


Current and Promising Approaches to Identify Horizontal Gene Transfer Events in Metagenomes

Gavin M. Douglas ^{1,*} and Morgan G. I. Langille¹

¹Department of Microbiology and Immunology, Dalhousie University, Halifax, Nova Scotia, Canada

*Corresponding author: E-mail: gavin.douglas@dal.ca.

Accepted: August 19, 2019

Abstract

High-throughput shotgun metagenomics sequencing has enabled the profiling of myriad natural communities. These data are commonly used to identify gene families and pathways that were potentially gained or lost in an environment and which may be involved in microbial adaptation. Despite the widespread interest in these events, there are no established best practices for identifying gene gain and loss in metagenomics data. Horizontal gene transfer (HGT) represents several mechanisms of gene gain that are especially of interest in clinical microbiology due to the rapid spread of antibiotic resistance genes in natural communities. Several additional mechanisms of gene gain and loss, including gene duplication, gene loss-of-function events, and de novo gene birth are also important to consider in the context of metagenomes but have been less studied. This review is largely focused on detecting HGT in prokaryotic metagenomes, but methods for detecting these other mechanisms are first discussed. For this article to be self-contained, we provide a general background on HGT and the different possible signatures of this process. Lastly, we discuss how improved assembly of genomes from metagenomes would be the most straight-forward approach for improving the inference of gene gain and loss events. Several recent technological advances could help improve metagenome assemblies: long-read sequencing, determining the physical proximity of contigs, optical mapping of short sequences along chromosomes, and single-cell metagenomics. The benefits and limitations of these advances are discussed and open questions in this area are highlighted.

Key words: horizontal gene transfer, lateral gene transfer, shotgun metagenomics, metagenome-assembled genomes, microbiome.

Introduction

Microbes are microscopic organisms that include prokaryotes (bacteria and archaea), viruses, fungi, and protists. These organisms, particularly prokaryotes and viruses, are known to rapidly adapt to novel abiotic and biotic environmental changes. The genetic bases for these adaptations have largely been identified by studying the genomes of isolated organisms of interest, which have greatly improved our understanding of the genetic bases of adaptations throughout microbial evolution (Parkhill et al. 2003; Etienne et al. 2013; Harding et al. 2017).

A substantial proportion of microbes in natural communities are currently uncultured (Amann et al. 1995; Martiny 2019) and acquiring isolate genomes for these organisms has proven difficult. These uncultured microbes have been studied through metagenomics approaches, which involve the sequencing of all, or an enriched set, of microbial genomes in a sample (Riesenfeld et al. 2004). Metagenomics sequencing (MGS) enables researchers to

investigate how environmental conditions shape the taxonomic and functional composition of natural microbial communities. Recently, shotgun MGS—unbiased high-throughput sequencing of all DNA in a sample—has become the predominant method of metagenomics profiling. MGS analyses have largely been gene-centric, meaning that the focus has been on the relative abundances of individual genes (and inferred pathway relative abundances) in a community. More recently, the focus has shifted toward generating metagenome-assembled genomes (MAGs) from this sequencing data (Parks et al. 2017; Stewart et al. 2018; Pasolli et al. 2019). An important challenge for either analysis approach is to determine whether genes hypothesized to be adaptive arose through gene gain mechanisms or alternatively were part of preexisting genetic variation.

New genes can be acquired through three processes: 1) gene duplication and diversification, 2) the gain of a de novo gene (e.g., in previously noncoding DNA), and 3) horizontal gene transfer (HGT). In addition to the gain of new genes,

microbes are known to adapt to new environments through gene loss. All of these processes will be described in detail below.

Although most MGS analyses are based on gene-centric methods, assembling MGS reads into genomes is one clear way to improve the identification of gene gain and loss events. The key challenge of this approach is that assembly errors can result in gene gain and loss events being falsely identified or missed. Several recent technological advancements could help overcome this barrier. These technologies include long-read sequencing, mapping the physical proximity and interactions of DNA fragments, optical mapping of short sequences along chromosomes, and single-cell metagenomics.

Herein, the approaches applied so far to detect gene gain and loss events in MGS data are reviewed. This review largely focuses on identifying HGT events because these events are of primary interest within the field and there have been several recent methodological advances in HGT-detection in the context of metagenomes. However, methods to identify alternative processes of gene gain and loss will first be discussed (fig. 1). The background and mechanisms of HGT are then described, followed by a comparison of the tools applied to detect HGT within assembled genomes and MGS data. Lastly, several recent technologies are described that could help address the issue of producing high-quality MAGs from MGS data.

Gene Gain and Loss Events

Although HGT is the predominant mode of gene gain studied in MGS data, several other mechanisms of genic gain and loss are important to consider: gene duplication, gene loss, and de novo gene birth (fig. 1). The relative importance of these processes, and HGT, in driving adaptive evolution in microbes remains unclear. However, profiling the occurrence and adaptive benefit of these events could help resolve this issue. These genic events will be described and possible methods for identifying them in MGS data will be discussed below.

Gene Duplication

Gene duplication has long been known to be an important process underlying adaptation to novel environments (Kondrashov 2012). In prokaryotes, gene duplication events are typically at the order of individual genes and operons, particularly of genes involved with transcription, metabolism, and defense (Gevers et al. 2004). Duplication of individual genes can similarly occur in eukaryotic genomes, but eukaryotes also commonly undergo larger genomic duplication events, such as whole-genome duplications (Aury et al. 2006). The vast majority of duplicate genes gain degenerative mutations and become nonfunctional pseudogenes (Lynch and Conery 2000); however, surviving genes can acquire

divergent or novel functions (Ohno 1970; Force et al. 1999). Regardless of the mechanism, gene duplicates are extremely common in nature: an estimated 7–41% of bacterial proteins are encoded by paralogs (Gevers et al. 2004). Duplicates can also be retained without diverging in function as well, which can provide higher protein expression levels (Schuster-Böckler et al. 2010; Kondrashov 2012) or can keep dosage levels proportional to other gene duplicates (Conant et al. 2014).

Mapping unassembled reads to gene family databases can be used to generate hypotheses about gene duplications, which can be evaluated with additional data. For example, high abundances of mercury resistance genes were observed in groundwater metagenomes dominated by *Rhodanobacter* (Hemme et al. 2016) and subsequent analyses identified putative duplicated mercury resistance operons within *Rhodanobacter* genomes (fig. 1A). However, in general, gene duplicates are difficult to identify in MGS data by characterizing the relative abundances of gene families within metagenomes. This challenge is largely due to the difficulty of distinguishing paralogous sequences from the same genome from orthologs across multiple genomes. In addition, most methods for detecting gene duplicates, and structural variants (SVs) in general, are intended to be applied to genome resequencing data of a single organism mapped against a reference genome (Ye et al. 2009; Rausch et al. 2012; Yavaş et al. 2014). Applying such approaches to identify SVs in a mixed community would likely result in widespread false inferences. The exceptions could be when a community is dominated by a small number of known species or if a subset of reads can be confidently binned into species. In these cases, mapping species-identified MGS reads to the appropriate reference genomes and then applying existing SV-detection methods would be reasonable. Although these methods are optimized for the human genome, previous work has shown that prokaryotic strain-level SVs can be accurately identified using a consensus of multiple tool outputs (Zojer et al. 2017).

No single pipeline is available for identifying gene duplicates in fragmented metagenomes, but numerous approaches could be leveraged by comparing MAGs with existing reference genomes. For instance, comparative genomics could be employed to identify clusters of homologous genes across genomes using reciprocal BLAST matching between assembled contigs and reference genomes of closely related taxa (*Drosophila* 12 Genomes Consortium et al. 2007; Hahn et al. 2007). Under this approach, orthologous genes are typically assumed to be reciprocal best-hits whereas other similar genes are putative paralogs (Ward and Moreno-Hagelsieb 2014). There are several methods available for summarizing species pan-genomes as well, such as Roary (Page et al. 2015) and panX (Ding et al. 2018), which could be useful for interpreting gene duplication patterns across genomes. In addition, although putative gene duplicates could be identified through comparison with known reference genomes it is possible that

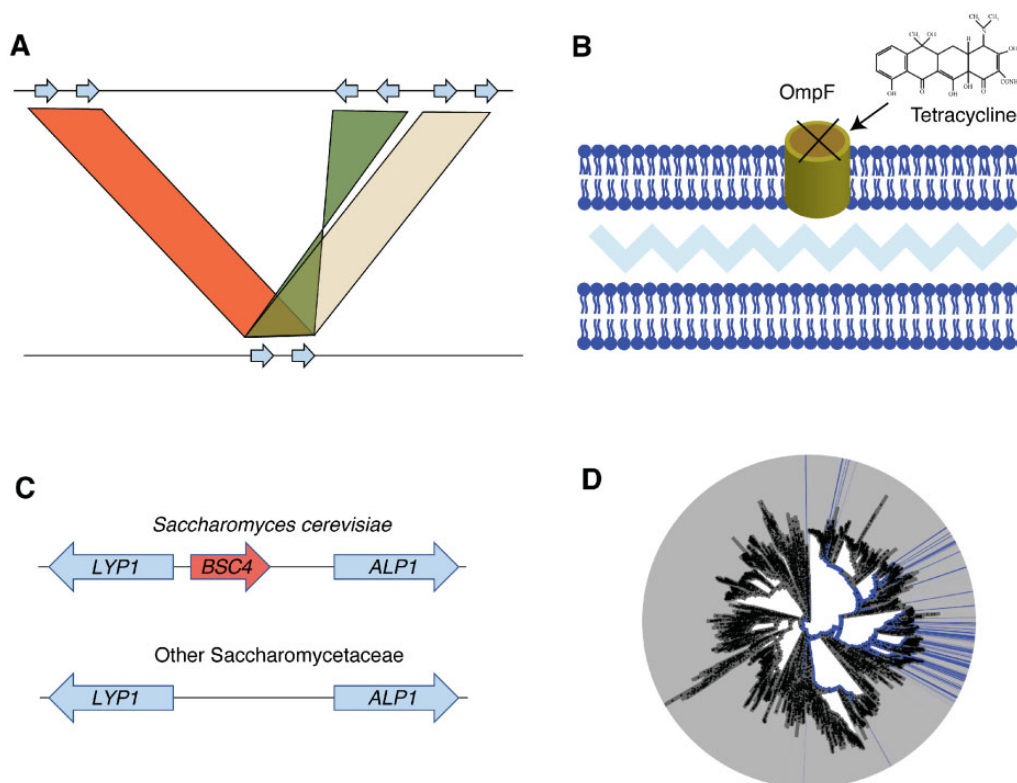


FIG. 1.—Examples of microbial gene gain and loss. (A) Illustration of operon duplications between two genomes. Arrows indicate genes and colored bars indicate different regions of homologous DNA shared between the two genomes. This simplified example is inspired by the mercury resistance operon duplications identified in *Rhodanobacter* genomes (Hemme et al. 2016). High levels of mercury resistance genes were reported in groundwater metagenomes dominated by *Rhodanobacter*, but genomic analyses were required to identify putative duplication events. (B) An example of adaptation through loss-of-function. Tetracycline (indicated by chemical structure) is largely imported through the OmpF porin in *Escherichia coli*. Deleting the gene encoding this porin allows for higher tetracycline tolerance (Thanassi et al. 1995). (C) Example of de novo gain of the *BSC4* gene in *Saccharomyces cerevisiae* compared with other Saccharomycetaceae (Cai et al. 2008). Simplified visualization of orthologous region across fungi demonstrates that *BSC4* is only present in *S. cerevisiae*. (D) Distribution of two KEGG orthologs (Kanehisa and Goto 2000) (K08928, K08929), which are responsible for anoxygenic photosystem II (M00597), that are broadly distributed across the prokaryotic tree likely due to horizontal gene transfer. The presence and absence of these gene families is indicated in blue and gray, respectively. Panel created with AnnoTree (Mendler et al. 2019).

alternative explanations such as gene loss or HGT could better account for this signature, which would need to be reconciled (see the Identifying HGT Events in Metagenomes section). It is also important to emphasize that misassembled MAGs might be especially susceptible to false inferences of gene duplication, especially if multiple closely related taxa are included in the same assembly.

Gene Loss

Gene loss is normally associated with decreased negative selection, but recently the importance of gene loss in adaptive evolution has been demonstrated (Hottes et al. 2013; Albalat and Cañestro 2016). Many adaptive loss-of-function (LOF) mutations knockout an individual protein, such as the deletion of the porin *ompF* locus in *Escherichia coli*, which grants tolerance to tetracycline by preventing its import (Thanassi et al. 1995) (fig. 1B). However, there are also cases of adaptive LOF

mutations disrupting regulatory networks, such as the knockout of genes in *Candida glabrata* required to synthesize nicotinic acid, which causes epithelial adhesion genes to be expressed and enables binding to murine renal cells (Domergue 2005). For segregating knockout mutations, there are existing tools for profiling strain-level variation, including single-nucleotide polymorphisms, relative to reference genomes (Nayfach et al. 2015; Scholz et al. 2016; Costea et al. 2017). The output of the identified mutations could be used with existing programs such as SnpEff (Cingolani et al. 2012), which predict the effect of mutations on protein-coding genes. One potential issue with this approach is that misaligned reads could result in many false LOF mutations being identified. Focusing analyses instead on MAGs could be a more straight-forward way to identify LOF mutations. However, these LOF mutations could result in genes not being annotated in assembled contigs if they are not identified to be open-reading frames. Also, large deletions would

result in genes missing entirely from MAGs. In this case, gene loss would need to be inferred by identifying orthologous regions of DNA using similar approaches as described above for identifying gene duplicates.

Alternatively, the absence of genes could be identified if they are annotated in closely related reference genomes. For example, the ribosomal proteins L9 and L1 were recently shown to be differentially absent within multiple MAGs from the poorly characterized candidate phyla radiation group (Brown et al. 2015). These ribosomal proteins are otherwise highly conserved in bacteria, so this finding indicates a shift in ribosome function within this lineage. These instances of gene loss were identified consistently across multiple MAGs, which provides additional evidence that this finding is not a technical artifact.

De Novo Gene Birth

De novo gene birth is caused by mutations that give rise to novel protein-coding regions, which fall into two categories. The first category is de novo genes arising in preexisting protein-coding regions, typically on the opposite strand or a separate reading frame, which is referred to as overprinting. This form of de novo gene birth was first described in the genome of the virus phiX174 (Weisbeek et al. 1977) and numerous other cases have since been identified in viruses (Sabath et al. 2012; Van Oss and Carvunis 2019) and bacteria (Ohno 1984; Hückler et al. 2018; Vanderhaeghen et al. 2018). One example is of the transcript *nog1* found on the reverse strand of the gene *citC* in *E. coli* (Fellner et al. 2015). Although the function of this gene remains unclear, it was shown that this gene likely encodes a protein, based on experimental evidence and the presence of a Shine–Dalgarno sequence upstream of the predicted start codon. In addition, knocking out *nog1* while maintaining the *citC* amino acid sequence resulted in decreased fitness relative to wild-type. Identifying such cases of overprinting is likely not feasible within MGS data sets alone and instead meta-transcriptomics data would be required to identify alternative transcripts originating from the same locus. One major challenge facing this approach would be distinguishing transcripts encoding proteins from antisense transcripts producing noncoding RNAs involved in gene regulation (Brantl 2007; Bao et al. 2015).

The second type of de novo gene birth occurs when protein-coding regions arise from noncoding DNA (Tautz and Domazet-Lošo 2011). This form of de novo gene birth is known to occur in eukaryotes, especially at low frequencies in a population, but these cases are largely nonadaptive and undergo pseudogenization (Tautz and Domazet-Lošo 2011). Nonetheless, there are many examples of such de novo genes conferring adaptive benefits (McLysaght and Guerzoni 2015; Schlötterer 2015). One example is of the *BSC4* gene in *Saccharomyces cerevisiae*, which is unique to that species (Cai et al. 2008) (fig. 1C). This gene encodes a protein involved

in the DNA repair pathway during stationary phase (Cai et al. 2008). Importantly, such examples of de novo gene birth are restricted to eukaryotes, likely due to the low proportion of noncoding DNA in prokaryotic genomes. Existing methods (Van Oss and Carvunis 2019) could be leveraged to identify these cases within high-quality eukaryotic MAGs. One possible approach would be to identify candidate de novo genes in a MAG that are homologous with noncoding DNA in all related taxa.

Signatures of HGT

HGT, also known as lateral gene transfer, is the transfer of genetic material outside parent–offspring inheritance. Importantly, HGT differs from the other processes discussed above because it enables the rapid transfer of genes between distantly related species and is thought to be especially important for the adaptation of microbes to novel environments (Ochman et al. 2000). In particular, there is a heightened interest in HGT due to concerns regarding the spread of antibiotic resistance genes in hospitals (Spellberg et al. 2008), livestock (Mathew et al. 2007), and waterways (Szczepanowski et al. 2009; Bengtsson-Palme et al. 2014). More generally, the importance of HGT for microbe niche specialization is demonstrated simply by the sparse distribution of key functions across life (Boucher et al. 2003). One representative example is of tetrapyrrole-based photosynthesis, which is patchily distributed across bacteria likely due to HGT (fig. 1D).

There are numerous mechanisms underlying HGT, which differ in frequency across taxa and depend on the genetic distance between the donor and acceptor genomes (Thomas and Nielsen 2005). The three main mechanisms are transformation, conjugation, and transduction. An understanding of these different mechanisms is important because they can be associated with distinct genomic signatures (e.g., specific sequences linked to a transfer mechanism), which can be corroborating evidence of HGT (reviewed in Zaneveld et al. [2008]).

Transformation refers to the uptake and integration of extracellular DNA and commonly occurs across prokaryotes. DNA uptake requires a cell to be in a physiological state known as competency, which typically involves the activity of 20–50 proteins (Thomas and Nielsen 2005). The transformation rate in an environment depends on both the proportion of competent cells and on the concentration of extracellular DNA, which greatly varies across environments. For instance, concentrations of extracellular DNA in marine sediments are typically three orders of magnitude higher than in marine water (Torti et al. 2015). Input DNA becomes single-stranded when translocated across the inner membrane and the translocated DNA can then undergo homologous recombination with similar sequences or be used as a source of nutrients (Finkel and Kolter 2001). Theoretically, the potential for these transfer events might be inferred by

identifying the presence of genes encoding proteins involved in this process (Zaneveld et al. 2008). For instance, there are many surface proteins involved in the uptake of extracellular DNA, some of which are similar to surface appendages such as type IV pili (Chen and Dubnau 2004). These proteins are well characterized in a small number of bacteria such as *Bacillus subtilis* and *Streptococcus pneumoniae* (Claverys et al. 2009) and could be used as markers for the potential to undergo transformation in these species. However, scanning for the genes encoding these proteins in the genomes of poorly characterized species would likely not yield accurate identification of the potential for DNA uptake.

Conjugation is a mode of unidirectional DNA transfer through cell-to-cell junctions, typically mediated by type IV secretion systems that only transfer single-stranded DNA (Zaneveld et al. 2008). Conjugation is the predominant mechanism of transfer of plasmids (Thomas and Nielsen 2005) (which are transferred as single-stranded and circularized copies) and is especially relevant for the spread of antibiotic resistance (Akiba et al. 1960; Kristiansson et al. 2011; Bengtsson-Palme et al. 2014). Many conjugative plasmids are self-transmissible and can either integrate into the host genome or remain autonomous in the cell (Yin and Stotzky 1997). Conjugal transfer is typically between closely related organisms but can also occur between distantly related taxa (Dahlberg et al. 1998; Lacroix and Citovsky 2016). The most well-known proteins involved in forming the cell–cell junction are the *tra* proteins, which are conserved in certain lineages (Lanka and Wilkins 1995). The origin of transfer (*oriT*) is an element within plasmid DNA that specifies where the relaxase protein binds (Grohmann et al. 2003). Relaxase binding promotes the formation of the relaxosome complex, which nicks the DNA at a conserved DNA motif called *nic* (Zaneveld et al. 2008). Identifying the presence of these proteins and DNA motifs in metagenomics data could be used to infer the potential for conjugation. However, the most straight-forward signature for potential conjugation in MGS data is the presence of plasmids.

There are several approaches that have been developed for identifying plasmid sequences in MGS data. One approach is to identify circular contigs in assembly graphs, as performed by the software Recycler (Rozov et al. 2017). Another major approach is to compare reads or contigs to a database of reference plasmids (Carattoli et al. 2014), which restricts researchers to previously identified plasmids. Lastly, differing k-mer profiles between chromosomal and plasmid DNA can be leveraged using machine learning approaches to identify novel plasmid sequences (Zhou and Xu 2010; Krawczyk et al. 2018). An especially promising tool using this approach is PlasFlow, which classifies chromosomal and plasmid contigs with a neural network trained on reference sequence k-mer content (Krawczyk et al. 2018). This tool performed substantially better than the other approaches described above and provides the added benefit of identifying linear plasmids,

which is not possible with assembly-based plasmid identification tools. These approaches for identifying plasmid sequences in MGS could be used to both establish that transferrable plasmids are present in the community and to identify genes contained on plasmids that could potentially be transferred.

The final major mechanism of HGT is transduction, which is the form of genetic transfer mediated through phage and can be categorized as generalized or specialized transduction (Yin and Stotzky 1997). Generalized transduction refers to the packaging of random DNA fragments from a bacterial genome into a phage capsid. This process can occur when a host cell is infected by either a virulent phage or a temperate phage undergoing a lytic cycle. In contrast, genes transferred through specialized transduction are integrated into the genomes of temperate phages when they are incorrectly excised from the host genome. Specialized transduction cannot involve virulent phages because integrating into the host genome as a prophage is a required step. Transferred genes are integrated along with the phage genome into new hosts (Yin and Stotzky 1997). Bacteriophage tropism is generally restricted to hosts within a single species (Koskella and Meaden 2013), although there are exceptions (Malki et al. 2015), which means that overall transduction is less common between distantly related organisms. It is also important to recognize that acquiring novel genes can enable phages to adapt to novel niches (Hatfull and Hendrix 2011). In addition, the intermediate stage of transduction within phages enables genes to rapidly evolve, which can result in novel beneficial functions if they are eventually acquired by a bacterial host (Comeau and Krisch 2005).

The genomic signatures of transduction have the most potential to be used to identify past transfer events. These signatures can be identified for specialized transduction, but not easily for generalized transduction because DNA transfer through the latter typically occurs through homologous recombination of randomly packaged bacterial DNA. In contrast, phages capable of specialized transduction typically integrate at a specific *attB* site within host genomes through recombination with an *attP* element in the phage genome (Canchaya et al. 2003; Zaneveld et al. 2008). In addition, only genes located nearby prophages in a bacterial genome will be transferred through specialized transduction. It has previously been established that genes transferred by specialized transduction can be identified by whether they are nearby phage-related sequences, such as phage integrases, or are nearby transfer RNAs (Williams 2002), which is a preferential integration site for certain temperate phages. Accordingly, the presence of prophage sequences nearby putatively transferred genes could be taken as corroborating evidence of specialized transduction.

Approaches to identify prophage sequences in isolate genomes typically rely on sequence similarity with known viral genes as well as identifying regions with viral genome characteristics (Akhter et al. 2012). These characteristics include

shorter protein lengths, shared transcription directionality of adjacent genes, and distinct k-mer profiles. Identifying viral sequences in MGS data is more complicated than in isolate genomes because the datatype is more diverse and fragmented, but nonetheless several tools have been developed for this purpose (Ren et al. 2017; Amgarten et al. 2018; Garretto et al. 2019). These tools have mainly been focused on identifying viral contigs in MGS data rather than large bacterial contigs containing a prophage, but there are exceptions. VirSorter is one approach that can explicitly identify prophages in MGS data (Roux et al. 2015). This tool scans contigs for the presence of viral genes and certain virus-like characteristics and reports the confidence that each gene, as well as the contig overall, is virus derived. PHASTER is a similar tool that identifies prophage regions in contigs based on the sequence similarity of open-reading frames with reference phage and prophage genes (Arndt et al. 2016). The output of these tools could be particularly useful for finding corroborating evidence of specialized transduction transfer events.

Although the three mechanisms discussed above are the best-studied modes of HGT, there are several other notable mechanisms. The foremost of these other mechanisms are gene transfer agents (GTAs), which are phage-like particles that transfer DNA between cells. At least five families of GTAs have been identified in prokaryotes (Lang et al. 2017). Although most of these GTAs have been identified in only a single species, homologs for the GTA family characterized in *Rhodobacter capsulatus* have been identified in numerous *Alphaproteobacteria* species (Lang and Beatty 2007). It remains unclear how widespread GTAs are in nature, but it has been suggested that they could represent a substantial proportion of diversity identified in the marine virome (Kristensen et al. 2010). Two additional mechanisms of genetic transfer related to conjugation are intercellular nanotubes (Dubey and Ben-Yehuda 2011) and membrane vesicles (Domingues and Nielsen 2017), which are conducted through direct cell–cell physical interactions and the release and uptake of extracellular vesicles, respectively. The importance of these mechanisms for HGT across prokaryotes also remains controversial (Ficht 2011; Gröll et al. 2018).

Identifying HGT Events in Metagenomes

Numerous approaches have been developed for detecting HGT in genomics data (table 1), which are largely intended to be applied to high-quality isolate genomes (Ravenhall et al. 2015). In contrast, although HGT is often hypothesized based on unassembled metagenomes, directly identifying putative HGT events is less common for this datatype (fig. 2A). For example, the higher relative abundances of antibiotic resistance genes and mobile elements, such as plasmids, in waterways downstream of wastewater treatment plants have been taken as putative evidence of the spread of antibiotic resistance through HGT (Szczepanowski et al. 2009; Bengtsson-

Palme et al. 2014). Another example is that HGT has been suggested to be potentially responsible for higher similarity in chromium resistance genes than in 16S ribosomal RNA gene sequences found in water samples with high concentrations of chromium (Parnell et al. 2010). These and other similar studies (Kurokawa et al. 2007; Rua et al. 2018) represent the most common approach of inferring putative HGT in metagenomes. This approach can be useful for generating hypotheses but does not provide direct evidence for specific HGT events.

In contrast, identifying potential HGT events in assembled contigs would be a source of direct evidence. The major challenge is that it is often only possible to assemble MGS reads into short contigs, which commonly represent a small fraction of complete genomes and can be enriched for assembly errors. Many existing methods for detecting HGT events in complete genomes cannot be applied to this fragmented data. This challenge is gradually being addressed as deeper sequencing read depths and improvements in related technologies have enabled higher-quality MAGs to be produced, as we discuss in the next sections. Nonetheless, methods specifically focused on poorly assembled MGS data have also been developed. These approaches for identifying putative HGT in both high and low-quality MAGs will be discussed below.

Composition-Based Approaches

The first major approach developed to detect HGT was based on comparing base and codon usage composition across genes within a genome (Médigue et al. 1991; Lawrence and Ochman 1997). This general approach is motivated by the findings that base composition and codon usage are largely homogenous within a genome (Sueoka 1961; Hildebrand et al. 2010). In addition, base composition is known to be linked to taxonomy: taxa within the same lineage tend to have genomes with similar GC-content and base composition overall (Sueoka 1961). Genomic regions that compositionally differ from background are referred to as “genomic islands.” Two popular tools for identifying these regions are GIST (Hasan et al. 2012) and IslandViewer (Bertelli et al. 2017) although many similar tools are also available (Langille et al. 2010; Lu and Leong 2016).

Composition-based approaches have been applied to MGS contigs (Hemme et al. 2010), but this is not typically performed because longer regions are thought to be required to accurately determine the background composition of the genome. Nonetheless, one composition-based method has been proposed specifically for MGS-assembled contigs (Tamames and Moya 2008). This method involves calculating the frequencies of k-mers within each gene in a contig to generate a vector of frequencies per gene. Pearson correlations between these vectors across genes are then calculated and these genes are clustered by these correlations to enable

Table 1
Approaches for Identifying Putative Cases of HGT in Metagenomes

Approach	MGS Specific ^a	Example Implementations
Identify outlier genomic regions based on DNA composition	No	GIST (Hasan et al. 2012); IslandViewer (Bertelli et al. 2017)
Identify outlier genes in contigs based on pairwise Pearson correlations of k-mer content	Yes	Described in Tamames and Moya (2008)
Identify genes in genomes displaying taxonomically discordant similarity to genes within a reference database	No	DarkHorse (Podell and Gaasterland 2007); HGTector (Zhu et al. 2014)
Identify genes in contigs displaying taxonomically discordant similarity to genes within a reference database	Yes	WAAFLE ^b and a method described in Tamames and Moya (2008)
Reconcile phylogenetic incongruencies between gene and species trees	No	AnGST (David and Alm 2011); RANGER-DTL (Bansal et al. 2018)
Identify putative donor and recipient transfer events within a given community based on a combined similarity and phylogenetic incongruency approach	Yes	MetaCHIP (Song et al. 2019)
Identify genomic hotspots of recombination between strains of a species	Yes	Described in Tyson et al. (2004) and Probst and Banfield (2018)

^aWhether an approach is intended specifically for shotgun metagenomics (MGS) data instead of isolate genomes.

^b<http://huttenhower.sph.harvard.edu/waafle>.

outlier genes to be identified (fig. 2B). The challenge of this method is that the choice of cutoff for distinguishing genes into clusters can have a large effect on the result and it is unclear what value should be used. This approach has been applied previously to fosmid clones corresponding to *Verrucomicrobia* to identify putative HGT events (Kielak et al. 2010).

There are two key limitations of the above compositional approaches. First, false positives can arise simply due to genomic variation in composition, such as variation in base composition related to distance from the replication terminus (Deschavanne and Filipinski 1995; Guindon and Perrière 2001). However, this issue is partially addressed in tools that test for differences in k-mer frequencies, which are thought to be more genome-specific (Karlín and Burge 1995), rather than GC-content and codon usage (Koski et al. 2001;

Wang 2001). The second major limitation is that ancient HGT events are difficult to detect because transferred genes eventually evolve (or “ameliorate”) to become similar to the rest of the genome (Lawrence and Ochman 1997, 1998). This issue implies that only recently acquired HGT events can be identified using compositional approaches.

Implicit Phylogenetic Approaches

Another common approach for detecting HGT is to identify genes with higher sequence similarity to homologs encoded by more distantly related taxa compared with close relatives. This class of approach is biased toward more ancient HGT events in contrast to the compositional approaches described above (Ragan et al. 2006). As above, variations on this approach have been implemented for complete genomes (Clarke et al. 2002; Ragan and Charlebois 2002; Podell and Gaasterland 2007). One example is HGTector, which compares protein sequences across genomes of varying phylogenetic distance and includes several improvements that make it resilient to technical and biological confounders (Zhu et al. 2014).

For short MGS-derived contigs, one taxonomic-assignment approach has been proposed that involves running BLASTX (Altschul et al. 1990) on genes within contigs against the GenBank nonredundant database and assigning taxonomy to the genes based on the best-hits (Tamames and Moya 2008). In this case, HGT events are identified when there is sufficient disagreement in the taxonomic assignment of genes within the same contig. WAAFLE (<http://huttenhower.sph.harvard.edu/waafle>; last accessed August 27, 2019) is a software that implements a similar approach (manuscript in prep). The WAAFLE pipeline involves identifying the most similar matches in a pan-genome database for each gene in a contig. The tool then determines whether the genes in the contig can be explained entirely by a single species or if multiple species are needed to account for the contig gene content. This latter case is taken as putative evidence of HGT. The major strength of this approach is that multiple similarity matches are retained per gene in a contig. This information enables a conservative taxonomic-assignment approach to be employed where contigs can be classified as a single species, even if that species is not the best-hit for each gene, which is intended to reduce the number of false positives.

Explicit Phylogenetic Approaches

Although the similarity-based HGT-detection methods described above use phylogenetic principles, they do not make direct use of phylogenetic trees to test for phylogenetic incongruencies. Testing for phylogenetic incongruencies refers to comparing a gene tree based on homologous sequences across taxa with the phylogenetic tree for those taxa. The growing number of prokaryotic genomes from pure cultures has enabled large-scale phylogenetic methods to be

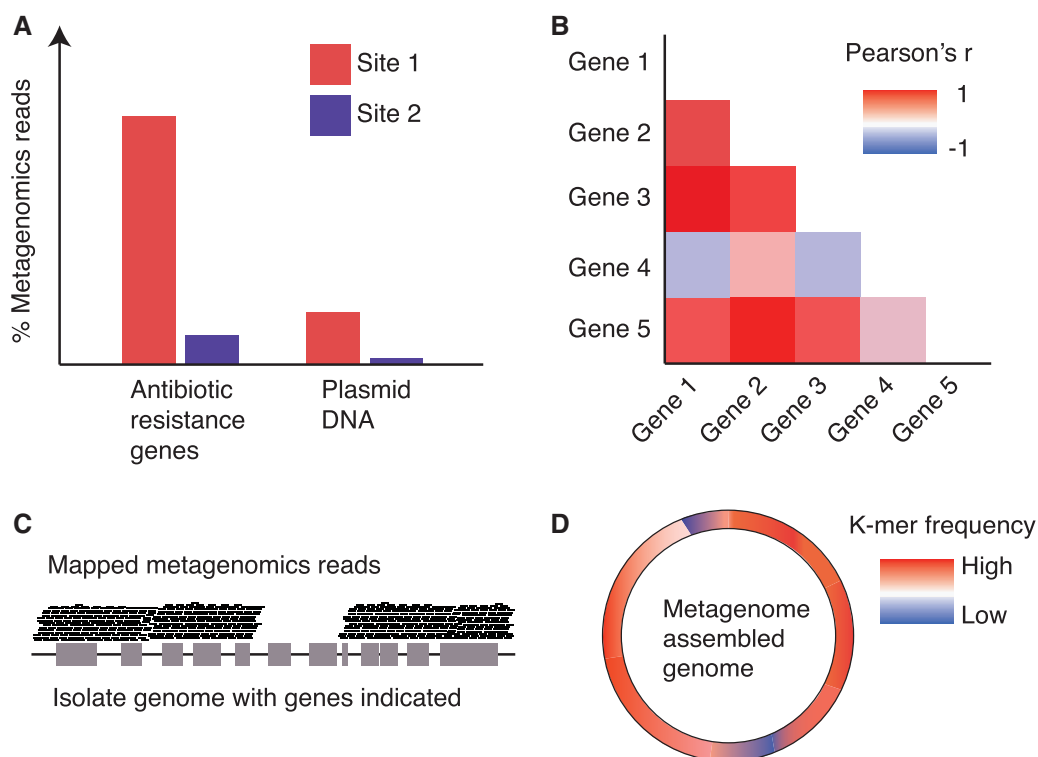


Fig. 2.—Key approaches to infer potential HGT in MGS data. (A) Identifying genes frequently transferred through HGT at differential relative abundance between two sites. One possible explanation for these observations is HGT, which is frequently hypothesized in the literature although this is not based on direct evidence. (B) Identifying outlier genes in short assembled contigs using a compositional approach (Tamames and Moya 2008). This approach involves tabulating k-mer frequencies within each gene and calculating the pairwise Pearson correlation between all genes within the contig. Outlier genes with atypical k-mer composition can then be identified, which are candidates for HGT (such as gene 4 in this example). (C) Isolate reference genomes have been used with MGS data on several occasions. One example usage is to map the metagenomics reads to existing reference genomes to identify genomic regions not found in the metagenome. HGT is one possible explanation for the absence of reads mapped to a particular region of the reference genome as shown in this example. (D) Generating high-quality MAGs allows any method for identifying HGT in isolate genomes to be applied. The example shown here is of detecting genomic regions that have divergent k-mer composition compared with the rest of the genome. Note that in this simplified example only one k-mer is being compared whereas typically the profile of many k-mers would be compared.

developed (Beiko et al. 2005; Puigbo et al. 2009). Similar approaches have been applied to MAGs in recent years as well (Guo et al. 2015; Soo et al. 2017). For example, genomes from a basal-branching clade known as Sericytochromatia within Cyanobacteria were assembled from multiple MGS data sets (Soo et al. 2017). These genomes were taxonomically classified based on their 16S ribosomal RNA gene sequences and placement within a supertree of reference genomes was performed based on concatenated protein sequences. These MAGs enabled standard tests for phylogenetic incongruency to be run to identify proteins that were phylogenetically dispersed differently from the supertree, including key genes involved in photosynthesis missing in Sericytochromatia. Despite such successes reconciling gene and species trees remains a nontrivial problem because disagreements in the tree can be biologically due to either gene duplication, HGT, or gene loss. Accordingly, this has been deemed the duplication-transfer-loss reconciliation problem and several methods have been developed to address this

issue (Karneva and Ward 2014). RANGER-DTL is one such method that compares gene and species trees and identifies the most likely positions on the species tree where either speciation, gene duplication, HGT, or gene loss have occurred (Bansal et al. 2018). A similar method is the software GLOOME (Cohen et al. 2010), which is a maximum-likelihood approach that can also be used to infer the position of gene gain and loss events across a phylogenetic tree, but does not consider duplication and speciation events.

Identifying Putative Gene Transfers in a Defined Community

The methods for detecting HGT presented above approach the problem from a range of perspectives, which can result in strikingly different inferences of HGT. One example is a comparison of a compositional and a sequence similarity approach that resulted in fewer than 5% of HGT events in agreement (Tamames and Moya 2008). This drastic difference and other

examples (Ragan et al. 2006) are at least partially due to the differing sensitivity of these tools for detecting HGT events of different ages. Importantly, these tools can be used in combination to yield more robust inferences (Omelchenko et al. 2003; Schoenfeld et al. 2013).

MetaCHIP is a recently published tool that is partially based on this idea, but intended specifically to identify HGT events between observed donor and recipient genomes in a natural community (Song et al. 2019). This tool first performs an all-against-all BLASTN of genes within assembled contigs from a given community. Potential HGT events are identified based on genes with best-hits in other taxonomic groups (e.g., in another family). False inferences due to duplicated regions of contigs are explicitly accounted for in order to reduce the false positive rate. A gene tree is then created for all genes on this short-list of putative HGT events. These trees are compared with a species tree based on 43 universal single-copy genes (USCGs) with the RANGER-DTL software to determine whether HGT, or a different mechanism, better accounts for any phylogenetic incongruencies.

Characterizing Strain Heterogeneity

Although the majority of HGT-detecting methods are focused on identifying transfers between different species, studying MAGs can also yield insight into population heterogeneity within a species. Because MAGs are based off the genomes of numerous bacterial cells in a community there is always some degree of genetic variation in the reads underlying MAGs. In addition, homologous recombination is known to occur between divergent strains, which can result in mosaic genomes with different gene blocks (Falush et al. 2001; Papke et al. 2004; Dunn et al. 2009). Due to the different assortment of gene blocks between closely related organisms, these recombination events are examples of recent HGT. Leveraging MAGs to assess intraspecies recombination is appealing because recombination hotspots can be readily identified with this datatype.

Key work in this area has focused on identifying large regions of homologous recombination within the genomes of *Leptospirillum* and *Ferroplasma* species originating from acid mine drainage sites (Tyson et al. 2004; Deneff et al. 2007; Deneff and Banfield 2012). These recombination blocks contain genes hypothesized to be needed for rapid adaptation to this extreme environment. Another recent example was that of a novel archaeon genome assembled from water samples of intermediate and deep aquifers (Probst and Banfield 2018). Not only was the complete genome of the taxon named *Candidatus "Forterrea multitransposorum"* assembled, but population variation of this taxon was also assessed by mapping reads to this assembled genome. Through this approach, the authors identified hotspots of homologous recombination occurring between members of the species. These examples highlight that because MAGs

represent a population rather than individual organisms, they can be leveraged to identify regions of recent HGT.

Leveraging Existing Reference Genomes

Rather than directly identifying HGT in metagenome sequencing data, inferences made from metagenomes have also been used to inform analyses on reference genomes (fig. 2C). For instance, putative HGT events in *Synechococcus* reference genomes were identified by mapping MGS reads to these genomes and identifying unmapped regions with divergent trinucleotide composition (Palenik et al. 2009). A different example focused on a peptides/nickel transport complex identified to be enriched in the gut metagenomes of lean individuals (Meehan and Beiko 2012). By placing the sequences of the individual gene families involved in this module into gene trees created from homologous genes in existing reference genomes, it was shown that the phylogenetic position of these genes greatly varied. The only potentially high contributor of all gene families involved in the module was the gut commensal *Faecalibacterium prausnitzii*. Evidence for rampant HGT of this peptides/nickel transport complex was found by focusing on this module within *Faecalibacterium prausnitzii* reference genomes. More generally, putative HGT events identified in metagenomes can help decrease the search space of gene families and reference genomes to help directly identify individual cases of HGT (Palenik et al. 2009; Meehan and Beiko 2012; Hemme et al. 2016; Llorens-Marès et al. 2017). This approach has proven useful but is highly dependent on existing reference genomes and does not take advantage of the potential to assemble metagenomes.

Current Barriers and Outlook

There are two main barriers to researchers detecting HGT in MAGs within their own data. The first barrier is the challenge of generating adequate quality MAGs, which is closely linked to the major goal in metagenomics of improving the quality of assemblies overall. This issue can best be addressed through several recent technological advances (see the Potential Avenues to Improve Metagenome Assemblies section). The second obstacle preventing researchers from detecting HGT events is that these analyses require substantial bioinformatics expertise. Determining which approach to use is nontrivial and will largely depend on the biological question. For example, different methods are available if researchers are interested specifically in identifying hotspots of recombination within a single species (Probst and Banfield 2018) or identifying HGT events between different species in a given community (Song et al. 2019). In addition, researchers' choices can be informed by what time-scale of HGT they are interested in investigating. However, even with a clear research question selecting a specific tool for these analyses can be problematic and so in practice comparing the output of several methods would likely be best. A robust evaluation of the performance

of these approaches is needed to better inform researchers' choice of tools. This evaluation is especially needed for tools applied to MGS-derived contigs and assemblies, which could be done by simulating HGT events within a defined set of genomes with a tool such as HgtSIM (Song et al. 2017). In addition, although existing HGT-detection methods can be applied to high-quality MAGs (fig. 2D), this may come at the cost of an unacceptable false positive rate because past evaluations of tool performances have largely been focused on isolate genomes with relatively little contamination.

One reassuring final point is that although different approaches to identify HGT in isolate genomes identify mainly nonoverlapping sets of genes (Ragan et al. 2006), the genes identified tend to be of similar functions (Lawrence and Roth 1996; Beiko et al. 2005; Cordero and Hogeweg 2009; Kanhere and Vingron 2009). Genes related to mobile elements, central intermediary metabolism, amino acid biosynthesis, and energy metabolism are enriched in gene sets identified as horizontally transferred (Jain et al. 1999; Beiko et al. 2005). In contrast, information-processing genes such as ribosomal proteins are less commonly identified as horizontally transferred (Beiko et al. 2005). There are exceptions to this rule, for example, genes related to translation have been found to be commonly horizontally transferred between bacteria, but not between kingdoms (Kanhere and Vingron 2009). These recurrent observations may be related to functions that rely on fewer gene families and regulatory partners being easier to transfer (Lawrence and Roth 1996; Aris-Brosou 2005). In addition, the widespread transfer of metabolism genes is likely related to strong selection for survival in novel environments with limited resources (Lawrence 2001) although the exact genes are environment-specific (Smillie et al. 2011). Identifying putative HGT events enriched for the above functional categories has been previously used as validation that an approach is working (Beiko et al. 2005), which would also be an important output to compare in future evaluations of HGT-detection approaches.

MAG Quality Control

The major challenge facing the identification of all the genic events described above is poor metagenome assemblies. This issue has recently been commented upon in the context of low-quality MAGs being added into public genome repositories (Shaiber and Eren 2019). Either composite assemblies of multiple taxa or incomplete genomes missing genes of interest could result in incorrect inferences of HGT. One extreme example is of the tardigrade genome, which was falsely identified as having 17% of genes acquired through HGT due to contaminant sequences within the assembly (Koutsovoulos et al. 2016). Such false inferences are more likely in metagenome assemblies compared with genome assemblies due largely to the challenge of distinguishing many organisms at different abundances (Ayling et al. 2019). Misassemblies can

also affect the detection of other genic events as well. For instance, repetitive regions of assemblies are difficult to resolve with current short-read sequencing (Chin et al. 2013), which can make duplication events difficult to identify. Due to these challenges, an understanding of the workflows for generating MAGs is needed. Here, we briefly outline the current approaches and issues in metagenome assembly to give the reader a starting point.

There are many metagenome assembly tools currently available, which are predominately based on De Bruijn graphs of overlapping k-mers (Vollmers et al. 2017; Ayling et al. 2019). The outputs of these tools are assembled contigs, which typically vary in length from ~500 bp to near-complete genomes. Some of the most popular freely available assembly tools are MetaSPAdes (Nurk et al. 2017), Ray Meta (Boisvert et al. 2012), Omega (Haider et al. 2014), IDBA-UD (Peng et al. 2012), and Megahit (Li et al. 2015). Choice of assembly tool can have a major influence on the resulting assembled contigs, and so careful consideration needs to be taken at this stage. An independent evaluation of these and other methods found that MetaSPAdes performed best overall with the caveat that it may not be appropriate for distinguishing highly similar genomes (Vollmers et al. 2017). However, no assembly tool performed best across all environments and it was suggested that the best choice of assembly tool depends on the study environment and research question.

Contig binning, where contigs from the same species or strain are grouped, is another key step when generating MAGs. Binning approaches typically group contigs based on sequence composition (e.g., GC or tetranucleotide content) and similar coverage of mapped reads (Ayling et al. 2019). The most popular freely available binning tools are CONCOCT (Aneberg et al. 2014), MaxBin2 (Wu et al. 2016), and MetaBAT (Kang et al. 2015). As above, the choice of binning software can have drastic effects on the resulting MAGs (Meyer et al. 2018). One partial solution to this issue is to run multiple binning tools and use the software Das Tool (Sieber et al. 2018) to identify the consensus output, which has been shown to produce high-quality bins (Meyer et al. 2018).

Evaluating the quality of MAGs is a crucial step once the final contig bins have been generated and guidelines for how to categorize MAGs based upon quality metrics have recently been established (Bowers et al. 2017). The two key metrics are completeness and contamination, which are based on the counts of USCGs identified in an assembly. Completeness is measured based on the proportion of USCGs identified in an assembly and contamination is defined as the proportion of USCGs found more than once in an assembly. Hard cutoffs for these metrics have been suggested for categorizing the overall quality of a MAG, for instance high-quality draft MAGs are defined as being >90% complete with <5% contamination (Bowers et al. 2017). CheckM (Parks et al. 2015) and BUSCO (Simão et al. 2015) are two tools that will estimate

the completeness and contamination of prokaryotic assemblies and BUSCO can also be used to evaluate eukaryotic assemblies. Determining strain heterogeneity, the degree of contamination due to different strains, within an assembly is also important, which can be measured using CheckM or alternatively custom methods to identify polymorphisms in an assembly (Pasolli et al. 2019). An assembly with high strain heterogeneity can still be useful but should be considered differently than an assembly of a single strain.

Importantly, when reporting gene gain and loss events in a set of MAGs it would be important to also report the estimated completeness and contamination within these MAGs. In particular, it would be important to establish within a given MAG that more gene gain and loss events were inferred than are expected given the two quality scores. Ideally, manual validation of inferred gene gain and loss events would also be performed upon assemblies. At minimum this validation would include visually assessing the read coverage across an assembly at the site of the inferred gene gain or loss event. In practice, manually validating many events in this way would not be feasible, but it could be performed for a representative set.

Potential Avenues to Improve Metagenome Assemblies

Several recent technologies have been developed which potentially could result in improved MAGs (fig. 3). These technologies include new long-read sequencing approaches, metagenomics chromosome conformation capture, barcoding reads from the same genomic fragment, and optical mapping of short sequences along genomes to inform assembly.

There are two promising long-read technologies currently available: Single-Molecule Real-Time (SMRT; Pacific Biosciences [McCarthy 2010]) and Oxford Nanopore (Mikheyev and Tin 2014) sequencing. SMRT sequencing involves binding a custom DNA polymerase with bound DNA to be sequenced at the bottom of a zero-mode waveguide (Slatko et al. 2018) (fig. 3A). Fluorescently labeled nucleotides are added to the growing chain, which enables each base to be identified as it is added. The key advantage of this approach is that long reads typically in the range of 10–15 kb (but ranging up to 50 kb or larger) can be produced (Slatko et al. 2018). SMRT sequencing reads generally contain 11–14% incorrect bases, but consensus sequences between overlapping sequences can be used for correction because the errors are random (Roberts et al. 2013). In contrast, short-read sequencing approaches result in nonrandom errors, which are more difficult to correct. This approach has been used mainly with hybrid approaches with short-read sequencing to make improved assemblies of isolate genomes (Koren et al. 2012). However, algorithms have been developed to error-correct the reads so that they can be used alone for high-quality assembly (Chin et al. 2013). Importantly, these

long reads are better able to sequence regions that are problematic with short-read approaches, including repetitive regions. New SMRT sequencing approaches are being developed, including circular consensus sequencing, which provides higher accuracy through repeated sequencing of the same circularized fragment of 500–2,500 nucleotides in length. This approach has been shown to result in improved assembled contigs compared with Illumina HiSeq sequencing on the same samples (Frank et al. 2016).

Nanopore sequencing refers to passing a strand of DNA through a nanopore and measuring the changing current, which differs depending on the bases passing through. This method also results in extremely long reads, typically in the range of 13–20 kb, and is rapidly improving in throughput (Tyson et al. 2018). The main down-side of this approach is the high error rate, which can range up to 40% of bases being incorrect (Laver et al. 2015). Nonetheless, this technology has been used to successfully assemble the *E. coli* K-12 mG1655 genome at 99.5% base accuracy by first making multiple-sequence alignments of nanopore reads to correct read errors (Loman et al. 2015). This technology is especially useful for resolving large repetitive regions and merging contigs derived from short-read data, as was recently demonstrated through improvements to the *Caenorhabditis elegans* reference genome (Tyson et al. 2018).

These evaluations of long-read sequencing technology are promising and both technologies will likely confer similar improvements to MAGs in the future. At least one example of improved quality of the continuity of MAGs has been demonstrated based on nanopore sequencing of a complex bio-reactor community (Arumugam et al. 2019). We expect that many more of these examples will be published as long-read sequencing becomes more widely available. However, despite this promising example, MGS data present many novel challenges and will take additional work to integrate into most bioinformatics pipelines for processing long-read data. For instance, both correcting read errors and resolving repetitive regions with nanopore sequencing would be considerably more challenging with MGS data due to the added complication of strain variation and homologous DNA between different species within the same community. MGS-specific software for processing long-read data is beginning to become available (Kolmogorov et al. 2019), but clear best practices remain to be determined.

Alternative approaches that could improve MAG quality are based on binning genomes prior to or independent of sequencing. One exciting development is Hi-C sequencing, which is an extension of chromosome conformation capture sequencing (Belton et al. 2012). This approach involves cross-linking DNA with formaldehyde, followed by digestion, and then religation so that interacting DNA fragments are ligated together. The novel addition in Hi-C is that a biotin-labeled nucleotide is incorporated at ligation junctions, which makes it much easier to purify out chimeric ligations and identify 3D

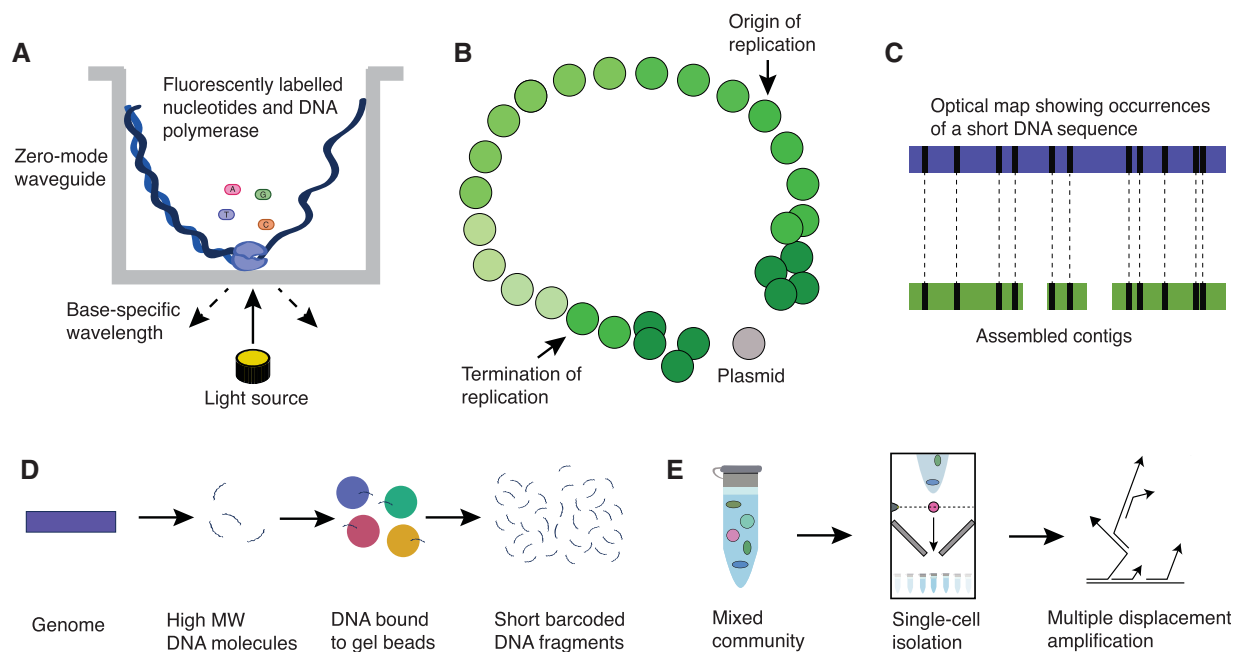


Fig. 3.—Promising technologies that could improve metagenome assembly. (A) Long-read sequencing as represented by single-molecule real-time (SMRT) sequencing, which takes place in a zero-mode waveguide. Fluorescently labeled nucleotides are added one at a time at the bottom of the well as the new strand of the input DNA is synthesized. The fluorescence of each added nucleotide is measured to determine the sequence. (B) Illustration of how relationships between contigs based on chromosome conformation capture can be visualized. This simplified illustration is based on the previously determined relationship between *Escherichia coli* contigs (Marbouty et al. 2014). The darker the shade of green, the higher the contact frequency of contigs. This visualization displays how the contig genomic ordering can be determined through chromosome conformation capture. The contig contact map can be used to improve the scaffolding step of the genome assembly. (C) Diagram illustrating principle of optical maps (blue) improving genomic assemblies (green). Solid black bars indicate occurrences of a short DNA sequence along the genome, which can be used to order contigs and correct assembly errors. (D) Simplified protocol for barcoding genomic fragments so that reads originating from the same high molecular weight (MW) DNA molecule can be identified. (E) Key steps required before single-cell sequencing. Individual cells need to be isolated using one of several techniques (e.g., flow cytometry as shown in this panel) and then whole-genome amplification is conducted using multiple displacement amplification. The small arrows indicate amplified regions.

interactions. This approach has mainly been used to identify long-range interactions, such as between enhancers and promoters (Ron et al. 2017) and to map genome conformational dynamics (Lieberman-Aiden et al. 2009). However, this method can also be exploited to improve genome assemblies by building probability maps of genetic interactions (Burton et al. 2013) (fig. 3B), based on the observation that intrachromosomal interactions are much more common than interchromosomal interactions, even at long distances (Lieberman-Aiden et al. 2009). As a proof-of-concept, this approach has been used to distinguish genomes within mock communities and contributed to the assembly of high-quality MAGs of bacterial, archaeal, and fungal genomes (Burton et al. 2014). In addition, Hi-C sequencing has been used to supplement MGS assembly of natural communities in river sediment (Marbouty et al. 2014) and cow rumen samples (Stewart et al. 2018).

Optical mapping is another approach that has been shown to improve genome assembly quality (Hastie et al. 2013). The most recent implementation of this approach is provided by the company BioNano (<https://bionanogenomics.com>; last accessed August 27, 2019). Their method involves annealing

fluorescent probes that bind specific short sequences in DNA. The DNA is then passed through narrow chambers that require a DNA molecule to be passed through in a straight line. The location of the fluorescent probes along genomic fragments is captured as DNA passes through the chamber. This approach has recently been applied to human samples to identify SVs and translocation events (Mak et al. 2016). Optical mapping data could potentially be integrated with MGS data to improve MAG qualities by assigning contigs to the same genome and to order and rearrange contigs within a genome (fig. 3C). However, this approach has not been applied to samples from natural communities and it remains unclear what challenges would be faced. A proof-of-concept of this technology applied to mock and natural communities is required for this approach to be evaluated properly.

Another promising approach involves barcoding reads from the same genome prior to MGS (fig. 3D). This technology, developed by 10x Genomics (<https://www.10xgenomics.com>; last accessed August 27, 2019), results in sets of reads that are derived from the same genomic fragment. This information is especially useful for resolving SVs and repeats and

for phasing variants (Bishara et al. 2015; Zheng et al. 2016). These “read clouds” could be leveraged in an analogous way to Hi-C sequencing data to produce improved MAGs. It was recently shown that this datatype can be used to generate high-quality MAGs from mock and natural communities with a barcode-aware assembler called Athena (Bishara et al. 2018). Further published work confirming this finding is required, but this promising result highlights that leveraging “read clouds” could be a straight-forward method for improving MAG qualities.

Lastly, single-cell metagenomics has recently been suggested as an improved approach for isolating individual genomes from mixed communities (Xu and Zhao 2018). This technique involves first isolating individual cells in a sample, extracting the DNA, and performing whole-genome amplification before conducting library preparation and sequencing (fig. 3E). The whole-genome amplification step can result in a high proportion of chimeric reads (Lasken and Stockwell 2007) and uneven read coverage (Xu and Zhao 2018), which can complicate genome assembly. Nonetheless, single-cell metagenomics has been performed successfully, especially when identifying phage and the corresponding bacterial host genomes (Labonté et al. 2015; Munson-Mcgee et al. 2018). In addition, single-cell metagenomics has been integrated with MGS on numerous occasions to improve MAGs (Dupont et al. 2012; Dong et al. 2016; Ji et al. 2017; Yu et al. 2017). Many questions remain regarding the best practices of single-cell metagenomics and the feasibility at high throughput (Xu and Zhao 2018), but this technology is an extremely promising approach to improve MAG quality.

Conclusions

Several bioinformatics approaches have been applied for identifying HGT events specifically in MGS data, but these approaches have not been extensively benchmarked and there are no clear best practices. These approaches have also been largely custom bioinformatics pipelines that are difficult to compare across studies, but there are several recently developed methods like WAAFL and MetaCHIP that are now available as stand-alone tools. In addition, currently there are limited bioinformatics approaches to identify gene loss, de novo genes, and gene duplications in MGS data sets. Although inferring these events will become easier as the quality of MAGs improves, there is still a need to develop methods to detect these events in poorly assembled MGS data sets. These methods are needed to better analyze existing MGS data sets and also to study communities with high richness (e.g., soil samples), for which it will likely remain unfeasible to sequence at sufficient coverage to produce many high-quality MAGs in the near future.

There are also many open questions about how to scan for HGT in assembled genomes. Currently, genomic context and potential transfer mechanisms are not directly integrated into

HGT-detection pipelines. Automatically identifying corroborating evidence for a transfer, such as the presence of nearby prophage sequences, could help identify recent HGT events, which has been previously argued (Zaneveld et al. 2008). The feasibility of such an approach and whether it would improve HGT event identification accuracy is unclear. Similarly, taxonomic-specific features such as DNA-uptake sequences in DNA imported through transformation (Davidsen et al. 2004) could directly be used to inform inferences.

Regardless of which bioinformatics approach is used, inferences of gene gain and loss in MGS data sets will continue to improve as higher-quality MAGs are produced. The promising technologies outlined above are still in the early stages of application and there remain many open questions. For instance, whether optical mapping can be accurately applied to mixed communities remains unclear. Additional validation of this approach in conjunction with MGS data is necessary to determine whether it would actually improve MGS assemblies. In addition, there are several limitations specific to individual technologies, such as long-read sequencing and Hi-C requiring larger biomass samples. Benchmarking of all of these promising methods is required on natural communities of varying richness, complexity, and biomass to evaluate whether these methods should be differentially applied depending on sample-type or whether general best practices could be developed using only a subset of approaches.

Acknowledgments

This manuscript was originally written as part of G.M.D.’s PhD comprehensive exam. We would like to thank all members of the exam committee for their time and feedback: Dr Robert Beiko, Dr Joseph Bielawski, Dr Brent Johnston, Dr John Rohde, Dr Andrew Stadnyk, and Dr Johan Van Limbergen. We would also like to thank Dr André Comeau, Casey Jones, Jacob Nearing, and Patrick Slaine who provided feedback on the manuscript. G.M.D. is supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Alexander Graham Bell Canada Graduate Scholarship (Doctoral) and M.G.I.L. is supported by an NSERC Discovery Grant.

Literature Cited

- Akhter S, Aziz RK, Edwards RA. 2012. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic Acids Res.* 40(16):e126.
- Akiba T, Koyama K, Ishiki Y, Kimura S, Fukushima T. 1960. On the mechanism of the development of multiple-drug-resistant clones of *Shigella*. *Jpn J Microbiol.* 4:219–227.
- Albalat R, Cañestro C. 2016. Evolution by gene loss. *Nat Rev Genet.* 17(7):379–391.
- Alneberg J, et al. 2014. Binning metagenomic contigs by coverage and composition. *Nat Methods.* 11(11):1144–1146.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. *J Mol Biol.* 215(3):403–410.

- Amann RL, Ludwig W, Schleifer K-H. 1995. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol Rev.* 59(1):143–169.
- Amgarten D, Braga LPP, da Silva AM, Setubal JC. 2018. MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Front Genet.* 9:304.
- Aris-Brosou S. 2005. Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. *Mol Biol Evol.* 22(2):200–209.
- Arndt D, et al. 2016. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44(W1):W16–W21.
- Arumugam K, et al. 2019. Annotated bacterial chromosomes from frame-shift-corrected long-read metagenomic data. *Microbiome* 7(1):61.
- Aury JM, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444(7116):171–178.
- Ayling M, Clark MD, Leggett RM. 2019. New approaches for metagenome assembly with short reads. *Brief Bioinform.* pii: bbz020.
- Bansal MS, Kellis M, Kordi M, Kundu S. 2018. RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics* 34(18):3214–3216.
- Bao G, Wang M, Doak TG, Ye Y. 2015. Strand-specific community RNA-seq reveals prevalent and dynamic antisense transcription in human gut microbiota. *Front Microbiol.* 6:896.
- Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci USA.* 102(40):14332–14337.
- Belton JM, et al. 2012. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods.* 58(3):268–276.
- Bengtsson-Palme J, Boulund F, Fick J, Kristiansson E, Joakim Larsson DG. 2014. Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. *Front Microbiol.* 5:648.
- Bertelli C, et al. 2017. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res.* 45(W1):W30–W35.
- Bishara A, et al. 2015. Read clouds uncover variation in complex regions of the human genome. *Genome Res.* 25(10):1570–1580.
- Bishara A, et al. 2018. High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat Biotechnol.* 36(11):1067–1080.
- Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* 13(12):R122.
- Boucher Y, et al. 2003. Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet.* 37:283–328.
- Bowers RM, et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol.* 35(8):725–731.
- Brantl S. 2007. Regulatory mechanisms employed by cis-encoded antisense RNAs. *Curr Opin Microbiol.* 10(2):102–109.
- Brown TB, et al. 2015. Unusual biology across a group comprising more than 15% of domain bacteria. *Nature* 523(7559):208–211.
- Burton JN, Liachko I, Dunham MJ, Shendure J. 2014. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3 (Bethesda)* 4(7):1339–1346.
- Burton JN, et al. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol.* 31(12):1119–1125.
- Cai J, Zhao R, Jiang H, Wang W. 2008. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* 179(1):487–496.
- Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brüssow H. 2003. Phage as agents of lateral gene transfer. *Curr Opin Microbiol.* 6(4):417–424.
- Carattoli A, et al. 2014. *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother.* 58(7):3895–3903.
- Chen I, Dubnau D. 2004. DNA uptake during bacterial transformation. *Nat Rev Microbiol.* 2(3):241–249.
- Chin CS, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods.* 10(6):563–569.
- Cingolani P, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6(2):80–92.
- Clarke GDP, Beiko RG, Ragan MA, Charlebois RL. 2002. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J Bacteriol.* 184(8):2072–2080.
- Claverys JP, Martin B, Polard P. 2009. The genetic transformation machinery: composition, localization, and mechanism. *FEMS Microbiol Rev.* 33(3):643–656.
- Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T. 2010. GLOOME: gain loss mapping engine. *Bioinformatics* 26(22):2914–2915.
- Comeau AM, Krisch HM. 2005. War is peace—dispatches from the bacterial and phage killing fields. *Curr Opin Microbiol.* 8(4):488–494.
- Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr Opin Plant Biol.* 19:91–98.
- Cordero OX, Hogeweg P. 2009. The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proc Natl Acad Sci USA.* 106(51):21748–21753.
- Costea PI, et al. 2017. metaSNV: a tool for metagenomic strain level analysis. *PLoS One* 12(7):e0182392.
- Dahlberg C, et al. 1998. Interspecies bacterial conjugation by plasmids from marine environments visualized by gfp expression. *Mol Biol Evol.* 15(4):385–390.
- David LA, Alm EJ. 2011. Rapid evolutionary innovation during an Archaeal genetic expansion. *Nature* 469(7328):93–96.
- Davidsen T, et al. 2004. Biased distribution of DNA uptake sequences towards genome maintenance genes. *Nucleic Acids Res.* 32(3):1050–1058.
- Denef VJ, Banfield JF. 2012. *In situ* evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science* 336(6080):462–466.
- Denef VJ, et al. 2007. Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* 446(7135):537–541.
- Deschavanne P, Filipinski J. 1995. Correlation of GC content with replication timing and repair mechanisms in weakly expressed *E. coli* genes. *Nucleic Acids Res.* 23(8):1350–1353.
- Ding W, Baumdicker F, Neher RA. 2018. panX: pan-genome analysis and exploration. *Nucleic Acids Res.* 46(1):e5.
- Domergue R, et al. 2005. Nicotinic acid limitation regulates silencing of *Candida* adhesins during UTI. *Science* 308(5723):866–870.
- Domingues S, Nielsen KM. 2017. Membrane vesicles and horizontal gene transfer in prokaryotes. *Curr Opin Microbiol.* 38:16–21.
- Dong H, et al. 2016. Single-cell-genomics-facilitated read binning of candidate phylum EM19 genomes from geothermal spring metagenomes. *Appl Environ Microbiol.* 82(4):992–1003.
- Drosophila* 12 Genomes Consortium, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450:203–218.
- Dubey GP, Ben-Yehuda S. 2011. Intercellular nanotubes mediate bacterial communication. *Cell* 144(4):590–600.
- Dunn KA, Bielawski JP, Ward TJ, Urquhart C, Gu H. 2009. Reconciling ecological and genomic divergence among lineages of *Listeria* under an ‘extended mosaic genome concept’. *Mol Biol Evol.* 26(11):2605–2615.
- Dupont CL, et al. 2012. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* 6(6):1186–1199.
- Etienne L, Hahn BH, Sharp PM, Matsen FA, Emerman M. 2013. Gene loss and adaptation to hominids underlie the ancient origin of HIV-1. *Cell Host Microbe* 14(1):85–92.

- Falush D, et al. 2001. Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proc Natl Acad Sci USA*. 98(26):15056–15061.
- Fellner L, et al. 2015. Evidence for the recent origin of a bacterial protein-coding, overlapping orphan gene by evolutionary overprinting. *BMC Evol Biol*. 15:283.
- Ficht TA. 2011. Bacterial exchange via nanotubes: lessons learned from the history of molecular biology. *Front Microbiol*. 2:179.
- Finkel SE, Kolter R. 2001. DNA as a nutrient: novel role for bacterial competence gene homologs. *J Bacteriol*. 183(21):6288–6293.
- Force A, et al. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151(4):1531–1545.
- Frank JA, et al. 2016. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci Rep*. 6:25373.
- Garretto A, Hatzopoulos T, Putonti C. 2019. virMine: automated detection of viral sequences from complex metagenomic samples. *PeerJ* 7:e6695.
- Gevers D, Vandepoel K, Simillion C, Van de Peer Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol*. 12:145–148.
- Grohmann E, Muth G, Espinosa M. 2003. Conjugative plasmid transfer in Gram-positive bacteria. *Microbiol Mol Biol Rev*. 67(2):277–301.
- Grüll MP, Mulligan ME, Lang AS. 2018. Small extracellular particles with big potential for horizontal gene transfer: membrane vesicles and gene transfer agents. *FEMS Microbiol Lett*. 365:fny192.
- Guindon S, Perrière G. 2001. Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes. *Mol Biol Evol*. 18(9):1838–1840.
- Guo J, et al. 2015. Horizontal gene transfer in an acid mine drainage microbial community. *BMC Genomics*. 16:496.
- Hahn MW, Han MV, Han SG. 2007. Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet*. 3(11):e197–2146.
- Haider B, et al. 2014. Omega: an overlap-graph de novo assembler for metagenomics. *Bioinformatics* 30(19):2717–2722.
- Harding T, Roger AJ, Simpson A. 2017. Adaptations to high salt in a halophilic protist: differential expression and gene acquisitions through duplications and gene transfers. *Front Microbiol*. 8:944.
- Hasan MS, et al. 2012. GIST: genomic island suite of tools for predicting genomic islands. *Bioinformatics* 8(4):203–205.
- Hastie AR, et al. 2013. Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex *Aegilops tauschii* genome. *PLoS One* 8(2):e55864.
- Hatfull GF, Hendrix RW. 2011. Bacteriophages and their genomes. *Curr Opin Virol*. 1(4):298–303.
- Hemme CL, et al. 2010. Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. *ISME J*. 4(5):660–672.
- Hemme CL, et al. 2016. Lateral gene transfer in a heavy metal-contaminated-groundwater microbial community. *MBio* 7(2):e02234-15.
- Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet*. 6(9):e1001107.
- Hottes AK, et al. 2013. Bacterial adaptation through loss of function. *PLoS Genet*. 9(7):e1003617.
- Hücker SM, et al. 2018. A novel short L-arginine responsive protein-coding gene (*laob*) antiparallel overlapping to a CadC-like transcriptional regulator in *Escherichia coli* O157: H7 Sakai originated by overprinting. *BMC Evol Biol*. 18(1):21.
- Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA*. 96(7):3801–3806.
- Ji P, Zhang Y, Wang J, Zhao F. 2017. MetaSort untangles metagenome assembly by reducing microbial community complexity. *Nat Commun*. 8:14306.
- Karneva OK, Ward NL. 2014. Reconciliation approaches to determining HGT, duplications, and losses in gene trees. In: Goodfellow M, Sutcliffe I, Chun J (eds.) *New Approaches to Prokaryotic Systematics*, Vol. 41 of *Methods in Microbiology*. Amsterdam, Netherlands: Academic Press, 183–199.
- Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28(1):27–30.
- Kang DD, Froula J, Egan R, Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*. 3:e1165.
- Kanhere A, Vingron M. 2009. Horizontal gene transfers in prokaryotes show differential preferences for metabolic and translational genes. *BMC Evol Biol*. 9:9.
- Karlin S, Burge C. 1995. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*. 11(7):283–290.
- Kielak A, et al. 2010. Phylogenetic and metagenomic analysis of *Verrucomicrobia* in former agricultural grassland soil. *FEMS Microbiol Ecol*. 71(1):23–33.
- Kolmogorov M, et al. 2019. metaFlye: scalable long-read metagenome assembly using repeat graphs. *bioRxiv*.
- Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc Biol Sci*. 279(1749):5048–5057.
- Koren S, et al. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol*. 30(7):693–700.
- Koskella B, Meaden S. 2013. Understanding bacteriophage specificity in natural microbial communities. *Viruses* 5(3):806–823.
- Koski LB, Morton RA, Golding GB. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol*. 18(3):404–412.
- Koutsovoulos G, et al. 2016. No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc Natl Acad Sci USA*. 113(18):5053–5058.
- Krawczyk PS, Lipinski L, Dziembowski A. 2018. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res*. 46(6):e35.
- Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. 2010. New dimensions of the virus world discovered through metagenomics. *Trends Microbiol*. 18(1):11–19.
- Kristiansson E, et al. 2011. Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. *PLoS One* 6(2):e17038.
- Kurokawa K, et al. 2007. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res*. 14(4):169–181.
- Labonté JM, et al. 2015. Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J*. 9(11):2386–2399.
- Lacroix B, Citovsky V. 2016. A functional bacterium-to-plant DNA transfer machinery of *Rhizobium etli*. *PLoS Pathog*. 12(3):e1005502.
- Lang AS, Beatty JT. 2007. Importance of widespread gene transfer agent genes in α -proteobacteria. *Trends Microbiol*. 15(2):54–62.
- Lang AS, Westbye AB, Beatty JT. 2017. The distribution, evolution, and roles of gene transfer agents in prokaryotic genetic exchange. *Annu Rev Virol*. 4(1):87–104.
- Langille MGI, Hsiao WWL, Brinkman F. 2010. Detecting genomic islands using bioinformatics approaches. *Nat Rev Microbiol*. 8(5):373–382.
- Lanka E, Wilkins BM. 1995. DNA processing reactions in bacterial conjugation. *Annu Rev Biochem*. 64:141–169.
- Lasken RS, Stockwell TB. 2007. Mechanism of chimera formation during the multiple displacement amplification reaction. *BMC Biotechnol*. 7(1):19.

- Laver T, et al. 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif.* 3:1–8.
- Lawrence JG. 2001. Catalyzing bacterial speciation: correlating lateral transfer. *Syst Biol.* 50(4):479–496.
- Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol.* 44(4):383–397.
- Lawrence JG, Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A.* 95(16):9413–9417.
- Lawrence JG, Roth JR. 1996. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* 143(4):1843–1860.
- Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31(10):1674–1676.
- Lieberman-Aiden E, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289–293.
- Llorens-Marès T, et al. 2017. Speciation and ecological success in dimly lit waters: horizontal gene transfer in a green sulfur bacteria bloom unveiled by metagenomic assembly. *ISME J.* 11(1):201–211.
- Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods.* 12(8):733–735.
- Lu B, Leong HW. 2016. Computational methods for predicting genomic islands in microbial genomes. *Comput Struct Biotechnol J.* 14:200–206.
- Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151–1155.
- Mak ACY, et al. 2016. Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics* 202(1):351–362.
- Malki K, et al. 2015. Bacteriophages isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. *Virology* 527:159–164.
- Marbouty M, et al. 2014. Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *Elife* 3:e03318.
- Martiny AC. 2019. High proportions of bacteria are culturable across major biomes. *ISME J.* 13(8):2125–2128.
- Mathew AG, Cissell R, Liamthong S. 2007. Antibiotic resistance in bacteria associated with food animals: a United States perspective of livestock production. *Foodborne Pathog Dis.* 4(2):115–133.
- McCarthy A. 2010. Third Generation DNA Sequencing: Pacific Biosciences' Single Molecule Real Time Technology. *Chem Biol.* 17(7):675–676.
- McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos Trans R Soc Lond B Biol Sci.* 370(1678):20140332.
- Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol.* 222(4):851–856.
- Meehan CJ, Beiko RG. 2012. Lateral gene transfer of an ABC transporter complex between major constituents of the human gut microbiome. *BMC Microbiol.* 12:248.
- Mendler K, et al. 2019. AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res.* 47(9):4442–4448.
- Meyer F, et al. 2018. AMBER: Assessment of Metagenome BinnERS. *Gigascience* 7(6):1–8.
- Mikheyev AS, Tin M. 2014. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour.* 14(6):1097–1102.
- Munson-Mcgee JH, et al. 2018. A virus or more in (nearly) every cell: ubiquitous networks of virus-host interactions in extreme environments. *ISME J.* 12(7):1706–1714.
- Nayfach S, et al. 2015. Automated and accurate estimation of gene family abundance from shotgun metagenomes. *PLoS Comput Biol.* 11(11):e1004573.
- Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27(5):824–834.
- Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial evolution. *Nature* 405(6784):299–304.
- Ohno S. 1984. Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. *Proc Natl Acad Sci USA.* 81(8):2421–2425.
- Ohno S. 1970. *Evolution by gene duplication.* New York: Springer.
- Omelchenko MV, Makarova KS, Wolf YI, Rogozin IB, Koonin EV. 2003. Evolution of mosaic operons by horizontal gene transfer and gene displacement *in situ.* *Genome Biol.* 4(9):R55.
- Page AJ, et al. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31(22):3691–3693.
- Palenik B, Ren Q, Tai V, Paulsen IT. 2009. Coastal *Synechococcus* metagenome reveals major roles for horizontal gene transfer and plasmids in population diversity. *Environ Microbiol.* 11(2):349–359.
- Papke RT, Koenig JE, Rodríguez-Valera F, Doolittle WF. 2004. Frequent recombination in a saltern population of *Halorubrum.* *Science* 306(5703):1928–1929.
- Parkhill J, et al. 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica.* *Nat Genet.* 35(1):32–40.
- Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25(7):1043–1055.
- Parks DH, et al. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol.* 2(11):1533–1542.
- Parnell JJ, et al. 2010. Functional biogeography as evidence of gene transfer in hypersaline microbial communities. *PLoS One* 5(9):e12919.
- Pasolli E, et al. 2019. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 176(3):649–662.
- Peng Y, Leung HCM, Yiu SM, Chin F. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28(11):1420–1428.
- Podell S, Gaasterland T. 2007. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol.* 8(2):R16.
- Probst AJ, Banfield JF. 2018. Homologous recombination and transposon propagation shape the population structure of an organism from the deep subsurface with minimal metabolism. *Genome Biol Evol.* 10(4):1115–1119.
- Puigbo P, Wolf YI, Koonin EV. 2009. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J Biol.* 8(6):59.
- Ragan MA, Charlebois RL. 2002. Distributional profiles of homologous open reading frames among bacterial phyla: implications for vertical and lateral transmission. *Int J Syst Evol Microbiol.* 52(Pt 3):777–787.
- Ragan MA, Harlow TJ, Beiko RG. 2006. Do different surrogate methods detect lateral genetic transfer events of different relative ages? *Trends Microbiol.* 14(1):4–8.
- Rausch T, et al. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28(18):i333–339.
- Ravenhall M, Škunca N, Lassalle F, Dessimoz C. 2015. Inferring horizontal gene transfer. *PLoS Comput Biol.* 11(5):e1004095.
- Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. 2017. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 5(1):69.
- Riesenfeld CS, Schloss PD, Handelsman J. 2004. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet.* 38:525–552.
- Roberts RJ, Carneiro MO, Schatz MC. 2013. The advantages of SMRT sequencing. *Genome Biol.* 14(6):2–5.

- Ron G, Globerson Y, Moran D, Kaplan T. 2017. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. *Nat Commun.* 8:2237.
- Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. *PeerJ.* 3:e985.
- Rozov R, et al. 2017. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* 33(4):475–482.
- Rua CPJ, et al. 2018. Microbial and functional biodiversity patterns in sponges that accumulate bromopyrrole alkaloids suggest horizontal gene transfer of halogenase genes. *Microb Ecol.* 76(3):825–838.
- Sabath N, Wagner A, Karlin D. 2012. Evolution of viral proteins originated de novo by overprinting. *Mol Biol Evol.* 29(12):3767–3780.
- Schlötterer C. 2015. Genes from scratch—the evolutionary fate of de novo genes. *Trends Genet.* 31(4):215–219.
- Schoenfeld TW, et al. 2013. Lateral gene transfer of Family A DNA polymerases between thermophilic viruses, Aquificae, and Apicomplexa. *Mol Biol Evol.* 30(7):1653–1664.
- Scholz M, et al. 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods.* 13(5):435–438.
- Schuster-Böckler B, Conrad D, Bateman A. 2010. Dosage sensitivity shapes the evolution of copy-number varied regions. *PLoS One* 5(3):e9474.
- Shaiber A, Eren AM. 2019. Composite metagenome-assembled genomes reduce the quality of public genome repositories. *MBio* 10(3):e00725–19.
- Sieber CMK, et al. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol.* 3(7):836–843.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–3212.
- Slatko BE, Gardner AF, Ausubel FM. 2018. Overview of next-generation sequencing technologies. *Curr Protoc Mol Biol.* 122(1):e59.
- Smillie CS, et al. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480(7376):241–244.
- Song W, Steensen K, Thomas T. 2017. HgtSIM: a simulator for horizontal gene transfer (HGT) in microbial communities. *PeerJ.* 5:e4015.
- Song W, Wemheuer B, Zhang S, Steensen K, Thomas T. 2019. MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. *Microbiome* 7(1):36.
- Soo RM, Hemp J, Parks DH, Fischer WW, Hugenholtz P. 2017. On the origins of oxygenic photosynthesis and aerobic respiration in *Cyanobacteria*. *Science* 355(6332):1436–1440.
- Spellberg B, et al. 2008. The epidemic of antibiotic-resistant infections: a call to action for the medical community from the Infectious Diseases Society of America. *Clin Infect Dis.* 46(2):155–164.
- Stewart RD, et al. 2018. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat Commun.* 9(1):870.
- Sueoka N. 1961. Compositional correlation between deoxyribonucleic acid and protein. *Cold Spring Harb Symp Quant Biol.* 26:35–43.
- Szczepanowski R, et al. 2009. Detection of 140 clinically relevant antibiotic-resistance genes in the plasmid metagenome of wastewater treatment plant bacteria showing reduced susceptibility to selected antibiotics. *Microbiology.* 155(Pt 7):2306–2319.
- Tamames J, Moya A. 2008. Estimating the extent of horizontal gene transfer in metagenomic sequences. *BMC Genomics.* 9:136.
- Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet.* 12(10):692–702.
- Thanassi DG, Suh GSB, Nikaido H. 1995. Role of outer membrane barrier in efflux-mediated tetracycline resistance of *Escherichia coli*. *J Bacteriol.* 177(4):998–1007.
- Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol.* 3(9):711–721.
- Torti A, Lever MA, Jørgensen BB. 2015. Origin, dynamics, and implications of extracellular DNA pools in marine sediments. *Mar Genomics.* 24:185–196.
- Tyson GW, et al. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428(6978):37–43.
- Tyson JR, et al. 2018. MiniION-based long-read sequencing and assembly extends the *Caenorhabditis elegans* reference genome. *Genome Res.* 28(2):266–274.
- Van Oss SB, Carvunis A-R. 2019. De novo gene birth. *PLoS Genet.* 15(5):e1008160.
- Vanderhaeghen S, Zehentner B, Scherer S, Neuhaus K, Ardern Z. 2018. The novel EHEC gene *asa* overlaps the TEGT transporter gene in antisense and is regulated by NaCl and growth phase. *Sci Rep.* 8(1):17875.
- Vollmers J, Wiegand S, Kaster A-K. 2017. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective—not only size matters! *PLoS One* 12(1):e0169662.
- Wang B. 2001. Limitations of compositional approach to identifying horizontally transferred genes. *J Mol Evol.* 53(3):244–250.
- Ward N, Moreno-Hagelsieb G. 2014. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? *PLoS One* 9(7):e101850.
- Weisbeek PJ, Borrias WE, Langeveld SA, Baas PD, Van Arkel GA. 1977. Bacteriophage phiX174: gene A overlaps gene B. *Proc Natl Acad Sci USA.* 74(6):2504–2508.
- Williams KP. 2002. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res.* 30(4):866–875.
- Wu YW, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32(4):605–607.
- Xu Y, Zhao F. 2018. Single-cell metagenomics: challenges and applications. *Protein Cell* 9(5):501–510.
- Yavaş G, Koyutürk M, Gould MP, McMahon S, LaFramboise T. 2014. DB2: a probabilistic approach for accurate detection of tandem duplication breakpoints using paired-end reads. *BMC Genomics.* 15:175.
- Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25(21):2865–2871.
- Yin X, Stotzky G. 1997. Gene transfer among bacteria in natural environments. *Adv Apply Microbiol* 45:153–212.
- Yu FB, et al. 2017. Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. *Elife* 6:e26580.
- Zaneveld JR, Nemergut DR, Knight R. 2008. Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns. *Microbiology.* 154(Pt 1):1–15.
- Zheng GXY, et al. 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol.* 34(3):303–311.
- Zhou F, Xu Y. 2010. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* 26(16):2051–2052.
- Zhu Q, Kosoy M, Dittmar K. 2014. HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. *BMC Genomics.* 15:717.
- Zojer M, et al. 2017. Variant profiling of evolving prokaryotic populations. *PeerJ.* 5:e2997.

Associate editor: Tal Dagan