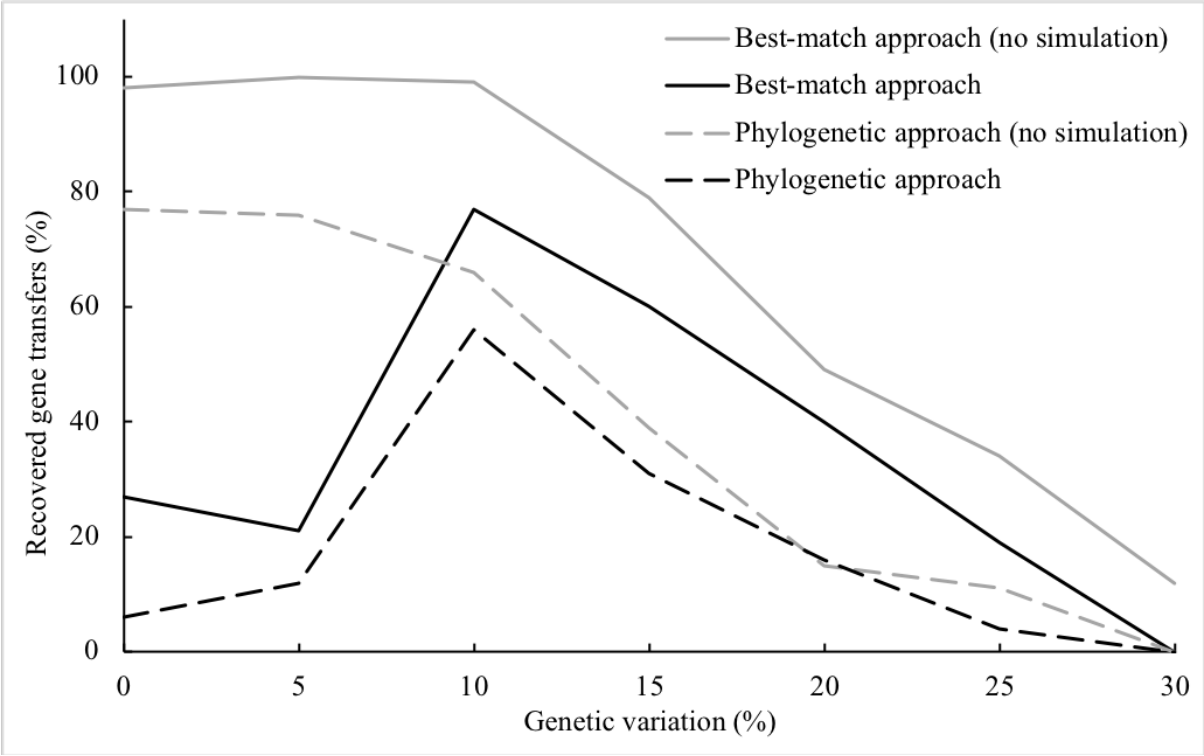
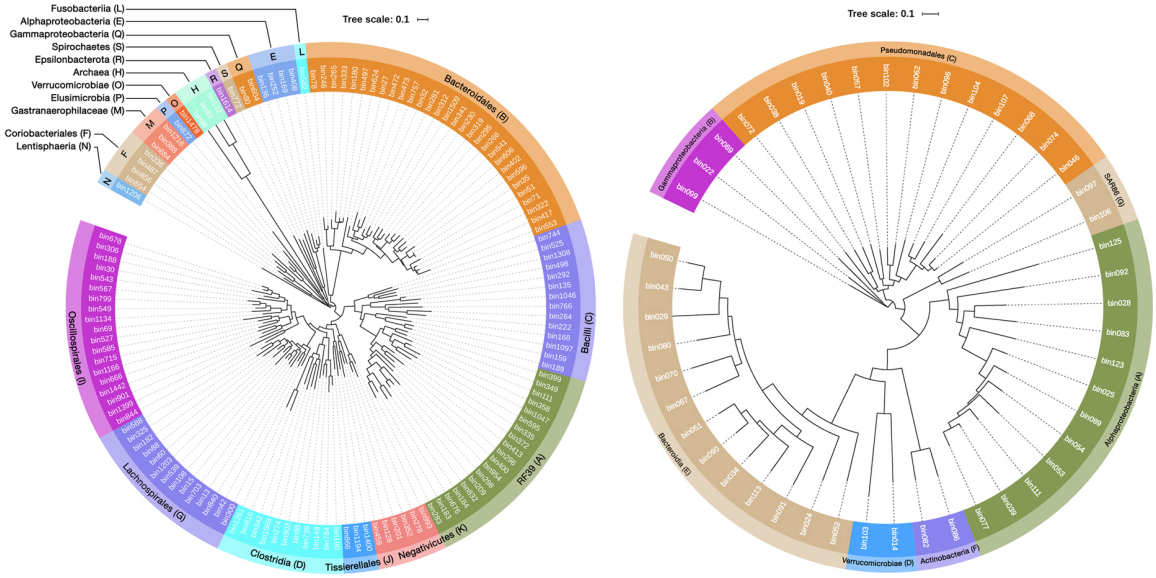


Figure 8. The percentage of recovered gene transfers by MetaCHIP after assembly of simulated reads and binning of genomes (simulation). For comparison, the results from original genomes (non-simulation) are also shown and are the same as in Figure 4.

Performance on real dataset

For the metagenomic dataset of free-living microorganisms in the North Sea, sequence assembly with metaSPAdes generated 315.33 Mbp of contiguous sequences ≥ 2500 bp

317 (35,190 contigs) and 69 genome bins were obtained, of which 47 had no contamination
318 detected with CheckM and with completeness higher than 40%. For the 1634 genome bins
319 obtained from the human gut dataset 138 were estimated to be contamination-free and more
320 than 40% complete (**Additional file 2**). The taxonomy of qualified genome bins was
321 analysed with GTDB-Tk (**Additional file 3**). The human gut and the North Sea bins were
322 clustered into 19 and 7 groups, respectively, after the automated clustering with MetaCHIP
323 and manual curation (**Figure 9**).



324

325 **Figure 9.** Grouping of the human gut and North Sea genome bins. Inner ring shows the bin
326 IDs, while the outer ring shows the cluster IDs and the lowest taxonomic assignment that
327 between the genomes in each cluster.

328

329 The best-match approach detected 623 gene transfers from the human gut genome bins and of
330 which 66 were also found by the phylogenetic approach. For the North Sea dataset, 208 and
331 28 gene transfers were detected by the two approaches, respectively. The direction of
332 predicted gene flows within the two communities were shown in **Figure 10.**

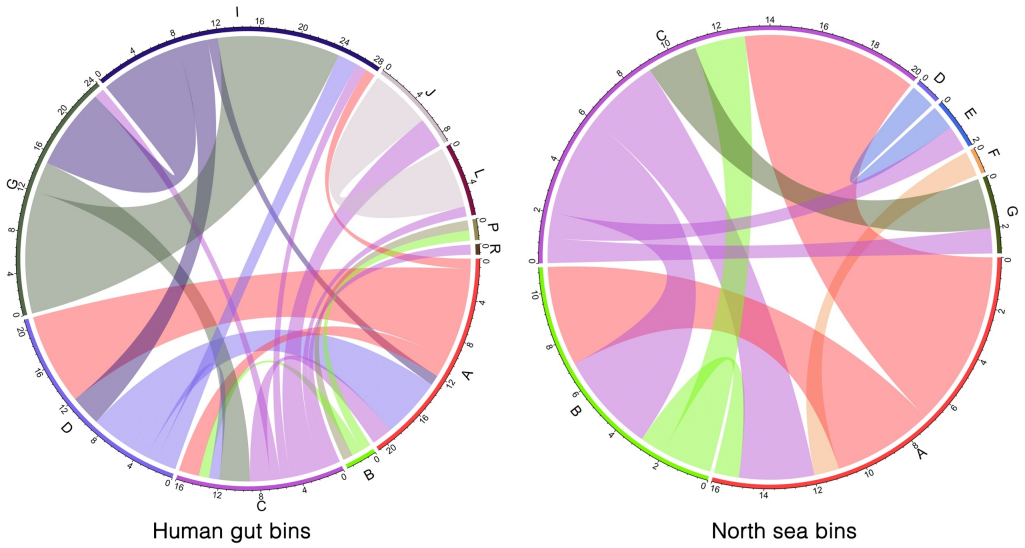


Figure 10. Predicted gene flow within the human gut and North Sea microbial communities. Bands connect donors and recipients, with the width of the band correlating to the number of HGTs and the colour corresponding to the donors.

We next performed a functional annotation of the genes identified in the HGT analysis based on the COG system (**Figure 11**). A COG category was considered to be enriched in the HGT dataset if its proportions in both the best-match and phylogenetic approaches were above the 75% percentile of its relative abundance across all input genomes. The results for the human gut dataset showed that genes subject to HGT were enriched for the COG categories of defence mechanisms (V), energy production and conversion (C) as well as posttranslational modification, protein turnover, chaperones (O) (**Figure 11**). The enrichment of category V was mainly due to individual COG functions related to antibiotic resistance (AR), which included ABC-type multidrug (COG1131) and antimicrobial peptide (COG1136) transport system, A/G-PRTase and related PRPP-binding proteins (COG0503) as well as acetyltransferases (COG0456). AR-related genes identified to be subject to HGT in the human gut microbiome together made up 12.5% of all predictions, while this value was only 3.7% for the North Sea dataset. The observation made here with MetaCHIP is consistent with previous observations and proposals that HGT is a dominant factor for the spread of AR in

the human gut microbiota [47-49]. For example, a variety of genes for ABC-type multidrug transport systems have been previously found to be often associated with transposable elements in gut microbiomes and this was postulated to facilitate their horizontal transfer [50].

COG categories preferentially subject to HGT between the free-living microorganisms in the North Sea include energy production and conversion (C) as well as amino acid transport and metabolism (E) (**Figure 11**). This observation is similar to a recent study on the inter-phylum HGTs among all available complete genome for free-living Archaea and Bacteria, where transferred genes most frequently also belonged to COG categories C and E [51].

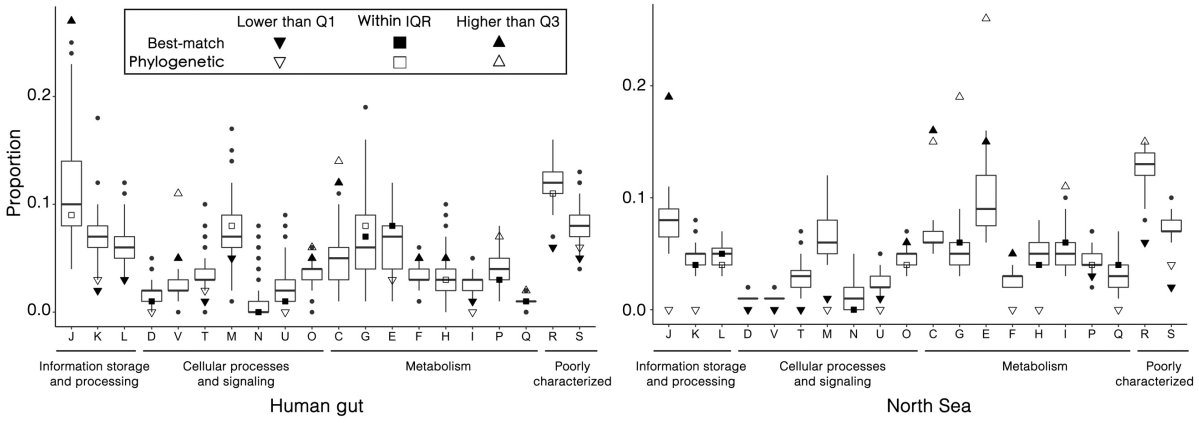


Figure 11. Relative proportion of COG functional categories for the input genome bins and predicted HGTs from human gut bins (left) and North Sea bins (right). The boxes in the plot are bound by the 25% to 75% quartile proportions with the thick line being the median value. Q1, Q3 and IQR refer to the 25%, 75% and interquartile range, respectively. The upper whisker refers to the largest observation less than or equal to upper Q3 + 1.5 * IQR, while the lower whisker refers to the smallest observation greater than or equal to Q1 - 1.5 * IQR. Letters on X-axis indicate COG categories: C (energy production and conversion), D (cell cycle control, cell division, chromosome partitioning), E (amino acid transport and metabolism), F (nucleotide transport and metabolism), G (carbohydrate transport and metabolism), H (coenzyme transport and metabolism), I (lipid transport and metabolism), J (translation, ribosomal structure and biogenesis), K (transcription), L (replication, recombination and repair), M (cell wall/membrane/envelope biogenesis), N (cell motility), O

(posttranslational modification, protein turnover, chaperones), P (inorganic ion transport and metabolism), Q (secondary metabolites biosynthesis, transport and catabolism), R (general function prediction only), S (function unknown), T (signal transduction mechanisms), U (intracellular trafficking, secretion, and vesicular transport), and V (defence mechanisms).

More than 50% of identified HGTs for the human gut and the North Sea bins had genetic variations of 25 ± 2.5% (**Table 2**). The best-match and phylogenetic approaches only detected 19% and 4%, respectively, of introduced gene transfers with this level of genetic variation in the simulated datasets (**Figure 8**) and hence the actual number of HGTs that occurred in the community are likely to be underestimated here. Interestingly, in both datasets only one transfer with genetic variations at around 15% or below were detected, for which we generally found high recovery rate in our simulations (**Figure 8**). This may indicate that HGT in these microbial communities does not involve a large number of recent transfers.

Table 2. Genetic variation of HGT identified by MetaCHIP's phylogenetic approach

Dataset	Approach	Genetic variation (± 2.5%)						
		0	5	10	15	20	25	30
Human gut	Best-match	0	0	0	2	71	434	116
	Phylogenetic	0	0	0	1	18	36	11
North Sea	Best-match	0	0	0	1	14	120	73
	Phylogenetic	0	0	0	0	3	16	8

Conclusion

Our tests of MetaCHIP showed that it can detect HGTs with various degree of genetic variation from microbial communities, but that prediction efficiency is affected by several factors. First, as transferred genes will undergo mutations in their new genome contexts, their detections will become difficult when the similarities between the donor and recipient genes fall below certain levels (Figure 3) [52]. Second, the detection of recent gene transfers (i.e. those with very little variation between donor and recipient) is largely affected by technical

limitations of metagenomic analyses. As current sequencing technologies and assemblers often failed to assemble long regions with high sequence similarity [36, 53], recent HGTs will not be captured effectively in the genomic context of the donor and recipient (**Figure 5**). This problem might be addressed in the near future by long-read sequencing technologies, such as PacBio's sequencing platform [54], when applied to metagenomic samples. Third, the successful detection of HGT from metagenomic dataset requires the reliable reconstruction of the organismal genome, in particular through genome binning, as mis-binned sequences (contamination) may introduce false positives in the HGT analysis and reliable organismal tree for phylogeny-based prediction of HGTs requires a certain degree of genome completeness (e.g. 40%) (**Figure 3**). Improvement of genome binning accuracy can be achieved either by incorporating more biological samples [1] or by combining the binning results from multiple binning programs [41], while the completeness of genome bins can be improved with higher sequencing depth. Despite these limitations, our analysis of simulated and real data with MetaCHIP shows that HGT can be detected from microbial communities with high degree of confidence to give new biological and ecological insights. However, the absolute numbers of HGTs that occur in the community might be underestimated given the limitations outlined above.

Availability and requirements

Project name: MetaCHIP

Project home page: https://github.com/songweizhi/MetaCHIP

Operating system(s): Linux and Mac

Programming language: Python

License: GNU AGPLv3

List of abbreviations

AR: antibiotic resistance

ARDB: antibiotic resistance genes database

COG: Clusters of Orthologous Groups

GTDB: Genome Taxonomy Database

HGT: horizontal gene transfers

NCBI: National Center for Biotechnology Information

SCG: single copy gene

Declarations

Funding

This research is funded by the Australian Research Council. Weizhi Song and Shan Zhang are funded by the China Scholarship Council.

Authors' contributions

WS, BW and TT developed the method. WS and BW designed and wrote the software components. WS, SZ and KS performed the analysis. WS wrote the manuscript. TT supervised the project. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Additional files

Additional file 1: Supplementary tables. (DOCX 48kb)

Additional file 2: The quality of input genome bins. (XLSX 15kb)

Additional file 3: Taxonomy classification of input genome bins. (XLSX 18kb)

References

1. Kang DD, Froula J, Egan R, Wang Z: **MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities.** *Peerj* 2015, **3**.
2. Lin H-H, Liao Y-C: **Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes.** *Scientific reports* 2016, **6:**24175.
3. Sangwan N, Xia F, Gilbert JA: **Recovering complete and draft population genomes from metagenome datasets.** *Microbiome 1* 2016, **4**.
4. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH: **Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes.** *Nature biotechnology* 2013, **31:**533.
5. Ji M, Greening C, Vanwonterghem I, Carere CR, Bay SK, Steen JA, Snape I: **Atmospheric trace gases support primary production in Antarctic desert surface soil.** *Nature 400* 2017, **552**.
6. Moitinho-Silva L, Vives C, Batani G, Esteves AI, Jahn MT, Thomas T, The I: **Díez- & Integrated metabolism in sponge-microbe symbiosis revealed by genome-centered metatranscriptomics.** *journal 1651* 2017, **11**.
7. Probst AJ, Ladd B, Jarett JK, Geller-McGrath DE, Sieber CM, Emerson JB, Klingl A: **Differential depth distribution of microbial function and putative symbionts**

through sediment-hosted aquifers in the deep terrestrial subsurface. *Nature Microbiology 1* 2018.

8. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Dodsworth JA: **Insights into the phylogeny and coding potential of microbial dark matter.** *Nature* 2013, **499:**431-437.
9. Dagan T, Artzy-Randrup Y, Martin W: **Dagan T, Artzy-Randrup Y, Martin W.. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci USA 105: 10039-10044.** *Proceedings of the National Academy of Sciences* 2008, **105:**10039-10044.
10. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation.** *Nature* 2000, **405:**299-304.
11. Hasan MS, Liu Q, Wang H, Fazekas J, Chen B, Che D: **GIST: Genomic island suite of tools for predicting genomic islands in genomic sequences.** *Bioinformation* 2012, **8:**203-205.
12. Langille MG, Brinkman FS: **IslandViewer: an integrated interface for computational identification and visualization of genomic islands.** *Bioinformatics* 2009, **25:**664-665.
13. Podell S, Gaasterland T: **DarkHorse: a method for genome-wide prediction of horizontal gene transfer.** *Genome Biology* 2007, **8:**1-18.
14. Zhu Q, Kosoy M, Dittmar K: **HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers.** *BMC genomics* 2014, **15:**717.
15. Bansal MS, Alm EJ, Kellis M: **Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss.** *Bioinformatics* 2012, **28:**i283-i291.
16. David LA, Alm EJ: **Rapid evolutionary innovation during an Archaean genetic expansion.** *Nature* 2011, **469:**93-96.
17. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification.** *BMC bioinformatics* 2010, **11:**119.
18. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Banfield JF: **Unusual biology across a group comprising more than 15% of domain Bacteria.** *Nature* 2015, **523:**208-211.
19. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Tyson GW: **Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life.** *Nature microbiology 1533* 2017, **2**.
20. Yuan C, Lei J, Cole JR, Sun Y: **Reconstructing 16S rRNA genes in metagenomic data.** *Solid-state Circuits Conference* 2015, **51:**1-3.
21. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW: **CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.** *Genome research* 2015, **25:**1043-1055.
22. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J: **Pfam: the protein families database.** *Nucleic acids research* 2013, **42:**D222-D230.
23. Haft DH, Selengut JD, White O: **The TIGRFAMs database of protein families.** *Nucleic acids research* 2003, **31:**371-373.
24. Eddy SR: **Accelerated profile HMM searches.** *PLoS computational biology e1002195* 2011, **7**.

25. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Molecular Biology & Evolution* 2013, **30:**772-780.
26. Price MN, Dehal PS, Arkin AP: **FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments.** *Plos One* 2010, **5**.
27. Jones E, Oliphant T, Peterson P, Joo HS, Fu CI, Otto M: **SciPy}: open source scientific tools for {Python}. & Bacterial strategies of resistance to antimicrobial peptides.** *Phil Trans R Soc B 0292* 2015, **371**.
28. Guha S, Mishra N: **Clustering data streams.** In *Data Stream Management.* Springer; 2016: 169-187
29. Darling AE, Jospin G, Lowe E, Matsen IV FA, Bik HM, Eisen JA: **PhyloSift: phylogenetic analysis of genomes and metagenomes.** *PeerJ* 2014, **2:**e243.
30. **Gtdb-Tk** [https://github.com/Ecogenomics/GtdbTk]
31. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P: **A proposal for a standardized bacterial taxonomy based on genome phylogeny.** *bioRxiv 256800* 2018.
32. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G: **De novo assembly and genotyping of variants using colored de Bruijn graphs.** *Nature genetics* 2012, **44:**226-232.
33. Pritchard L, White JA, Birch PR, Toth IK: **GenomeDiagram: a python package for the visualization of large-scale genomic data.** *Bioinformatics* 2006, **22:**616-617.
34. Contreras-Moreira B, Vinuesa P: **GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis.** *Applied & Environmental Microbiology* 2013, **79:**7696-7701.
35. Mantel N: **The detection of disease clustering and a generalized regression approach.** *Cancer research* 1967, **27:**209-220.
36. Song W, Steensen K, Thomas T: **HgtSIM: a simulator for horizontal gene transfer (HGT) in microbial communities.** *Peerj* 2017, **5**.
37. McElroy KE, Luciani F, Thomas T: **GemSIM: general, error-model based simulator of next-generation sequencing data.** *Bmc Genomics* 2012, **13:**1-9.
38. Bolger AM, Lohse M, Usadel B: **Trimmomatic: A flexible read trimming tool for Illumina NGS data.**
39. Peng Y, Leung HC, Yiu SM, Chin FY: **IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth.** *Bioinformatics* 2012, **28:**1420-1428.
40. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA: **metaSPAdes: a new versatile metagenomic assembler.** *Genome Research* 2017, **27**.
41. Song W, Thomas T: **b). Binning_refiner: improving genome bins through the combination of different binning programs.** *Bioinformatics* 2017, **33:**1873-1875.
42. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Mende DR: **A human gut microbial gene catalogue established by metagenomic sequencing.** *nature 59* 2010, **464**.
43. Wemheuer B, Wemheuer F, Hollensteiner J, Meyer FD, Voget S, Daniel R: **The green impact: bacterioplankton response toward a phytoplankton spring bloom in the southern North Sea assessed by comparative metagenomic and metatranscriptomic approaches.** *Frontiers in microbiology 805* 2015, **6 SRC - BaiduScholar**.
44. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of molecular biology* 1990, **215:**403-410.

45. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science (New York, NY)* 1997, **278:**631-637.
46. Liu B, Pop M: **ARDB—antibiotic resistance genes database.** *Nucleic acids research* 2008, **37:**D443-D447.
47. von Wintersdorff CJ, Penders J, van Niekerk JM, Mills ND, Majumder S, van Alphen LB, Savelkoul PH, Wolffs PF: **Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer.** *Frontiers in microbiology* 2016, **7:**173.
48. Duranti S, Lugli GA, Mancabelli L, Turroni F, Milani C, Mangifesta M, Ventura M: **Prevalence of antibiotic resistance genes among human gut-derived bifidobacteria.** *Applied and environmental microbiology* 2017, **83:**e02894-02816.
49. Reyes A, Semenkovich NP, Whiteson K, Rohwer F, Gordon JI: **Going viral: next-generation sequencing applied to phage populations in the human gut.** *Nature Reviews Microbiology 607* 2012, **10**.
50. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, Taylor TD, A. DN: **Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes.** 2007, **14:**169-181.
51. Caro-Quintero A, Konstantinidis KT: **Inter-phylum HGT has shaped the metabolism of many mesophilic and anaerobic bacteria.** *The ISME journal* 2015, **9:**958.
52. Boto L, London B: **Horizontal gene transfer in evolution: facts and challenges.** *Proceedings of the Royal Society of Biological Sciences* 2010, **277:**819-827.
53. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions.** *Nature Reviews Genetics* 2012, **13:**36-46.
54. Rhoads A, Au KF: **PacBio sequencing and its applications.** *Genomics, proteomics & bioinformatics* 2015, **13:**278-289.

Supplementary Materials

Table S1 The selected 20 genomes

Class	Strain	abbreviation	*BioProject ID*
Alphaproteobacteria	*Acidiphilium multivorum* AIU301	AAM	60101
	Ketogulonigenium vulgarum WSH 001	AKV	161161
	Mesorhizobium australicum WSM2073	AMAU	47287
	Methylocapsa acidiphila B2	AMAC	72841
	Methyloferula stellata AR4	AMS	165575
	Rhodovibrio salinarum DSM 9154	ARS	84315
	Roseobacter litoralis Och 149	ARL	19357
	Sphingobium japonicum UT26S 1	ASJ	19949
	Starkeya novella DSM 506	ASN	37659
	Tistrella mobilis KA081020 065	ATM	76349
Betaproteobacteria	*Alicycliphilus denitrificans* K601	BAD	50751
	Dechlorosoma suillum PS	BDS	37693
	Gallionella capsiferriformans ES 2	BGC	32827
	Herbaspirillum seropedicae SmR1	BHS	47945
	Nitrosospira multiformis ATCC 25196	BNM	13912
	Ramlibacter tataouinensis TTB310	BRT	16294
	Sideroxydans lithotrophicus ES 1	BSL	33161
	Snodgrassella alvi wkB2	BSA	167602
	Sulfuricella denitrificans skB26	BSD	170011
	Tetrathiobacter kashmirensis WT001	BTK	67337

Table S2 Distribution of transferred genes

Recipient	Genes transferred to the recipient genome
BAD	AAM_03063, AKV_01007, AMAC_01196, AMAU_02632, AMS_01785, ARL_01226, ARS_03086, ASJ_00197, ASN_04073, ATM_00526
BDS	AAM_00175, AKV_00943, AMAC_00215, AMAU_02085, AMS_01465, ARL_00634, ARS_02444, ASJ_02645, ASN_02939, ATM_01561
BGC	AAM_00176, AKV_01272, AMAC_01576, AMAU_00617, AMS_02653, ARL_00442, ARS_00175, ASJ_01651, ASN_01773, ATM_03442
BHS	AAM_00195, AKV_01273, AMAC_01674, AMAU_05963, AMS_03303, ARL_01008, ARS_03177, ASJ_01197, ASN_02864, ATM_02182
BNM	AAM_00209, AKV_00282, AMAC_02914, AMAU_02414, AMS_03378, ARL_00629, ARS_00690, ASJ_00320, ASN_04445, ATM_03557
BRT	AAM_00308, AKV_02353, AMAC_03303, AMAU_00830, AMS_01655, ARL_01194, ARS_00958, ASJ_00902, ASN_03136, ATM_02773
BSA	AAM_00473, AKV_01812, AMAC_00998, AMAU_01212, AMS_00937, ARL_02642, ARS_01255, ASJ_00103, ASN_03987, ATM_00724
BSD	AAM_00571, AKV_00359, AMAC_00685, AMAU_02488, AMS_02811, ARL_03413, ARS_02706, ASJ_00412, ASN_03327, ATM_03491
BSL	AAM_00702, AKV_00288, AMAC_00418, AMAU_01255, AMS_00102, ARL_03240, ARS_00911, ASJ_01392, ASN_00060, ATM_00976
BTK	AAM_00862, AKV_02380, AMAC_02518, AMAU_04187, AMS_01716, ARL_03996, ARS_02511, ASJ_03214, ASN_02198, ATM_02037

Table S3 Relative abundance of the 20 genomes for metagenomics data simulation

Genome	Relative abundance		
	Replicate 1	Replicate 2	Replicate 3
AAM	226	640	134
AKV	669	315	16
AMAC	731	8	261
AMAU	507	292	201
AMS	555	332	113
ARL	184	554	262
ARS	380	290	330
ASJ	399	221	380
ASN	246	324	430
ATM	338	465	197
BAD	468	336	196
BDS	651	69	280
BGC	456	173	371
BHS	652	339	9
BNM	938	44	18
BRT	97	363	540
BSA	478	400	122
BSD	332	116	552
BSL	497	375	128
BTK	345	549	106

Table S4 Contamination and completeness of selected 20 genomes according to CheckM

Genome	Completeness (%)		Contamination (%)	
	Lineage specific	Universal	Lineage specific	Universal
AAM	100	100	0	0
AKV	95.53	97.67	0.3	0
AMAC	99.58	100	0.31	0
AMAU	100	100	0.39	0
AMS	99.48	100	0	0
ARL	99.36	97.67	0.32	0
ARS	99.57	100	1.74	0
ASJ	99.66	100	0	0
ASN	100	100	0.47	0
ATM	89.55	100	0.25	0
BAD	99.85	100	0.46	0
BDS	100	100	0.12	0
BGC	99.53	100	1.02	0
BHS	99.87	100	0.5	0
BNM	99.52	100	0.71	0
BRT	100	100	0.43	0
BSA	99.57	100	0.85	0
BSD	98.82	100	0.47	0
BSL	99.53	100	0.47	0
BTK	94.31	97.67	1	0
Average	96.69	99.65	0.49	0