Microbiome

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

--Manuscript Draft--

Manuscript Number:	MBIO-D-18-00332R1
Full Title:	MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic tree approaches
Article Type:	Methodology
Funding Information:	Australian Research Council — Professor Torsten Thomas China Scholarship Council (201508200019) — Mr. Weizhi Song China Scholarship Council (201708200017) — Ms Shan Zhang
Abstract:	Background: Metagenomic datasets provide an opportunity to study horizontal gene transfer (HGT) on the level of a 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 with various degrees of genetic divergence from metagenomic datasets. The results also indicated that the detection of very recent gene transfers (ie. those with low levels of genetic divergence) from metagenomics datasets is largely affected by the read assembly step. Comparison of MetaCHIP with a previous analysis on soil bacteria showed a high level of consistency for the prediction of recent HGTs and revealed a large number of additional non-recent gene transfers, which can provide new biological and ecological insight. Assessment of MetaCHIP's performance on real metagenomic datasets confirmed the role of HGT in the spread of genes related to antibiotic resistance in the human gut microbiome. Further testing also showed that functions related to energy production and conversion as well as carbohydrate transport and metabolism are frequently transferred among free-living microorganisms. 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. MetaCHIP is implemented in Python and freely available at: https://github.com/songweizhi/MetaCHIP.
Corresponding Author:	Torsten Thomas University of New South Wales AUSTRALIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of New South Wales
Corresponding Author's Secondary Institution:	
First Author:	Weizhi Song
First Author Secondary Information:	
Order of Authors:	Weizhi Song Bernd Wemheuer Shan Zhang Kerrin Steensen Torsten Thomas
Order of Authors Secondary Information:	

Response to Reviewers:	MicrobiomeSydney, 4th of February 2019 Dear Madame, Dear Sir. Thanks for the assessment of our manuscript MBIO-D-18-00332 entitled “MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic tree approaches". We have carefully considered the reviewers’ comments and have revised our manuscript accordingly. Please find below a response and attached a revised version of our manuscript. We thank you for your consideration of our manuscript and please do no hesitate to contact me with any further questions or comments. Sincerely, Prof. Torsten Thomas Director, Centre for Marine Bio-Innovation Future Fellow, School of Biotechnology and Biomolecular Sciences The University of New South Wales, Sydney, NSW 2052, Australia Ph + 61 (0) 2 938 53467, Mob +61 (0) 410159525, Fax + 61 (0) 2 938 51779 Email: t.thomas@unsw.edu.au Reviewer reports: Reviewer #1: [Summary] The authors describe MetaCHIP, a method for detecting putative horizontal gene transfer events from metagenomic datasets. The software combines two approaches; a blast "top-hit" based approach followed by confirmation with Ranger-DTL. [Major comments] 1) The authors select 10 genomes from two diverse classes of bacteria: Alphaproteobacteria and Betaprotobacteria. While a good start, the authors should start at the species or genus level then work their way backward to class level differences. This is relevant as the majority of horizontal gene transfers likely occur between closely related bacteria; this test would capture MetaCHIP's performance on transfer between more distantly related genomes (easier) to more closely related genomes (harder). For recent work in this area please see: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004408 Response: Thanks for this suggestion, MetaCHIP’s performance has now been assessed on simulated transfers on the genus and class levels and the results have been added to the manuscript. They generally show a good performance for the detection of HGTs that had level of genetic divergence less than 15%. a. The set of 10 randomly selected genes should be bootstrapped as some genes will be more difficult to detect in foreign genomes than others. Response: The analysis has been redone with 10 bootstraps for the 100 selected genes transferred between genomes from different classes and genera. 2) The authors assume a single gene is the unit of genetic transfer. There are previous studies that provide evidence that genes nor domains are the unit of genetic transfer. For reference, see https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0004524. This is an important point, even it subtle on the surface. Partial matches make methods like blast "best hit" approaches and phylogenetic approaches more challenging due to the effect on partial alignments with unknown homology boundaries on both approaches. a. E.g. this will have implications on the approach the authors use for determining "reconstruction" of gene transfer events from assembly: "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]. "

Response:
We believe this comment is due to a misunderstanding of when we apply the criterion of looking at the flanking regions. This criterion was only applied to evaluate if a simulated HGT was recovered in the genome assembly (see Methods) to allow us to fairly assess the performance of MetaCHIP (i.e. provide a prediction rate for the HGT that were actually assembled). This criterion is not part of the normal pipeline of MetaCHIP.

We do appreciate that a single gene is not necessarily the only unit of genetic transfer and therefore provide the user with a graphic outputs (as shown in Figure 2) for the genomic regions subject to a putative HGT. This allows the user to evaluate if the transfer goes beyond the boundaries of a gene.

3) What is Table 1 measuring? It is not fully clear. The values for the Precision and Recall should be indicated and clarified.

Response:
Table 1 shows the quality of the refined genome bins we reconstructed from simulated metagenomic datasets. The definition of precision and recall has now been added to the manuscript.

4) The vast majority of presented results focus on sensitivity, or the # of recovered transfers. Table 1 is the only table that includes False Positives, and it is unclear what is exactly being measured. If I understand correctly, it is measuring the precision of the genome binner, not MetaCHIP's HGT detection. This needs to be corrected:

a. MetaCHIP should be ran on non-metagenomic data (full genomes and sequenced reads from full genomes) with previously documented HGTs. There exist a multitude of prior datasets with predicted HGTs. For reference, see:
http://darkhorse.ucsd.edu/search.shtml.

Response:
MetaCHIP’s performance has now been assessed on 2094 complete bacterial genomes and compared to their previously described HGTs (Smillie et al. (2011) Nature, 480:241). This showed generally a high consistency between MetaCHIP and the previous study.

b. The authors should compare MetaCHIP to previously reported HGTs in metagenomic samples. For reference, see:
https://www.nature.com/articles/nature10571.

Response:
Please see the previous comment. We used the data suggested by the reviewer here to benchmark MetaCHIP.

c. Table 2: Most of the putative HGTs that MetaCHIP identifies are at >20% variation, and the authors state there are even likely more. So these are ancient HGT events? Or are is it that you are confident that these are transferred genes, but the source is likely unknown? Please clarify, especially with respect to alternative technical and biological explanations.

Response:
It is possible that the lack of recent HGTs identified from the human gut dataset is due to the loss of real sources during the assembly and binning processes. We acknowledged this limitation in our conclusion.

[Minor comments]
1) The authors propose clustering via: "The produced SCG protein tree is then converted into a distance matrix and clustered using the Nearest Point Algorithm implement in SciPy [27]". However, the sequences are already clustered if they are in a

phylogenetic tree (cut the tree horizontally to get any clustering of k clusters). Please clarify.

Response:
The grouping of input genomes has now been changed to involve the recently released, phylogenetically calibrated GTDB (Genome Taxonomy Database) (Parks et al. (2018) Nature Biotechnology). This has the advantage that the user can now simply specify at which taxonomic level (e.g. class, order or family) HGTs should be analysed, rather than manually defining phylogenetic groups. The concern raised above therefore does not apply anymore.

2) All figures are low resolution, though some seem lower res than others, and the text in Figure 9 is illegible.

Response:
Figures in the manuscript have been updated to high resolution.

[Software comments]

1) I had some difficulty both tracking down and installing all of the dependencies on my workstation; a utility installation script, container, or bioconda integration would have greatly simplified the process.

Response:
MetaCHIP can now be installed with pip. A detailed manual is also now made available.

2) The code is a collection of disjoint scripts; I'd encourage the authors to consider encapsulating as a workflow. For reference see: SnakeMake.

Response:
The scripts have been encapsulated with python's 'subparsers' function. The best-match and phylogenetics approaches as well as the script for input file preparation have now been provided as submodules of MetaCHIP.

3) Hard-coded, non-default (BLAST) parameter choices should be described/highlighted in the README.

Response:
Customized BLASTN output format has been added to the README of MetaCHIP's github repository.

Reviewer #2: The manuscript "MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and explicit phylogenetic tree approaches" by Song et al. describes the identification of horizontal gene transfer in genome and metagenome sequence data based on gene nucleotide BLAST searches and phylogenetic tree predictions of the corresponding peptide sequences.

The performance of the analysis method was evaluated using full genome sequences and simulated shotgun genome and metagenome sequence data of randomly selected alpha- and betaproteobacterial genomes combined with randomly inserted genes from the same genome set. As real datasets, the method was tested on two published datasets from human gut and marine seawater metagenomics projects.

Validation test results were further analyzed with respect to different sequence similarities between predicted HGT genes and their genomic source, different sequencing depth of the simulated genome shotgun sequence data, and binning and assembly efficiencies. For the real datasets, the authors evaluated the efficiency of assembly, contamination and completeness of genome bins, genome bin cluster types and HGT prediction. The authors interpret large fractions of predicted HGT genes with functions in antibiotic resistance from the human gut and with energy production and amino acid transport and metabolism from the marine datasets - which is in agreement

	with previous publications - as indications for successful prediction of HGT events. While the authors carefully validate their analysis method, the specific advantage of the new method remains unclear to me. Although the authors combine to separate methods (similarity search and phylogenetic predictions) the results from these analyses are not combined into a single, potentially more accurate prediction. Response: We believe that this comment is based on a misunderstanding. The output of MetaCHIP is indeed a combination of the best-match and the phylogenetic analysis and we show that this has overall high accuracy. In addition, the method depends to a large degree on the applied methods for metagenome binning and assembly, both of which are also affected by HGT. Response: We do acknowledge this limitation in our conclusion. Finally, there is no indication of any significant new findings that would result from the new method over previous analyses, as the authors merely use it to confirm previously published research findings. Response: We think it is important to “validate” a new method by comparing its output to previous findings. To show how MetaCHIP goes beyond previous works, we have also now analysed over 2000 bacterial genomes and found 2414 non-recent gene transfers that a previous analysis (Smillie et al. 2011 Nature, 480:241) could not identify.
Additional Information:	
Question	**Response**
Is this study a clinical trial?<hr>A clinical trial is defined by the Word Health Organisation as 'any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes'.	No

MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic 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 a 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 with various degrees of genetic divergence from metagenomic datasets. The results also indicated that the detection of very recent gene transfers (ie. those with low levels of genetic divergence) from metagenomics datasets is largely affected by the read assembly step. Comparison of MetaCHIP with a previous analysis on soil bacteria showed a high level of consistency for the prediction of recent HGTs and revealed a large number of additional non-recent gene transfers, which can provide new biological and ecological insight. Assessment of MetaCHIP's performance on real metagenomic datasets confirmed the role of HGT in the spread of genes related to antibiotic resistance in the human gut microbiome. Further testing also showed that functions related to energy production and conversion as well as carbohydrate transport and metabolism are frequently transferred among free-living microorganisms.

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. MetaCHIP is implemented in Python and freely available at: https://github.com/songweizhi/MetaCHIP.

Keywords: Metagenomics, horizontal gene transfer, HGT identification, 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]. The 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 horizontal gene transfer (HGT) within communities of uncultured microorganisms.

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) for HGT prediction. Explicit phylogenetic approaches are employed by Ranger-DTL [15] and AnGST [16], which predict HGTs through the reconciliation of gene trees with corresponding species trees.

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 within a microbial community, while DarkHorse [13] requires suitable references genomes to predict HGTs, which are often not available for uncultured microorganisms.

We therefore 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 for HGT detection (see above). Its inputs are the sequence files of a set of genomes or genome bins derived from metagenomic data as well as their taxonomic classifications. The recently developed GTDB-Tk tool [17], which is based on the phylogenetically-calibrated Genome Taxonomy Database (GTDB) [18], is recommended for taxonomic classification of input genomes. Input genomes are initially grouped by MetaCHIP according to their taxonomic classifications at user-specified rank (e.g. class, order, family or genus).

Best-match approach

Open reading frames (ORFs) are predicted from input genomes with Prodigal v2.6.3 [19] and an all-against-all BLASTN [20] search is performed among all predicted ORFs. The BLASTN results are first filtered with user-defined alignment length (e.g. 200 bp) and coverage cut-offs (e.g. 75%). 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 [21], Velvet [22], SPAdes [23], IDBA [24]) will produce "bubbles" for sequence regions with sequencing error, but high similarity [25]. The resolution of such bubbles may produce two contigs with overlapping sequences at the end of the contigs. This duplication could be falsely considered in the HGT analysis and to avoid this putative HGT candidates located at contigs end with high-similarity (>95%) are disregarded. In addition, putative HGT candidates located on contigs, which had

95% of their full-length matching with a longer contig, were disregarded, as these contigs are likely artificial duplicates of the assembly process.

To further corroborate the predicted HGT candidates, their flanking sequences within user-defined length (e.g. 10 Kbp) are extracted from the annotation files. A pairwise BLASTN is performed between each pair of flanking regions. Plots for the genomic regions are generated with GenomeDiagram [26] and provided for visual inspection (**Figure 2**).

Phylogenetic approach

A phylogenetic approach is used to further corroborate the results given by the best-match approach and to provide information on the direction of gene flow. For each pair of genes, which were identified as putative HGT by the best-match approach, a protein tree is generated using the genes used for the HGT analysis in the best-match approach and all orthologs from the two groups, from which the paired genes come from. Amino acid sequences are aligned with MAFFT v7.310 [27] and followed by the removal of columns represented by <50% of proteins and/or with an amino acid consensus of <50%. A protein tree is then constructed using FastTree v2.1.10 [28] with default parameters.

A "species" tree is then generated to compare to the gene tree. As the 16S rRNA gene, which is the most commonly used phylogenetic and taxonomic marker of bacterial and archaeal organisms, is often missing in genome bins [29-31], we build a phylogenetic tree for all input genomes using the protein sequences of 43 universal single-copy genes (SCGs) used by CheckM [32]. Predicted protein sequences for the input genomes are searched for the PFAM v31.0 [33] and TIGRFAM v14.0 [34] hmm profiles of these SCG proteins using HMMER v3.1b2 [35]. Protein sequences for each hmm profile are then individually aligned using

HMMER and concatenated into a multiple sequence alignment (MSA). Columns represented by <50% of genomes and/or with an amino acid consensus <25% are removed and a phylogenetic tree is built using FastTree [28]. A subtree, which includes only the genomes relevant to the particular genes analysed is extracted with preserved branch length using ETE v3.1.1 [36]. The reconciliation between each pair of protein tree and "species" subtree is performed using Ranger-DTL v2.0 with dated mode. Briefly, Ranger-DTL predicts HGTs by performing a duplication-transfer-loss (DTL) reconciliation between a protein family phylogeny and its corresponding organismal phylogeny [15].

To assess how reliable SCG protein trees are to reconstruct organismal phylogenies from partial genome bins, we selected 20 alpha- and beta-proteobacterial genomes (see below) and divided each of them into 100 contigs with equal length. Next, 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 Mantel tests [37].

Assessment of MetaCHIP on simulated datasets

MetaCHIP's performance was first assessed on simulated datasets at different taxonomic levels. To assess its performance at a low taxonomic level, ten genomes from species of the genus *Sphingobium* (donor group) and *Sphingomonas* (recipient group) within the family *Sphingomonadaceae* were selected (see **Table S1**), while for class level transfers, 10 alphaproteobacterial (donor group) and betaproteobacterial (recipient group) genomes were chosen (see **Table S2**). Ten genes (with at least two orthologs in the recipient group) from each of the 10 donor genomes were selected and randomly transferred into the 10 recipient genomes

with different levels of genetic divergence (0, 5, 10, 15, 20, 25 and 30%) using HgtSIM [38]. The six-frame stop codon sequence "TAGATGAGTGATTAGTTAGTTA" was added to the two ends of transferred genes to facilitate correct gene prediction. This process was bootstrapped 10 times and donor and mutated recipient genomes from each bootstrap were used directly as inputs into MetaCHIP to assess its performance at class and genus levels.

Sequencing reads were also simulated from the 10 alphaproteobacterial and 10 mutated betaproteobacterial genomes for each level of genetic divergence from one of the 10 bootstraps. Sequencing reads for each level of genetic divergence were simulated three times with different abundance profiles (**Table S3**) using GemSIM [39].

As the reconstruction of genes involved in HGT are highly affected by sequencing depth or the assembler used [38], 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 divergence. The paired-end reads were quality filtered using Trimmomatic v0.36 [40] with a quality cut-off of 20 and a sliding window of 6 bp. Reads from the three replicates were combined and then assembled with IDBA_UD v1.1.1 [24] or metaSPAdes v3.9.0 [23] 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 [38]. 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 divergence. The BLASTN results were then filtered with an identity cut-off of >98% and a coverage cut-off of >98% for the transferred genes.

Metagenome binning was performed with MetaBAT v0.32.5 [1] and MyCC v2017 [2] and the results were refined with Binning_refiner v1.2 [41]. Bin completeness and contamination were assessed with CheckM v0.9.7 [32]. The correlations between the genome bins and the reference genomes were obtained by running pairwise BLASTN searches. The correlations between MetaCHIP predicted HGTs and the known simulated gene transfers were determined by running pairwise BLASTN searches with identity and coverage cut-off of >98%.

Assessment of MetaCHIP on a dataset with previously described HGTs

MetaCHIP's performance was also assessed on 2094 full bacterial genomes, which were previously analysed for HGTs using blocks of nearly identical DNA (> 99% identity, over 500bp) in distantly related genomes (16S rRNA gene similarity less than 97%) [42]. The 2094 bacterial genomes were downloaded from the NCBI RefSeq database and their taxonomy was determined using GTDB-Tk v0.1.6 [17]. HGT events were then analysed with MetaCHIP at genus level. BLASTN search with a 100% identity and coverage cut-off were used to compare MetaCHIP predicted HGTs with previously identified, transferred DNA blocks. COG annotation of predicted HGTs was performed by running RPS-BLAST [20] against the COG database [43].

Assessment of MetaCHIP on real metagenomic dataset

Genome bins derived from metagenomic datasets for microbiomes from human guts [1, 44] and seawater samples taken in the North Sea [45] were used to assess the performance of MetaCHIP on real metagenomic datasets. For the human gut dataset, genome bins previously produced by MetaBAT [1] were used directly here after removing contigs shorter than 2000 bp. For the North Sea dataset, all sequencing reads were quality filtered with Trimmomatic as previously described [45] and assembled using metaSPAdes v3.9.1 [46]. Binning was

performed as described above. CheckM v0.9.7 [32] was subsequently used to assess the quality of genome bins. The SCG protein tree of these bins as well as COG annotation of predicted HGTs were performed as described above and antibiotic resistance related COGs were retrieved from the antibiotic resistance genes database (ARDB; April 2018) [47].

Results and discussion

Performance on simulated datasets

MetaCHIP requires a SCG protein tree of all input genomes for the phylogenetic approach. We therefore first assessed how reliable the reconstruction of a SCG-based phylogeny is for incomplete genome bins. 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.

MetaCHIP's performance was first assessed by introducing defined HGTs in genomes at genus and class levels. MetaCHIP had a high recovery rate of artificially introduced HGTs up to levels of genetic divergence of 10% at both taxonomic levels (**Figure 4**). A steady decline in recovery was observed with higher levels of genetic divergence. Detection of between-class HGTs became unsuccessful at 30% divergence, while at genus level, the detection threshold was reached at around 20% divergence. This performance is consistent with previous findings and algorithms that showed the difficulties of detecting HGT between closely related taxa (e.g. genera of the same family) [48, 49]. Nevertheless, no less than 40% of between-genera HGTs with genetic divergence less than 15% can be detected by MetaCHIP. The phylogenetic analysis predicted the correct directions of gene flow in more than 81% of cases for transfers

between classes at all divergence levels and in more than 86% of cases between genera with genetic divergence less than 15% (**Figure 4**).

We next evaluated how different assemblers and sequencing depths influence the recovery of class-level HGTs with different levels of genetic divergence. We also introduced realistic sequencing errors into the read dataset before assembly. When no mutation was introduced to the transferred genes, more transferred genes were recovered by metaSPAdes than with IDBA_UD. For 5% genetic divergence, both assemblers 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 divergence levels between 10-30% **(Figure 5)**. MetaSPAdes was therefore used for the assembly of metagenomic reads with no genetic divergence, while IDBA_UD was selected for datasets with the other levels of genetic divergence. For gene transfers with no genetic divergence, the recovery rate for metaSPAdes assemblies was highest with a sequencing depth of 11.33x, beyond which it declined. For the 5% genetic divergence, 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.

Based on these choices of coverage and assembler we next binned genomes from the simulated datasets. The precision (defined as how pure a bin is) and recall (defined as how complete a bin is) of the genome bins for all divergence groups were calculated with evaluate.py from the MyCC package [2]. The results showed that their overall precision and recall were not lower than 99.73% and 89.49%, respectively (**Table 1**).

Table 1. The quality of refined genome bins reconstructed from simulated metagenomic datasets at different level of genetic divergence of introduced HGTs

Genetic divergence (%)	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 divergence, 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 divergence greater than 5%, no less than 73.7% of transferred gene copies were found in both the donor and recipient genome bins (**Figure 6**).

By applying MetaCHIP to the genome bins, 26% of the 100 introduced gene transfers were recovered by the best-match approach for the 0% genetic divergence and 9 of them were validated by the phylogenetic approach (**Figure 7**), which accounts for 86.7% and 30%, respectively, of the gene transfers that actually exist in the genome bins. For a 5% genetic divergence, 93.8% of introduced gene transfers that were found in the bins were also identified by the best-match approach and 81.3% of them were validated by the phylogenetic approach. The best recovery rates were obtained when the genetic divergence is 10%, where at least 74% of introduced gene transfers were recovered by the best-match approach and 69% of them were validated by the phylogenetic approach, which accounted for 91.4% and 85.2% of all binned gene transfers, respectively. A steady decline in the ability of MetaCHIP to detect HGT was also observed with higher genetic divergence (**Figure 7**).

Performance on dataset with previously described HGTs

We next benchmarked MetaCHIP's performance against a previous large-scale study that analysed HGT in 2,235 bacterial genomes [42]. Only 2,094 of the genome from [42] were available for download from the NCBI RefSeq database (**Additional file 2**) and they were subsequently grouped by MetaCHIP into 664 genera. MetaCHIP identified 10,255 HGTs (**Additional file 3 and 4**), of which 2,694 (26.3%) showed a genetic divergence less than 1% i.e. represent recent transfers (**Figure 8**). We compared MetaCHIP's prediction with previously described HGTs only for the soil isolates, as metadata for the other isolate types were missing or incomplete (see Supplementary Table 5 from [42]). 433 of the 2,694 recent gene transfers fulfil the criteria that they were between genomes with 16S rRNA gene similarity less than 97% (a filter implemented in [42]). The previous analysis identified 368 HGTs with no more than 1% genetic divergence involving soil isolates and 248 (77.2%) of them overlap with MetaCHIP's predictions, which showed relatively high consistency between the two approaches.

Beside the 433 recent HGTs, MetaCHIP also identified 2414 gene transfers with genetic divergence higher than 1% involving the genome of soil isolates (**Figure 8**). Recent (i.e. genetic divergence ≤ 1%) and non-recent (i.e. genetic divergence > 1%) HGTs predicted by MetaCHIP and all genes for the 368 soil genomes were then annotated using the COG system. COG categories were considered to be enriched in the HGT dataset if their proportion were above the 75% percentile of the relative abundance across all input genomes. The results revealed that COG categories enriched for recent HGTs are different to those enriched in the non-recent HGTs. For example, COG category C (energy production and conversion), E (amino acid transport and metabolism), I (lipid transport and metabolism) and L (replication, recombination and repair) were only enriched in the non-recent HGTs, while categories K (transcription), P (inorganic ion transport and metabolism) and U (intracellular trafficking,

secretion, and vesicular transport) were enriched in recent HGTs (**Figure 9**). This observation was missed by the previous analysis [42] and shows that MetaCHIP can provide new biological and ecological insights into the HGT of microbial communities.

Performance on real metagenomic datasets

Finally, we assessed MetaCHIP's performance on two real metagenomic datasets: one for free-living seawater microorganisms in the North Sea [50] and the other for the human gut microbiome [44]. For the metagenomic dataset of seawater microorganisms, sequence assembly with metaSPAdes generated 315.33 Mbp of contiguous sequences ≥ 2,500 bp (35,190 contigs) and 69 genome bins were obtained, of which 37 had no contamination as detected with CheckM and completenesses higher than 40%. For the 1,634 genome bins obtained from the human gut dataset described in [44], 138 were estimated to be contamination-free and more than 40% complete. The taxonomy of qualified genome bins was determined with GTDB-Tk (**Additional file 5**). The human gut and the seawater bins were taxonomically grouped into 29 and 16 orders, respectively (**Figure S1**). The best-match approach detected 560 gene transfers for the human gut genome bins and of which 113 were also found by the phylogenetic approach. For the seawater dataset, 121 and 32 gene transfers were detected by the two approaches, respectively. The direction of predicted gene flows within the two communities were shown in **Figure 10.** Not surprisingly, the number of HGT detected in any given group is proportional to the number of genome bins it contained (**Figure S1** and **Figure 10**). One exception however is the order *Bacteroidales* from the human gut dataset, where only four HGTs were detected in its 30 genome bins. High rates of HGT within the order *Bacteroidales* have been previously described [51], but our results indicate that this does not apply to HGTs of this order with other taxonomic groups. A genome bin of the *Chitinophagales* from the North Sea dataset was also found to have all its 10 HGTs with genomes from the order *Flavobacteriales* (**Figure 10**),

which is consistent with previous observation that HGT are more likely to occurr between those two closely related taxa [52].

We next performed a functional annotation of the genes identified in the HGT analysis based on the COG system as described above. 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), translation, ribosomal structure and biogenesis (J) as well as nucleotide transport and metabolism (F) (**Figure 11**). The enrichment of defence mechanisms (V) was mainly due to 15 HGTs involving genes with functions related to ABC-type multidrug (COG1131) and antimicrobial peptide (COG1136) transport systems. This observation is consistent with previous observations and proposals that HGT is a dominant factor for the spread of AR in the human gut microbiota [53-55]. 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 [56]. COG categories preferentially subject to HGT between the free-living microorganisms in the North Sea include energy production and conversion (C), carbohydrate transport and metabolism (G) as well as translation, ribosomal structure and biogenesis (J) (**Figure 11**). This observation is similar to a recent study on HGTs among all available complete genomes for free-living Archaea and Bacteria, where transferred genes most frequently also belonged to COG categories C and G [57].

More than 40% of identified HGTs for the human gut and the seawater bins had genetic divergence 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 divergence in the simulated datasets (**Figure 7**) and hence we predict that the actual numbers of HGTs that

occurred in the community are likely to be underestimated here. Interestingly, only one transfer with genetic divergence at less than 15% divergence was detected, for which we generally found high recovery rate in our simulations (**Figure 7**). This may indicate that HGT in these microbial communities does not involve a large number of recent transfers or that the actual donors were not recovered in the genome bins due to assembler limitations [38] or the removal of low-quality genome bins.

Table 2. Genetic divergence of HGT identified by MetaCHIP from the human gut and North Sea datasets

Dataset	Approach	Genetic divergence (± 2.5%)						
		0	5	10	15	20	25	30
Human gut	Best-match	0	0	0	1	79	406	74
	Phylogenetic	0	0	0	0	18	78	17
North Sea	Best-match	0	0	0	0	8	78	35
	Phylogenetic	0	0	0	0	4	14	14

Conclusion

Our development and tests of MetaCHIP showed that the tool can detect HGTs with various degree of genetic divergence from microbial community data, but that prediction efficiency is affected by several factors. First, as transferred genes will undergo mutations in their new genomic contexts, their detections will become difficult when the similarities between the donor and recipient genes fall below certain levels (**Figure 3**) [58]. 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 [38, 59], 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 [60], when applied to metagenomic samples. Third, the successful detection of HGT from metagenomic dataset requires the reliable reconstruction of the organismal genomes, 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 HGTs can be detected from microbial community data 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.

List of abbreviations

AR: antibiotic resistance

ARDB: antibiotic resistance genes database

COG: Clusters of Orthologous Groups

GTDB: Genome Taxonomy Database

HGT: horizontal gene transfer

NCBI: National Center for Biotechnology Information

SCG: single-copy gene

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

https://github.com/songweizhi/MetaCHIP

Competing interests

The authors declare that they have no competing interests.

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 and TT wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

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* 2016, **4:**8.

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. 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.
6. Probst AJ, Ladd B, Jarett JK, Geller-McGrath DE, Sieber CM, Emerson JB, Anantharaman K, Thomas BC, Malmstrom RR, Stieglmeier M: **Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface.** *Nature microbiology* 2018, **3:**328.
7. Moitinho-Silva L, Díez-Vives C, Batani G, Esteves AI, Jahn MT, Thomas T: **Integrated metabolism in sponge–microbe symbiosis revealed by genome-centered metatranscriptomics.** *The ISME journal* 2017, **11:**1651.
8. Ji M, Greening C, Vanwonterghem I, Carere CR, Bay SK, Steen JA, Montgomery K, Lines T, Beardall J, van Dorst J: **Atmospheric trace gases support primary production in Antarctic desert surface soil.** *Nature* 2017, **552:**400.
9. Dagan T, Artzy-Randrup Y, Martin W: **Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution.** *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. **Gtdb-Tk v0.1.6** [https://github.com/Ecogenomics/GtdbTk]
18. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P: **A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.** *Nature Biotechnology* 2018, **36**.
19. 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.
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of molecular biology* 1990, **215:**403-410.
21. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y: **SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.** *Gigascience* 2012, **1:**18.

22. Zerbino D, Birney EJGr: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Research* 2008, **18:**gr. 074492.074107.
23. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA: **metaSPAdes: a new versatile metagenomic assembler.** *Genome research* 2017**:**gr. 213959.213116.
24. 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.
25. 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.
26. 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.
27. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Molecular Biology & Evolution* 2013, **30:**772-780.
28. Price MN, Dehal PS, Arkin AP: **FastTree: computing large minimum evolution trees with profiles instead of a distance matrix.** *Molecular biology and evolution* 2009, **26:**1641-1650.
29. 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.
30. 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* 2017, **2**.
31. Yuan C, Lei J, Cole JR, Sun Y: **Reconstructing 16S rRNA genes in metagenomic data.** *Solid-state Circuits Conference* 2015, **51:**1-3.
32. 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.
33. 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.
34. Haft DH, Selengut JD, White O: **The TIGRFAMs database of protein families.** *Nucleic acids research* 2003, **31:**371-373.
35. Eddy SR: **Accelerated profile HMM searches.** *PLoS computational biology* 2011, **7:**e1002195.
36. Huerta-Cepas J, Dopazo J, Gabaldón T: **ETE: a python Environment for Tree Exploration.** *BMC bioinformatics* 2010, **11:**24.
37. Mantel N: **The detection of disease clustering and a generalized regression approach.** *Cancer research* 1967, **27:**209-220.
38. Song W, Steensen K, Thomas T: **HgtSIM: a simulator for horizontal gene transfer (HGT) in microbial communities.** *PeerJ* 2017, **5:**e4015.
39. McElroy KE, Luciani F, Thomas T: **GemSIM: general, error-model based simulator of next-generation sequencing data.** *Bmc Genomics* 2012, **13:**1-9.
40. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics* 2014, **30:**2114-2120.
41. Song W, Thomas T: **Binning_refiner: improving genome bins through the combination of different binning programs.** *Bioinformatics* 2017, **33:**1873-1875.

42. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ: **Ecology drives a global network of gene exchange connecting the human microbiome.** *Nature* 2011, **480:**241.
43. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278:**631-637.
44. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T: **A human gut microbial gene catalogue established by metagenomic sequencing.** *nature* 2010, **464:**59.
45. Wemheuer B, Wemheuer F, Hollensteiner J, Meyer F-D, 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* 2015, **6:**805.
46. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA: **metaSPAdes: a new versatile metagenomic assembler.** *Genome Research* 2017, **27**.
47. Liu B, Pop M: **ARDB—antibiotic resistance genes database.** *Nucleic acids research* 2008, **37:**D443-D447.
48. Adato O, Ninyo N, Gophna U, Snir S: **Detecting horizontal gene transfer between closely related taxa.** *PLoS computational biology* 2015, **11:**e1004408.
49. Langille MG, Brinkman FS: **Bioinformatic detection of horizontally transferred DNA in bacterial genomes.** *F1000 biology reports* 2009, **1**.
50. 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**.
51. Coyne MJ, Zitomersky NL, McGuire AM, Earl AM, Comstock LE: **Evidence of extensive DNA transfer between bacteroidales species within the human gut.** *MBio* 2014, **5:**e01305-01314.
52. Bolotin E, Hershberg R: **Horizontally Acquired Genes Are Often Shared between Closely Related Bacterial Species.** *Frontiers in microbiology* 2017, **8:**1536.
53. 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.
54. 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.
55. 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* 2012, **10:**607.
56. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, Takami H, Morita H, Sharma VK, Srivastava TP: **Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes.** *DNA research* 2007, **14:**169-181.
57. Caro-Quintero A, Konstantinidis KT: **Inter-phylum HGT has shaped the metabolism of many mesophilic and anaerobic bacteria.** *The ISME journal* 2015, **9:**958.
58. Boto L, London B: **Horizontal gene transfer in evolution: facts and challenges.** *Proceedings of the Royal Society of Biological Sciences* 2010, **277:**819-827.
59. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions.** *Nature Reviews Genetics* 2012, **13:**36-46.

60. Rhoads A, Au KF: **PacBio sequencing and its applications.** *Genomics, proteomics & bioinformatics* 2015, **13:**278-289.

Additional files

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

Additional file 2: Metadata of downloaded 2094 genomes. (162 kb)

Additional file 3 MetaCHIP identified 10255 HGTs from the 2094 genomes. (825 kb)

Additional file 4 Nucleic acid sequences MetaCHIP identified 10255 HGTs (FASTA 17.2 mb)

Additional file 5 Taxonomic classification of the human gut and North Sea genome bins. (XLSX 18kb)

Figure titles and legends

Figure 1. Workflow of MetaCHIP.

Figure 2. Example output for the flanking regions of an identified HGTs. Genes encoded on the forward strand are displayed in light blue, and genes coded on the reverse strand are displayed in light green. The name of genes predicted to be HGT are highlighted in blue, large font with pairwise identity given in parentheses. Contig names are provided at the left bottom of the sequence tracks and numbers following the contig name refer to the distances between the gene subject to HGT and either the left or right end of the contig. Red bars show similarities of the matched regions between the contigs based on BLASTN results.

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.

Figure 4. MetaCHIP's performance on the recovery of HGTs introduced between genomes at class and genus levels. 'BM' and 'PG' refer to recovered gene transfers after the 'best-match

approach' and 'phylogenetic approach', respectively. 'RD' shows the recovery of predictions with the 'right direction' of gene flow.

Figure 5. The effect of sequencing depth on the recovery of introduced gene transfers with different assemblers and different levels of genetic divergence.

Figure 6. The percentage of recovered gene transfers during assembly and binning.

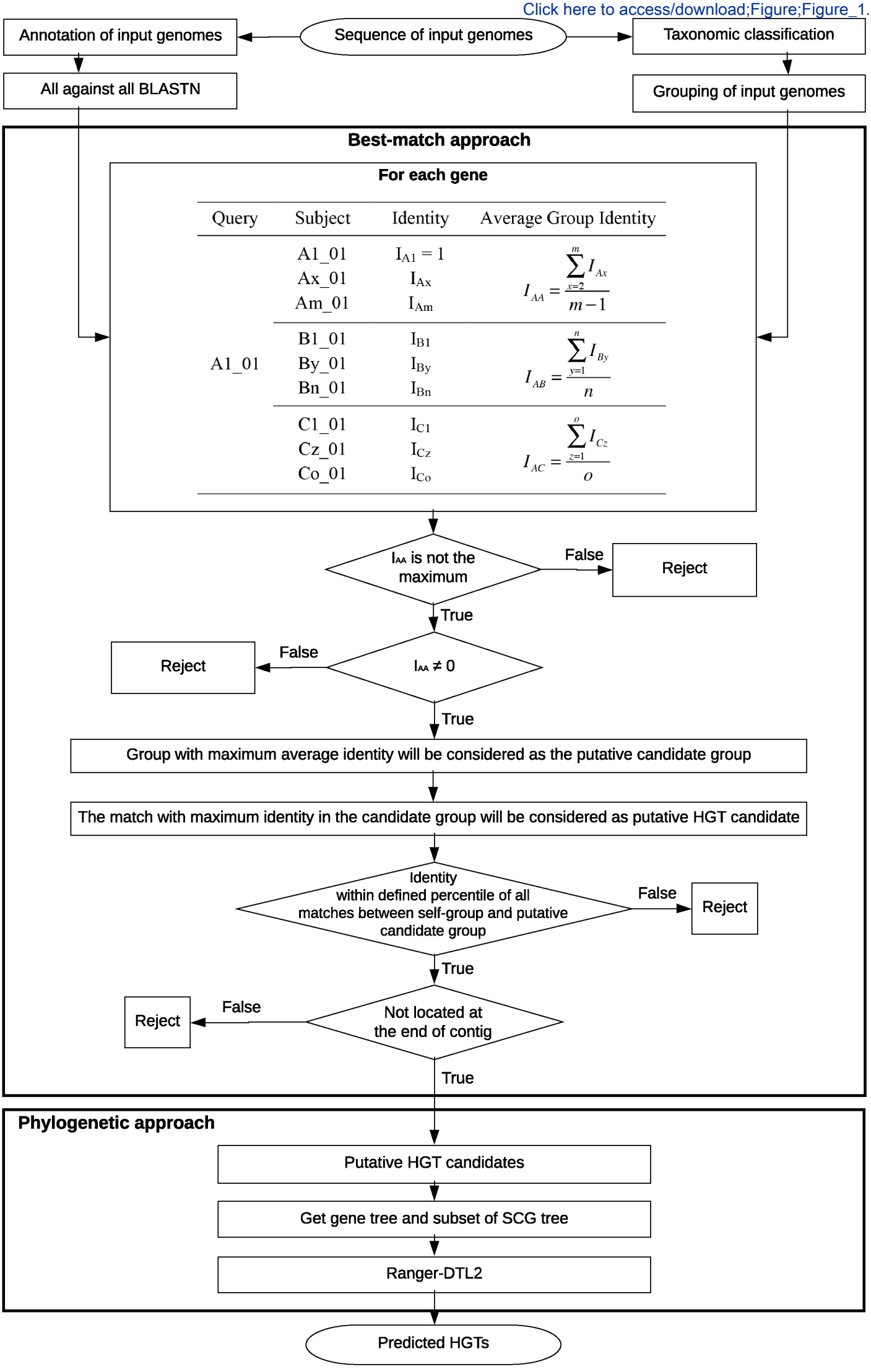
Figure 7. The percentage of recovered gene transfers by MetaCHIP after assembly of simulated reads and binning of genomes (simulation). For comparison, the results from original genomes (no reads simulation) are also shown and are the same as in Figure 4.

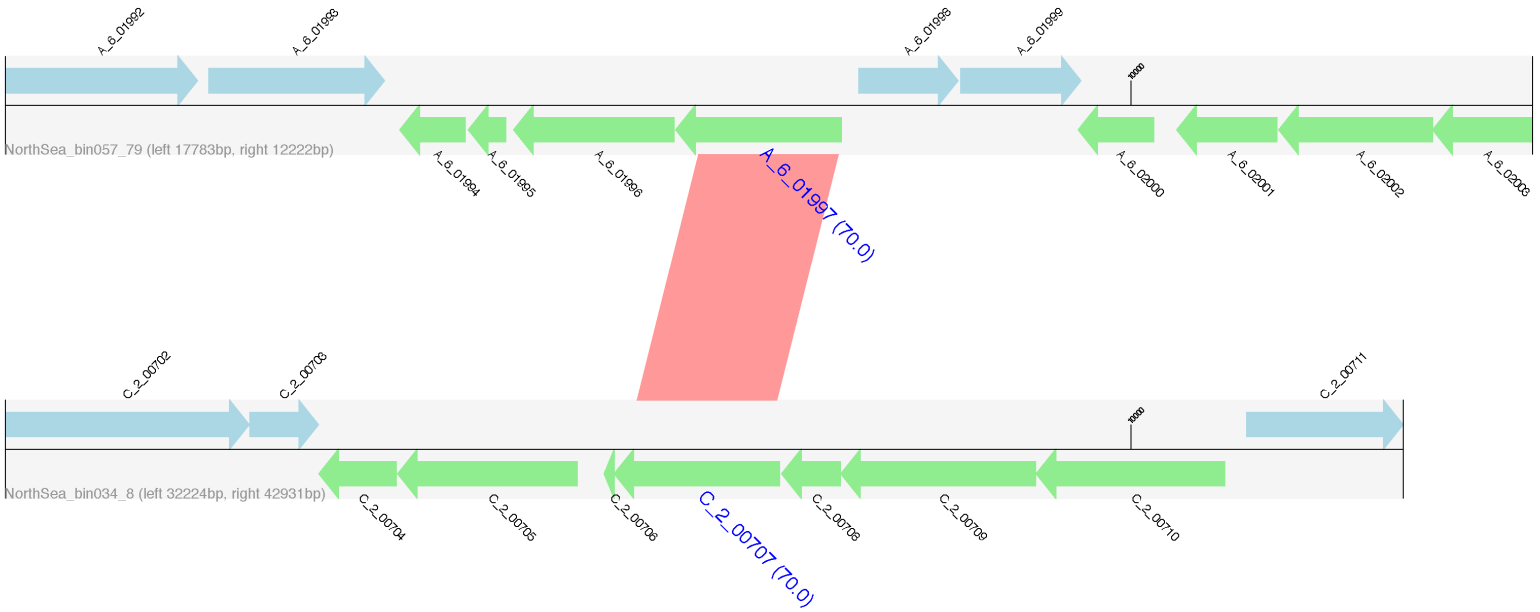
Figure 8. Genetic divergence of MetaCHIP identified HGTs from the 368 genomes of soil isolates.

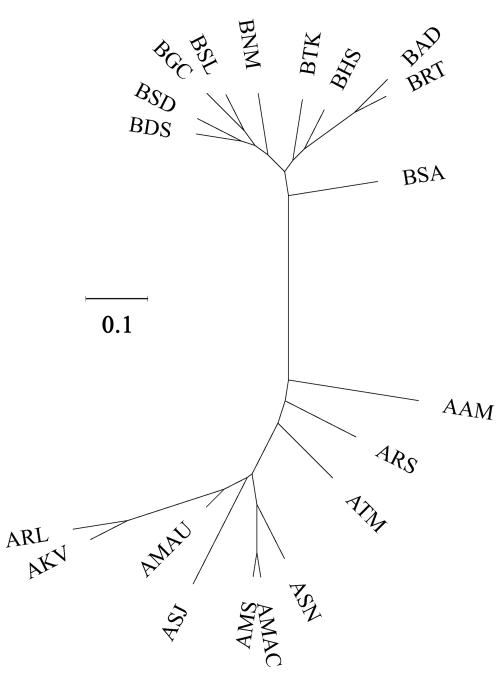
Figure 9. Relative proportion of COG functional categories for the 2094 genomes as well as MetaCHIP predicted recent (with genetic divergence ≤ 1%) and non-recent (with genetic divergence > 1%) HGTs. The boxes in the plot are bound by the 25% to 75% quartile proportions with the thick line being the 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).

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.

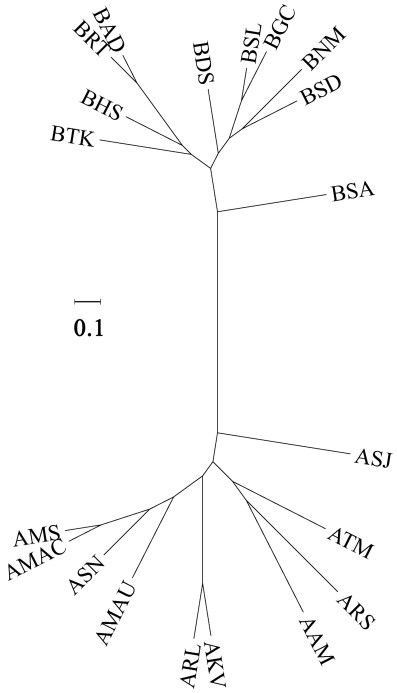
Figure 11. Relative proportion of COG functional categories for the input genome bins and predicted HGTs from human gut bins. The boxes in the plot are bound by the 25% to 75% quartile proportions with the thick line being the median value. Q1, Q3 and IQR refer to the 25%, 75% and interquartile range, 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.



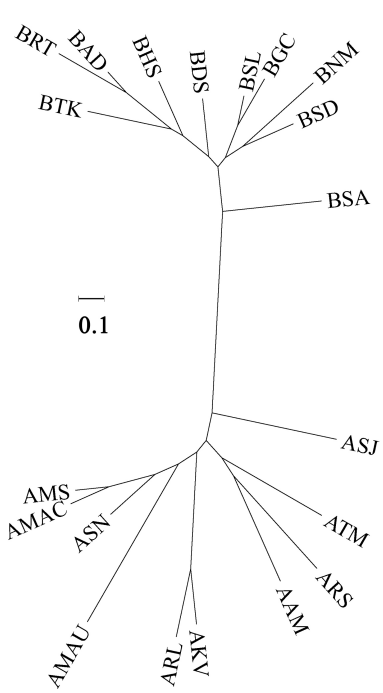




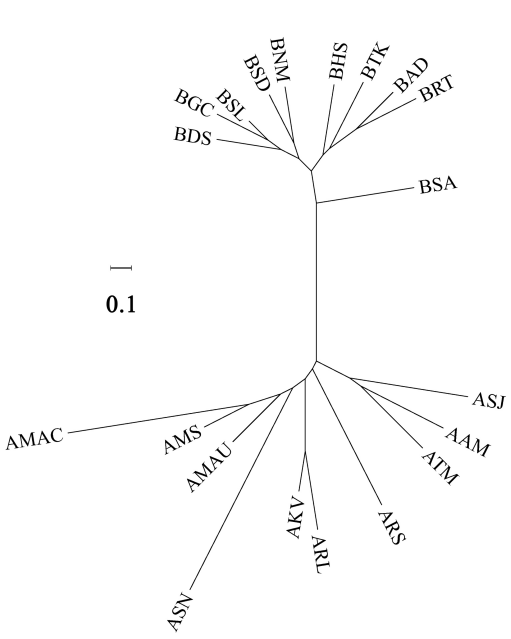
16S rRNA



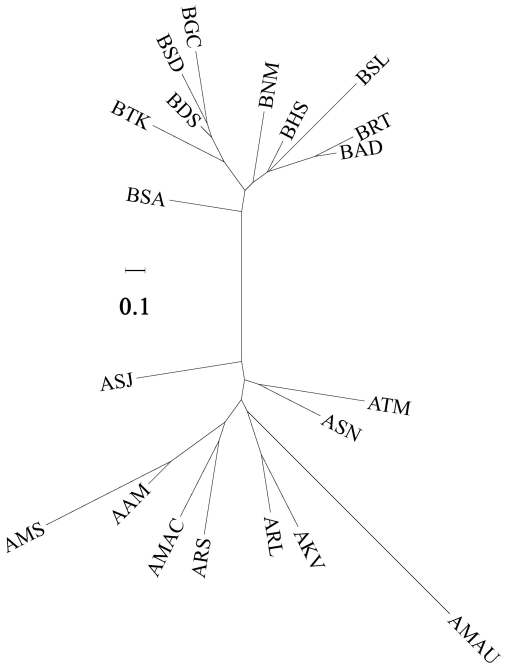
Completeness 100%
Similarity 94.22



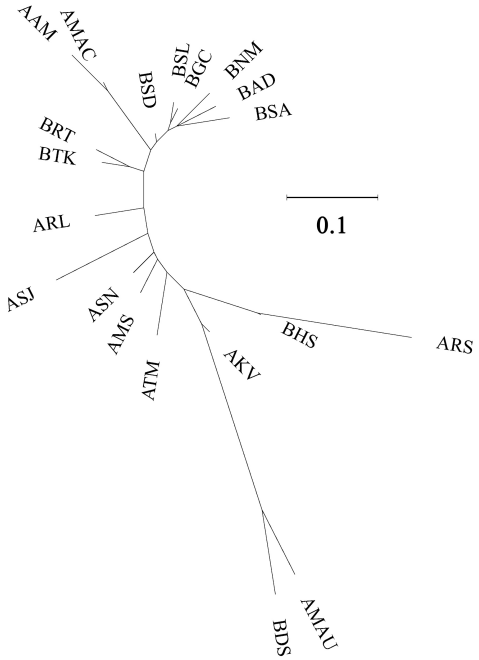
Completeness 80%
Similarity 91.29



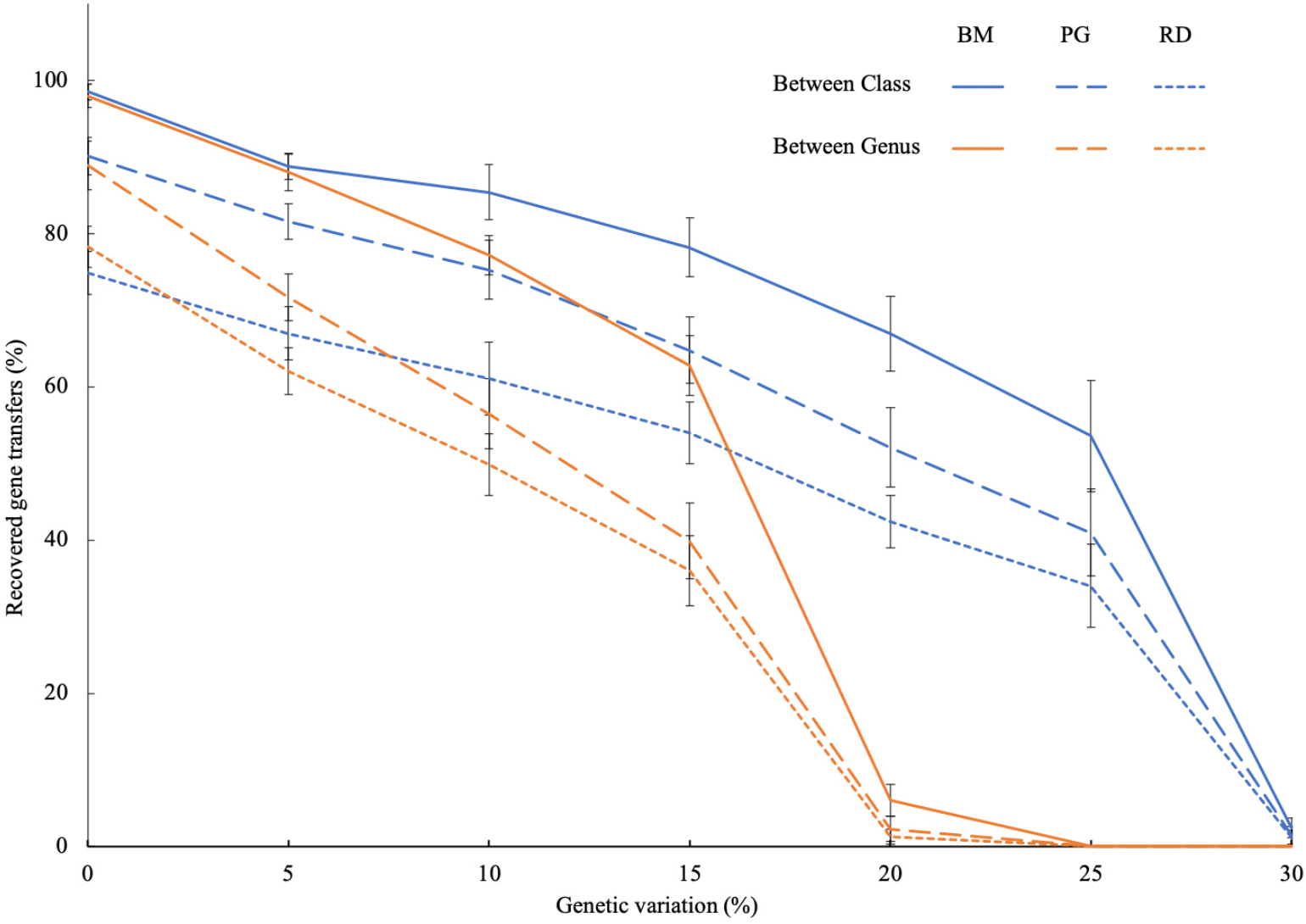
Completeness 60%
Similarity 84.93

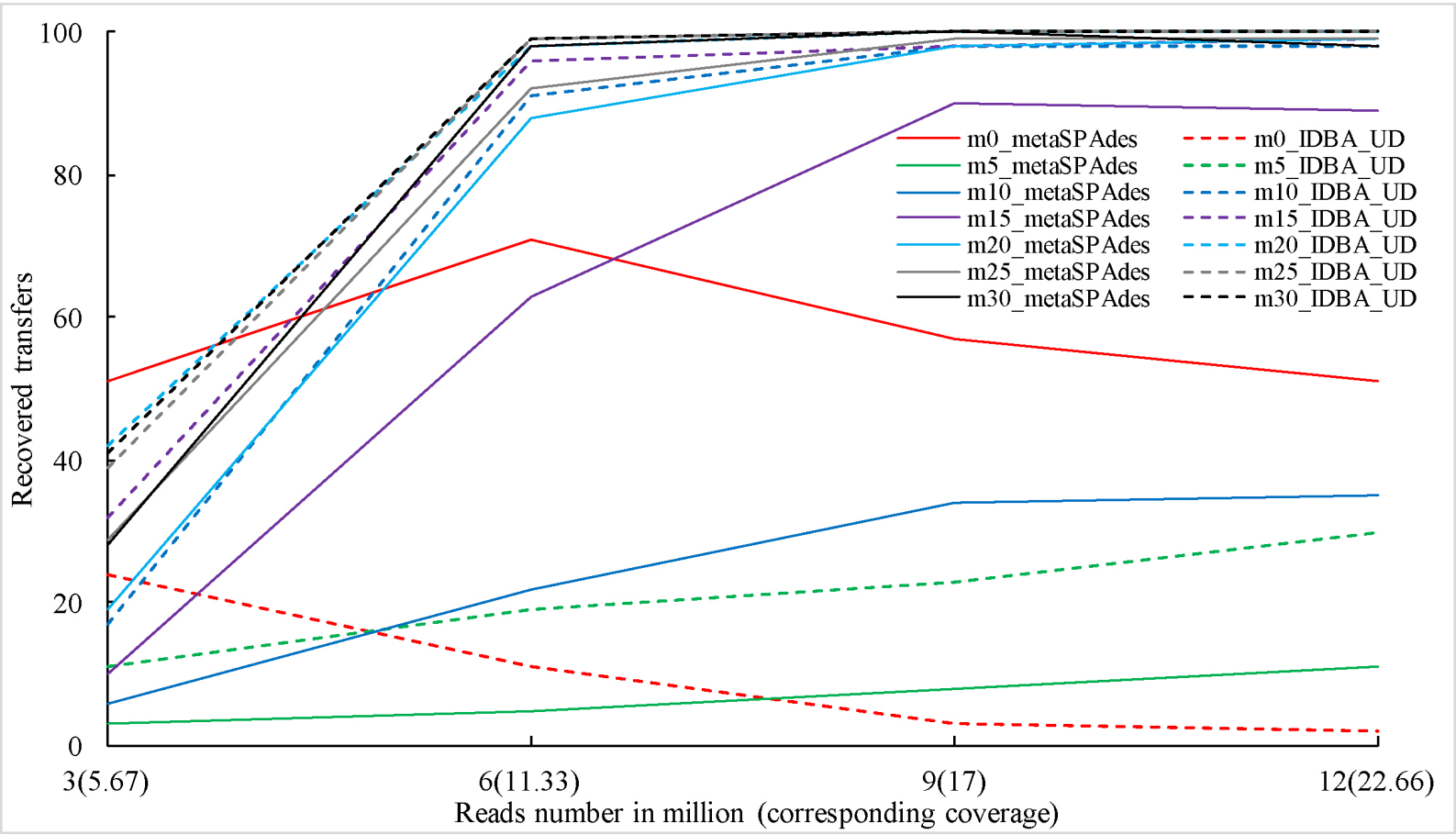


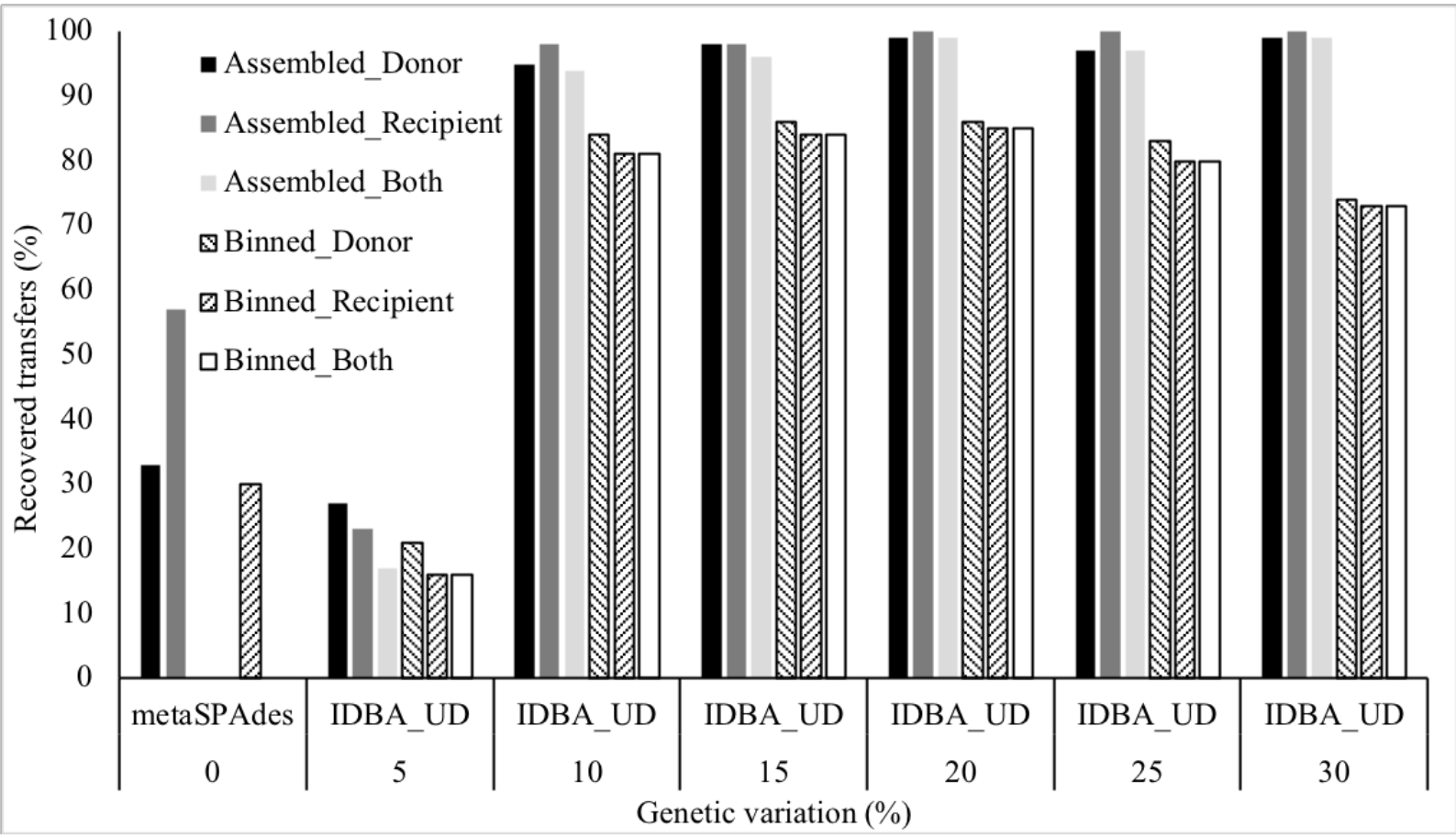
Completeness 40%
Similarity 72.64

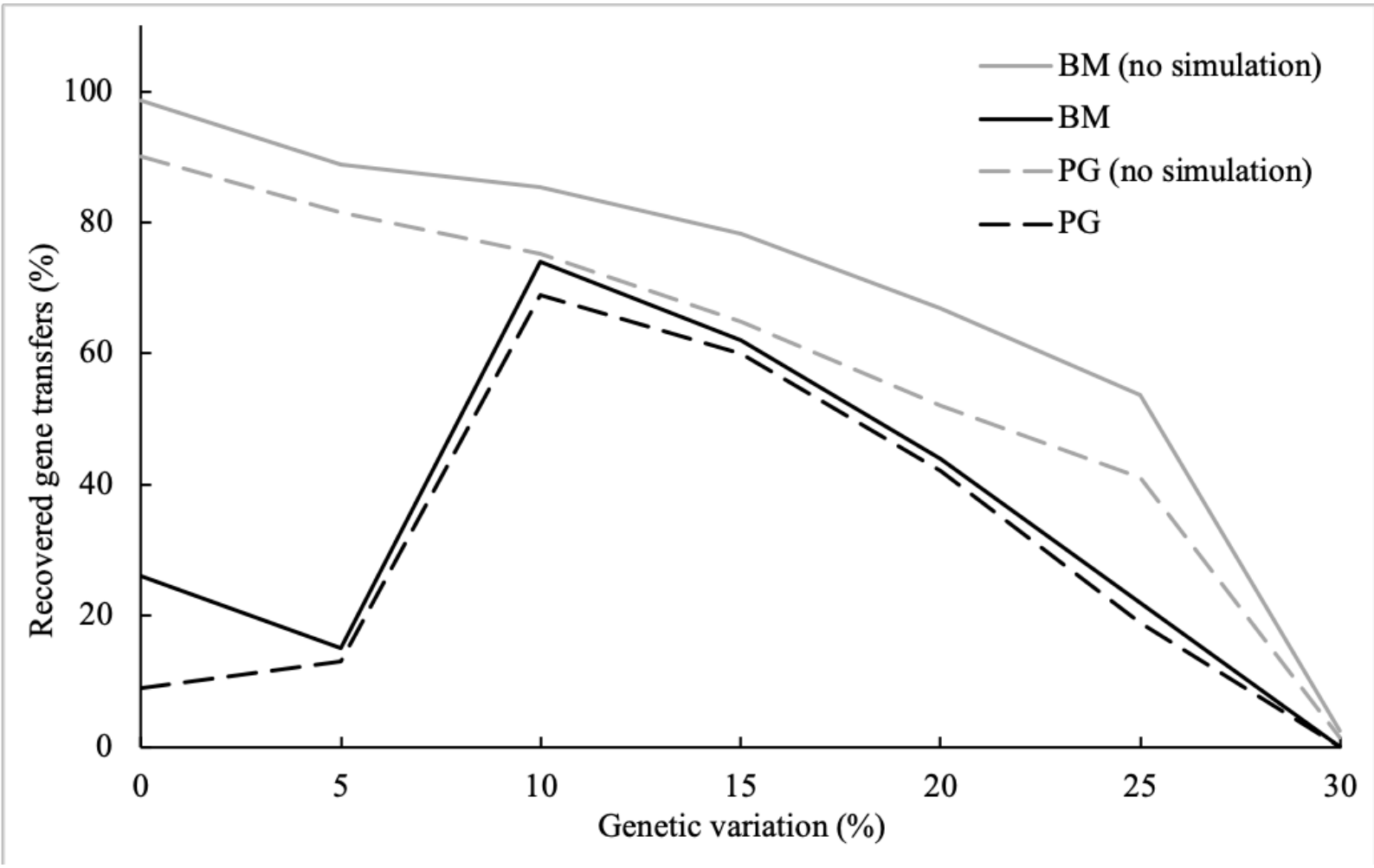


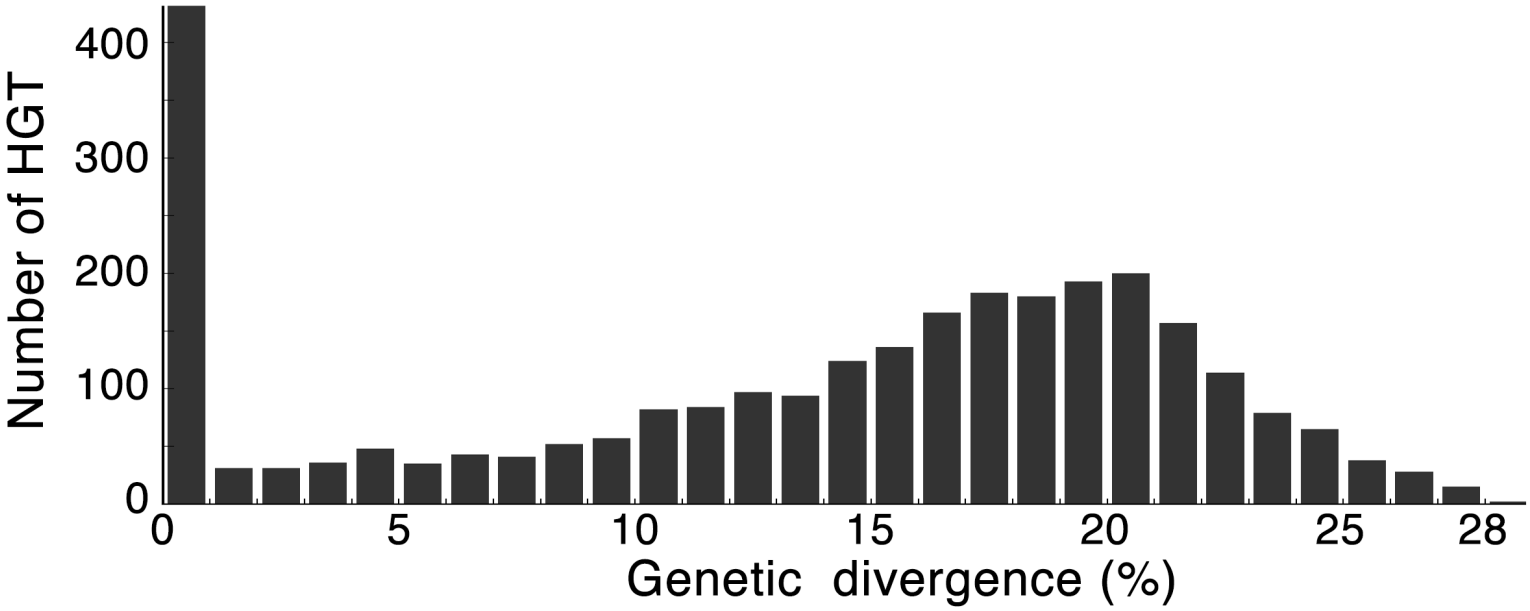
Completeness 20%
Similarity 0.72

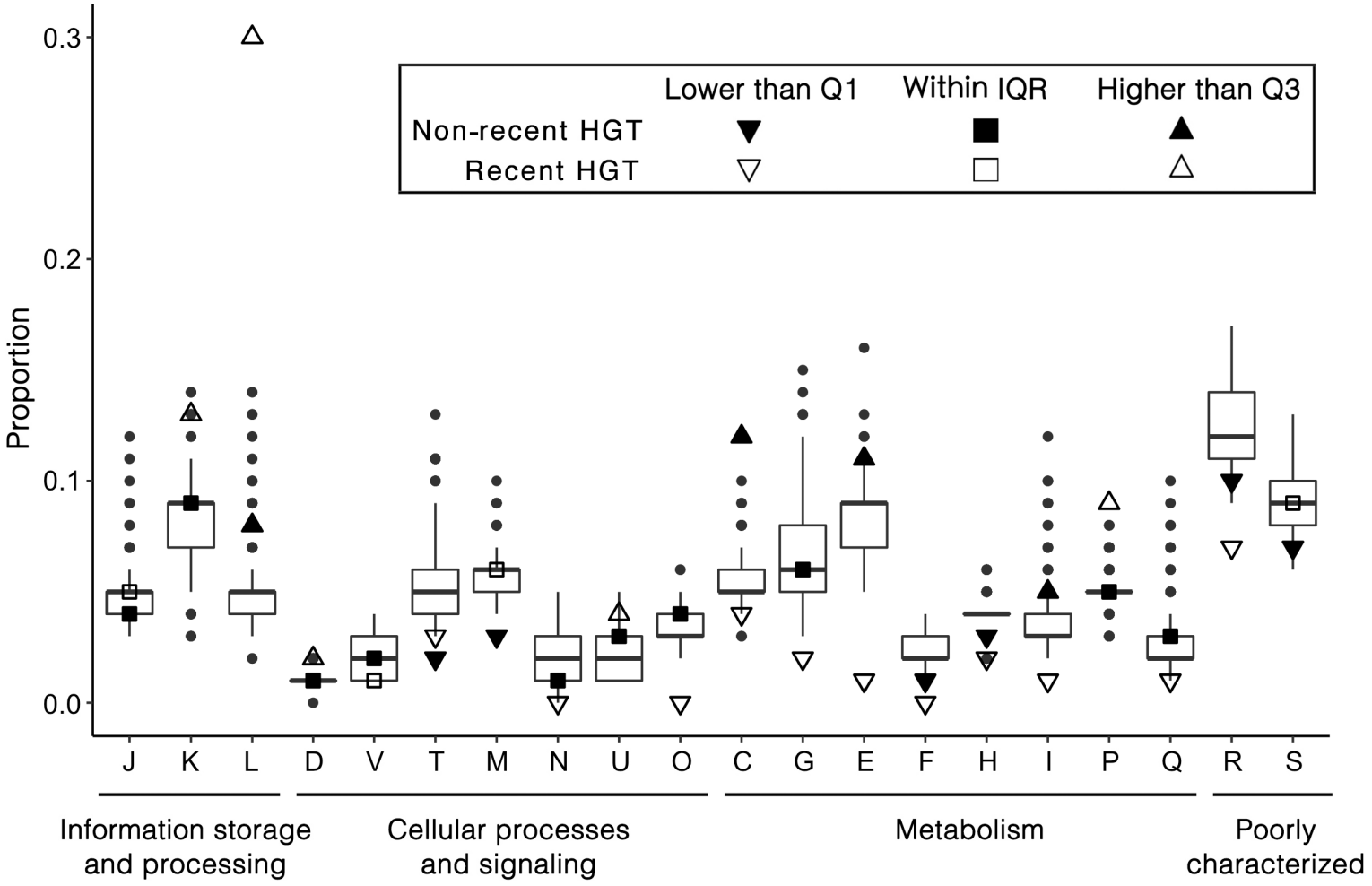


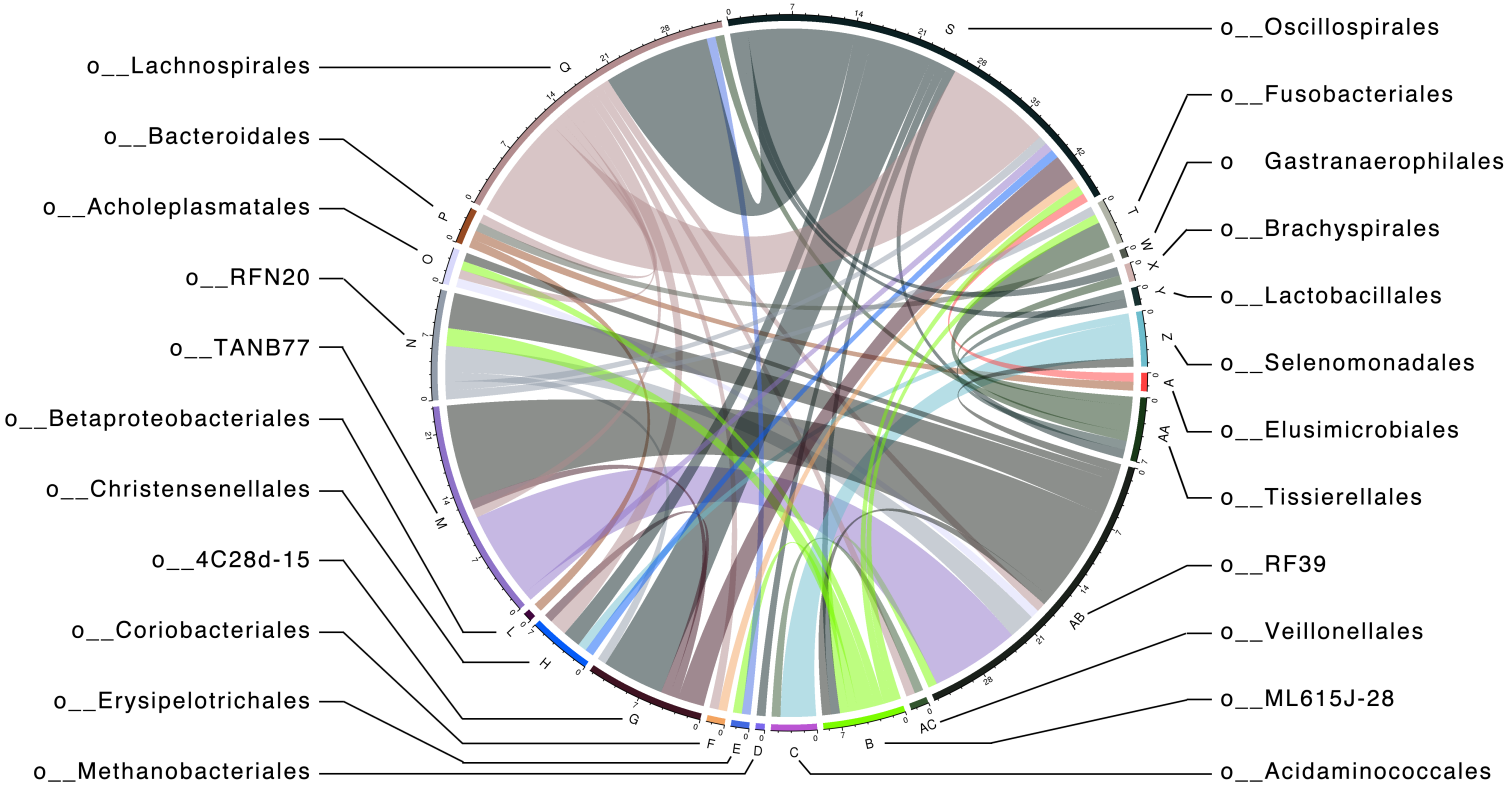




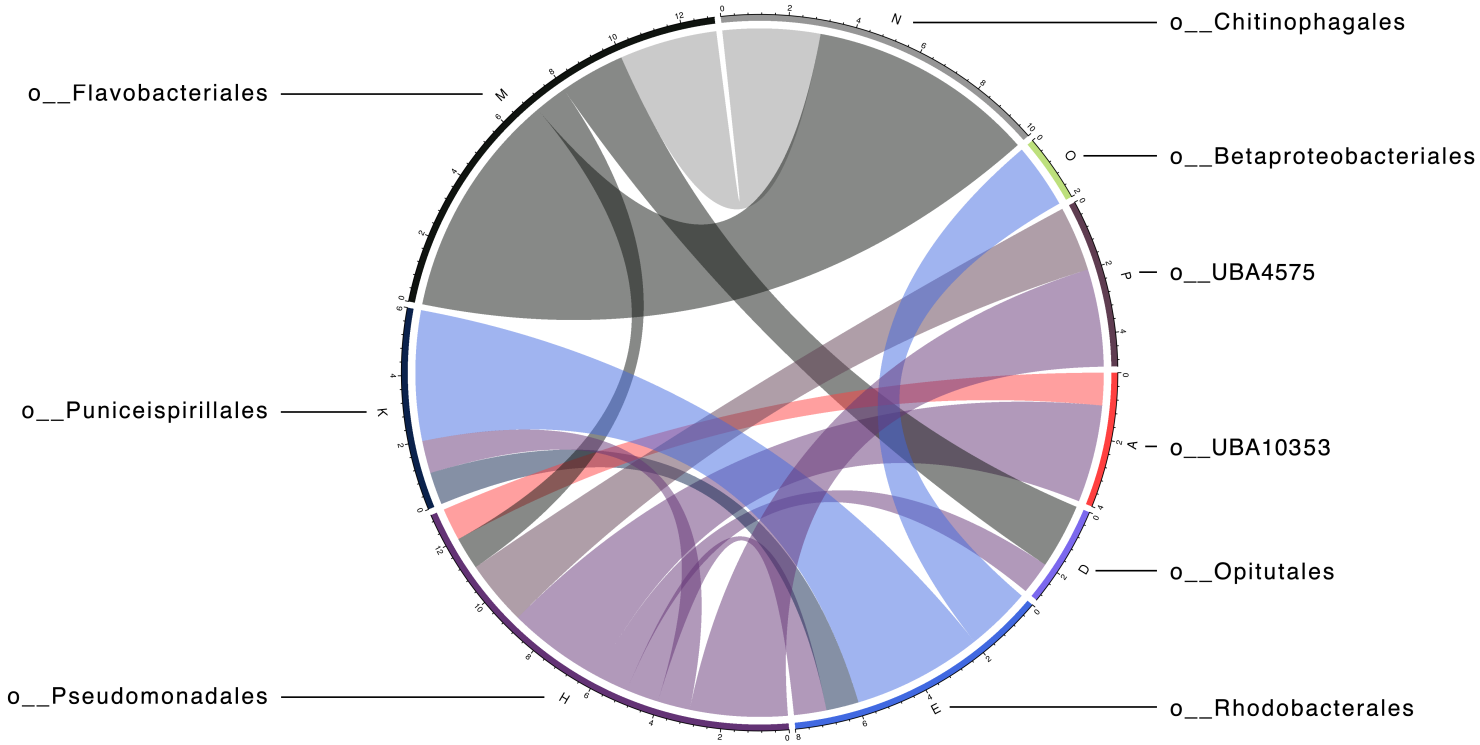








Human gut



North Sea