High resolution shotgun metagenomics: the more data, the better?

Julien Tremblay[1#] and Charles W Greer[1]

[1]: Energy Mining and Environment Research Centre, National Research Council Canada, Montreal, QC, Canada H4P-2R2

[#]: Corresponding author - julien.tremblay@nrc-cnrc.gc.ca

Running title: shotgun metagenomics data load input

Abstract

In shotgun metagenomics (SM), the state of the art bioinformatic workflows are referred to as high resolution shotgun metagenomics (HRSM) and require intensive computing and disk storage resources. The increase in data output of the latest iteration of high throughput DNA sequencing systems can allow for unprecedented sequencing depth at a minimal cost and will require adaptations in HRSM workflows architecture. Such a strategy is to generate so-called shallow SM datasets that contain fewer sequencing data per sample as compared to the more classic high coverage sequencing. While shallow sequencing is a promising avenue for SM, detailed benchmarks using real data are lacking. In this case study, we took two public SM datasets one moderate and the other massive in size and subsampled each dataset at various levels to mimic shallow sequencing datasets of various sequencing depths. Our results suggest that shallow SM sequencing is a viable avenue to obtain sound results regarding microbial structures and that high depth sequencing does not bring additional elements for ecological interpretation. One area, however, where ultra-deep sequencing and maximizing the usage of all data was undeniably beneficial was in the generation of metagenomic amplified genomes (MAGs). We finally include a proof of concept analysis showing that alpha diversity is the main driver of gut microbiome structure and demonstrate that this conclusion can be reached using shallow SM, validating this method as a viable and sound option for HRSM analyses.

Keywords

Shotgun metagenomics; bioinformatics; high performance computing

Introduction

DNA sequencing costs have decreased dramatically in recent years. With the introduction of Illumina's short reads-based NovaSeq system, the cost of sequencing has reached a new low mark. In the field of metagenomics, it is challenging to estimate the required output of the generated sequences needed to get a satisfactory level of coverage and is particularly true for complex environments, like soil and animal guts. To date, sequencing technology could hardly reach the saturation of such complex environments and common wisdom in estimating sequencing output for such environments was simply to generate the largest possible amount of sequence within budget constraints. However, the amount of data generated by the latest iteration of sequencing systems (i.e. Illumina's NovaSeq; Oxford Nanopores's Promethion) has reached a point where it can now far exceed computational capacity in shotgun metagenomics (SM) workflows [1] like what is loosely referred to as high-resolution shotgun metagenomics (HRSM). In this type of workflow, raw sequence libraries are usually controlled for quality, trimmed and *de novo* co-assembled. Quality controlled reads are then mapped back to the co-assembly in order to estimate contigs and gene abundance to ultimately generate abundance matrices and metagenome-assembled genomes (MAGs) [2]. The most critical aspect of this type of workflow is probably the *de novo* co-assembly of all the relevant sequence libraries generated for a given project. This step ideally requires what is referred to as a large compute node: a computer node usually equipped with tens of cores and large amounts or Random Access Memory

(RAM) to adequately process all short reads into a data structure to perform a *de novo* co-assembly of many samples at once. While some *de novo* assembly software are written to handle multiple compute nodes (i.e. MetaHipMer [3], Ray Meta [4]) through distributed-memory systems paradigms such as MPI, they are not the most practical to implement and use because of their inherent configuration complexity and do not necessarily generate the most accurate assemblies [5]. Moreover, the most widely used and arguably some of the best *de novo* assembly software are written as a single node solution (for instance metaSPAdes [6,7], MEGAHIT [6]). The question of what assembling software package is the most performant is currently the subject of debate [8] and is beyond the scope of this study. Although the analysis pipeline used for this study supports using metaSPAdes, we ended up using MEGAHIT for our analyses because it was the only viable option to process this objectively large dataset in an acceptable amount of time. It is to be noted that in order to circumvent the issue of having enough RAM resource to perform a large multi-sample co-assembly, some workflows instead favor performing de novo assembly for each sample (for instance, see [9]). This approach, however, prevents the optimal analysis of end results as it generates significant redundancy in assembled contigs and MAGs, making it intractable to directly compare abundance between samples. In contrast, co-assembling all samples together has the advantage of creating one single reliable baseline on which all samples can be easily compared/analyzed and increases the power of segregating contigs during MAGs generation.

Here we investigated the end results of a typical HRSM workflow using the largest public Illumina Novaseq6000 dataset available at the moment of writing

(PRJNA588513 [10]). This dataset holds 912 paired-end (2 x 150 bp) human gut microbiome shotgun metagenomic sequence libraries representing 12,389 giga-bases (Gb) of raw sequence data for a total of 5.6 terabytes (TB) of compressed fastqs. Co-assembling an amount of sequence data of this magnitude is not achievable on the vast majority of existing large compute node hardware as it would require unrealistic - or at least hardly accessible - amounts of both RAM and compute time. To circumvent this limitation, we emulated various level of shallow sequencing by iteratively subsampling this sequence data to up to a total of 2,927 Gb (3.2 Gb / library) which saturated our largest compute node (i.e. one '4 x Intel Xeon E7-8860v3 @ 2.20GHz; 3 TB RAM; 64 cores' node) and dissected the end results (taxonomic profiles, alpha- beta- diversity and MAGs). We performed the same exercise with a more modest dataset consisting of 18 NovaSeq6000 libraries (2 x 150 bp; 583 Gb) from samples obtained from Antarctic soil environments (PRJNA513362 [9]).

Results

Experimental design. A large-scale shotgun metagenomic sequencing project consisting of 912 gut microbiome samples collected from individuals across six provinces of China was recently published [10]. These samples were sequenced on numerous lanes of a NovaSeq6000 system and yielded a total of 12,232 Gb representing 5.7 TB of compressed fastqs. To our knowledge, this dataset is one of, if not the largest SM dataset to be made publicly available as part of a single project and provides an opportunity to determine if generating that much data was necessary to obtain meaningful results or validate hypotheses. As sequencing systems keep getting

more performant in generating data, this becomes an increasingly important question as processing 5.7 TB of raw SM data is not a trivial endeavour and will generate countless intermediate files, inflating the storage and compute resources requirements to properly analyze the end results. This maked it unsuitable for a HRSM workflow as is (i.e. without subsampling or reducing the raw reads input load). With the current trends in sequencing technology development, it is not unreasonable to expect a similar amount of sequencing data being routinely generated in the near future for any given SM projects.

We therefore downloaded the raw fastq files related to this project and processed them with an iterative subsampling strategy in order to determine if smaller subsets of this dataset would be sufficient to reach sound biological conclusions. Like the vast majority of research units, we did not have a compute node with enough RAM to perform a de novo co-assembly for the entirety of such a large dataset. The best resources we had access to at the moment of writing were 3 TB RAM 64 core nodes. Therefore, for each dataset, we adopted a strategy (Fig. 1) in which we performed quality control of all the raw sequence data files to trim adapters and remove sequencing artefacts and contaminants which yielded what we will refer to as quality controlled reads. We iteratively subsampled the quality-controlled fastqs to obtain 0.1, 0.5, 1, 4, 8 and 12 million of sequencing clusters (*i.e.* one sequencing cluster represents two reads of 150 bp each) for each sample which amounts to approximately 0.03, 0.14, 0.27, 1.07, 2.14 and 3.21 Gb per library, respectively (Table I). We then executed the remaining part of our HRSM workflow to perform the *de novo* co-assembly, mapping of quality controlled reads on the co-assembly to generate contigs and gene abundance

matrices, computation of alpha and beta diversity, assignation of both taxonomic lineages and KEGG orthologs (KO) and MAGs generation. We favored the MEGAHIT software to perform the *de novo* co-assembly because of its speed, its capacity to handle relatively large sequencing datasets and because it is designed to be executed on a single compute node. We performed the same *in silico* experiment with a SM sequence dataset of lower magnitude obtained from natural extreme Antarctic environments (PRJNA513362 [9]). Although this latter dataset is not as massive as the one described above, it is nonetheless substantial and typical of a recent SM dataset. Additionally, for this latter dataset, our compute resources could support processing all the fastqs totalling 523.16 Gb.

The amount of resources required for de novo assembly increases almost linearly with the number of input reads used for co-assembly. From Table I and Figure S1, we show that the amout of consumed RAM increases linearly with the number of input reads fed into the assembler, which suggests that the internal memory data structure of kmers still does not saturate as we reach the maximum memory capacity of the compute node during the assembly process. This also implies that this SM data, even though it is very large, contains a level of complexity for which kmer saturation is not reached with a 3 TB RAM node. Given the near perfect linear correlation between amount of bp and consumed RAM, significantly more memory would be required to take into account all the sequencing data (12 Tb) from this dataset which we estimated to be approximately 10 TB of RAM.

De novo co-assembly contiguity is positively correlated with the number of input sequences used for co-assembly. There is a clear linear correlation between the

quantity of bases used to perform the co-assembly and the contiguity of the generated assembly (Table I; Fig. S1). Simply put, the more the reads fed into the co-assembly, the more and longer contigs (and genes) are obtained. The number of quality controlled reads mapping on the co-assembly is a metric that can inform on the quality of the assembly. For the human gut dataset, the percentages of properly aligned reads were lower for the 0.1M cluster / library workflow (Fig. S2) while inputs from 0.5M onward showed an average mapped reads rate of approximately 90%. The number of reads mapped from the arm (i.e. all reads mapped) workflows - that is, mapping all the quality controlled reads from the dataset on the subsampled co-assemblies (see experimental design in Fig. 1) - showed high a mapping rate (> 90%) suggesting that while only a subset of reads were used to generate the co-assemblies, the generated contigs managed to catch the vast majority of the complete libraries. A similar trend is observed for the smaller Antarctic soil dataset, but with the saturation inflection point being reached at the 4M clusters onwards.

Beta diversity (Bray-Curtis dissimilarity index) comparison between various co-assemblies suggests that the amount of input of reads in the co-assembly does affect the overall population structure (Fig. 2). The Spearman correlations (Mantel tests between Bray-Curtis dissimilarity matrices) between low (0.1M, 0.5M and 1M clusters) and high (4M, 8M and 12M (and complete dataset for the Antarctic project)) input configurations were consistently showing lower values compared to high vs high configurations. The contigs and gene richness diversity index were found to be good indicators of each participant's microbiota composition and is visually highlighted in the PCoAs (Fig. 2A lower panel) in which participants cluster according to their diversity

quintiles in essentially identical patterns across all workflow configurations. For the Antarctic dataset, the workflow configuration that showed the lowest correlations with the others was the 0.1M - arm design with Spearman *r* statistics values of approximately 0.85. Accordingly, the PCoAs of this dataset gave nearly identical patterns for all workflows except the 0.1M - arm design which is significantly different (Fig. 2B lower panel).

Observed contigs and observed genes indexes as a function of sequencing efforts (Fig. 3) suggest that saturation is reached at the 4M subsampled clusters for the human gut dataset (Fig. 3A) while both metrics are still in the exponential phase for the Antarctic dataset (Fig. 3B).

Correlation at the taxonomic and functional level was also assessed and followed similar trends to what was found in the beta diversity comparisons of Figure 2 with low read inputs showing lower correlation against high read input configurations (Fig S3). Even though workflows with more input sequences are associated with a higher number of recovered taxa (Fig. S3 - barplots and Venn diagrams), these “rare” taxa only account for a minor fraction of the total reads (Fig. S3 lower right panels) and the quasi-totality of reads are associated with taxa common to all workflows. For each of the human gut and antarctic datasets, relative abundance profiles of a selection of some of the most overall abundant taxa (Fig. 4A;C) and KOs were generated to further validate their consistency through all workflows. For the human gut microbiome dataset, it is generally the case except for contigs assigned to the genus *Prevotella* that show significant variation in the 0.1M, 0.5M and 1M cluster workflows (Fig. 4A). Interestingly, the same sequencing cluster input loads processed with the arm method correct these

variations and bring them to the same stability level as the other workflow configurations. Important ecological trends like the overrepresentation of *Coproccus* and prevalence of *Bacteroides* and *Megamonas* in participants harboring a high diversity microbiota are consistent across all workflows. Except for the 0.1M clusters workflow, functional profiles were in general similar between other input workflows (Fig. S4). Significantly differentially abundant KOs were determined between low diversity (quintile #1) and high diversity (quintile #5) groups and their abundance profiles were similar between workflows as shown by the selected KOs K01593, K11444, K18143, K22225, K22607 that were found to be more abundant in low diversity participants and K00178, K16149 that were more abundant in high diversity participants. As shown in Fig. 4B, the abundance of these KOs are nearly identical for all participants in all workflows.

For the Antarctic dataset, a similar trend is observed with the relative abundance of selected taxa higher in low input workflows (0.1M, 0.5M and 1M), but with no correction by the arm method (Fig. 4C). This is also observed for KO relative abundance as illustrated by the selected KOs relative abundance through all workflows in Fig. 4D. In this dataset, ecological trends are consistent across workflows that contain 4M sequencing clusters or more, but lower input workflows show taxonomic and functional abundance profiles significantly different from the complete dataset.

The relationship between sequencing input and MAGs generation yield and quality was also investigated and showed that the yield, % contamination and completeness are positively correlated with the amount of input sequences in each workflow (Fig. 5). Moreover, the arm workflows consistently generated higher yield MAGs of better quality compared to the standard workflows.

Discussion

Shallow sequencing vs high depth sequencing.

Shallow shotgun metagenomic (SSM) sequencing has been the subject of only a few studies so far [11–13] and focused mainly on the correlation between SSM and 16S rRNA amplicon sequencing. They were done with artificial datasets or executed with reads-based bioinformatics methods, avoiding the *de novo* co-assembly process inherent to HRSM workflows. Here, we adopted an approach where we used real SM datasets of two very different environments and sequencing depths to get insights on 1) how much data is actually needed to obtain reliable microbial ecology results and 2) the consequences on computational resources required to analyze ever expanding SM datasets - a topic often overlooked and poorly considered during planning of a shotgun metagenomic project. We performed this study using a state of the art high-resolution shotgun metagenomic methodology, which compared to other methods such as reads-based and sample-centric assembly methods (*i.e.* performing a separate de novo assembly of each sequence library) is arguably the bioinformatic method that allows the generation of the most comprehensive and meaningful end results from raw SM datasets, including the generation of beta- alpha- diversity metrics, full length genes and MAGs [2]. In that regard, the human gut microbiome dataset selected for this study (PRJNA588513) was of particular interest, as it is one of the first massive publicly available SM dataset (12,389 Gb of raw sequence data) enabled by the Illumina NovaSeq platform generated as part of a single finite project and offers a glimpse of what kind of sequencing output could be routinely achieved in terms of sequencing

output in the near future for a metagenomic project. At first glance, obtaining lots of sequencing data for a SM project might seem like a good thing, but since a sequencing dataset of that magnitude cannot be readily used in its entirety in a HRSM workflow (i.e. storage of intermediate files; RAM requirements for co-assembly) on the vast majority of HPC systems, how to extract the most of it has yet to be explored. In order to shed some lights on these questions, we favored an experimental design where the complete dataset was randomly subsampled at various loads to mimic various sequencing depths.

From our results with the human gut dataset, it is clear that the 0.1M clusters dataset (or 0.03 Gbases / sample; total of 24.43 Gbases) was not enough to achieve a good correlation with the results of other subsampled analyses. However, because this dataset had 912 samples, sub-sampling slightly higher, at levels as low as 0.5M clusters / sample (0.14 Gbases / sample; total of 130.25 Gbases) resulted in acceptable correlations with larger subsampled datasets (Figs. S3-4). Even though the number of bases at 0.5M clusters sampling is objectively low on a per-sample basis, the corresponding total amount of bases can be considered adequate for capturing accurate population structure metrics. Moreover, even if the variability of the microbial diversity in each sample is highly dispersed (*i.e.* alpha diversity results in Fig. 3A), the fact that pooling 912 samples of 0.14 Gb each is enough to obtain a sound co-assembly and corresponding downstream results suggest that there is a redundant core of microbes common and abundant enough to many samples so that this allows for the pooling of only a tiny fraction of each library to obtain a decent quality co-assembly and consequently accurate downstream results. The estimations of required reads for SM

projects are usually considered on a per-sample basis, but this information should ideally be combined with the total number of samples that are part of the project. Accordingly, the number of required bases per sample should probably not be calculated on a per-sample basis, but on the total amount of bases required to obtain a reasonable co-assembly for the investigated biological system. In the case of the human gut dataset investigated here, it could be argued that 130 (0.5M clusters / sample) or 244 Gb (1M clusters / sample) of input data used for the co-assembly gave end results that were overall very similar to the results of the largest subsampled dataset of 2,927 Gb (12M clusters). In more practical terms, this suggests that this set of 912 samples could have been sequenced on one or two lanes of NovaSeq6000 S4 and give very similar results for a fraction of the cost of the original study. Similar trends were observed with the Antarctic SM dataset. In this case however, since the number of samples was much lower (i.e. 18) and highly variable from one another, the lowest subsampling level that gave the results mostly similar to the total dataset was approximately 20 Gbases (4M clusters / sample or 1.12 Gbases / sample).

Compute resources

In the field of SM, the question of how many reads should be generated per sample has always been critical and the subject of continuing discussions. Currently, the most common accepted answer is probably along the lines of: the more the better - depending on the available funds of course. However, with the significant increase in

sequencing throughput seen recently with the Illumina NovaSeq platform, we have reached a tipping point where computational capabilities required to perform a sound HRSM workflow are not sufficient to integrate all of the generated data anymore. In the current study it was simply not possible to perform a co-assembly of 12 Tb worth of SM data. By extrapolating from the relation between RAM as a function of input data (Table I; Fig. S1), co-assembling the human gut microbiome dataset would require approximately 10 TB of RAM and 900 hours (37.5 days) of continuous compute real time, which would translate into 57,600 core•hour (or 6.57 core•year). On most governmental and academic HPC systems, resources are usually allocated on a yearly basis and have to be carefully managed. In that regard, completely processing such a large dataset would inevitably represent a major sink in consumption of allocated resources, leaving few resources for the data processing of other projects.

MAGs

The one area where having more data unquestionably improved end results was for MAGs generation (Fig. 5). For both Human gut and Antarctic datasets investigated

here, the number and quality of MAGs increased with the number of reads used for the co-assembly. Moreover, MAGs generation was the only area where using the arm (all reads mapped) workflow undoubtedly benefited end results metrics (i.e. yield and quality of MAGs). There are a number of software packages to generate MAGs, here we presented results generated with MetaBAT2, essentially because the other implemented method in our workflow, Maxbin2, could not complete even after more than 200 hours of runtime on a 16 cores compute node for the human gut dataset.

Conclusion

The main driver that prompted us to perform this *in silico* study is the foreseeable inability to use all of the reads generated by the most recent sequencing platforms for a given shotgun metagenomics project in a HRMS type of workflow. We anticipate that even our largest memory compute nodes will hardly keep up with the amount of sequences to consider all of the reads in the *de novo* assembly process inherent to HRSM workflows. Disk storage also becomes a concern as the total file size of the sequencing library files and the intermediate data that has to be transiently stored on rapid disk storage is significant. For instance, for the human gut microbiome 12M clusters arm workflow, the total compressed file size including intermediate files (*i.e.* filtered fastqs, bam, co-assembly, abundance matrices) was approximately 20 TB. Therefore, the always increasing output of sequencing data is a significant concern for the stress imposed even on large high performance computing systems. As access for HPC resources becomes increasingly competitive (see for instance 2021 Resource Allocations Competition Results), fewer resources can be allocated to an ever

increasing pool of research needs. This underscores the importance of refining strategies for analyzing SM data and finding alternative advantageous ways to make good use of all the reads generated by latest sequencing technologies. In this regard, a type of workflow of potential interest would be to assemble each sample individually and combine the assembled contigs to generate a single consensual assembly. This workflow would have the advantage of eliminating the limitations in memory as de novo assembling a single library at a time should not require sizable amounts of RAM. The execution of such a workflow would however require an accepted computational method to merge or combine multiple assemblies together, which to the best of our knowledge does not exist. Software reported to perform this type of merging task are usually not maintained anymore and not suitable for modern large metagenomic assemblies [14–16] and target single genome assemblies [17–19]. At first glance, long reads (PacBio, Oxford Nanopores) can also seem attractive to replace short-reads in the objective of obtaining more contiguous co-assemblies, but performing a multi-library co-assembly of long reads data type also requires enormous amounts of RAM and compute time (personal observations), especially if reads need to be corrected prior to be assembled, as it is often the case with error-prone long reads data types.

Overall the findings reported here suggest that shallow sequencing, up to a certain level, allows reaching similar conclusions that could be reached with deep sequencing. The only exception to this is if maximizing the yield of high quality MAGs is a primary outcome or if there is a particular interest in rare functions or taxa. These conclusions are prone to have significant impacts on the planning of shotgun metagenomic projects as - in light of the results presented here - the number of

sequences per sample does not need to be overly abundant to reach sound conclusions of a biological system.

Methods

Bioinformatics. Sequencing libraries were processed in ShotgunMG, our metagenomics bioinformatics pipeline [20–22] that was developed over the GenPipes workflow management system [23]. Sequencing adapters were first removed from each read and bases at the end of reads having a quality score <30 were cut off (Trimmomatic v0.39;[24]) and scanned for sequencing adapters contaminants reads using BBDUK (BBTools v38.1) [25] to generate quality controlled (QC) reads. The QC-passed reads from each sample were co-assembled using MEGAHIT v1.2.9 [6] on a 3 terabytes of RAM compute node with iterative kmer sizes of 31, 41, 51, 61, 71, 81, 91, 101, 111, 121 and 131 bases. MetaSPAdes [7] was also considered for co-assembly, but could not complete even after several days of computing for the lowest input human gut dataset. *Ab initio* gene prediction was performed by calling genes on each assembled contig using Prodigal v2.6.3 [26]. Assignment of KEGG orthologs (KO) was done by using DIAMOND Blastp v2.0.8 [27] to compare each predicted gene amino acids sequence of the co-assembly against the KEGG GENES database [28,29] (downloaded on 2020-03-23). COG orthologs were assigned using RPSBLAST (v2.10.1+) [30] with the CDD training sets (ftp.ncbi.nlm.nih.gov/pub/mmdb/cdd/little_endian/). The QC-passed reads were mapped (BWA mem v0.7.17; [31]) against contigs to assess quality of metagenome assembly and to obtain contig abundance profiles. Alignment files in bam format were

sorted by read coordinates using samtools v1.9 [32], and only properly aligned read pairs were kept for downstream steps. Each bam file (containing properly aligned paired-reads only) was analyzed for coverage of contigs and predicted genes using bedtools (v2.23.0; [33]) using a custom bed file representing gene coordinates on each contig. Only paired reads both overlapping their contig or gene were considered for gene counts. Coverage profiles of each sample were merged to generate an abundance matrix (rows = contig, columns = samples). Taxonomic lineage assignment to each contig was performed using CAT v5.2.3 [34] with the following key parameters (*f*=0.5; *r*=1). Taxonomic summaries and beta diversity metrics were computed with microbiomeutils v0.9.4 [35]. Alpha diversity metrics were obtained with RTK v0.93.2 [36]. MAGs were generated using MetaBAT (v2.12.1) [37] using an abundance matrix generated with the jgi_summarize_bam_contig_depths software [37] with the —minContigLength 1000 —minContigDepth 2 and —minContigIdentity 97 parameters. The quality of obtained MAGs was assessed with CheckM v1.1.3 [38]. MaxBin2 [39] was also considered, but could not be completed even after several days of computing the lowest input human gut dataset.

Functional analyses. For each predicted gene, the best hit having at least an e-value ≤ 1e-10 against the KEGG genes database was kept as the KEGG representative of that gene. Similarly, COG representative of each gene corresponded to the best hit COG hit having at least an e-value ≤ 1e-10. Each assigned KEGG gene may be associated with a KEGG ortholog (KO). For the analyses of figures 4 and S4, read

counts all the genes from the human gut or antarctic gene abundance matrix that were assigned with the same KO were summed to generate a KO abundance matrix. For each library, each KO aggregated value was divided by the total read abundance of the *recA* gene (COG0468) to obtain a normalized value for each KO.

Statistics and figures were generated using R v4.1.2 for which the code is available on github (https://github.com/jtremblay/shotgunmg_paper).

Availability of source code and requirements

ShotgunMG - http://jtremblay.github.io/shotgunmg.html

The ShotgunMG pipeline wrapper code and Python, Perl and R scripts that are being called by ShotgunMG are available here:

https://bitbucket.org/jtremblay514/nrc_pipeline_public/src/1.3.0/

https://bitbucket.org/jtremblay514/nrc_tools_public/src/1.3.0/

External software packages module install scripts are available here:

https://bitbucket.org/jtremblay514/nrc_resources_public/src/1.3.0/

A Docker image built on the CentOS 7 operational system which contains all necessary

modules for full pipeline functionality is available for testing/evaluation purposes and running small datasets (https://cloud.docker.com/u/julio514/repository/docker/julio514/centos).

Availability of supporting data and materials

Human gut microbiome dataset: PRJNA588513

Antarctic soil dataset: PRJNA513362

An example of the complete list of commands with all parameters for each package used in our analysis pipeline for both the human gut and antarctic dataset and all the key results generated during this study, including de novo assemblies, genes and contigs abundance matrices, functional and taxonomic annotations and MAGs are available on zenodo.org : https://doi.org/10.5281/zenodo.6349279.

Declarations

List of abbreviations

Tb = Tera-base; TB = Terabyte; Gb = Giga-base; GB = Gigabyte; HPC = High Performance Computing; bp = base-pairs; MB = Megabyte; SM = Shotgun metagenomics; SSM = Shallow Shotgun Metagenomics; HRSMG = High resolution shotgun metagenomics. KO = KEGG ortholog; arm = all reads mapped.

Competing Interest

The authors declare that they have no competing interests.

Author contributions

JT planned the experimental design, wrote the software, analyzed the data and wrote the manuscript. CWG analyzed results and edited the manuscript.

Acknowledgments

We wish to acknowledge Compute Canada for access to both the Waterloo University (Graham system) infrastructure. We acknowledge Shared Services Canada for access to the General Purpose Scientific Cluster (GPSC).

Figure legends

Figure 1. Experimental design used to investigate the effects of the number of reads input on the end results of two shotgun metagenomics datasets sequenced on a NovaSeq6000 system. For each dataset, a common core workflow was performed where each library was trimmed according to their quality profile and filtered for common contaminants. Quality controlled libraries were subsampled at 0.1, 0.5, 1, 4, 8 and 12 million clusters (i.e. times two for the number of reads). Subsampled quality controlled libraries were then co-assembled and analyzed for downstream analyses for the standard workflow. In the ‘all reads mapped’ (arm) workflow, all of the quality controlled reads were mapped against their corresponding co-assembly to estimate contig and gene abundance and investigate how it affects end results. The Antarctic

dataset was small enough so that the whole dataset could be assembled and processed at once.

Figure 2. Heatmaps of Spearman correlation coefficients (Mantel test) of normalized (TMM method from edgeR) contigs abundance Bray-Curtis (beta diversity) dissimilarity matrices between each workflow for the A) human gut and B) Antarctic datasets. Higher coefficient means higher correlation between distance matrices. The lower panels show PCoA plots of beta-diversity of each dataset at each sequencing depth. For the human gut dataset, samples were found to cluster according to their diversity level (i.e. diversity quantiles) with high diversity individuals clustering together and low diversity individuals clustering together as well. Samples were colored by sampling location for the Antarctic dataset.

Figure 3. Observed contigs and genes indices binned by quintiles for the A) human gut and B) Antarctic datasets. To improve ease of data visualization, each data point was binned in a quintile. The data points for a given quintile correspond to the same data points from one boxplot to another. Diversity indices were computed from raw contigs or genes abundance tables using RTK v0.93.2.

Figure 4. Taxonomic and functional profiles of selected taxa and KOs for the human gut and Antarctic datasets for each of the described workflows. A) Taxonomic summary of selected taxa for each workflow for quintiles #1 (participants with low diversity microbiota) and #5 (participants with high diversity microbiota). B) Abundance profiles of a selection of significantly differentially abundant KOs for each workflow for quintiles #1 and #5. KOs were identified by using a one-way anova between samples of quintiles #1 and #5 followed by a Bonferroni correction. KOs having a corrected p-value < 0.01 and a 8 times fold-change between the two quintiles were selected. C) Taxonomic summary of selected taxa for the Siegfried Peak and University Valley locations. D) Abundance profile of selected KOs for these same two locations. The Antarctic dataset was unreplicated and was therefore not conducive for feature selection by a statistical method. As a consequence, a selection of the most abundant KOs is displayed in D). Each point represents the aggregated normalized KO count (see methods) of a sample.

Figure 5. Histograms of MAGs quality assessment metrics for each workflow of the A) human gut and B) Antarctic datasets. MAGs were generated with MetaBAT2 v2.12.1 and Completion % and contamination % were obtained with CheckM v1.1.3.

Figure S1. Statistics on the various co-assemblies generated on a A) per Gb scale and B) per million sequencing clusters scale.

Figure S2. Statistics of aligned reads on co-assembly references for the A) human gut and B) Antarctic datasets. arm = all reads mapped workflow.

Figure S3. Taxonomic coverage analyses for the A) human gut and B) Antarctic datasets. Top left panel: Spearman correlation coefficients of Bray-Curtis dissimilarity

matrices computed from relative abundance taxonomic summary tables of each workflow. The Higher coefficient means higher correlation between taxonomic summaries. Top right panel: Number of unique taxonomic lineages observed as a function of each workflow's input data load. Lower left panel: Venn diagram of each unique taxonomic lineage as a function of data input load workflows. Lower right panel: Number of reads associated with each category listed in the Venn diagram.

Figure S4. KEGG orthologs (KO) coverage analyses for the A) human gut and B) Antarctic datasets. Top left panel: Spearman correlation coefficients of Bray-Curtis dissimilarity matrices computed from normalized KO abundance tables of each workflow. The Higher coefficient means higher correlation between KO abundance profiles. Top right panel: Number of unique KO observed as a function of each workflow's input data load. Lower left panel: Venn diagram of each unique KO as a function of data input load workflows. Lower right panel: Number of reads associated with each category listed in the Venn diagram.

References

1. Breitwieser, F. P., Lu, J. & Salzberg, S. L. A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform.* **20**, 1125–1136 (2019).
2. Chao Yang, Debajyoti Chowdhury, Zhenmiao Zhang, William K. Cheung, Aiping Lu, Zhaoxiang Bian, Lu Zhang. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput. Struct. Biotechnol. J.* **19**, 6301–6314 (2021).
3. Georganas, E. *et al.* Extreme Scale De Novo Metagenome Assembly. *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis* (2018) doi:10.1109/sc.2018.00013.
4. Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F. & Corbeil, J. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* **13**, R122 (2012).
5. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation-a benchmark

of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).

6. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
7. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
8. Meyer, F. *et al.* Critical Assessment of Metagenome Interpretation - the second round of challenges. *bioRxiv* (2021) doi:10.1101/2021.07.12.451567.
9. Coleine, C. *et al.* Metagenomes in the Borderline Ecosystems of the Antarctic Cryptoendolithic Communities. *Microbiol Resour Announc* **9**, (2020).
10. Sun, Y. *et al.* Population-Level Configurations of Gut Mycobiome Across 6 Ethnicities in Urban and Rural China. *Gastroenterology* vol. 160 272–286.e11 (2021).
11. Hillmann, B. *et al.* Evaluating the Information Content of Shallow Shotgun Metagenomics. *mSystems* (2018) doi:10.1128/mSystems.00069-18.
12. Xu, W. *et al.* Characterization of Shallow Whole-Metagenome Shotgun Sequencing as a High-Accuracy and Low-Cost Method by Complicated Mock Microbiomes. *Front. Microbiol.* **0**, (2021).
13. Snipen, L., Angell, I.-L., Rognes, T. & Rudi, K. Reduced metagenome sequencing for strain-resolution taxonomic profiles. *Microbiome* **9**, 1–19 (2021).
14. Scholz, M., Lo, C.-C. & Chain, P. S. G. Improved assemblies using a source-agnostic pipeline for MetaGenomic Assembly by Merging (MeGAMerge) of contigs. *Sci. Rep.* **4**, 6480 (2014).

15. Vicedomini, R., Vezzi, F., Scalabrin, S., Arvestad, L. & Policriti, A. GAM-NGS: genomic assemblies merger for next generation sequencing. *BMC Bioinformatics* **14 Suppl 7**, S6 (2013).
16. Soto-Jimenez, L., Estrada, K. & Sanchez-Flores, A. GARM: Genome Assembly, Reconciliation and Merging Pipeline. *Current Topics in Medicinal Chemistry* vol. 14 418–424 (2014).
17. Tang, L., Li, M., Wu, F.-X., Pan, Y. & Wang, J. MAC: Merging Assemblies by Using Adjacency Algebraic Model and Classification. *Front. Genet.* **0**, (2020).
18. Wences, A. H. & Schatz, M. C. Metassembler: merging and optimizing de novo genome assemblies. *Genome Biol.* **16**, 207 (2015).
19. Lin, S.-H. & Liao, Y.-C. CISA: contig integrator for sequence assembly of bacterial genomes. *PLoS One* **8**, e60843 (2013).
20. Liu, J. *et al.* Long-Term Land Use Affects Phosphorus Speciation and the Composition of Phosphorus Cycling Genes in Agricultural Soils. *Front. Microbiol.* **0**, (2018).
21. Tremblay, J. *et al.* Chemical dispersants enhance the activity of oil- and gas condensate-degrading marine bacteria. *ISME J.* **11**, 2793–2808 (2017).
22. Tremblay, J, NRC pipeline public code repository. ttps://bitbucket.org/jtremblay514/nrc_pipeline_public. (Bitbucket)
23. Bourgey, M. *et al.* GenPipes: an open-source framework for distributed and scalable genomic analyses. *Gigascience* **8**, (2019).
24. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114 (2014).

25. BBMap. *SourceForge* https://sourceforge.net/projects/bbmap/.

26. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

27. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).

28. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

29. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, (2019).

30. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

31. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

32. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

33. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* vol. 26 841–842 (2010).

34. von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **20**, 1–14 (2019).

35. Tremblay, J. *microbiomeutils: Python utility to generate distance matrices, perform PCoAs and generate taxonomic summaries using simple tab-separated feature tables*. (Github).

36. Saary, P., Forslund, K., Bork, P. & Hildebrand, F. RTK: efficient rarefaction analysis of large datasets. *Bioinformatics* vol. 33 2594–2595 (2017).

37. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

38. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043 (2015).

39. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2015).

Table I. Assembly statistics for each data input loads of both human gut and Antarctic shotgun metagenomic datasets.

Number of clusters (M)	Number of bases (Gb) / sample	Number of bases (Gb)	Mem (GB)	Compute time (hrs)	N50	Max contig length	Number of assembled bases	Number of contigs	Number of genes	gt10kb	gt20kb	gt40kb	gt80kb	gt160kb	gt320kb	gt640kb	gt1Mb
Human gut																	
0.1	0.03	24.43	25	10:09:39.00	3,672	203,220	787,465,679	277,959	918,037	9,861	2,670	492	50	3	0	0	0
0.5	0.14	130.25	101	12:00:31.00	3,962	296,162	1,917,013,874	649,225	2,242,645	25,987	6,999	1,432	211	25	0	0	0
1	0.27	244.34	202	49:57:50.00	4,026	411,438	2,646,482,912	888,205	3,097,246	36,209	10,009	2,082	316	26	3	0	0
4	1.07	977.19	806	116:57:29.00	4,071	549,470	4,624,054,834	1,542,016	5,427,163	63,450	17,869	4,090	734	88	4	0	0
8	2.14	1,953.41	1611	184:41:41.00	4,165	549,492	5,955,714,998	1,961,647	6,980,671	82,488	23,784	5,597	1,064	144	16	0	0
12	3.21	2,927.52	2414	225:05:31.00	4,229	549,546	6,836,771,270	2,234,441	8,005,820	96,332	28,024	6,834	1,365	187	20	0	0
Estimated for all reads	*13.41*	*12,232.00*	*9998*	*948:05:31.00*	-	-	-	-	-	-	-	-	-	-	-	-	-
Antarctic																	
0.1	0.03	0.50	2	00:10:31.00	1,394	38,038	5,167,606	3,319	5,909	20	5	0	0	0	0	0	0
0.5	0.14	2.52	6	01:28:46.00	2,092	87,764	101,375,698	48,610	123,818	912	176	12	1	0	0	0	0
1	0.28	5.04	11	03:07:59.00	2,059	89,591	309,165,148	153,452	386,103	1,311	274	32	2	0	0	0	0
4	1.12	20.15	26	09:04:57.00	4,002	666,973	1,038,288,457	348,473	1,228,815	13,271	4,212	1,185	297	31	4	1	0
8	2.24	40.31	38	14:09:00.00	4,148	428,970	1,601,062,499	520,149	1,893,811	19,941	6,246	1,975	527	106	11	0	0
12	3.36	60.46	51	18:41:55.00	4,564	658,365	2,000,521,439	622,427	2,358,452	27,726	8,677	2,446	601	121	9	1	0
All data	29.06	523.16	432	59:15:41.00	6,227	1,015,264	4,976,503,623	1,308,831	5,740,456	80,886	27,735	8,809	2,374	417	61	6	2

gt10kb: number of contigs greater than 10 kb; gt20kb: number of contigs greater than 20 kb;
gt40kb: number of contigs greater than 40 kb; gt80kb: number of contigs greater than 80 kb;
gt160kb: number of contigs greater than 160 kb; gt320kb: number of contigs greater than 320 kb;
gt640kb: number of contigs greater than 640 kb; gt1Mb: number of contigs greater than 1 Mb.

Figure 1

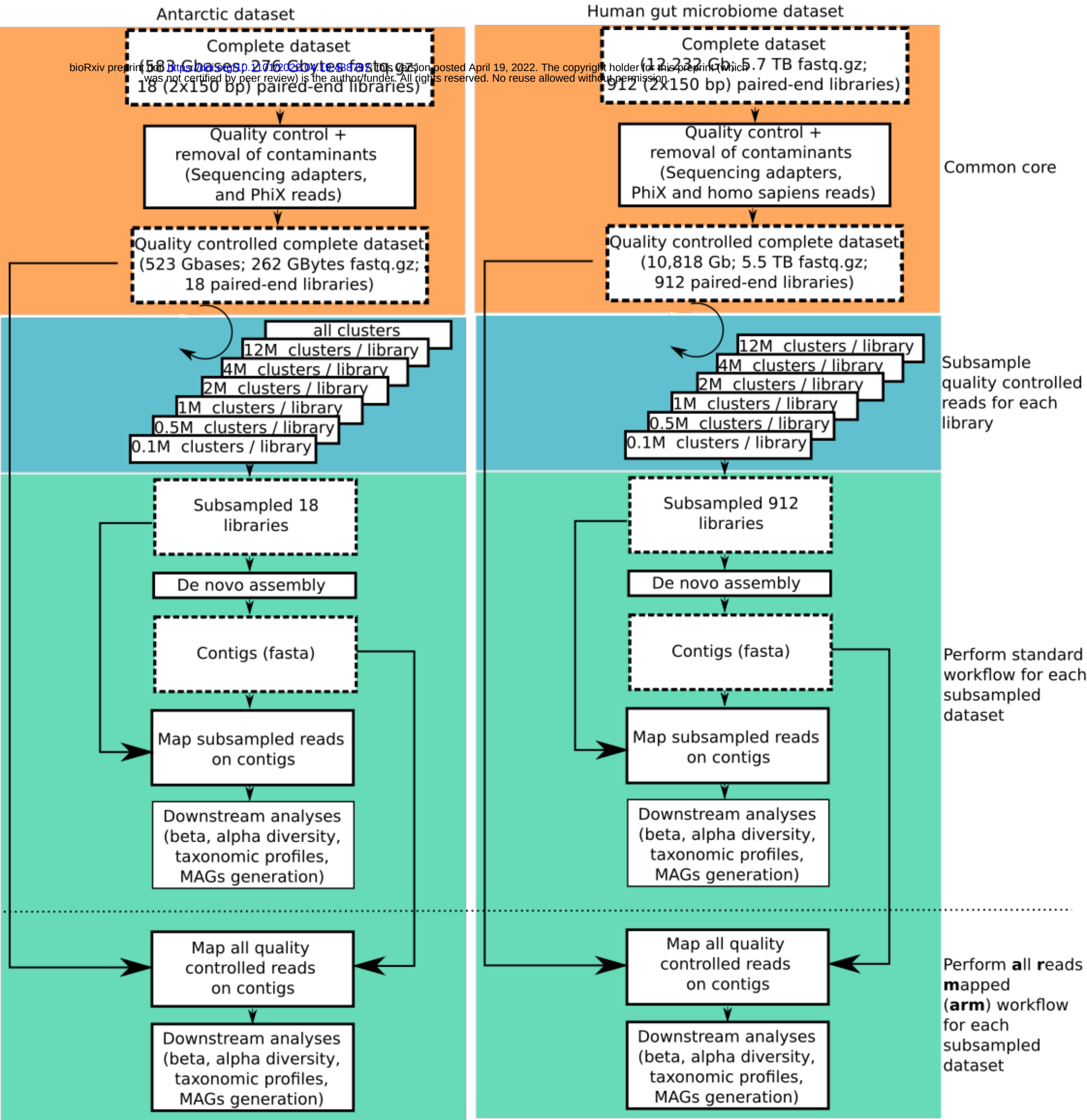
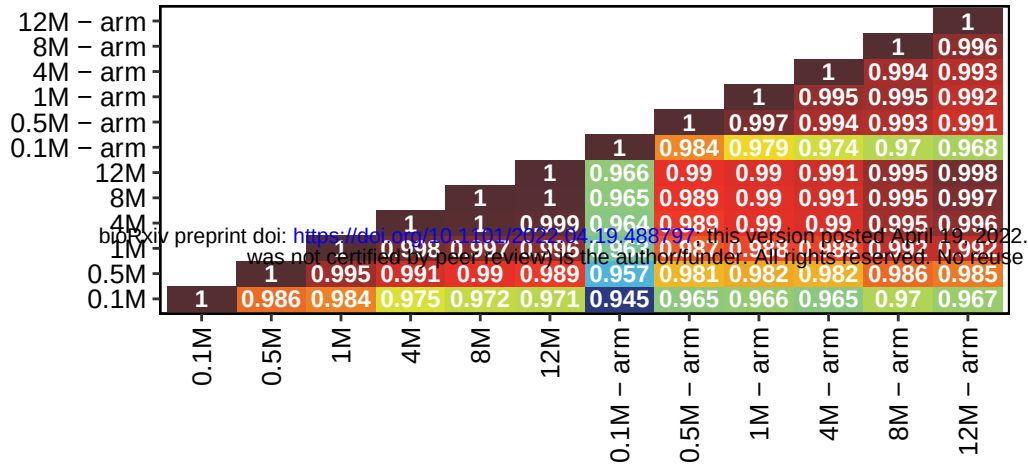
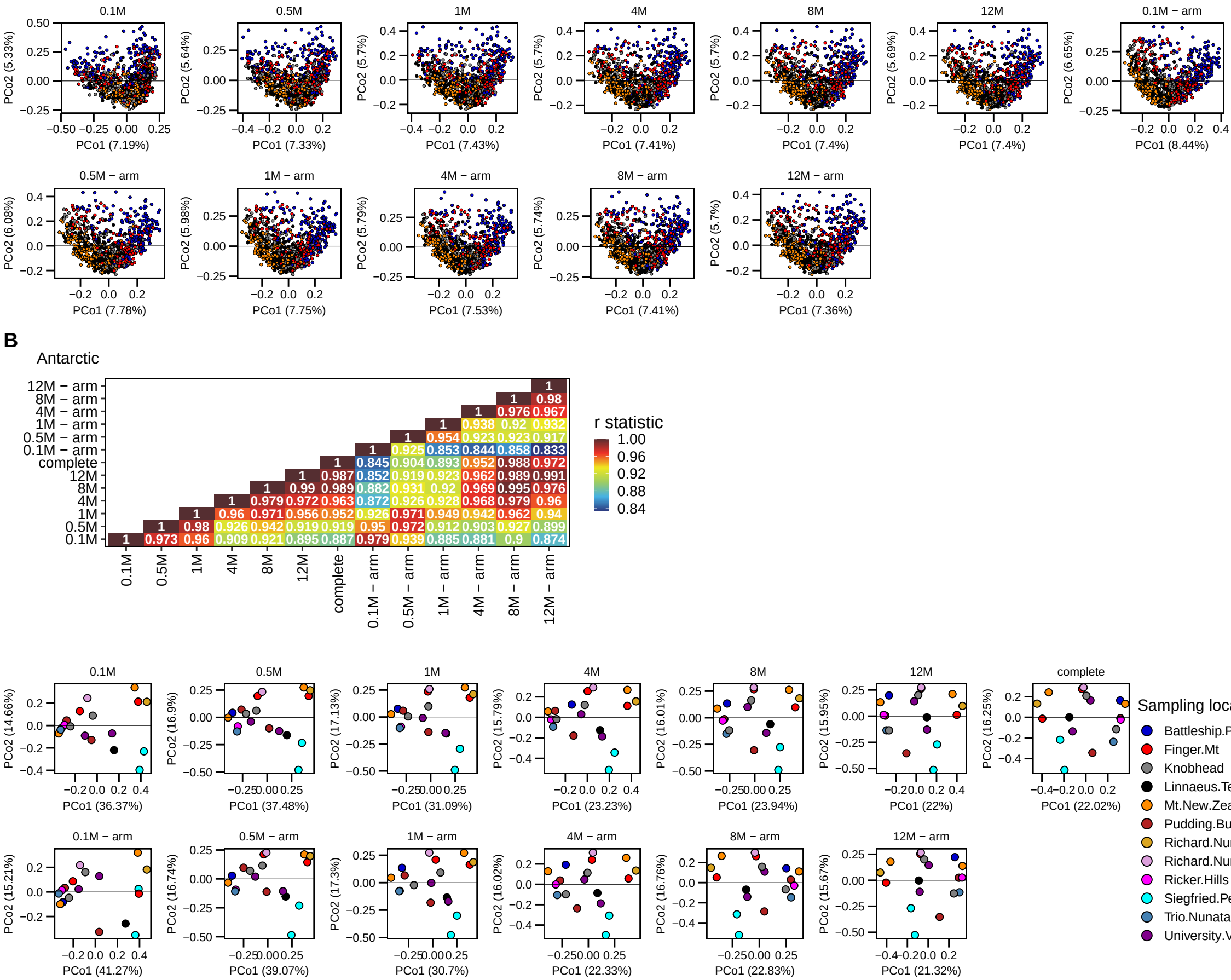


Figure 2

A

Human gut

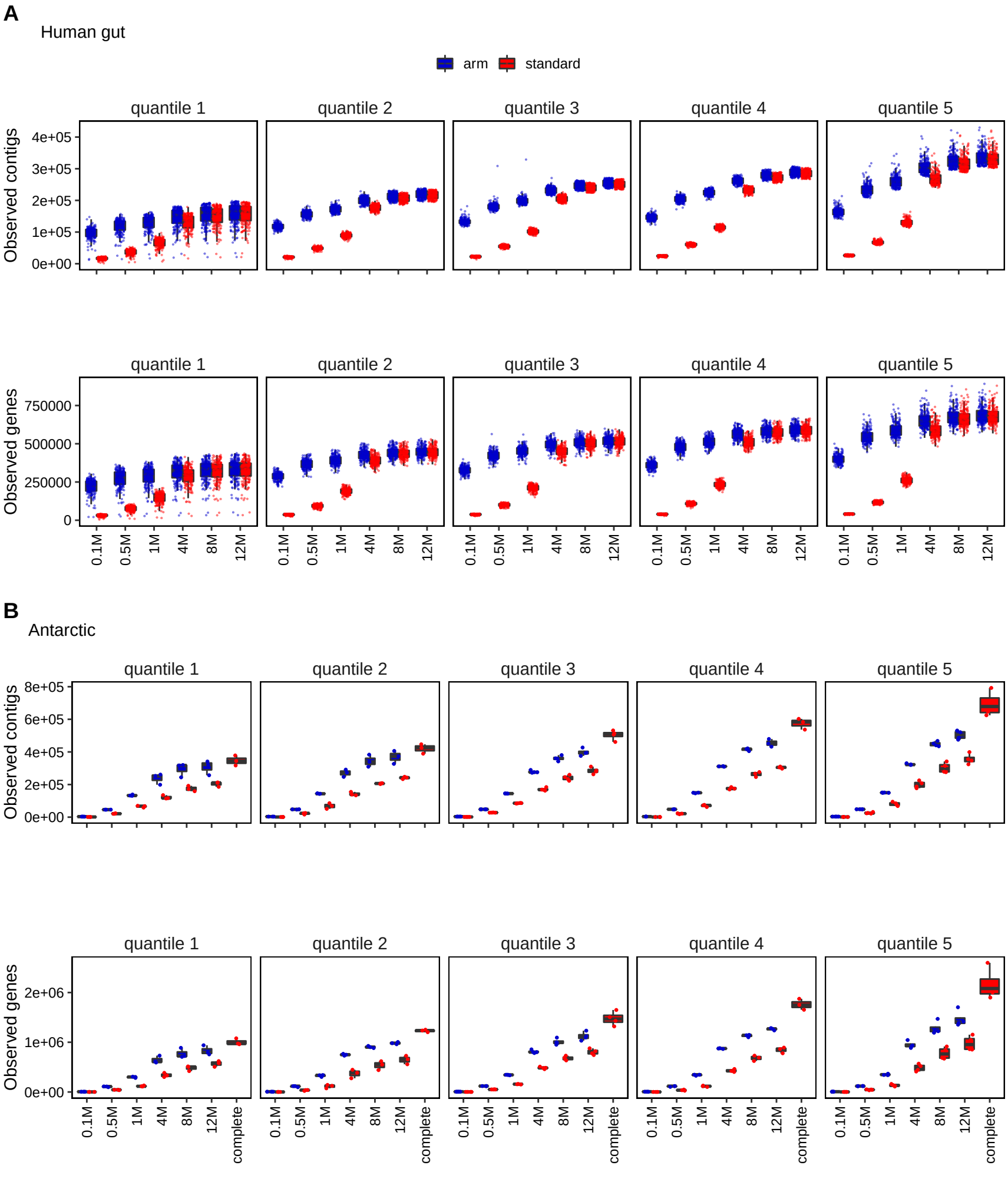




B

Antarctic

Figure 3



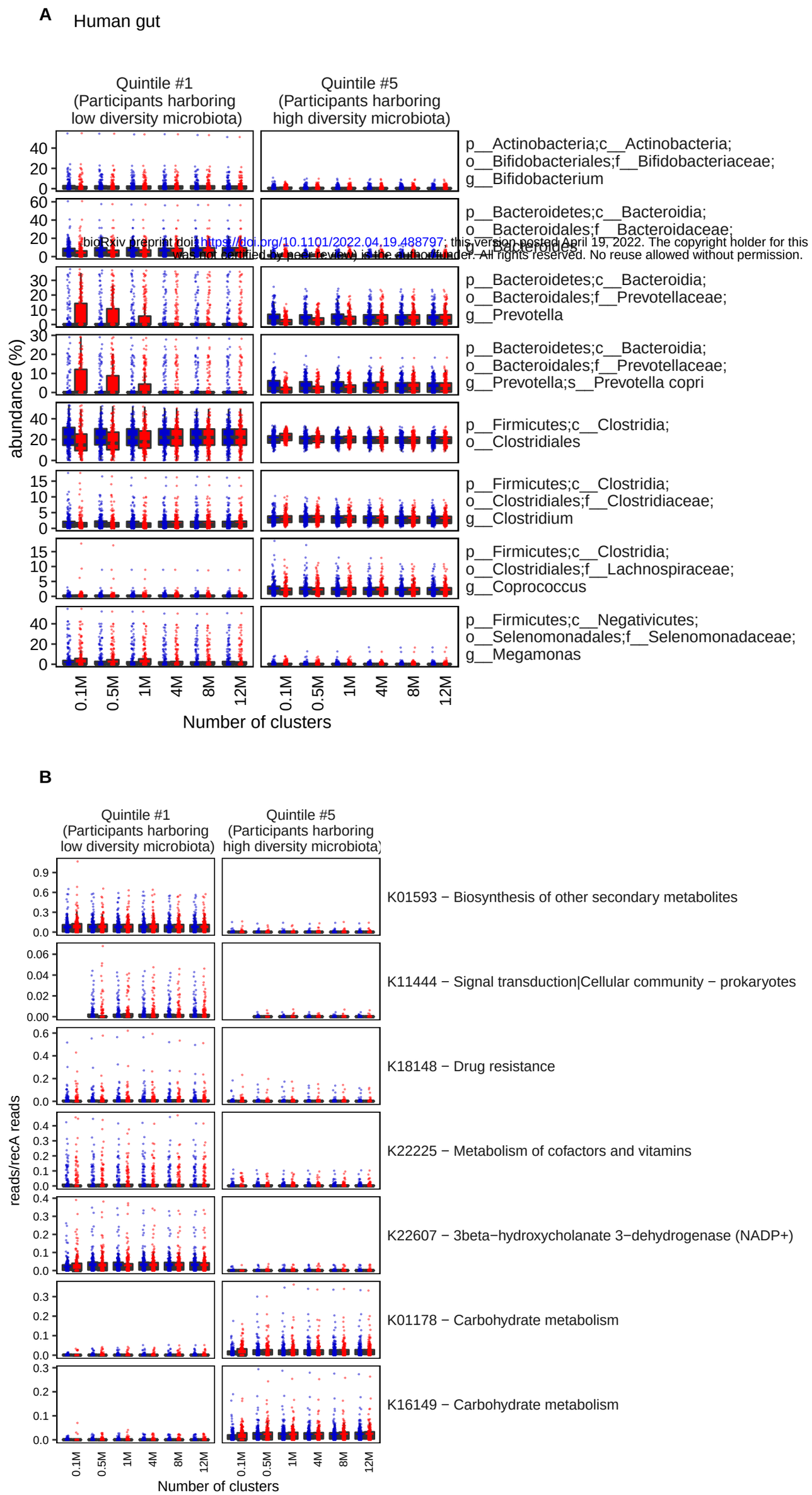
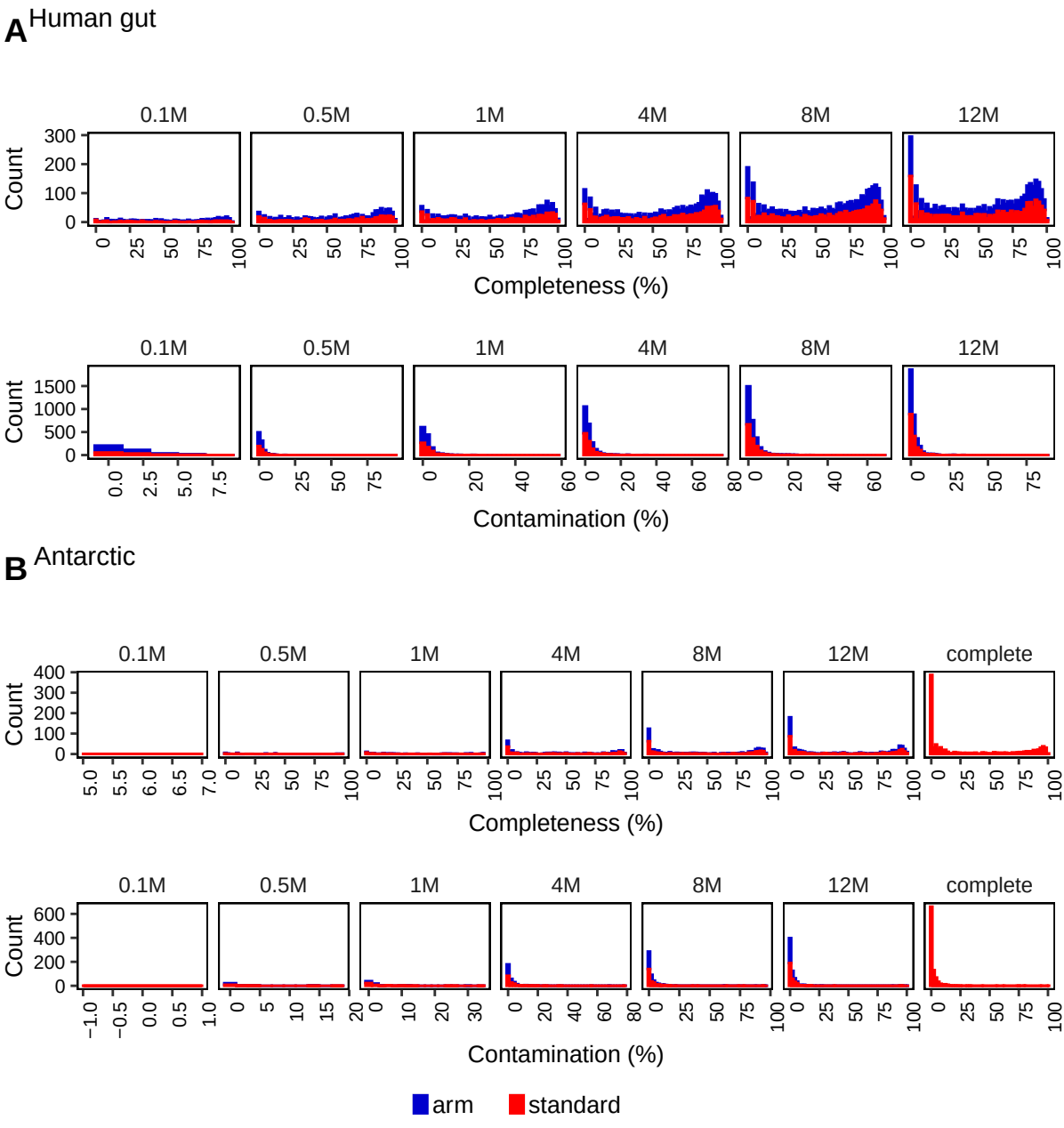


Figure 5


A
Human gut
0.1M
0.5M
1M
4M
8M
12M
Count
Completeness (%)
Contamination (%)
B
Antarctic
complete
arm
standard