Article

hgtseq: A Standard Pipeline to Study Horizontal Gene Transfer

Simone Carpanzano 1, Mariangela Santorsola 1, nf-core community ‡ and Francesco Lescai *

Department of Biology and Biotechnology, University of Pavia, 27100 Pavia, Italy
* Correspondence: francesco.lescai@unipv.it
‡ nf-core community members are available at acknowledgments.

Abstract: Horizontal gene transfer (HGT) is well described in prokaryotes: it plays a crucial role in evolution, and has functional consequences in insects and plants. However, less is known about HGT in humans. Studies have reported bacterial integrations in cancer patients, and microbial sequences have been detected in data from well-known human sequencing projects. Few of the existing tools for investigating HGT are highly automated. Thanks to the adoption of Nextflow for life sciences workflows, and to the standards and best practices curated by communities such as nf-core, fully automated, portable, and scalable pipelines can now be developed. Here we present nf-core/hgtseq to facilitate the analysis of HGT from sequencing data in different organisms. We showcase its performance by analysing six exome datasets from five mammals. Hgtseq can be run seamlessly in any computing environment and accepts data generated by existing exome and whole-genome sequencing projects; this will enable researchers to expand their analyses into this area. Fundamental questions are still open about the mechanisms and the extent or role of horizontal gene transfer: by releasing hgtseq we provide a standardised tool which will enable a systematic investigation of this phenomenon, thus paving the way for a better understanding of HGT.

Keywords: horizontal gene transfer; exome; sequencing; microbiome; bioinformatic pipelines

1. Introduction

While whole-genome sequencing (WGS) is becoming the sequencing strategy of choice thanks to the progressively lower prices, whole exome sequencing (WES) remains one of the leading technologies for studying coding regions of large genomes, and for clinical applications in humans. Most projects are usually carried out by aligning the raw reads to the host genome, while unmapped reads are discarded: the discovery by Samuels and colleagues [1] that unmapped reads in human sequencing data contain a certain number of microbial reads, defined by the authors as “a lost treasure”, changed somehow the perspective, and raised the question whether horizontal gene transfer (HGT) occurs in humans.

Horizontal gene transfer refers to the non-vertical transmission of genetic material between different organisms, allowing for the acquisition of novel traits, including antibiotic resistance, capacity to rapidly adapt to new environments or the use of new food sources [2]. HGT is relevant to the variation of both prokaryotic and eukaryotic genomes [3], resulting in some evolutionary trajectories which can be best described by reticulated networks rather than a typical phylogenetic tree [4–7].

HGT is a crucial evolutionary force in archaea and bacteria, enabling such organisms to acquire new genetic functions via transduction, conjugation and direct transformation by exogenous DNA [8,9]. The proportion of genes with signatures of HGT into eukaryotic nuclear genomes is usually considered lower than that in prokaryotes [3]. Recently, a work by Li et al. [10] (not peer yet reviewed) suggests that putative HGTs are also widespread, however, in eukaryotes. Different and more complex mechanisms underlie HGT events in eukaryotes due to the presence of a nucleus, the separation of somatic and germine tissues, and genome surveillance mechanisms which prevent or minimise the HGT material
integration. The evolutionary and functional significance of the prokaryotes-to-eukaryotes HGT events continues to be debated [11] and the molecular mechanisms driving these events are not entirely clarified. However, the transfer of material from chloroplasts and mitochondria to their eukaryotic hosts [12] is known to exist as “nuclear mitochondrial transfers” (NUMTs) or “nuclear plastid transfers” (NUPTs). The proposed molecular events underlying these phenomena [13,14] might offer some understanding of HGTs in higher organisms. Wolbachia, an obligated endosymbiotic bacterium living in the germline of its animal, insect and nematode hosts is another example of HGT between prokaryote endosymbionts and their eukaryotic hosts. The proximity of this endosymbiont to the germ cells of its hosts makes it an ideal donor for heritable HGT to eukaryotic genomes. Several hypotheses have been proposed for the proximity-driven transfer of material, but robust evidence is still missing. Moreover, the constant exposure of multicellular organisms to microbial species (parasites, endosymbionts, or the microbiome) challenges the capacity to distinguish between true HGTs, detectable in genome sequencing data from the host, and contaminants acquired during sample collection in large sequencing projects. Hongseok Tae et al. [15] analysed 150 datasets from the 1000 Genomes Project, and their results indicated that each sequencing centre had specific contaminating signatures defined as “time stamps”. More importantly, DNA extraction kits have been proven to play an important role in affecting the results of HGT analyses, since some DNA extraction methods appear to be prone to pull down specific DNA sequences from selected microbial genera. This issue has been reported to especially impact the discovery of low abundance microbial species [16]. A confirmation of this finding has been provided by Salter and colleagues [17], who conducted an in-depth study on the presence of contaminants related to ordinary extraction kits, after detecting sequences of microbial species in negative controls during PCR. They provided evidence of a specific bacterial signature associated with each of them. Taking advantage of this information, we curated a list of genera to be flagged as potential contaminants and therefore to be handled carefully or filtered out entirely during the analysis.

These large sequencing studies are a key focus of interest, because they provide a window into potential HGT events in humans. Here, horizontal gene transfer would take place mostly in somatic cells: in this case the newly acquired sequences are not transmitted to the next generation but could, however, induce mutations contributing to cancer or autoimmune diseases [18,19]. Riley and colleagues [20] reported the integration of bacterial sequences into samples of human acute myeloid leukaemia as well as into human adenocarcinoma samples from the stomach. Cancer is known to be characterised by genomics instability, which might suggest tumour cells to be more permissive to HGT events: several studies have been carried out to confirm this phenomenon [21]. The indication, however, that other sequencing studies, carried out in control individuals, display a similar presence of microbial-derived sequences in unmapped reads suggests that HGT events might be more widespread than initially thought.

The measure of rates and patterns of HGT loci relies on the methods used to identify HGT candidates in genomes and transcriptomes [22]. Parametric methods, including gmos [23], GeneMates [24], ShadowCaster [25] and nearHGT [26], are aimed at finding genetic loci distinguished from their new host genome either by GC content, oligonucleotide spectrum, DNA structure modelling, or based on their genomic context. Explicit and implicit phylogenetic methods compare tree topologies and analyse distances between genomes [27], respectively. Examples of these are ShadowCaster and HGT-Finder [28]. These methods have developed rapidly in the last years, fuelled by the availability of large datasets of complete eukaryote genomes from next-generation sequencing (NGS). However, many of these either lack implementation in standard automated workflows or cannot carry out the analyses from the raw data up to all the downstream HGT steps. In many cases they do not scale with increasingly large datasets or cannot be easily implemented in any environment [29]. To date, few highly automated pipelines for revealing HGT events are available. These include the Daisy pipeline by Trappe et al. [30] and its updated version.
DaisyGPS [31] for the detection of HGTs directly from NGS reads. These are mapping-based tools and require sequencing data from the HGT host organism and reference sequences of both host and donor genomes. Baheti et al. [32] developed the HGT-ID workflow, which detects the viral integration sites in the human genome. This workflow is fully automated from the raw reads to the viral integration site detection and downstream visualisation. It is, however, limited to the analysis of viral HGT. Other existing methods for identifying viral integration sites are VirusSeq software [33], ViralFusionSeq [34], VirusFinder [35], and VirusFinder2 [36], but these are also limited to the investigation of viral integration events. Finally, methods for identification of non-human sequences in both genomic and transcriptomic libraries of human tissues include PathSeq [37] and its updated implementation GATK PathSeq [38].

Despite the availability of these solutions, we believe there is still a need for a user-friendly automated pipeline, based on a few fundamental features: accessibility, reproducibility, portability, and scalability [39]. Accessibility is a fundamental value, to ensure the widest range of researchers can use the pipeline, in an easy and understandable way. Reproducibility, which is at the basis of the scientific method, requires a version-controlled and documented approach to coding, which ensures the same results are always produced with the same inputs. Portability and scalability are now increasingly important, and can be translated into the capacity to execute the pipeline in any computing environment without major changes to the code, and to scale well with the size of the input data.

The introduction of Nextflow [40], an increasingly popular workflow framework, allows the implementation of the above-mentioned values. Nextflow is now largely adopted within research and industry contexts in life science, because of its features and flexibility. This domain-specific language relies heavily on container engines, in particular Docker and Singularity, which ensure all software dependencies are resolved regardless of the environment where the pipeline is executed; this also means that the code can be transported anywhere seamlessly. Most importantly, one community of Nextflow users, called nf-core [41], offered the opportunity to converge around specific standards both in terms of coding style and conventions, as well as best practices for the use of bioinformatic tools widely adopted in life sciences and beyond. There has been an investment in the modularity of the pipeline’s composition, which allows code-sharing among those implementing the same tools in different workflows, thanks to the adherence to these best practices. Similarly, Biocontainers are used [42] in most cases. This ensures an even wider standardisation of containers, their design and version-control.

In this context, and in line with these values and standards, we hereby present hgtseq, a fully automated pipeline for the detection and analysis of microbial sequences in unmapped reads from host organisms, and we investigate potential evidence of horizontal gene transfer events.

2. Results

The hgtseq pipeline has been developed using 21 modules from the shared nf-core repository, and four “local” modules, as described in Materials and Methods. A total of 329 commits have been made to the pipeline repository, excluding contributions to shared repositories, in order to compose the pipeline. The release 1.0.0 (https://doi.org/10.5281/zenodo.7244735 accessed on 24 October 2022) accepts either a FASTQ with raw paired-end reads from Illumina sequencing as input, or an already aligned paired-end BAM file (Figure 1).

Raw reads are first trimmed for quality and to remove Illumina adapters: the resulting high-quality reads are aligned to the host genome, which is defined as the reference by its identifier in the iGenomes repository (https://support.illumina.com/sequencing/sequencing_software/igenome.html accessed on 24 October 2022) for seamless download, and via NCBI taxonomic identifier. Pre-aligned BAM files are then processed in parallel to extract two categories of reads, via their SAM bitwise flags. With bitwise flag 13, we extract reads classified as paired, which are unmapped and whose mate is also unmapped
(i.e., both mates unmapped). With bitwise flag 5 we extract reads classified as paired, which are unmapped but whose mate is mapped (i.e., only one mate unmapped in a pair). In both cases we use flag 256 to exclude non-primary alignments. Both categories are classified using Kraken2, as described in Methods. The second category, i.e., unmapped reads whose mate is mapped, provides the opportunity to infer the potential genomic location of an integration event, if confirmed. To this end, the pipeline uses information available for the properly mapped mate in the pair: for this category of reads, the pipeline parses the genomic coordinates of the mate from the BAM file and merges them with the unmapped reads classified by Kraken2.

Finally, host-classified reads are filtered out and the data are used to generate Krona plots and an HTML report with RMarkdown, as described in Methods (Figure 1). A full set of QC metrics is available for all steps of the pipeline, and an additional QC report is compiled using MultiQC.

The infrastructure provided by the nf-core community allows the user to access the most important information about the usage and the configuration parameters via a website (https://nf-co.re/hgtseq accessed on 24 October 2022): these pages are automatically updated from the github repository of the pipeline, and render the markdown documentation together with a JSON schema of the pipeline parameters. This website also offers a “launch” button, prominently displayed at the top of the landing page: this feature allows a user to configure all parameters on the page, and to generate a simple JSON file to be passed at the command line, thus simplifying the execution at the terminal. Alternatively, the launch feature allows the user to access the Nextflow Tower launchpad (https://cloud.tower.nf accessed on 24 October 2022) and manage the execution on the cloud through a graphical interface.

We have tested the pipeline on six different datasets, downloaded as described in Materials and Methods: we have verified that the pipeline can be run seamlessly in either an on-premise HPC (slurm scheduler), or on the cloud (Amazon AWS). In our system, the execution of all steps on a human exome (COPD dataset) completed in just about 3 h (Figure 2A), generating a total number of 150 jobs (Figure 2B). This corresponds to a total of 154.3 CPU/hours and a total memory usage of 1136.87 GB. The dataset generated a total I/O of 3742.28 GB read and 2389.91 GB written. Previously developed tools reported runtimes ranging between 4.3 (HGT-ID) to 23.4 h (VirusFinder2): However, a direct comparison is difficult, given the differences in methods, computing environments and datasets used.
The hgtseq pipeline includes a few well-known compute intensive tasks, such as BWA-MEM for the alignment, Kraken2 for taxonomic classification and Qualimap for the QC of the BAM files, as described in the Methods section. BWA showed the ability to handle well multi-threading in different datasets (four of them are shown in Figure 2C), while Kraken2 mostly uses one core during the classification process. In terms of memory usage, as expected, both Kraken2 and Qualimap use the most amount of memory among all processes in this pipeline, reaching 60 GB of memory (Figure 2D), particularly in datasets from human and dog. The memory usage of Qualimap is shown to be related to the size of the BAM files, whereas Kraken2 memory usage shows dependency on the chosen database for the analysis.

The execution on AWS, using the AWS-batch API which Nextflow supports, allowed confirmation of the portability and reproducibility of the code.

The samples included in the six datasets that we analysed for testing ranged between 49,516,660 in the Bos taurus cohort to 198,822,100 in the Canis lupus familiaris dataset (Table S1). As described above, we analysed two categories of unmapped reads separately (i.e., depending on whether both mates or only one mate of the pair was unmapped): human datasets showed a proportion of unmapped reads classified to microbial genera, compared to the total number of reads in the sample, ranging between 1.04 × 10^{-5} (human CHIKV cohort, both mates unmapped) to 2.44 × 10^{-6} (human COPD dataset, both mates unmapped—Figure 3 and Table S1). Similar ranges were observed in the other species: the proportion in Bos taurus ranged between 8.5 × 10^{-5} (both mates unmapped)
to $8.67 \times 10^{-5}$ (only one mate unmapped); in Canis lupus familiaris it was $3.42 \times 10^{-8}$ for both mates unmapped reads and $7.32 \times 10^{-7}$ for unmapped reads having their mate mapped. The proportion in Macaca fascicularis was $1.16 \times 10^{-5}$ for both mates unmapped and $2.14 \times 10^{-5}$ for only one mate unmapped reads, while in Mus musculus we observed a proportion of $4.91 \times 10^{-6}$ for both mates unmapped and $6.21 \times 10^{-6}$ for only one mate unmapped. An example list of genera potentially involved in HGT identified in the COPD dataset has been provided within Supplementary Materials (Table S2).

![Krona plots showing proportion of reads assigned to microbial genera in different datasets](image)

**Figure 3.** Proportion of reads assigned to microbial genera in different datasets, compared to the total number of reads in the sample. The figure illustrates the proportion of unmapped reads assigned to microbial genera in each of the test datasets. The reads are categorised by type, with “both” indicating those where both mates in the pair are unmapped, and “single” those where only one mate in the pair is unmapped.

We then set out to understand the proportion of potential contaminants in these results, by assigning a flag to the classified reads, depending on whether the genus of their assignment belonged to the list of genera we compiled, as described in Materials and Methods. We assigned the category “flagged” if the genus was included in the list, but belonged to genera which might include species known to be associated with humans. We assigned the category “blacklisted” if the genus belonged to any other genera in the list (Table S3). We also stratified this analysis using the classification score we calculated in the final step of the pipeline analysis. This analysis (Figure S1) shows that, in most of the datasets, the proportion of classified reads belonging to potentially contaminant genera is higher among those reads with a lower classification score (Mus musculus, Macaca fascicularis, human in the CHIKV dataset, Canis lupus familiaris). This cannot, however, be observed in all cases, as the human COPD dataset shows a higher proportion of flagged genera in reads with higher classification scores, and the results in the Bos taurus dataset show an unexpectedly high proportion of flagged reads overall.

In order to explore these results, we then analysed in particular the classified reads belonging to the category single-unmapped (i.e., those unmapped reads whose mate in the pair is mapped). The vast majority of the reads have been assigned to bacterial organisms, and we therefore compared the krona plots generated on those reads from each of the
different datasets we processed, after filtering out the blacklisted genera (Figure S2). A qualitative inspection of the plots shows clear differences between the datasets of Bos taurus and Mus musculus, while a number of similarities can be observed in primates (human COPD and CHIKV, and Macaca fascicularis).

In light of the release of a telomere-to-telomere sequence of the human reference, we finally analysed the COPD human dataset by using the T2T-CHM13 reference [43] and compared the results with those obtained by using the GRCh38 reference (Table S4). We observed a reduction in the total number of unmapped reads (−1.13% for the category single-unmapped and −0.91% for the category both-unmapped). A small reduction in the amount of microbial classified reads can be recorded, particularly among the single-unmapped category (−1.46%). The impact of the T2T-CHM13 reference that we could measure is likely limited by the differences between the design of the exome capture in the COPD project and by the nature of the additional sequences included in the new reference.

3. Discussion

Since the first identification of a significant presence of non-human sequences in human data [1], and a large-scale follow-up in different sequencing data [15], this intriguing topic has been poorly investigated. Horizontal gene transfer events between bacteria and eukaryotes have been described, and several molecular mechanisms controlling these events have been proposed [2]: while those phenomena in plants and invertebrates are documented in detail, much less is known about horizontal gene transfer events in mammals, which might explain the initial findings of Samuels and colleagues. The available data support the existence of horizontal gene transfer events in mammals, including in humans, and call for a renewed effort in investigating this phenomenon in a systematic way, exploiting the wealth of sequencing data already available in a variety of species and conditions. The main obstacle to this is therefore represented by the availability of tools that enable these analyses at scale, with any type of data and in the widest variety of computing environments available to researchers. Here, we documented our work towards releasing a standard pipeline for investigating the presence of microbial DNA in mammals, characterised by ease of use, portability and scalability.

The first critical choice we made in order to achieve this goal was to develop the code using Nextflow. The nature of this domain-specific language, agnostic to the code or the software being used, makes it both flexible and capable of accommodating a wide range of choices in terms of tools for the pipeline. Nextflow supports most schedulers for on-premise high-performance computing clusters, and all cloud infrastructure vendors: a pipeline developed with this workflow engine can therefore be run virtually anywhere, with just little changes to the configuration files. The second fundamental choice we made was to develop the pipeline within the framework of nf-core: besides the standards and best practices, the community also developed an infrastructure which allows the code to be continuously tested, reviewed by a large number of people, and shared among different pipelines. Common choices are usually discussed among scientists from different disciplines and institutions. Additionally, a strong focus on documentation and on the availability of a website to configure pipeline parameters and launch the pipelines on the cloud, makes the nf-core infrastructure a best-in-class choice for life science applications to be used at scale. Hgtseq benefits from this crucial environment and facilities, and therefore has the potential to push this area of research forward. The pipeline accepts both FASTQ and BAM/CRAM files as input: this was a strategic choice, meant to allow the execution of hgtseq downstream to a wide range of routine pipelines, particularly in human studies. In this way, we aimed to target research groups who primarily study human or other host genomes, but who do not look for non-host reads in their data. With this in mind, we hope to offer additional opportunities to tap into the wealth of data available, which can be used to answer the fundamental questions still open. Additionally, the modularity of the pipeline has been designed to provide further opportunities to include selected features
from solutions implemented in other valuable tools, thus expanding its capabilities in further releases.

In order to demonstrate the functionality of this pipeline, we chose a wide array of species to run the analyses on. In addition to the two different human datasets, we decided to use other mammals whose genome is well curated, such as Bos taurus, Mus musculus and Canis lupus familiaris. We also chose Macaca fascicularis, in order to provide an additional insight of the pipeline performance on sequencing data from another primate. For all of these species, we chose to analyse exome data: this was meant to mirror the initial analyses by Samuels and colleagues [1], but also to test the capability of hgtseq to detect this phenomenon within the coding regions of other mammalian species.

In terms of performance, we have been able to demonstrate that the pipeline can handle a wide variety of sequencing data, and can process datasets with more than 100 million reads per sample in about 3 h. The speed of the analysis will naturally depend on the computing environment where it is run, and we provide here an indication on the resources needed in terms of CPUs and memory by the most resource-consuming tools in this pipeline, such as BWA, Kraken2 and Qualimap. Users will be able to plan their analyses based on this information. Changes to the configuration can be easily introduced by creating new config files for the infrastructure to be used at the time of analysis.

The analysis we present here has the limited scope of testing the computational performance of the pipeline that we introduce, and of showing its basic functionality. Within the limited size of the datasets we used, we have been able to confirm that the presence of microbial sequences in exome data is not limited to human exomes, but exists to a similar extent also in other mammalian species. In particular, in this work we are able to confirm the findings of Tae et al. [15] in terms of abundance of microbial species in this kind of data. However, when evaluating these findings, it is quite crucial to assess the extent of a bias in measuring the abundance of unmapped reads classified to microbial taxonomies. A source of bias could be a potential contamination of the sequencing data. We have provided two ways to control for this within the pipeline: firstly, we have curated a list of genera which have been reported over the years to be known contaminants of DNA extraction kits. This list, accompanied by a review of the literature used as a source, can be embedded into the analyses and can be used to filter the results. For analysing human data, we further categorised this list by highlighting those microbial genera which might include species known to be associated with a human host: the user can therefore decide whether those results should be filtered out, or marked for a closer inspection. Additionally, we coded an evaluation score into the results report in order to further stratify the strength of the Kraken2 taxonomic assignment of each read: this score provides a second tool to filter the results and to fine tune the stringency of the analysis. In our tests, we have shown that by using this score we can control the abundance of genera flagged for contamination. We have also highlighted the extent of variability in sequencing quality among different datasets.

In conclusion, we believe there is an urgent need to systematically investigate the extent of horizontal gene transfer phenomena in mammals. There are still several open questions to be answered, about the molecular mechanisms controlling these events, or about the extent to which mammalian genomes are affected by them, and what the source of the microbial material might be. While a much larger analysis is needed to address these questions, we are offering the scientific community a standardised, portable and scalable tool to carry out a systematic investigation of these events in a wide range of hosts, thus paving the way for new potential breakthroughs in this area of research.

4. Materials and Methods

4.1. Datasets

Six exome datasets from different species were chosen to test the pipeline in a variety of conditions. Two datasets from human exomes were downloaded: 9 samples from a Chinese population affected by chronic obstructive pulmonary disease (COPD), available from SRA (BioProject accession: PRJNA785331); and 12 samples from Brazilian individuals
with Chikungunya virus (CHIKV) infection, available from SRA (BioProject accession: PRJNA680041). Additionally, we used 10 samples from a population of whole exome sequences from 69 male founder mice (BioProject accession: PRJNA603006), 12 samples from a dataset of Macaca fascicularis exomes (BioProject accession: PRJNA685531), 10 samples from an exome dataset of Bos taurus (BioProject accession: PRJNA479823) and 10 samples of Poodle breed from a project of exome sequencing in three dog breeds (Canis lupus familiaris, BioProject accession: PRJNA385156).

4.2. Alignment and Read Extraction

When FASTQ files are used as input, the pipeline first uses TrimGalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ accessed on 24 October 2022) to trim the reads by quality and to remove any adapters (default options, plus ‘–illumina’). Trim Galore combines both Cutadapt (https://github.com/marcelm/cutadapt/ accessed on 24 October 2022) and Fastqc (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ accessed on 24 October 2022) to perform both types of trimming in the same step. These processed reads are then mapped to the host reference genome using BWA [44], with default parameters. These steps are automatically skipped if the user provides pre-aligned BAM files as input. According to standard practice, the aligned BAM files are then sorted by chromosome coordinates and indexed.

Using ‘samtools view’ [45], we extract two different types of reads pairs identified by their SAM bitwise flag: each read type is then processed separately, according to their different properties. Specifically, to extract the first category of data we apply the command ‘samtools view-b-f 13-F 256’ to select those reads with bitwise flag 13 and, therefore, extract those in a pair where both mates do not map to the host reference (flag properties: read paired, read unmapped, mate unmapped); we define them as both_unmapped. To extract the second category of data, we apply the command ‘samtools view-b-f 5-F 256’ to select reads with a bitwise flag 5, i.e., to extract those unmapped reads in a pair whose corresponding mate is mapped (flag properties: read paired, read unmapped); we define them as single_unmapped. As can be seen in these commands, by applying the option ‘–F’ in both cases we use the bitwise flag 256 to exclude any non-primary alignments.

A local module named parseoutputs, which contains two combined commands ‘samtools view | cut -f 1,3,8’, allows the extraction from the single_unmapped reads of mapping information of their mapped mate: sample name, chromosome and exact position of the mate. These data are merged with the taxonomic classification of the unmapped reads in the RMarkDown report, for further analyses of potential integration sites.

4.3. Reads Classification

In order to perform the taxonomic classification of the unmapped reads extracted as described above, the pipeline uses Kraken2 [46], a k-mers-based algorithm that provides a high-accuracy classification by matching each k-mer to the lowest common ancestor. Reads are first converted into a FASTQ format to accommodate Kraken2 input requirements.

We built a custom database with all available libraries in Kraken2 (archea, bacteria, plasmid, viral, fungi, plant, protozoa, human) by using the command ‘kraken2-build–download-library’. Additionally, we downloaded from ENSEMBL the genome of other host species to be used in this analysis, such as Canis familiaris, Bos taurus, Macaca fascicularis and Mus musculus. We used an in-house Python script to reformat the FASTA header of the genome sequences, according to the Kraken2 specifications. Finally, we added the sequences to the Kraken2 library with the command ‘kraken2-build–add-to-library’, and built the database. We then ran Kraken2 on the two categories of unmapped reads with default settings, and collected the default output (report and reads assignment).

4.4. Reporting

The QC report is generated by MultiQC [47], by aggregating data from multiple sources within the pipeline, and in particular FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
The analysis report is generated via a specific process in the pipeline, which aggregates results from all samples and executes a parameterised RMarkdown [49] to create tables and plots.

In order to provide a measure of the evidence that each read classification is based upon, we have added a classification score to the analysis that generates the report on the datasets. The R code parses the k-mer information from the Kraken2 output, and counts: (a) the total number of k-mers analysed in each read (tot_kmers), (b) the sum of all k-mers supporting the classified taxonomic ID, which is chosen by Kraken2 as the one with the highest number of supporting k-mers (max_kmers), (c) the total number of classified k-mers in the read, i.e., with taxonomic ID not zero (sum_class_kmers). A classification score (taxid_score) is then provided as the ratio max_kmers/tot_kmers; an additional score is also provided as the ratio max_kmers/sum_class_kmers, i.e., indicating the proportion of the assigned taxonomic ID over other classified taxonomic IDs.

Each genera is flagged as a potential contaminant, depending on its membership of a list compiled according to literature [15–17]: membership is reported either as flagged, if the genus includes species potentially associated with humans, or blacklisted otherwise (Table S3).

Besides this RMarkDown report, the pipeline generates two multi-layered interactive pie charts made with Kronatools [50]. For convenience, we wrote a local module that groups together all samples to make only one chart from each of the two read categories. Using the command ‘ktImportTaxonomy-t 3’ this tool parses only the taxa from each read.

### 4.5. Pipeline Composition

In order to compose the pipeline, we have used Version 2 of the domain-specific language (DSL2) Nextflow [40]. This version of the workflow language allows the combination of multiple independent scripts (defined as modules), each defining functions, processes and/or workflows. The nf-core community has released standard guidelines [41], identifying modules as atomic scripts, where one process uses one single tool or tool function. This approach makes it possible to share Nextflow modules within the community across multiple users and pipelines. In the process of designing the pipeline, the modules for general use have therefore been contributed to a shared nf-core repository first (krana, kraken2, samtools utilities, bamtools stats), and they have subsequently been included in the hgtseq repository. Those scripts specifically designed to carry out tasks within hgtseq have been committed to the hgtseq repository as “local” modules. Each module uses standardised containers, curated by the community Biocontainers (https://biocontainers.pro accessed on 24 October 2022).

We have used the pipeline template developed by nf-core and the nf-core tools (https://nf-core/tools/ accessed on 24 October 2022), released to facilitate pipeline composition according to community standards. These include continuous integration (CI) tests run using the tool pytest, which are triggered by github actions at commits push or pull requests. An extensive array of CI tests is being run, using three containers/package managers (docker, singularity, conda) on the two key use cases of the hgtseq pipeline, i.e., with FASTQ input, and with BAM input.

Additionally, the design includes a full-size test run at pipeline release on AWS and uses the human COPD samples (BioProject accession: PRJNA785331) as input.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms232314512/s1.
Author Contributions: Conceptualisation and methodology: F.L. and S.C.; Software and validation: S.C., F.L. and the nf-core community wrote and reviewed the code; Formal analysis: F.L. and S.C.; Writing, original draft, reviewing and editing: S.C., F.L. and M.S.; Resources: the nf-core community provided the infrastructure to facilitate access to documentation and graphical user interfaces. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the institutional funding “FRG-Fondo Ricerca & Giovani”, provided by the University of Pavia via the Department of Biology and Biotechnology.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code is available at https://github.com/nf-core/hgtseq. The release 1.0.0 of the pipeline can also be referenced with https://doi:10.5281/zenodo.7244735.

Acknowledgments: A full list of members is available at https://nf-co.re/community.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Samuels, D.C.; Han, L.; Li, J.; Quanghu, S.; Clark, T.A.; Shyr, Y.; Guo, Y. Finding the Lost Treasures in Exome Sequencing Data. Trends Genet. 2013, 29, 593–599. [CrossRef] [PubMed]
2. Husnik, F.; McCutcheon, J.P. Functional Horizontal Gene Transfer from Bacteria to Eukaryotes. Nat. Rev. Microbiol. 2018, 16, 67–79. [CrossRef] [PubMed]
3. Keeling, P.J.; Palmer, J.D. Horizontal Gene Transfer in Eukaryotic Evolution. Nat. Rev. Genet. 2008, 9, 605–618. [CrossRef] [PubMed]
4. Gabaldón, T. Patterns and Impacts of Nonvertical Evolution in Eukaryotes: A Paradigm Shift. Ann. N. Y. Acad. Sci. 2020, 1476, 78–92. [CrossRef] [PubMed]
5. Mallet, J.; Besansky, N.; Hahn, M.W. How Reticulated Are Species? Bioessays 2016, 38, 140–149. [CrossRef]
6. Soucy, S.M.; Huang, J.; Gogarten, J.P. Horizontal Gene Transfer: Building the Web of Life. Nat. Rev. Genet. 2015, 16, 472–482. [CrossRef]
7. Swithers, K.S.; Soucy, S.M.; Gogarten, J.P. The Role of Reticulate Evolution in Creating Innovation and Complexity. Int. J. Evol. Biol. 2012, 2012, 418964. [CrossRef]
8. Lacroix, B.; Citovsky, V. Beyond Agrobacterium-Mediated Transformation: Horizontal Gene Transfer from Bacteria to Eukaryotes. In Agrobacterium Biology; Gelvin, S.B., Ed.; Current Topics in Microbiology and Immunology; Springer International Publishing: Cham, Switzerland, 2018; Volume 418, pp. 443–462. ISBN 978-3-030-03256-2.
9. García-Vallvé, S.; Romeu, A.; Palau, J. Horizontal Gene Transfer in Bacterial and Archaeal Complete Genomes. Genome Res. 2000, 10, 1719–1725. [CrossRef] [PubMed]
10. Li, K.; Yan, F.; Duan, Z.; Adelson, D.L.; Wei, C. Widespread of Horizontal Gene Transfer Regions in Eukaryotes. bioRxiv 2022. [CrossRef]
11. Danchin, E.G.J. Lateral Gene Transfer in Eukaryotes: Tip of the Iceberg or of the Ice Cube? BMC Biol. 2016, 14, 101. [CrossRef]
12. Dunning Hotopp, J.C. Grafting or Pruning in the Animal Tree: Lateral Gene Transfer and Gene Loss? BMC Genom. 2018, 19, 470. [CrossRef] [PubMed]
13. Ku, C.; Nelson-Sathi, S.; Roettger, M.; Garg, S.; Hazkani-Covo, E.; Martin, W.F. Endosymbiotic Gene Transfer from Prokaryotic Pangenomes: Inherited Chimerism in Eukaryotes. Proc. Natl. Acad. Sci. USA 2015, 112, 10139–10146. [CrossRef] [PubMed]
14. Dunning Hotopp, J.C.; Clark, M.E.; Oliveira, D.C.S.G.; Foster, J.M.; Fischer, P.; Muñoz Torres, M.C.; Giebel, J.D.; Kumar, N.; Ishmael, N.; Wang, S.; et al. Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes. Science 2007, 317, 1753–1756. [CrossRef] [PubMed]
15. Tae, H.; Karunasena, E.; Bavarva, J.H.; McIver, L.J.; Garner, H.R. Large Scale Comparison of Non-Human Sequences in Human Sequencing Data. Genomics 2014, 104, 453–458. [CrossRef] [PubMed]
16. Laurence, M.; Hatzis, C.; Brash, D.E. Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes. PLoS ONE 2014, 9, e97896. [CrossRef] [PubMed]
17. Salter, S.J.; Cox, M.J.; Turek, E.M.; Calus, S.T.; Cookson, W.O.; Moffatt, M.F.; Turner, P.; Parkhill, J.; Loman, N.J.; Walker, A.W. Reagent and Laboratory Contamination Can Critically Impact Sequence-Based Microbiome Analyses. BMC Biol. 2014, 12, 87. [CrossRef] [PubMed]
18. Rubinstein, M.R.; Wang, X.; Liu, W.; Hao, Y.; Cai, G.; Han, Y.W. Fusobacterium Nucleatum Promotes Colorectal Carcinogenesis by Modulating E-Cadherin/β-Catenin Signaling via Its FadA Adhesin. Cell Host Microbe 2013, 14, 195–206. [CrossRef]
19. Robinson, K.M.; Sieber, K.B.; Hotopp, J.C. A Review of Bacteria-Animal Lateral Gene Transfer May Inform Our Understanding of Diseases like Cancer. PLoS Genet. 2013, 9, e1003877. [CrossRef] [PubMed]
20. Riley, D.R.; Sieber, K.B.; Robinson, K.M.; White, J.R.; Ganesan, A.; Nourbakhsh, S.; Dunning Hotopp, J.C. Bacteria-Human Somatic Cell Lateral Gene Transfer Is Enriched in Cancer Samples. PLoS Comput. Biol. 2013, 9, e1003107. [CrossRef] [PubMed]
21. Akimova, E.; Gassner, F.J.; Greil, R.; Zaborsky, N.; Geisberger, R. Detecting Bacterial–Human Lateral Gene Transfer in Chronic Lymphocytic Leukemia. *JMS* 2022, 23, 1094. [CrossRef]

22. Shikov, A.E.; Malovichko, Y.V.; Nizhnikov, A.A.; Antonets, K.S. Current Methods for Recombination Detection in Bacteria. *Int. J. Mol. Sci.* 2022, 23, 6257. [CrossRef] [PubMed]

23. Domazet-Lošo, M.; Domazet-Lošo, T. Gmos: Rapid Detection of Genome Mosaicism over Short Evolutionary Distances. *PLoS ONE* 2016, 11, e0166602. [CrossRef]

24. Wan, Y.; Wick, R.R.; Zobel, J.; Ingle, D.J.; Inouye, M.; Holt, K.E. GeneMates: An R Package for Detecting Horizontal Gene Co-transfer between Bacteria Using Gene-Gene Associations Controlled for Population Structure. *BMC Genom.* 2020, 21, 658. [CrossRef] [PubMed]

25. Sánchez-Soto, D.; Agüero-Chapin, G.; Armijos-Jaramillo, V.; Perez-Castillo, Y.; Tejera, E.; Antunes, A.; Sánchez-Rodríguez, A. ShadowCaster: Compositional Methods under the Shadow of Phylogenetic Models to Detect Horizontal Gene Transfers in Prokaryotes. *Genes* 2020, 11, E756. [CrossRef] [PubMed]

26. Adato, O.; Ninyo, N.; Gophna, U.; Snir, S. Detecting Horizontal Gene Transfer between Closely Related Taxa. *PLoS Comput. Biol.* 2015, 11, e1004095. [CrossRef] [PubMed]

27. Ravenhall, M.; Skunca, N.; Lassalle, F.; Dessimoz, C. Inferring Horizontal Gene Transfer. *PLoS Comput. Biol.* 2015, 11, e1004095. [CrossRef] [PubMed]

28. Nguyen, M.; Ekstrom, A.; Li, X.; Yin, Y. HGT-Finder: A New Tool for Horizontal Gene Transfer Finding and Application to Aspergillus Genomes. *Toxins* 2015, 7, 4035–4053. [CrossRef]

29. de Carvalho, M.O.; Loreto, E.S. Methods for Detection of Horizontal Transfer of Transposable Elements in Complete Genomes. *Genet. Mol. Biol.* 2012, 35, 1078–1084. [CrossRef]

30. Trappe, K.; Marschall, T.; Renard, B.Y. Detecting Horizontal Gene Transfer by Mapping Sequencing Reads across Species Boundaries. *Bioinformatics* 2016, 32, 595–604. [CrossRef]

31. Seiler, E.; Trappe, K.; Renard, B.Y. Where Did You Come from, Where Did You Go: Refining Metagenomic Analysis Tools for Horizontal Gene Transfer Characterisation. *PLoS Comput. Biol.* 2019, 15, e1007208. [CrossRef]

32. Bahetti, S.; Tang, X.; O’Brien, D.R.; Chia, N.; Nelson, H.; Boughey, J.C.; Wang, L.; Goetz, M.P.; Kocher, J.-P.A.; et al. HGT-ID: An Efficient and Sensitive Workflow to Detect Human-Viral Insertion Sites Using next-Generation Sequencing Data. *BMC Bioinf.* 2018, 19, 271. [CrossRef]

33. Chen, Y.; Yao, H.; Thompson, E.J.; Tannir, N.M.; Weinstein, J.N.; Su, X. VirusSeq: Software to Identify Viruses and Their Integration Sites Using next-generation Sequencing of Human Cancer Tissue. *Bioinformatics* 2013, 29, 266–267. [CrossRef]

34. Li, J.-W.; Wan, R.; Yu, C.-S.; Co, N.N.; Wong, N.; Chan, T.-F. ViralFusionSeq: Accurately Discover Viral Integration Events and Reconstruct Fusion Transcripts at Single-Base Resolution. *Bioinformatics* 2013, 29, 649–651. [CrossRef]

35. Wang, Q.; Jia, P.; Zhao, Z. VERSE: A Novel Approach to Detect Virus Integration in Host Genomes through next-generation Sequencing Dat. *PLoS ONE* 2013, 8, e64465. [CrossRef] [PubMed]

36. Wang, Q.; Jia, P.; Zhao, Z. VERSE: A Novel Approach to Detect Virus Integration in Host Genomes through Reference Genome Customization. *Genome Med.* 2015, 7, 2. [CrossRef]

37. Kostic, A.D.; Ojesina, A.I.; Pedamallu, C.S.; Jung, J.; Verhaak, R.G.W.; Getz, G.; Meyerson, M. PathSeq: Software to Identify or Discover Viruses by Deep Sequencing of Human Tissue. *Nat. Biotechnol.* 2011, 29, 393–396. [CrossRef]

38. Walker, M.A.; Pedamallu, C.S.; Ojesina, A.I.; Bullman, S.; Sharpe, T.; Whelan, C.W.; Meyerson, M. GATK PathSeq: A Customizable Computational Tool for the Discovery and Identification of Microbial Sequences in Libraries from Eukaryotic Hosts. *Bioinformatics* 2018, 34, 4287–4289. [CrossRef]

39. Watten, L.; Wilm, A.; Göke, J. Reproducible, Scalable, and Shareable Analysis Pipelines with Bioinformatics Workflow Managers. *Nat. Methods* 2021, 18, 1161–1168. [CrossRef]

40. Di Tommaso, P.; Chatzou, M.; Floden, E.W.; Barja, P.P.; Palumbo, E.; Notredame, C. Nextflow Enables Reproducible Computational Workflows. *Nat. Biotechnol.* 2017, 35, 316–319. [CrossRef]

41. Ewels, P.A.; Pelletz, A.; Filling, S.; Patel, H.; Alneberg, J.; Wilm, A.; Garcia, M.U.; Di Tommaso, P.; Nahnsen, S. The Nf-Core Framework for Community-Curated Bioinformatics Pipelines. *Nat. Biotechnol.* 2020, 38, 276–278. [CrossRef]

42. da Veiga Leprevost, F.; Grüning, B.A.; Alves Aflitos, S.; Röst, H.L.; Vaudel, M.; Moreno, P.; Gatto, L.; Weber, J.; et al. BioContainers: An Open-Source and Community-Driven Framework for Software Standardization. *Bioinformatics* 2017, 33, 2580–2582. [CrossRef]

43. Nurk, S.; Koren, S.; Rhiie, A.; Rautaiinen, M.; Bzikadze, A.V.; Mikheneko, A.; Vollger, M.R.; Altemose, N.; Uralsky, L.; Gershman, A.; et al. The Complete Sequence of a Human Genome. *Science* 2022, 376, 44–53. [CrossRef]

44. Li, H.; Durbin, R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 2009, 25, 1754–1760. [CrossRef]

45. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. 1000 Genome Project Data Processing Subgroup the Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 2009, 25, 2078–2079. [CrossRef]

46. Wood, D.E.; Lu, J.; Langmead, B. Improved Metagenomic Analysis with Kraken 2. *Genome Biol.* 2019, 20, 257. [CrossRef]

47. Ewels, P.; Magnusson, M.; Lundin, S.; Källér, M. MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report. *Bioinformatics* 2016, 32, 3047–3048. [CrossRef]
48. Okonechnikov, K.; Conesa, A.; García-Alcalde, F. Qualimap 2: Advanced Multi-Sample Quality Control for High-Throughput Sequencing Data. *Bioinformatics* 2015, 32, 292–294. [CrossRef]

49. Xie, Y.; Dervieux, C.; Riederer, E. *R Markdown Cookbook*, 1st ed.; Chapman & Hall/CRC The R Series; CRC Press/Taylor & Francis: Boca Raton, FL, USA, 2021; ISBN 978-0-367-56383-7.

50. Ondov, B.D.; Bergman, N.H.; Phillippy, A.M. Interactive Metagenomic Visualization in a Web Browser. *BMC Bioinform.* 2011, 12, 385. [CrossRef]