THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Constructing genetic exchange communities among bacteria and archaea

Yingnan Cong

Bachelor in Management and Master in Management Science and Engineering

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2016

Institute for Molecular Bioscience

Abstract

Lateral genetic transfer (LGT) refers to several processes by which microbes can take up, maintain and often integrate into their own genome genetic material from other organisms. It is widely accepted that LGT plays an important role in the evolution of microbial genomes, and in the ability of these organisms to adapt and exploit new ecological opportunities. Computational methods have been applied to detect LGT since the 1990s. Most classical approaches to inferring LGT follow the steps of delineating sets of orthologous sequences, multiple sequence alignment, phylogenetic inference and then finding incongruities between the topology of this tree and a reference tree. Most of these steps are computationally hard, so these methods lack scalability for analysis of very large datasets. With the ongoing development of new sequencing technologies in recent years, more and more sequences are becoming available for study, necessitating the development of new methods to detect LGT on large datasets. With access to lateral events, we can generate LGT networks in which nodes represent DNA carriers such as genomes or plasmids, and edges represent LGT events. By analysis of these networks, we can delineate genetic exchange communities (GECs), groups of organisms that have transferred genetic material amongst themselves, and study their properties. This thesis has three aims: 1) design and implement a method to detect LGT with high efficiency and effectiveness which can identify directionality of transfer; 2) apply this new method on empirical datasets to evaluate its performance, and build LGT networks based on the detections; and 3) analyse the LGT networks and identify genetic exchange communities.

In Chapter 2, we develop an alignment-free method to detect LGT, based on term frequency – inverse document frequency (TF-IDF). TF-IDF is a concept from text mining, originally used to find the key words in a document. We treat genomes as documents and use *k*-mers (fixed-size short reads) to represent words. The genomes are arranged into groups, usually according to recognised biological relationships. If, in a sequence, we find a series of *k*-mers (separated from each other by no more than a gap of size *G*) that are infrequent within its own group, but frequent in a different group, then this segment is judged as lateral, with direction of transfer from the latter (donor) group into that (recipient) sequence. We tested this method on simulated datasets varying *k*, *G* and rates of nucleotide replacement within-group, between-group and post-LGT. We find that in many biologically relevant cases, the method performs effectively (precision and recall above 85%); it performs better if *k* is between 25 and 45, between-group distance is large, and within-group distance is small. We also compare our TF-IDF method with ALFY, another alignment-free method for LGT detection, on both simulated and empirical datasets (seven *Staphylococcus aureus* genomes). On the simulated datasets, TF-IDF exhibits slightly lower recall but much greater precision than ALFY. On the empirical dataset, TF-IDF finds all LGT events inferred by ALFY, as well as some other areas of interest including likely lateral regions containing antibiotic-resistance genes. TF-IDF runs much faster than ALFY on large datasets, but in the current implementation can be memory-intensive. These results establish TF-IDF as a competitive method for inferring LGT.

In Chapter 3, I apply TF-IDF to three empirical datasets (genomes of 27 *Escherichia coli* and *Shigella*; 110 enteric bacteria; and 143 bacteria and archaea) to investigate its performance on datasets of different evolutionary breadth. We study the dependence of the method on *k* and *G*, and identify optimal parameters for a range of realistic scenarios. We observe an abundance of lateral transfers among groups of *Escherichia coli* and *Shigella*, and found indications of more-ancient transfers, which are otherwise difficult to detect. In the enteric bacteria dataset, most of the LGT signal comes from exchanges between *E. coli* and *Shigella*, but we could nonetheless recognise a lower rate of LGT with the other groups (except *Yersinia*). Few LGT events could be inferred between different phyla in the prokaryote dataset, as expected. We map these lateral regions to genes, and use enrichment tests to determine which biological process annotations are over- or under-represented among these lateral genes.

In LGT networks, regions in which most nodes are interconnected with each other represent potential biological communities that exchange genetic material. In Chapter 4 we define cliques in LGT networks as genetic exchange communities (GECs). We are interested in the taxonomic and physiological nature of these GECs, and whether their members share common environments. Finding cliques (or near-cliques) in networks is an NP-hard problem; however, there exist several good heuristic methods for this, many of which are implemented in the software package GrAPPA). In this chapter we use GrAPPA to identify GECs in the datasets we studied in Chapter 3. By varying the parameter values of TF-IDF, we can identify phyla or classes that persist as members of GECs, and which are more transient in this sense. We then apply enrichment tests to identify the biological processes that underlie these GECs.

Overall, this project has introduced new capabilities, generated new understanding and opened new perspectives in our understanding of LGT among bacteria and archaea. Using the TF-IDF method we can detect LGT in large genome-scale datasets, and for the first time systematically infer the directionality of transfer. The concept of GEC sheds new light on the processes behind lateral transfer, and will allow researchers to better understand the mechanisms and conditions behind LGT.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Yingnan Cong, Yao-ban Chan, Mark A. Ragan. A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF (2016) *Scientific Reports* 6:30308. DOI: 10.1038/srep30308.

Yingnan Cong, Yao-ban Chan, Mark A. Ragan. Exploring lateral genetic transfer among microbial genomes using TF-IDF (2016) *Scientific Reports* 6:29319. DOI: 10.1038/srep29319.

Publications included in this thesis

Yingnan Cong, Yao-ban Chan, Mark A. Ragan. A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF (2016) *Scientific Reports* 6:30308.
– incorporated as Chapter 2.

Contributor	Statement of contribution
Author Yingnan Cong (Candidate)	Designed experiments (70%) Wrote the paper (60%) Figures and tables (100%)
Author Mark A. Ragan	Designed experiments (20%) Wrote and edited paper (30%)
Author Yao-ban Chan	Designed experiments (10%) Wrote and edited paper (10%)

Yingnan Cong, Yao-ban Chan, Mark A. Ragan. Exploring lateral genetic transfer among microbial genomes using TF-IDF (2016) *Scientific Reports* 6:29319.
– incorporated as Chapter 3.

Contributor	Statement of contribution
Author Yingnan Cong (Candidate)	Designed experiments (70%) Wrote the paper (60%) Figures and tables (100%)
Author Mark A. Ragan	Designed experiments (25%) Wrote and edited paper (30%)
Author Yao-ban Chan	Designed experiments (5%) Wrote and edited paper (10%)

Contributions by others to the thesis

No contributions by others.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

Acknowledgements

It is amazing how time flies when I am doing PhD study. I would like to gratefully acknowledge a number of people who have supported me in this work and in preparing this thesis.

Firstly, I would gratefully thank my supervisor, Professor Mark Ragan, for offering me such a challenging project when I was exploring a new research field and new institute. He is one of the most knowledgeable scholars I have ever met. He taught me so much on scientific writing and biological knowledge in this field. His help and guidance throughout my PhD is greatly appreciated, and will never be forgotten.

I would like to thank Dr Yao-ban Chan, now is a lecturer in the University of Melbourne, for being a supervisor, mentor and friend. It was my fortunate to work closely with him at the University of Queensland (UQ). I am very grateful for his insightful suggestions for my research and I appreciate his friendship.

Special gratitude is devoted to Dr Leanne Haggerty, who was my co-supervisor for a short period. Although we worked together for only four months, her help and concern always led me to move on.

I would like to thank all members of Ragan Group for providing me a comfortable environment to do the research.

I also gratefully thank Dr Amanda Carozzi, Ms Olga Chaourova and Ms Cody Mudgway for their administrative support throughout my candidature.

The PhD project was funded by China Scholarship Council (CSC) and UQ International Scholarship (UQI) for stipends and tuition fees. The stipend was partially funded by the James S. McDonnell Foundation.

Keywords

lateral genetic transfer, phylogeny, term frequency and inverse document frequency, genetic exchange communities, bacteria, archaea

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 080109, Pattern Recognition and Data Mining, 70%

ANZSRC code: 060102, Bioinformatics, 20%

ANZSRC code: 010402, Biostatistics, 10%

Fields of Research (FoR) Classification

FoR code: 0801, Artificial Intelligence and Image Processing, 70%

FoR code: 0601, Biochemistry and Cell Biology, 20%

FoR code: 0104, Statistics, 10%

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Description
ALF	Artificial Life Framework (Dalquen *et al*. 2012)
ALFY	Alignment-free local homology method (Mirjana *et al*. 2011)
AMD	Advanced Micro Devices, Inc.
BA	Bacteria and Archaea
BLAST	Basic Local Alignment Search Tool
BP	Biological Process (section of the Gene Ontology)
CC	Clonal complex
CPU	Central Processing Unit
DNA	Deoxyribonucleic acid
EB	Enteric bacteria
ECS	*Escherichia coli* and *Shigella*
F84	Felsenstein (1984) model of nucleotide replacement
FDR	False discovery rate
G	Threshold for merging two lateral segments
GEC	Genetic exchange community
GO	Gene Ontology
GrAPPA	Graph Algorithms Pipeline for Pathway Analysis (software)
G+C	Guanine-cytosine content
HGT	Horizontal gene transfer
HYK85	Hasegawa-Yano-Kishino (1985) model of nucleotide replacement
IDF	Inverse Document Frequency
k	Length of *k*-mer
K_r	Distance of DNA sequences (measured by number of substitutions per site)

LGT	Lateral genetic transfer
MLST	Multi-Locus Sequence Typing
MRP	Matrix Representation with Parsimony (supertree method)
Mb	Megabyte(s)
NGS	Next-generation sequencing
NP-hard	Non-deterministic polynomial-time hard
nt	nucleotide
PB	Petabyte(s)
R	Matrix saves shared *k*-mers between sequences and groups
RAM	Random-access memory
RNA	Ribonucleic acid
TB	Terabyte(s)
TF	Term Frequency
t-test	Student's test

Chapter 1 Introduction

Many bacteria and archaea can acquire genetic material not only *vertically* from their ancestor, but also *laterally* from other DNA carriers or the environment, and incorporate it into their genome. This process is referred to as lateral genetic transfer (LGT) or horizontal gene transfer (HGT). LGT is widespread among bacteria and archaea, where it has contributed fundamentally to functional innovation including antibiotic resistance, virulence and diverse metabolic capabilities. In addition, it is increasingly recognised that LGT has affected the genomes of many eukaryotic microbes. Although LGT plays an important role in the evolution of microbes, how to detect, delineate and analyse LGT systematically are still open issues.

In this research, I adopted and modified *term frequency* - *inverse document frequency* (TF-IDF), one of the most successful ideas in text mining, to develop a new algorithm and software for LGT detection. I carried out rigorous tests on this alignment-free approach with synthetic datasets and with empirical datasets. Based on these tests, I conclude that TF-IDF can detect LGT with high efficiency and effectiveness, and will be scalable to much larger datasets.

Thereafter I used TF-IDF to detect LGT events and infer LGT networks for three empirical datasets. In these networks, the nodes represent bacterial or archaeal genomes, and the edges connect pairs of genomes between which one or more LGT events has been inferred. Mapping these LGT events to annotated genes, I investigated how many source groups have contributed *via* LGT to each gene, and what biological functions have been affected.

Finally, I took a more-formal approach to describe the sharing of genetic material within groups of genomes. Densely connected regions within LGT networks can be considered to reveal genetic exchange communities (GECs). Up to now, it has mostly not been possible to infer the direction of lateral transfer. However, with TF-IDF the edges are directed, *i.e.* for each recipient genome the TF-IDF method identifies a source group. This has allowed me to explore more-precise definitions of GECs for empirical genome-sequence datasets.

1.1 Lateral genetic transfer

The first indication that genetic material could be exchanged between different bacteria dates to 1911, when it was observed that the human paratyphoid bacillus could take on the agglutination properties of calf paratyphoid bacillus during passage through a calf[1]. In 1928, Griffith[2] showed that non-virulent strains of *Streptococcus pneumoniae* could acquire virulence from a heat-stable substance extracted from a virulent strain. This substance was subsequently identified as DNA[3]. LGT did not gain much public attention until the 1950s, when bacteria with multi-drug resistance began to emerge globally[4]. In recent decades, evidence has accumulated showing that LGT is an important source of evolutionary novelty in microbes[5-9].

Three main mechanisms of LGT are recognised: transformation[10], transduction[11] and conjugation[12]. In transformation, an organism takes up DNA from its environment directly. Conjugation is the process by which DNA is physically transferred through a cell-to-cell junction, carried on a plasmid. In transduction, DNA is acquired through infection by a bacteriophage. Conjugation frequently mediates gene transfer in proteobacteria, and is notable in the spread of determinants for antibiotic resistance.

Under strong selection from antibiotics such as methicillin[13], β-lactams or fluoroquinolones[14], the antibiotic-resistance phenotype can spread rapidly within a population of bacteria, sometimes crossing a genus or species boundary. Although LGT is an important process in the spread of drug-resistance genes, in laboratory studies almost any type of gene can be affected by LGT[15]. Particularly with the development of next-generation sequencing (NGS) technologies in last decade, it has become ever-more apparent that LGT is responsible for functional innovations far beyond antibiotic resistance[16-18]. Therefore, to understand the history and processes of microbial evolution it is essential to identify LGT events based on large NGS datasets.

It is widely considered that a phylogenetic approach, in which the topology of a gene-family (or protein-family) tree is compared with that of a reference (species or genome) tree, is the gold-standard approach to discovering instances of LGT. Thus by default, genes are often treated as the unit of LGT, and the term *xenologs* has been introduced for genes whose distribution has come about by LGT[19,20]. However, the segment of DNA taken up into a host cell, and later recombined or inserted into its genome, does not have to coincide with a whole gene or any particular type of region within a gene, for example a region that encodes a structural domain of a protein[21]. DNA regions ranging from seven nucleotides in length[22] up to an entire chromosome greater than 3 Mb[23] have been identified as being of lateral origin. More generally, portions of genes, multi-gene clusters, pathogenicity islands and transposable elements can be involved in LGT events[24]. Once a DNA segment has been successfully transferred and integrated into the recipient genome[25], the integrated segment can be overwritten by subsequent LGT events, making the genome a mosaic. LGT may have happened at any point in the past, and over time the resulting lateral region will evolve to become indistinguishable from the non-lateral parts, a process known as amelioration[26]. These two phenomena, overwriting and amelioration, make it difficult to infer precisely the locations and origins of lateral events.

1.2 Methods for detecting LGT

Detection necessarily forms the first step toward studying LGT or GECs. The recognised mechanisms of LGT (transformation, transduction, conjugation) introduce exogenous DNA segments of different lengths. A subset of this DNA may avoid the host defences and become stabilised in the new host organism by maintenance on a plasmid or other extrachromosomal element, or by incorporation into the host genome. For consistency with most of the literature, in this thesis I consider only regions of lateral origin that have been incorporated into the host genome and have subsequently become frequent in a population. If the newly introgressed DNA is very similar to the sequence it replaces, as for example would typically happen if the source and target organisms are close relatives, it may be difficult or impossible to identify that region as lateral. Similarly, as introduced above, overwriting and amelioration also complicate discovery. Various approaches have been used to detect regions of lateral origin; the most-general and most-commonly encountered methods are based on sequence composition or on phylogenetic inference.

Sequence composition methods build on the observation that some properties tend to vary only within a limited range within a particular genome. A region of that genome atypical for that property might be suspected to be of exogenous origin. For instance, a stretch of DNA with 60% G+C inside an otherwise 40% G+C genome might be suspected to have arisen by LGT from an exogenous, 60% G+C genome[27]. The same applies to other features such as codon usage[26], *k*-mer frequency distributions, profiles corresponding to hidden Markov models, protein domains and so on.

The phylogenetic method is usually applied *via* a series of steps. Many variants have been introduced, but the key steps are as follows: the genes or proteins in a dataset are sorted into putatively orthologous families; for each family, the sequences are multiply aligned, a tree is inferred and its topology is compared with that of a reference (species or genome) tree, for example a ribosomal RNA sequence tree or a supertree. Incongruent structures (*e.g.* bipartitions[16,28]) are discovered, and the extent of disagreement is assessed, *e.g.* as a normalised count of subtree pruning and regrafting operations needed to reconcile the family and species trees[29]. LGT events lie in these incongruences[30].

Although these approaches have been applied widely to infer LGT, their limitations are also obvious: they are not scalable to very large datasets, or may be scalable only under simplifying assumptions and heuristics; they usually do not give the direction of transfer; and they cannot detect LGT between closely related genomes (neighbouring leaves on a tree). To address these problems, in this thesis I introduce and modify TF-IDF to infer LGT.

1.3 Term frequency-inverse document frequency (TF-IDF)

The term frequency (TF) concept was first introduced in the late 1950s[31,32] to find key words of a document or categorise the texts automatically. The content of a document can be subdivided into terms (words), each of which conveys part of the information in the document. The TF statistic quantifies the importance of each word in that article. There are three main ways to calculate this statistic (Table 1.1):

Table 1.1 Three formats of TF calculation

Name	Meaning
Binary weight (b_w)	If the term appears in a document this is denoted 1, otherwise 0.
Raw term frequency (TF_w)	The number of times a term appears in a document
Augmented normalised term frequency ($0.5 + 0.5\frac{TF_w}{\max(TF_w)}$)	TF_w is normalised by maximum TF_w to avoid the effects of very long documents; the value lies between 0.5 to 1.0.

Although TF is useful in showing that some terms are more representative (occur at higher frequencies) than others, it cannot retrieve more information about terms in a whole corpus. For example, two terms might occur at the same frequency in one document, but one of these terms also appears frequently in other documents in that corpus, whereas the other appears only rarely in the other documents. In such a case, the two terms carry different quantities of information about that specific document. For this reason, inverse document frequency (IDF) was introduced to calculate more accurately the importance of a term to a document[33]. The typical calculation of IDF is $idf_w = \log(N/D_w)$, where N is the number of documents within a corpus, and D_w is the number of documents containing term w. So the full TF-IDF statistic is calculated by multiplying the two [34]:

$$tf \times idf = tf_w \times \log\left(\frac{N}{D_w}\right)$$

where $tf_w = TF_w \times D_w$ is the number of times of a term appears in the corpus.

As one of the most famous ideas in information retrieval, TF-IDF has been successfully applied to extract relevant documents associated with a query, sort documents into categories, make decisions, mine emotions and in many other applications[31,32,35-39]. The form of TF-IDF to measure term weights accurately has been proven using information theory[40]. If a word contains more information for a specific query in a document, the TF-IDF value will be larger; and if a word appears in fewer documents, the TF-IDF value will again be larger.

Although TF-IDF has enjoyed considerable success in text mining, it cannot be applied to LGT detection directly. In its classical application, TF-IDF counts occurrences of words within a document and within a corpus. To improve accuracy and recall, such statistics may instead be calculated within specific categories, *e.g.* articles about biomedicine or published in *Nature*. No identical concepts exist naturally in molecular sequences, so modifications are necessary. In my thesis and published articles, we replaced classical words with *k*-mers (also referred to as *n*-grams[41,42]), which are (usually short) substrings of fixed size in a sequence. For example, sequence *S* = ACGTTA can be decomposed into three *4*-mers, ACGT, CGTT and GTTA. A single sequence (gene, protein or genome) is treated as an article, and the corpus (in computational linguistics, a set of articles) is replaced by a set of such sequences, for example a dataset of microbial genomes. The analogue of a category of documents is a group of sequences, which in Chapter 3 was defined by criteria relevant to the problem *e.g.* taxon or multi-locus sequence type. These correspondences between TF-IDF and my application to LGT can be summarised as follows (Table 1.2):

Table 1.2 Correspondence between classical TF-IDF, and TF-IDF for LGT detection.

Classical TF-IDF	TF-IDF on LGT detection
Word in an article	Fixed size short reads (*k*-mers) in a sequence
Article	Sequence (*e.g.* a genome)
Corpus (a set of articles)	Dataset (*e.g.* a number of genomes)
Category (articles with similar topics)	Phyletic group (a group of sequences with a common relevant property, *e.g.* distinct common ancestor)

In this application, I calculate TF and IDF differently than in classical TF-IDF. My approach is designed to find LGT that has not yet spread (vertically or laterally) to all members of the recipient group. For example, I might seek to find DNA regions that have introgressed laterally into one or a few genomes of *Escherichia coli* but not into enteric bacteria more generally. If this is the case, *k*-mers characteristic of the lateral segments should be relatively infrequent in genomes of enteric bacteria, but common among genomes of the donor group. Thus I calculate and use the TF and IDF statistics separately in two stages, not multiplicatively as in the original TF-IDF application case.

For the IDF component, the classical approach finds how many documents within a corpus contain a specific word. If that word appears in few documents, it is weighted more highly in the document in which it occurs (and in later extensions to classical IDF, in all documents where it occurs). For LGT inference, however, the goal of the IDF step is to determine whether a sequence contains segments from groups other than it own. Thus I count the occurrences of all *k*-mers shared between a sequence (the target sequence) and each external group (in the row of the *R* matrix). If this number is significantly greater than the average of all entries in the R matrix, there may have been LGT events between the target sequence and that external group, and the sequence is taken forward for further analysis in the TF component.

In classical TF, if a word occurs at high frequency in a document, it is potentially highly representative of that document. Classically, the TF value of this word is multiplied by its IDF value (above), yielding the weight of that word in a document. For LGT analysis, we carry out the corresponding operation within segments of a sequence. If *k*-mers within a segment (of a sequence) are less frequent than the average frequency of all unique *k*-mers in that group (the group containing the target sequence), the segment is inferred as lateral. I present full details of the TF-IDF approach for LGT detection in Chapter 2, including pseudocode (in the corresponding Supplementary Material).

1.4 LGT networks

Following Darwin, the evolutionary history of biological entities (originally species, but more recently families of genes or proteins) can be summarised as a phylogenetic tree. However, widespread LGT events in microbes complicate the tree by adding lateral branches (Figure 1.1). For this reason, phylogenetic networks have been introduced as an alternative to trees.

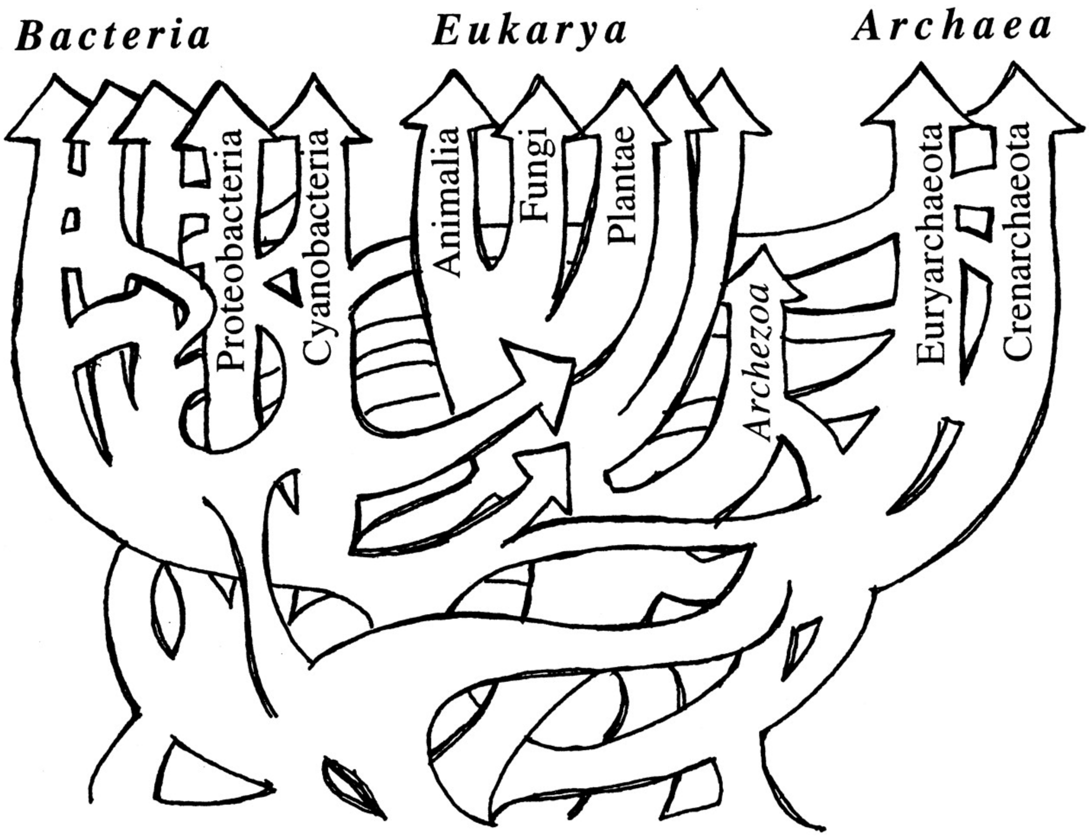


Figure 1.1 A phylogenetic tree, reticulated by lateral branches to form a network. From W. Ford Doolittle (1999).

Since the mid-1700s, networks have been used to describe real or imagined patterns among biological organisms[43]. In the 1990s, networks started to be applied to depict the molecular recombination history of sequences[44] and this become mature in the following decades[45-48]. In a phylogenetic network, nodes represent DNA carriers (usually organism genomes, but sometimes viruses or plasmids) and edges connect pairs of nodes between which transfer has been inferred. Where there is only vertical transmission, this network is a standard phylogenetic tree. Alternatively, we might consider only the subset of nodes and edges involved in LGT, and these would form an LGT network.

An early evolutionary network, depicting the combined lateral and vertical gene evolutionary history of bacteria and archaea, was constructed in 2005 by analysis of 165 microbial genomes[49]. This network displays small-world properties, implying that genetic material can be transferred amongst all of these organisms, across taxonomic boundaries, *via* a small number of LGT events: the original donor and final recipient might have no phylogenetic, ecological or geographical relationship but communicate through highly connected nodes (hubs) in the network. In the same year Beiko *et al.*[16] used a network representation to identify "highways" of sharing among 144 fully sequenced genomes of bacteria and archaea. Dagan *et al.*[50] reported a network describing the combined vertical and lateral evolutionary history of 22 archaebacterial and 159 eubacterial genomes, and they analysed this network to reveal details of community structure. Fondi and Fani reconstructed a network on 5,030 proteins associated with antibiotic resistance from 122,482 plasmid-encoded amino acid sequences[51]. A small number of these nodes account for most of the connections (*i.e.* are hubs), indicating that a few microorganisms (and/or their plasmids) interact with a large spectrum of microbes. These hubs thus play a pivotal role in the sharing of antibiotic resistance throughout the bacterial world, and the authors propose that focusing on them could prevent the dissemination of antibiotic resistance in certain habitats.

The largest genetic network published so far is that of Halary *et al.*[52], linking 45,845 DNA sequences from environmental metagenome projects, and 532,682 protein sequences encoded in microbial, phage and plasmid genomes. The nodes in the network are individual sequences, and the edges derive from reciprocal best BLAST hits. By adjusting the threshold requirement (BLAST identity), these authors argue that plasmids are key vectors of genetic exchange among bacterial chromosomes. Kloesges *et al.*[53] reconstructed a network among 329 sequenced proteobacteria, reporting that in proteobacteria most LGT takes place between closely related species (those with high sequence similarities) and that this transfer is frequently mediated by conjugation, *i.e.* by plasmids. In 2011, a directed network of recent LGT events was reconstructed among 657 microbes[54]; age and directionality of LGT were inferred from G+C content.

1.5 Genetic exchange communities (GECs)

The term *exchange community* was first introduced in connection with LGT in 2003[17] and described as “a collection of organisms that can share genes by (LGT), but need not be in physical proximity”. In 2011, Skippington and Ragan recommended a somewhat different definition[55]: “a GEC is a set of entities, each of which has over time both donated genetic material to, and received genetic material from, every other entity in that GEC, *via* a path of lateral transfer”. Their intention was twofold: to disentangle the definition from issues such as spatial, temporal or phylogenetic proximity, and to enable a graph-based framework for “thinking precisely and operationally about GECs”. They also argued that GECs do not exist *a priori* in nature, but instead are “constructed dynamically by organisms through diverse physical, chemical and biological interactions with their environment and with each other”[55]. In an LGT graph, potential GEC structures include cliques (sets of nodes, each of which reaches all others directly), para-cliques (near-cliques, falling short of being a clique by lacking a defined, usually small, number of edges, transitively closed sets (sets of nodes, each of which can reach all the others) and paths (series of nodes connected by edges). Abstracting ideas about GECs in this way can help us think rigorously and ask well-defined questions about the flow of genetic determinants across the biosphere[55]. The first GEC reconstructed from empirical data[56] showed that all five phyletic groups recognised within the *E. coli – Shigella* clade are members of a single common GEC.

In this thesis, I rigorously define dense structures (cliques) in LGT graphs as GECs. Delineating GECs can help us understand what kinds of microbes exchange genetic material frequently, and what kinds of biological processes are frequently involved in LGT events. Considering the GECs that exist, and those that apparently do not, may reveal opportunities for and barriers to LGT. This information will help us understand how microbes evolve, and how they contribute to the biosphere.

1.6 Objectives

Genetic exchange communities have been defined as sets of entities which exchange (donate and receive) genetic material over time *via* LGT. In order to find GECs among the very large datasets being built up for infectious disease microbes, or GECs occurring at broad scale across the microbial biosphere, we need a method that can be applied to very large multi-genome datasets quickly, accurately and in a highly scalable manner. TF-IDF has proven to be an efficient, effective and highly scalable tool to discover patterns in very large datasets of documents, such as traditional texts, web pages or output from social media. This study aimed to modify TF-IDF so it can detect LGT within large microbial genome datasets; construct LGT networks; find GECs, and identify the biological processes associated with these GECs.

1.7 Thesis outline

The following chapters introduce how I implemented and tested a method based on TF-IDF for LGT detection, and how I identified GECs by reconstructing LGT networks. I first modified the classical TF-IDF idea to make it suitable for LGT detection. Specifically, the TF part was modified from finding high-frequency terms in one article, to finding significantly low-frequency *k*-mers in a genome; and the IDF part was modified from finding low-frequency terms in a corpus, to finding significantly high-frequency *k*-mers in other phyletic groups. The new method was tested rigorously and compared with ALFY[57], another alignment-free method for LGT detection, using simulated datasets and a small empirical dataset. I varied sequence lengths, evolutionary distances between phyletic groups and between sequences within-group, and the deletion rate of each sequence to test the performance of the method under a range of biologically relevant situations. I was able to show that TF-IDF works well when the boundaries of groups are clear, and sequences within a group are similar to each other. TF-IDF exhibits slightly lower recall but much greater precision than ALFY. I also applied TF-IDF on a dataset of seven *Staphylococcus aureus* genomes; previous studies have demonstrated LGT to exist between two of these strains, TW20 and MRSA252[57,58]. My TF-IDF method finds all the LGT events inferred by ALFY, and further identifies some other genomic regions to be of interest; these regions encode mobile elements and antibiotic-resistance genes. The TF-IDF method runs much faster than ALFY on large datasets. The details are presented in Chapter 2.

In Chapter 3, I applied my TF-IDF method to three empirical genome datasets (27 *Escherichia coli* and *Shigella*; 110 enteric bacteria; and 143 bacteria and archaea) to investigate its performance in a variety of real-life situations, and to identify instances of LGT. I studied how the method depends on word length and gap size, and found optimal (or nearly optimal) parameters for a range of biologically realistic scenarios. These results provide guidance for selecting word length and gap size in real applications.

I observed an abundance of lateral transfers among groups within the *E. coli – Shigella* dataset. Some of these indicated more-ancient transfers, which are otherwise difficult to detect. Using the reference tree structure, I was able to distinguish (in part) between multiple recent transfers from different source groups, versus a single transfer from their common ancestor. In the enteric bacteria dataset, most of the LGT signal arises from exchanges between *E. coli* and *Shigella*, I observed a lower frequency of exchange involving the remaining groups (with the exception of *Yersinia*). Few LGT events were inferred among phyla in the prokaryote dataset, as expected. I carried out enrichment tests on Gene Ontology (GO) terms to see what biological processes are over- or under-represented in inferred LGT events; this analysis showed that a wide range of processes, not only those involved in antibiotic resistance or recombination, are over-represented. I reconstructed several LGT networks and showed that they manifest complex subgraph structures (GECs).

In Chapter 4, I further studied the LGT networks inferred for the three datasets used in Chapter 3. Using different criteria, I identified GECs within these networks and identified lateral genes that underlie these GECs. Core GEC structures are those that tend to be stable as *k* is varied, and encompass most of the LGT events in a dataset. Different frequencies of LGTs make up various densities of the network, and specific bacterial species tend to be represented among the "hubs", while others are relatively silent to LGT. The enriched biological processes within GECs are accordant to the whole LGT network, although the actual rank of specific terms may be slightly different. This reveals GECs plays a key role in microbial evolutionary processes, which are affected by LGT.

Finally, in Chapter 5 I briefly summarise the main research results, and identify important directions and challenges that remain for future work. These include optimisation of the data structure underlying TF-IDF to reduce memory usage, and deployment on a distributed computation platform (*e.g.* Hadoop) to improve its capability and efficiency for processing larger datasets. More microbial genomes will be sequenced and should be studied for patterns of LGT, which will increase our fundamental understanding of LGT and our knowledge of their importance in specific environments. Currently, nodes in the inferred LGT networks and GECs represent groups of sequences (or genomes), but for many applications we instead want to understand lateral relationships involving individual sequences (or genomes). Bringing the TF-IDF approach to bear on individual sequences may require changes or refinement at the level of algorithms, data management, parameter-value optimisation, graphical representation, and the definition of GECs.

This project is the first to apply TF-IDF to molecular or genomic datasets to identify LGT, indeed to our knowledge the first to apply TF-IDF to any biological question. Directional and multi-source transfers were systematically studied for the first time. The TF-IDF method is fast, efficient, and sufficiently scalable to detect evidence of LGT in large multi-genome datasets, which provides us the opportunity to construct LGT networks of microbes on a large scale. These LGT networks are not uniform across genomes, and potential GECs can be recognised that reveal what kinds of bacteria frequently exchange genetic information, and the genes and biological processes that have been affected by LGT.

1.8 References

1 Schmitt, R. M. Zur Variablilität der Enteritis-bakterien. *Zeitschr Infektionskrankh parasit Krankh Hyg Haustiere* **9**, 188 (1911).

2 Griffith, F. The significance of pneumococcal types. *J Hyg (Lond)* **27**, 113-159 (1928).

3 Avery, O. T., Macleod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III. *J Exp Med* **79**, 137-158 (1944).

4 Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Reviews* **74**, 417-433 (2010).

5 Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284**, 2124-2129 (1999).

6 Martin, W. Mosaic bacterial chromosomes: a challenge on route to a tree of genomes. *BioEssays* **21**, 99-104, doi:10.1002/(Sici)1521-1878(199902)21:2<99::Aid-Bies3>3.0.Co;2-B (1999).

7 Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304, doi:10.1038/35012500 (2000).

8 Doolittle, W. F. & Bapteste, E. Pattern pluralism and the Tree of Life hypothesis. *Proc Natl Acad Sci U S A* **104**, 2043-2049, doi:10.1073/pnas.0610699104 (2007).

9 Goldenfeld, N. & Woese, C. Biology's next revolution. *Nature* **445**, 369, doi:10.1038/445369a (2007).

10 Chen, I. & Dubnau, D. DNA uptake during bacterial transformation. *Nat Rev Microbiol* **2**, 241-249, doi:10.1038/nrmicro844 (2004).

11 Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* **3**, 711-721, doi:10.1038/nrmicro1234 (2005).

12 Chen, I., Christie, P. J. & Dubnau, D. The ins and outs of DNA transfer in bacteria. *Science* **310**, 1456-1460, doi:10.1126/science.1114021 (2005).

13 Grundmann, H., Aires-de-Sousa, M., Boyce, J. & Tiemersma, E. Emergence and resurgence of meticillin-resistant *Staphylococcus aureus* as a public-health threat. *Lancet* **368**, 874-885, doi:10.1016/S0140-6736(06)68853-3 (2006).

14 Barlow, M. What antimicrobial resistance has taught us about horizontal gene transfer. *Methods Mol Biol* **532**, 397-411, doi:10.1007/978-1-60327-853-9_23 (2009).

15 Sorek, R., Zhu, Y., Creevey, C. J., Francino, M. P., Bork, P., & Rubin, E. M. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**, 1449-1452, doi:10.1126/science.1147112 (2007).

16 Beiko, R. G., Harlow, T. J. & Ragan, M. A. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A* **102**, 14332-14337, doi:10.1073/pnas.0504068102 (2005).

17 Jain, R., Rivera, M. C., Moore, J. E. & Lake, J. A. Horizontal gene transfer accelerates genome innovation and evolution. *Mol Biol Evol* **20**, 1598-1602, doi:10.1093/molbev/msg154 (2003).

18 Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature Genetics* **36**, 760-766, doi:10.1038/ng1381 (2004).

19 Gray, G. S. & Fitch, W. M. Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*. *Mol Biol Evol* **1**, 57-66 (1983).

20 Hellmuth, M. & Wieseke, N. From sequence data including orthologs, paralogs, and xenologs to gene and species trees. In Pontarotti, P. (Ed.) *Evolutionary Biology. Convergent Evolution, Evolution of Complex Traits, Concepts and Methods,* 373-392 (Springer, 2016).

21 Chan, C. X., Darling, A. E., Beiko, R. G. & Ragan, M. A. Are protein domains modules of lateral genetic transfer? *PLoS ONE* **4**, e4524, doi:10.1371/journal.pone.0004524 (2009).

22 Denamur, E., Lecointre, G., Darlu, P., Tenaillon, O., Acquaviva, C., Savada, C., Sunjevaric, I., Rothstein, R., Elion, J., Taddri, F., Radman, M. & Matic, I. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* **103**, 711-721 (2000).

23 Lin, C. H., Bourque, G. & Tan, P. A comparative synteny map of Burkholderia species links large-scale genome rearrangements to fine-scale nucleotide variation in prokaryotes. *Mol Biol Evol* **25**, 549-558, doi:10.1093/molbev/msm282 (2008).

24 Ragan, M. A. & Beiko, R. G. Lateral genetic transfer: open issues. *Philos Trans R Soc Lond B Biol Sci* **364**, 2241-2251, doi:10.1098/rstb.2009.0031 (2009).

25 Lawrence, J. G. & Retchless, A. C. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. *Methods Mol Biol* **532**, 29-53, doi:10.1007/978-1-60327-853-9_3 (2009).

26 Lawrence, J. G. & Ochman, H. Amelioration of bacterial genomes: Rates of change and exchange. *J Mol Evol* **44**, 383-397, doi:10.1007/Pl00006158 (1997).

27 Daubin, V., Lerat, E. & Perriere, G. The source of laterally transferred genes in bacterial genomes. *Genome Biol* **4** (2003).

28 Zhaxybayeva, O., Hamel, L., Raymond, J. & Gogarten, J. P. Visualization of the phylogenetic content of five genomes using dekapentagonal maps. *Genome Biol* **5**, R20, doi:10.1186/gb-2004-5-3-r20 (2004).

29 Hallett, M. T. & Lagergren, J. Efficient algorithms for gene transfer problems. *Proc Fifth Ann Int Conf Computat Biol,* 149-156 (ACM).

30 Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: The complexity hypothesis. *Proc Natl Acad Sci U S A* **96**, 3801-3806, doi: 10.1073/Pnas.96.7.3801 (1999).

31 Luhn, H. P. A statistical approach to mechanized encoding and searching of literary information. *IBM J Res Dev* **1**, 309-317 (1957).

32 Luhn, H. P. The automatic creation of literature abstracts. *IBM J Res Dev* **2**, 159-165 (1958).

33 Jones, K. S. A statistical interpretation of term specificity and its application in retrieval. *J Docum* **28**, 11-21 (1972).

34 Salton, G. & Buckley, C. Term-weighting approaches in automatic text retrieval. *Inform Process Manag* **24**, 513-523, doi: 10.1016/0306-4573(88)90021-0 (1988).

35 Salton, G. *The SMART Retrieval System; Experiments in Automatic Document Processing*. (Prentice-Hall, 1971).

36 Salton, G. *New Approaches to Automatic Document Processing*. (Cornell University, Dept of Computer Science, 1971).

37 Salton, G. Developments in automatic text retrieval. *Science* **253**, 974-980, doi:10.1126/Science.253.5023.974 (1991).

38 Salton, G. & McGill, M. J. *Introduction to Modern Information Retrieval*. (McGraw-Hill, 1983).

39 Salton, G. & McGill, M. J. The SMART and SIRE experimental retrieval systems. in *Readings in Information Retrieval* (eds Spärck Jones, K. & Willett, P.), 381-399 (Morgan Kaufmann Publishers Inc., 1997).

40 Robertson, S. Understanding inverse document frequency: on theoretical arguments for IDF. *J Doc* **60**, 503-520, doi:10.1108/00220410560582 (2004).

41 Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D. & Lai, J. C. Class-based n-gram models of natural language. *Comput Linguist* **18**, 467-479 (1992).

42 Cavnar, W. B. & Trenkle, J. M. N-gram-based text categorization. *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 161-175 (1994).

43 Ragan, M. A. Trees and networks before and after Darwin. *Biol Direct* **4**, 43; doi:10.1186/1745-6150-4-43 (2009).

44 Hein, J. A heuristic method to reconstruct the history of sequences subject to recombination. *J Mol Evol* **36**, 396-405 (1993).

45 Gusfield, D., Eddhu, S. & Langley, C. Efficient reconstruction of phylogenetic networks with constrained recombination. *Proc IEEE Comput Soc Bioinformatics Conf* **2**, 363-374 (2003).

46 Gusfield, D. Optimal, efficient reconstruction of root-unknown phylogenetic networks with constrained and structured recombination. *J Comput Syst Sci* **70**, 381-398 (2005).

47 Song, Y. S. & Hein, J. Constructing minimal ancestral recombination graphs. *J Comput Biol* **12**, 147-169 (2005).

48 Tofigh, A., Hallett, M. & Lagergren, J. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans Comput Biol Bioinformatics* **8**, 517-535 (2011).

49 Kunin, V., Goldovsky, L., Darzentas, N. & Ouzounis, C. A. The Net of Life: reconstructing the microbial phylogenetic network. *Genome Res* **15**, 954-959 (2005).

50 Dagan, T., Artzy-Randrup, Y. & Martin, W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci U S A* **105**, 10039-10044, doi:10.1073/pnas.0800679105 (2008).

51 Fondi, M. & Fani, R. The horizontal flow of the plasmid resistome: clues from inter-generic similarity networks. *Environ Microbiol* **12**, 3228-3242, doi:10.1111/j.1462-2920.2010.02295.x (2010).

52 Halary, S., Leigh, J. W., Cheaib, B., Lopez, P. & Bapteste, E. Network analyses structure genetic diversity in independent genetic worlds. *Proc Natl Acad Sci U S A* **107**, 127-132, doi:10.1073/pnas.0908978107 (2010).

53 Kloesges, T., Popa, O., Martin, W. & Dagan, T. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Mol Biol Evol* **28**, 1057-1074, doi:10.1093/molbev/msq297 (2011).

54 Popa, O., Hazkani-Covo, E., Landan, G., Martin, W. & Dagan, T. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res* **21**, 599-609, doi:10.1101/gr.115592.110 (2011).

55 Skippington, E. & Ragan, M. A. Lateral genetic transfer and the construction of genetic exchange communities. *FEMS Microbiol Rev* **35**, 707-735, doi:10.1111/j.1574-6976.2010.00261.x (2011).

56 Skippington, E. & Ragan, M. A. Evolutionary dynamics of small RNAs in 27 *Escherichia coli* and *Shigella* genomes. *Genome Biol Evol* **4**, 330-345, doi:10.1093/gbe/evs001 (2012).

57 Domazet-Lošo, M. & Haubold, B. Alignment-free detection of horizontal gene transfer between closely related bacterial genomes. *Mob Genet Elements* **1**, 230-235, doi:10.4161/mge.1.3.18065 (2011).

58 Holden, M. T., Lindsay, J. A., Corton, C., Quail, M.A., Cockfield, J. D., Pathak, S., Batra, R., Parkhill, J., Bentley, S. D. & Edgeworth, J. D. Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant *Staphylococcus aureus*, sequence type 239 (TW). *J Bacteriol* **192**, 888-892, doi:10.1128/JB.01255-09 (2010).

Chapter 2 A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF

Lateral genetic transfer plays a key role in the evolution of microbes. Inferring instances of LGT from genome-sequence data is one of the most challenging issues in this research area. The traditional methods for LGT detection usually search for regions of sequence with atypical features (*e.g.* G+C content or codon usage), or incongruence between gene trees and a reference tree. With the development of next-generation sequencing technologies, the number of sequences available for research has grown dramatically in the last decade. However, the traditional methods are not scalable to data of this size. New, efficient and effective methods for LGT inference are needed for phylogenetic and phylogenomic research.

To tackle this issue, I introduced term frequency – inverse document frequency (TF-IDF), one of the most successful methods in text mining, to develop an algorithm for LGT detection. I modified the original TF-IDF to make it applicable to molecular sequences. DNA sequences are decomposed into short reads of fixed size (*k*-mers). A sequence region composed of *k*-mers that appear infrequently in their own group of sequences, but frequently in an external group of sequences, are inferred as lateral, with the direction of transfer from the latter (donor) group into the former (recipient) sequence. The method has been implemented as software and is downloadable from Github. The method has been tested using both simulated datasets, and (in this chapter) one empirical dataset. In the following chapters it will be applied to more empirical datasets. The results show that TF-IDF can be an effective and efficient method for LGT detection on large sequence datasets.

Results presented as a publication

The work reported in this chapter has been published as a research article in the Nature Publishing Group journal *Scientific Reports*, volume 6, article 30308 (publication date 25 July 2016). Supplementary material for this publication follows the main text. Supporting datasets are assigned DOI numbers http://dx.doi.org/10.14264/uql.2016.483-487 and are available for download from UQ eSpace.

A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF

Yingnan Cong[1], Yao-ban Chan[2], Mark A. Ragan[1*]

[1]Institute for Molecular Bioscience and ARC Centre of Excellence in Bioinformatics, The University of Queensland, St Lucia, Brisbane, QLD 4072, Australia

[2]School of Mathematics and Statistics, The University of Melbourne, Parkville, Melbourne, VIC 3010, Australia

Email addresses:

YC: y.cong@uq.edu.au
YBC: yaoban@unimelb.edu.au
MAR[*]: m.ragan@uq.edu.au

Abstract

Lateral genetic transfer (LGT) plays an important role in the evolution of microbes. Existing computational methods for detecting genomic regions of putative lateral origin scale poorly to large data. Here, we propose a novel method based on TF-IDF (Term Frequency-Inverse Document Frequency) statistics to detect not only regions of lateral origin, but also their origin and direction of transfer, in sets of hierarchically structured nucleotide or protein sequences. This approach is based on the frequency distributions of *k*-mers in the sequences. If a set of contiguous *k*-mers appears sufficiently more frequently in another phyletic group than in its own, we infer that they have been transferred from the first group to the second.

We performed rigorous tests of TF-IDF using simulated and empirical datasets. With the simulated data, we tested our method under different parameter settings for sequence length, substitution rate between and within groups and post-LGT, deletion rate, length of transferred region and *k* size, and found that we can detect LGT events with high precision and recall. Our method performs better than an established method, ALFY, which has high recall but low precision. Our method is efficient, with runtime increasing approximately linearly with sequence length.

2.1 Introduction

Many microbes can acquire DNA from their environment and incorporate it into their genome *via* processes of lateral genetic transfer (LGT; also known as horizontal gene transfer, HGT)[1]. Circumstantial evidence for LGT was first reported more than a century ago[2], and the phenomenon gained widespread attention in the 1950s with the emergence and spread of multi-drug resistance in bacteria[3]. With the uptake of genome sequencing over the last two decades, it has become increasingly clear that LGT plays a central role in the evolution of microbial genomes[1,4-6]. LGT not only contributes to the spread of antibiotic resistance, but is also responsible for a range of metabolic innovations involving carbon and nitrogen metabolism, ion transport and other core processes[7], which in turn can define microbial physiology and thus ecosystem function.

The recognised mechanisms of LGT (transformation, transduction and conjunction) can introduce exogenous regions of very different lengths, from short fragments to large chromosomal blocks[8]. Recombination need not be constrained by gene boundaries[9], and there is little evidence to suggest that entire genes, or structurally based regions within genes, are privileged units of transfer[10,11]. In any event, genomic regions of lateral origin can be overwritten, wholly or in part, by subsequent LGT events. Thus microbial genomes can become mosaics, with regions of different lengths reflecting the history of LGT events, transfer mechanisms and donors in each lineage. Further, over time, sequence regions of lateral origin will evolve to become indistinguishable from the non-lateral background, a process known as amelioration[12].

This complex biology presents challenges for the detection and delineation of genomic regions of lateral origin. As typically applied, approaches based on the topological comparison of inferred phylogenetic trees implicitly take genes (gene families) as the unit of analysis. Extensions that test for recombination breakpoints are computationally intensive, yet fail to identify the specific lineage(s) affected by transfer and/or subsequent overwriting. Directionality of transfer can also be difficult or impossible to determine by any phylogenetic approach. More broadly, computational methods are differentially sensitive to the extent of amelioration[13,14]. Considerable scope thus remains for the development of new methods that are sensitive, directional, scalable, informative on individual genomes or lineages, and do not require the units of analysis to be delineated *a priori*.

Alignment-free approaches to detect LGT at genome level have been developed in recent years. ALFY (ALignment-Free local homologY)[15,16] uses K_r[17] based on *shustrings* (SHortest Unique subSTRINGS) to calculate pairwise evolutionary distances between genomes, which can then serve as input into a neighbor-joining algorithm[18] to compute a phylogenetic tree. Then ALFY compares the generated tree with a reference, inferring topological incongruence as instances of LGT.

Another alignment-free method for LGT detection is based on the so-called purity measure[19]. This is a concept from text mining, and is used to detect unusual regions of a string without recourse to domain knowledge. If most substrings of string *x*, which is itself a substring of string *T*, appear with the same frequency as *x*, then the purity value of *x* is high, *i.e.* subpatterns in *x* occur infrequently in *T* outside whole occurrences of *x*, as would be expected if *x* had arisen by LGT. Both of these alignment-free methods use suffix trees[20] for scalability on large sequence datasets. However, they consider only one target sequence (although ALFY incorporates a pairwise comparison between query and multiple subject sequences) and do not take into account any natural group structure of the dataset, whether taxonomic (a hierarchy of species, genera etc.), ecological or otherwise.

In this paper, we propose a novel alignment-free method for LGT detection based on concepts from TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF is a numerical statistic from document analysis that reflects the importance of a word (term) to a document within a collection or corpus, by comparing the frequency of a word in a document with its occurrence in other documents.

Term frequency (TF) is used to indicate the topic of a document[21]. The TF of term *t* in document *D* is simply the raw frequency of *t* in *D*, denoted by *tf(t,D)*. The inverse document frequency (IDF)[22] is used to distinguish a word from the prevalent vocabulary in the corpus. If *t* appears in D_t articles, then its IDF is $idf(t)=D^*/D_t$, where D^* is the number of all documents in the corpus. Thus a high IDF indicates that the term appears infrequently, and as such carries more importance for a specific article. Salton and Buckley combined the TF and IDF statistics into a single statistic that is widely used as a weighting factor in text mining and information retrieval[22-24].

Here we apply concepts from TF-IDF to develop an algorithm to detect LGT events in microbial genomes. Using simulated datasets, we test this algorithm and compare its performance with ALFY on sets of sequences of different length, from the size of a single gene (1000 nucleotides) up to 300-fold longer, and evaluate its performance over *k*-mer length and a biologically relevant range of values for parameters including substitution rate between groups, within groups and post-LGT. We find that with appropriate parameter values, the algorithm performs with good precision and recall; furthermore, runtime increases approximately linearly with sequence length, and in most cases TF-IDF performs much better than ALFY[15]. We also apply this method to an empirical dataset composed of seven *Staphylococcus aureus* genomes**,** and recover putative regions of lateral origin that correspond to genes involved in transport, antibiotic resistance, pathogenicity and virulence. Our results are comparable with those found with ALFY, and include two genomic regions independently confirmed by Holden *et al.*[25].

2.2 Results

2.2.1 Performance with different parameter values

As described in Methods, we varied branch length at three stages of the simulation process (variation between groups, variation within groups, and variation post-LGT) and examined the effect on precision and recall. The results are shown in Figures 2.1-2.4 for simulations under the HYK85[26] model of sequence change; the corresponding plots for F84[27] are in the Supplementary file. Since TF-IDF does not detect LGT between sequences within a group, for the comparison we ignore such regions that are detected by ALFY; and if an atypical region is equally predicted in several sequences of potential donor groups, we treat this result as a single prediction for the calculation of precision and recall.

Figure 2.1 shows that when variation between groups is less than 0.05, the average distance accumulated between groups is less than 15%; at this degree of between-group similarity, the precision of our TF-IDF method is low (less than 50%) because the high similarity makes lateral regions harder to distinguish in the recipient group. Precision increases to a high level when variation between groups is above 0.1. Recall is high throughout (approximately 90%) and is less affected by variation; however, at the shortest sequence length examined here (1000), some simulated LGT segments are less than 50 nt in length, too short to contain enough information to make them distinct. As a consequence, recall is significantly lower for this sequence length only.

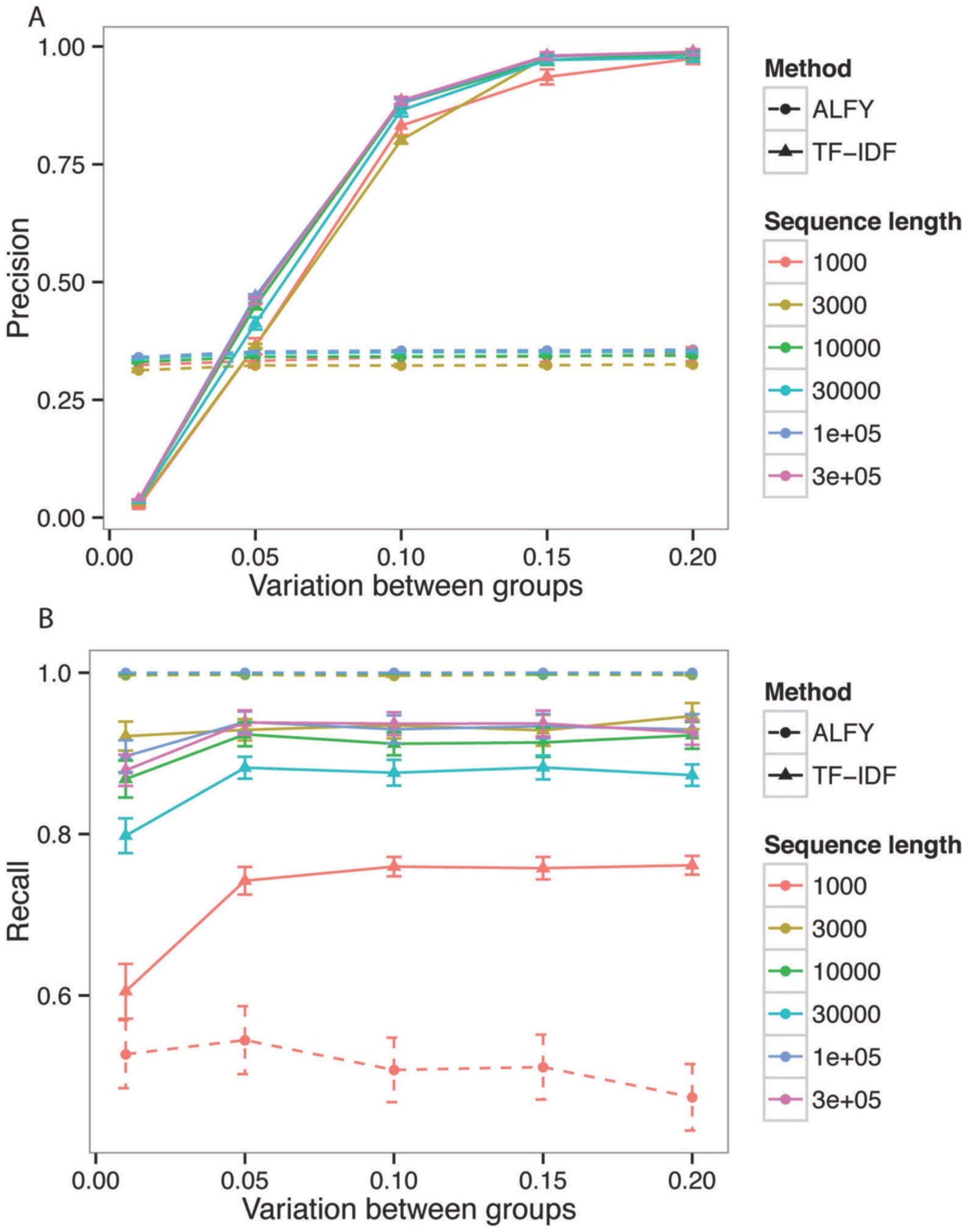


Figure 2.1. Performance of TF-IDF with variation between groups. Precision (A) increases with variation between groups. Recall (B) is not substantially affected by variation between groups. Variation within groups is 0.01, variation post-LGT is zero, and deletion is zero. Error bars are 2× standard error.

The precision of ALFY is low, around 0.35, and stable across all branch lengths, but its recall is high. There are two reasons for this. Firstly, ALFY cannot infer the direction of transfer, and may correctly predict one transfer from donor to recipient, but then (erroneously) predict it again from recipient to donor, effectively halving its precision. In the accompanying article[28] we compare TF-IDF with another directional LGT inference approach[29] applied to genome-scale empirical data. Secondly, ALFY predicts all most-similar regions as lateral transfers without using a threshold to determine if the similarity is significant or not. As such, it is apparent that ALFY is a useful tool for determining areas which should be further studied for transferred segments, but as a stand-alone detector of LGT it is inferior to TF-IDF. For sequences of length 1000 nt, ALFY's default sliding window size is too large, leading to reduced performance.

Figure 2.2 shows the effect of variation within groups on precision and recall. Here, the precision of TF-IDF increases as variation increases. As above, the sequences must be sufficiently dissimilar for the TF statistic to support a decision of LGT. Recall is high, and stable when the sequence length is ≥ 3000 nt. Again, at sequence length 1000, some short LGT events (< 50 nt) are ignored, resulting in decreased recall. The precision of ALFY is stable for variation above 0.005, but again low. TF-IDF shows greater stability and better performance than ALFY in almost all cases, and increasingly outperforms it as the variation increases. As in Figure 2.1, ALFY displays better recall than TF-IDF at sequence lengths greater than 1000 nt, but the gap is not large. When the variation within groups is low and the sequence length is short (1000 nt), ALFY again fails to detect most LGT events, leading to extremely low recall (see Supplementary file).

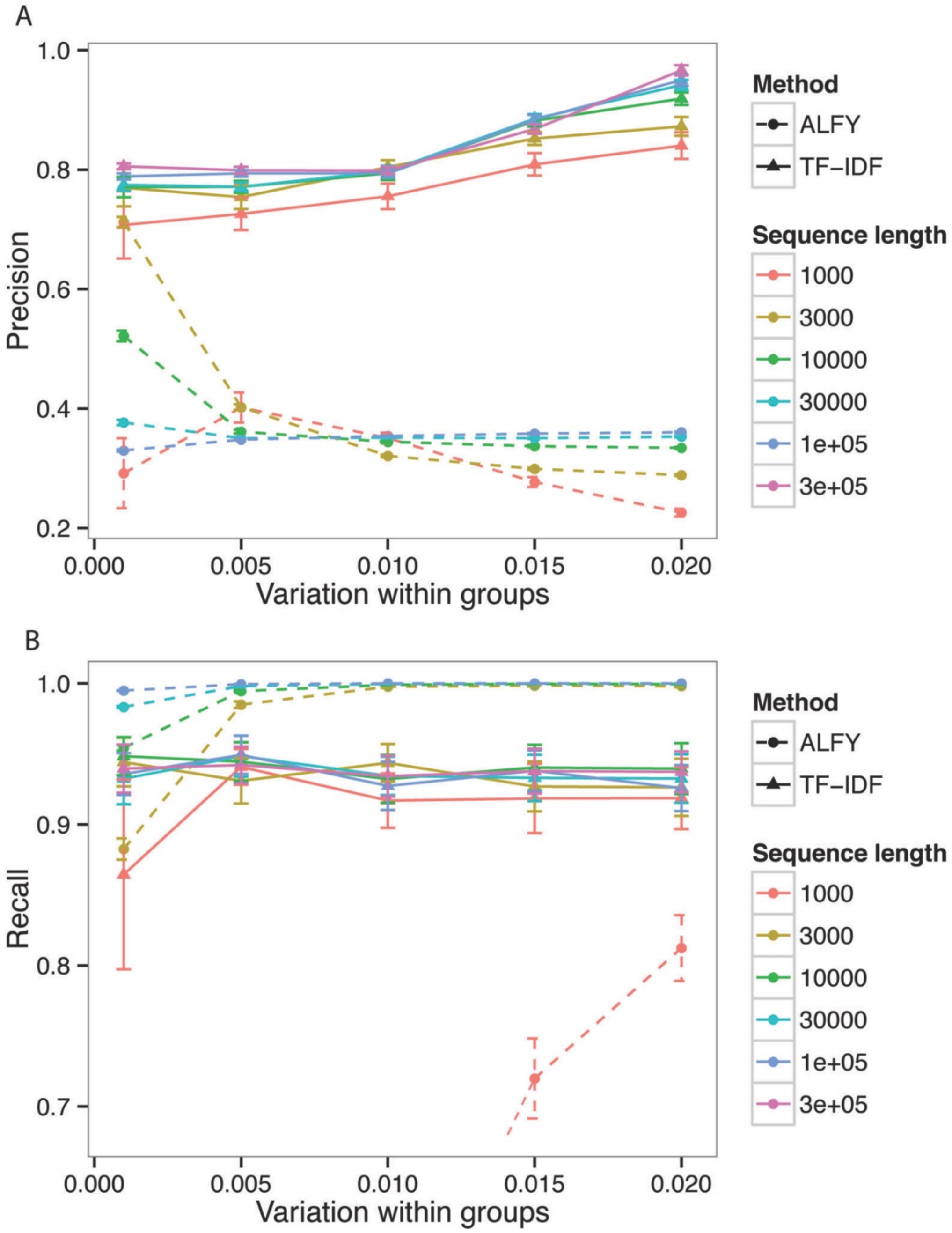


Figure 2.2. Performance of TF-IDF with variation within groups. Precision (A) increases with variation within groups, while recall (B) is essentially unchanged. Variation between groups is 0.1, variation post-LGT is zero, and deletion is zero. Error bars are 2× standard error.

Figure 2.3 shows the performance of TF-IDF against variation post-LGT and deletion rate for sequences of length 300,000 nt. Plots for other sequence lengths are similar in nature and can be found in the Supplementary file. As variation increases, both precision and (especially) recall decrease substantially, as substitutions progressively obscure the regions of lateral origin. When the branch length post-LGT reaches 0.05 (*i.e.* one nucleotide in ten is expected to have changed, as this is a two-level tree), almost all *k*-mers (for *k* = 40) have been changed, whether in lateral regions or not. In this case, all *k*-mer based methods, including TF-IDF, will fail (and indeed, even alignment-based methods will struggle).

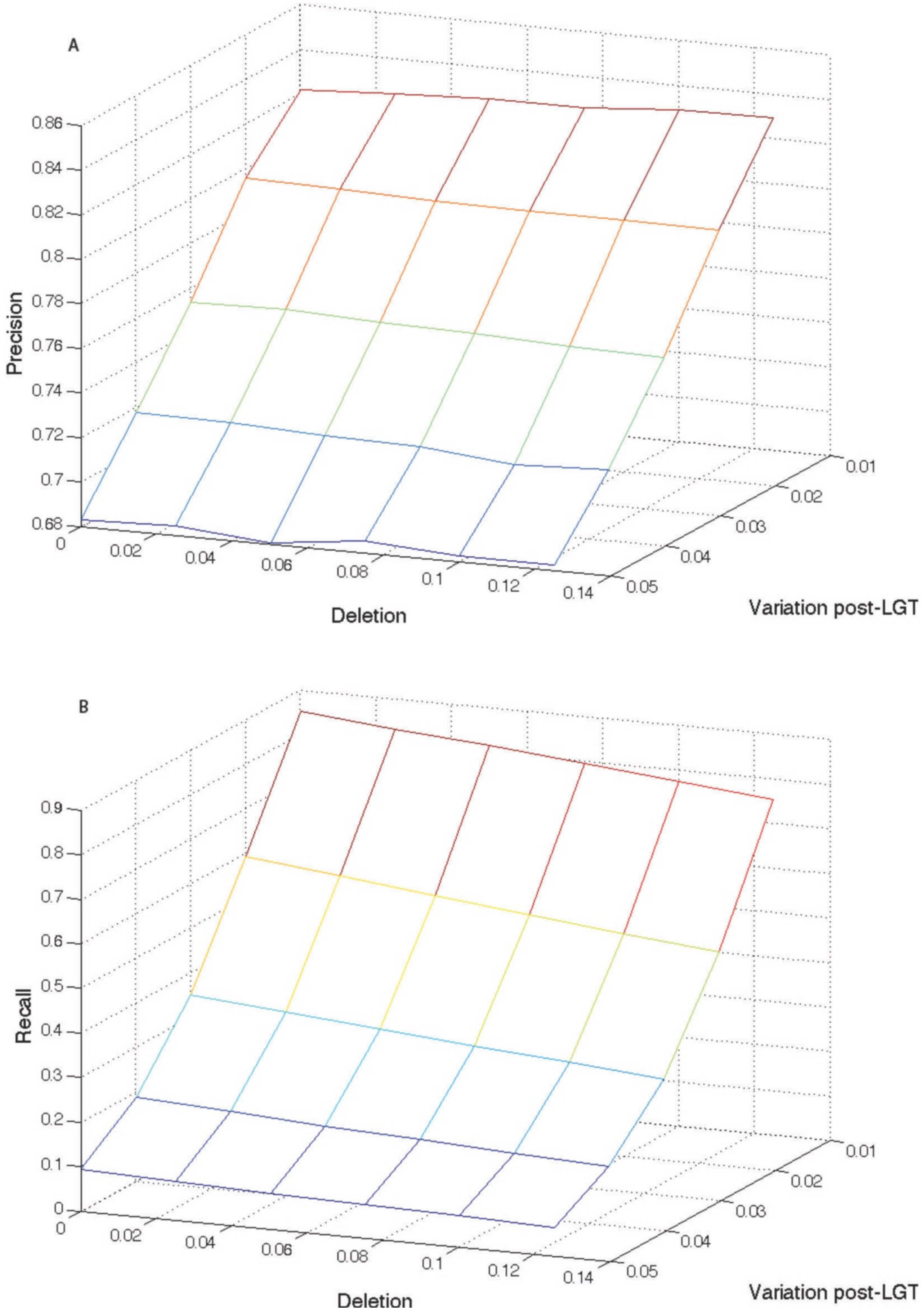


Figure 2.3. Performance of TF-IDF with variation post-LGT and deletion. Precision (A) decreases with variation post-LGT, but is unaffected by deletion. Recall (B) decreases greatly with variation post-LGT and slightly with deletion. Variation between groups is 0.1, and variation within groups is 0.01. Sequence length is 300,000 nt.

As the amount of deletion increases, precision remains stable and recall decreases slightly. Deletion can move an LGT segment within a sequence, or delete part (or parts) of it. Moving an LGT region does not change its *k*-mers, so this will not affect the performance of TF-IDF. Deletions within a lateral region affect only the immediately adjacent *k*-mers, with little effect on precision unless the region becomes so fragmented that *k*-mer counts are reduced to the point where they are ignored by TF-IDF, degrading the recall.

Precision and recall increase slightly with sequence length, but length does not appear to interact substantially with the substitution-rate parameters. Since there is no interaction between variation post-LGT and deletion (Figure 2.3), we can fix one of these parameters and vary the other. Figure 2.4 shows that for TF-IDF and ALFY, both precision and recall decrease as variation post-LGT increases. The precision of ALFY is worse than that of TF-IDF, but its recall is higher and more stable. When deletion is varied (Figure 2.5), precision is stable except at sequence length 1000, while recall decreases slightly for TF-IDF. As before, TF-IDF is more precise than ALFY, whereas ALFY exhibits higher recall (except at sequence length 1000).

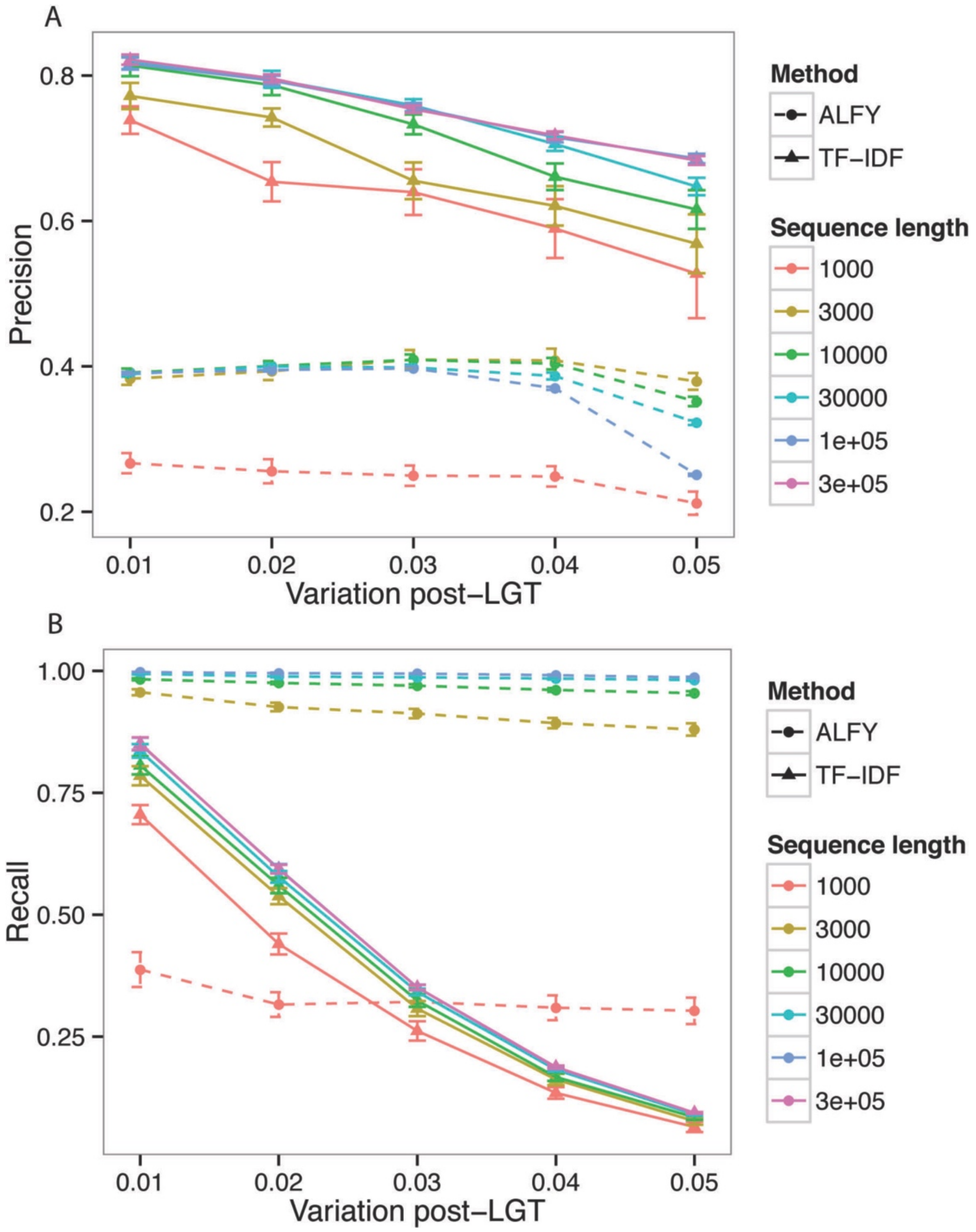


Figure 2.4. Performance of ALFY with variation post-LGT. Precision (A) and recall (B) decrease with variation post-LGT. Variation between groups is 0.1, variation within groups is 0.01, and deletion is zero. Error bars are 2× standard error.

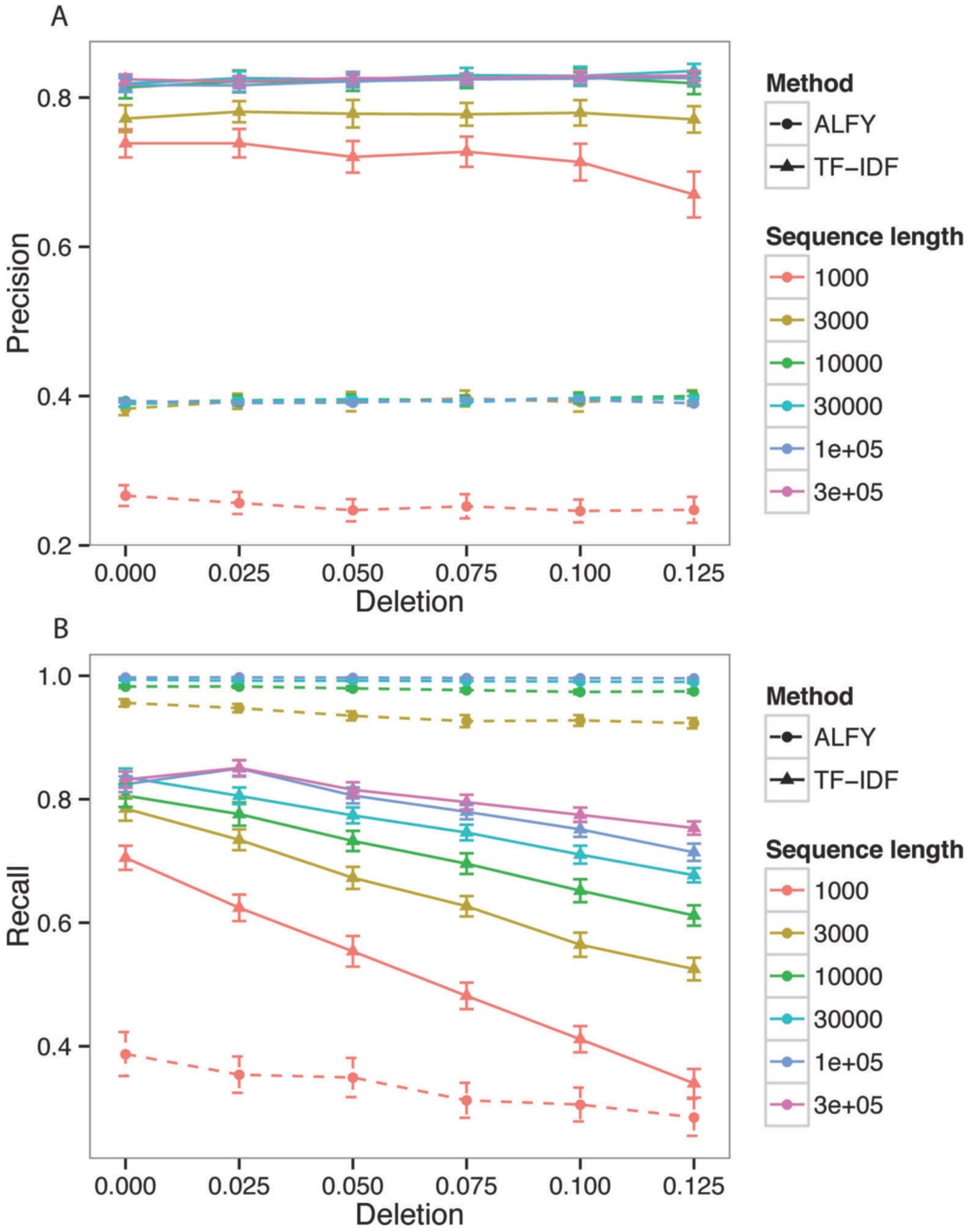


Figure 2.5. Performance of ALFY with deletion. Precision (A) is stable with deletion. Recall (B) decreases with deletion. Variation between groups is 0.1, variation within groups is 0.01, and variation post-LGT is 0.01. Error bars are 2× standard error.

2.2.2 *k*-mer size

k-mer size also affects the performance of TF-IDF. As shown in Figure 2.6, precision increases with *k*, but recall decreases. This effect is roughly consistent for every sequence length we examined. The two plots indicate that in this simulation, precision and recall are best balanced at *k* = 40. Indeed, in our experience (as shown and unpublished) *k* = 40 is a useful default setting, in the absence of conditions that argue otherwise. However, if LGT is sufficiently obscured by substitution such that nearly all *k*-mers are unique, TF-IDF will not be able to find sets of *k*-mers that appear frequently in distant groups, and no LGT will be predicted. In such cases, shorter *k* may give better performance. Note that larger *k* imposes a greater memory cost, and more computational time is spent indexing unique *k*-mers.

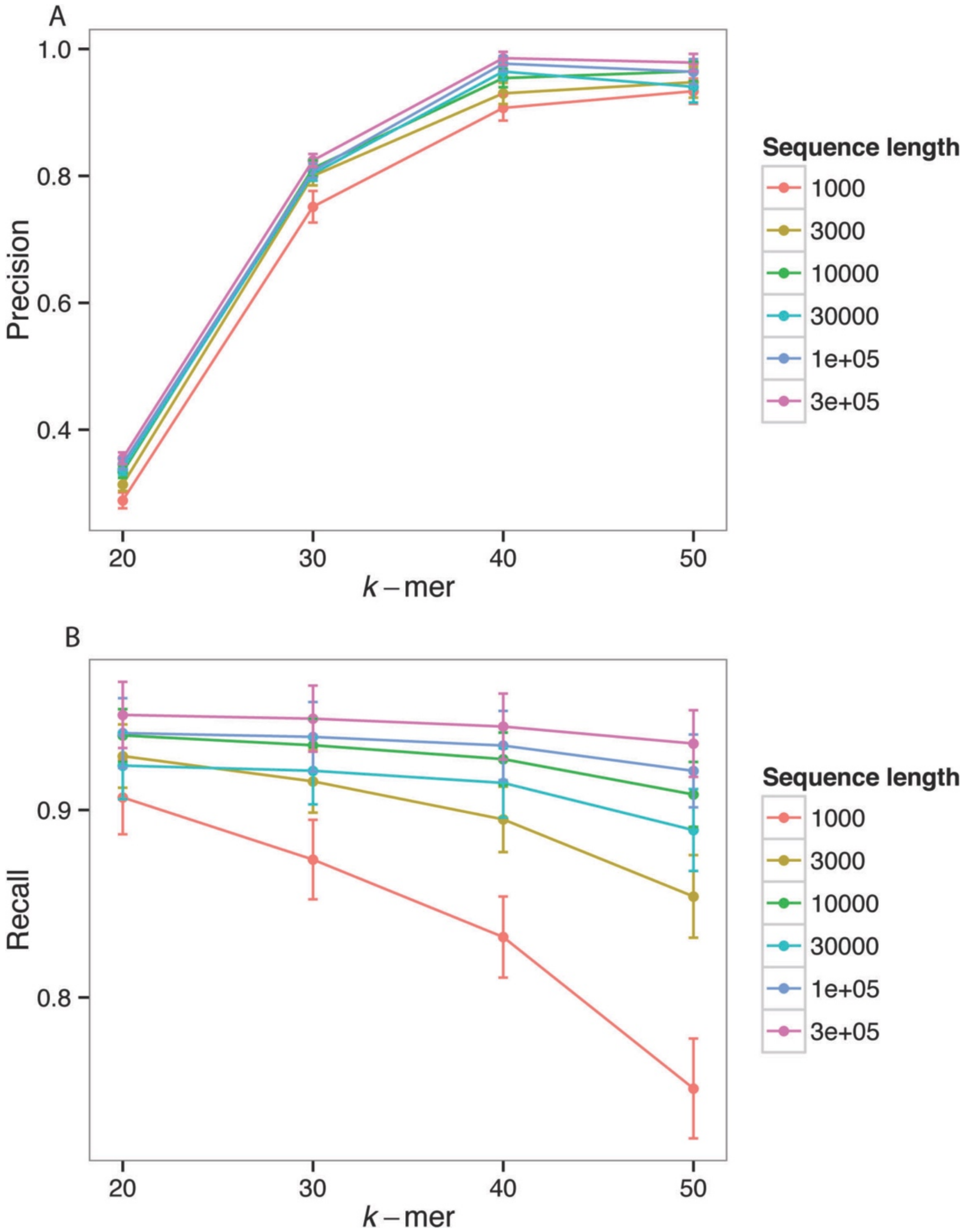


Figure 2.6. Performance of TF-IDF with *k*-mer size. Precision (A) increases with *k*, while recall (B) decreases with *k*. Error bars are 2× standard error.

2.2.3 Computation time

Figure 2.7 compares computation time (walltime) for various sequence lengths *L* for ALFY and TF-IDF. All experiments were done on a virtual machine with a single AMD Opteron 2.3-GHz processor and 256 GB memory. As noted below, TF-IDF is expected to scale as $O(nL \log U)$, where *U* is the number of unique *k*-mers in the dataset. *U* is highly dependent on variation at all levels of the simulation, which also leads to variation of time consumption in each experiment; if the final sequences are sufficiently dissimilar, we expect U to increase as the number of possible k-mers in the dataset, *i.e.* as nL. Thus, we expect the time to have an $O(L \log L)$ dependence on L, and this is verified in Figure 2.7; the slope of the fitted line is 1.07. For ALFY, the time consumption is $O(n^2L)$ for detecting LGTs between all sequences in a dataset. In a dataset with tens of sequences or more, ALFY will take much longer than TF-IDF, and this is shown in Figure 2.7.

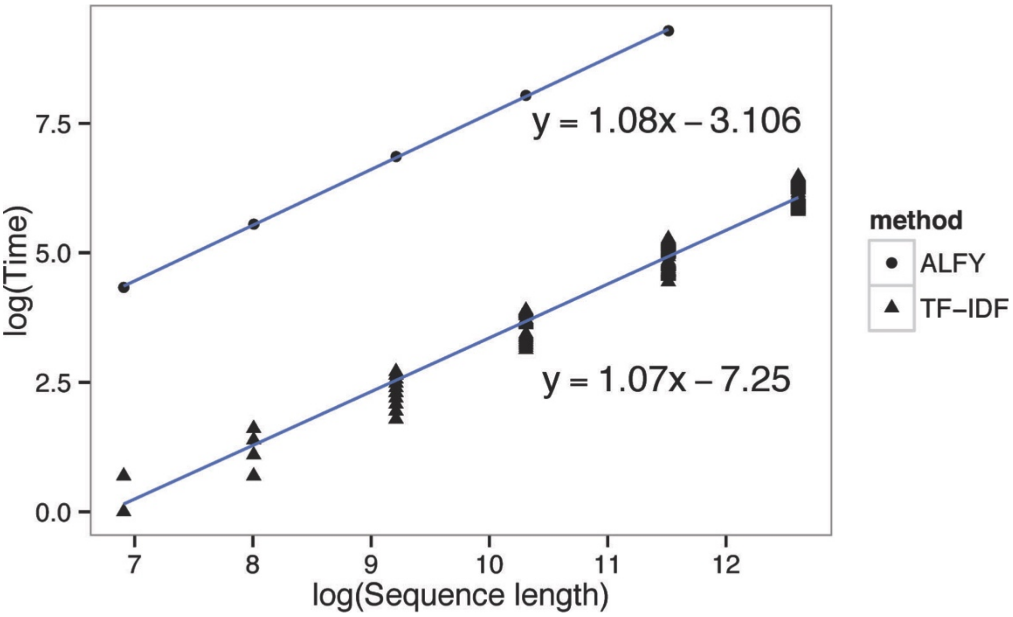


Figure 2.7. Log-log plot of sequence length against walltime**.** The base of log is e.

Figure 2.8 shows how walltime depends on U. As above, we expect time divided by L to have a linear relationship with $\log U$, and this is clearly shown.

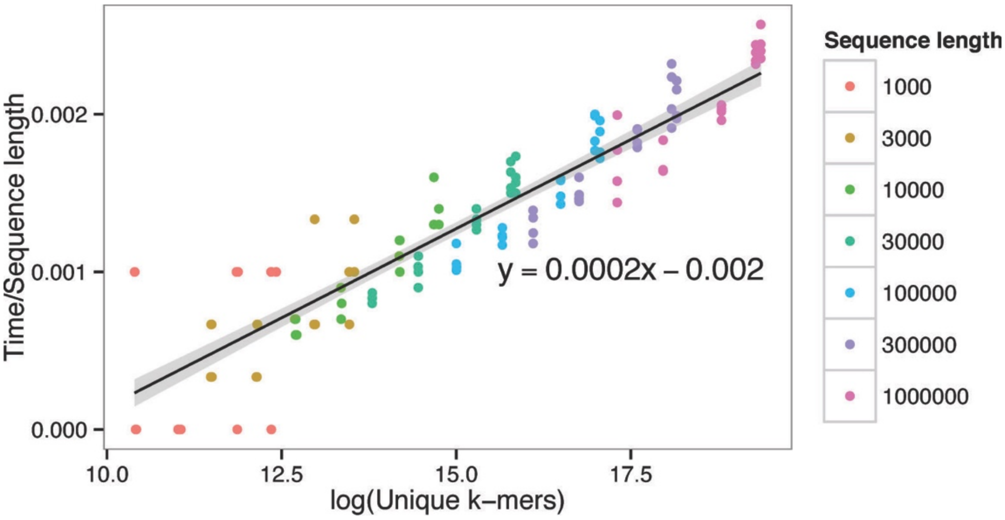


Figure 2.8. *Log U* against time divided by sequence length. The slope of the regression line is 0.0002, and the grey area is the 95% confidence interval. The base of log is e.

2.2.4 Analysis of an empirical dataset

We also tested our algorithm on an empirical dataset that had previously been examined by the developers of ALFY[15]. We used a subset of their dataset, seven genomes of *Staphylococcus aureus*, because this dataset contains strong group information (six genomes from Clonal Complex 8 (CC8) and one multi-drug resistant strain from CC30, *S. aureus* MRSA252) and showed LGT in their analysis. We investigate potential LGT into *S. aureus* TW20, a member of CC8, from MRSA252.

Setting $k = 40$, we identify 1421 regions of TW20 as of lateral origin. Many of these are short and, in this simple example (where the donor group is of size 1, reducing the efficacy of the IDF component) potentially due to noise; but 173 are of length ≥ 2000, 52 of ≥ 4000 and 20 of ≥ 6000 nt (Table 2.1). It is unclear how to optimise selection of the length threshold, but setting it at ≥ 2000 nt we infer as lateral 35.6% of the genome, which incorporates 66.7% (4/6) of the TW20 penicillin-binding genes, and ≥ 50% (*i.e.* > 1.5-fold over-representation) of the annotated genes encoding efflux proteins (2/4), metalloproteinases and -peptidases (3/3), permeases (31/45) and uptake proteins (2/4), types of functions known to be mobilised by LGT[11,30]. For details see Supplementary Table S1. By contrast, hypothetical proteins, which might be expected to show no bias for or against lateral origin, are not enriched at any of the length thresholds mentioned above. Ribosomal proteins, which are not expected to be lateral (Jain *et al.*, 1999), are rarely represented in our lateral regions (8/60). Phage proteins are not represented in our detected lateral regions; recalling that our approach can discover LGT only *within* the dataset, these results might accurately reflect the history of genetic relationships among these seven genomes. Scope remains for further analysis with other empirical data, and with different settings for *k* and gap size.

Table 2.1. Summary of regions in the *Staphylococcus aureus* TW20 genome (GenBank NC_017331.1) inferred as lateral by TF-IDF. Numbers in the top row refer to the length ranges of segments selected for analysis.

	2000-3999		4000-5999		6000+		2000+	
	No.	%	No.	%	No.	%	No.	%
LGT regions	121	8.5	32	2.3	20	1.4	173	12.2
Mean size (nt)	2797	----	4782	----	29600	----	6263	----
Median size (nt)	2786	----	4727	----	10496	----	3112	----
Nucleotides	338413	11.1	153009	5.0	592007	19.5	1083429	35.6
Proteins[1]	405	14.6	169	6.1	515	18.5	1071	39.2
Hypothetical proteins	116	14.3	38	4.7	157	19.3	311	38.3

[1] Protein-coding genes fully or partially contained within a region inferred as lateral by TF-IDF.

Both our TF-IDF method and ALFY identify most of the genomic region from 2.80-0.42 Mb (TF-IDF) or 2.8-0.5 Mb (ALFY) as lateral (Figure 2.9); this region includes two transposons, SCC elements and genes encoding methicillin and penicillin resistance. Robinson and Enright[31] hypothesised that the methicillin resistance, at least, had been transferred from CC30 into a CC8 background as part of a large chromosomal replacement. The region from 1.75-1.80 Mb includes the transposon Tn*554*[25], which encodes resistance to erythromycin and spectinomycin. A region from 2.11-2.15 Mb incorporating a number of annotated phage genes was likewise identified. Regions identified as lateral by TF-IDF but not by ALFY include 1.06-1.17 Mb (transport protein genes) and 2.64-2.65 Mb (a transporter and a member of the TetR family of regulatory proteins, which control the expression of genes involved in multidrug resistance and pathogenicity).

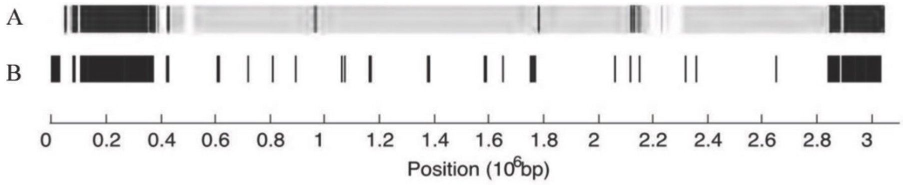


Figure 2.9. Comparison of TF-IDF and ALFY with an empirical dataset. Both A and B represent the genome of *Staphylococcus aureus* TW20. A shows the result of ALFY analysis[15]; regions inferred to have been transferred from MRSA252 are represented in black, while regions homologous between TW20 and USA300.TCH15156 are shown in grey. B shows the result of TF-IDF analysis. TF-IDF can infer LGT only from outside the target group, so no region is in grey. Both plots were generated from analysis of the seven *S. aureus* genome dataset.

2.3 Discussion and Conclusion

We have developed a fast alignment-free method to infer LGT events. Our method is based on TF-IDF, one of the most important methods used in information retrieval. TF-IDF has been widely applied in search engines, document classification and related applications including relevance decision-making. Here we apply TF-IDF to sequence analysis for the first time, treating a sequence or genome as an article and each *k*-mer as a word. Using simulated datasets, we show that TF-IDF can effectively find LGT events with good precision and recall, outperforming ALFY in most biologically realistic situations. We also analyse an empirical dataset and show that TF-IDF finds essentially all regions identified by ALFY as of lateral origin. TF-IDF further detects other regions that, based on annotated gene content, may also have arisen *via* LGT. Our method is alignment-free and scales very well in both length and number of sequences, *i.e.* to many entire genomes. It is worth noting that in each simulated dataset, all sequences share the same length and group size. For the empirical dataset, the group sizes and lengths of the seven *S. aureus* are of the same magnitude. For this reason, we did not normalise the count of *k*-mers in the IDF step. However, in other empirical datasets the sequence length and group size may vary greatly, and normalisation might be considered[28].

Our method is purely data-driven, its performance relying strongly on sequence and group information in the dataset. In our simulations, when sequences are relatively similar within-group (variation 0.005-0.02) and relatively dissimilar between-group (variation > 0.1), group boundaries are clear, and the precision and recall of our algorithm is high. When speciation is modest (< 0.05), within-group divergence high (< 0.1) or LGT events obscured by subsequent evolution (> 0.02), TF-IDF loses precision in inferring LGT events.

In the accompanying article[28] we apply this method to larger empirical datasets. TF-IDF could further be applied to environmental data, e.g. to study the flow of genetic material in communities and across the biosphere. We anticipate that significant scope remains for further algorithmic and implementational improvements.

2.4 Methods

2.4.1 Notation

Here we establish some notation. We start with a dataset of n sequences, each of length L. For empirical datasets (and for some approaches to simulation) the length may vary among sequences; in those cases we use L to denote the average length. The sequences in the dataset are divided into m groups corresponding to closely related genomes (*e.g.* belonging to the same clonal group, species or genus). We denote each sequence as $S_{i,j}$, where i = 1,2,…,n is the number of the sequence in the dataset and j = 1,2,…,m is the number of the group to which the sequence belongs. The number of sequences in group j is denoted by h_j.

Our method proceeds by comparing substrings (words) of a fixed length k, called k-mers. We encode each sequence as a frequency vector of k-mers, counting only those k-mers that actually appear in the sequence, and denoting the number of unique k-mers appearing in the dataset by U. In general, U is much smaller than 4^k, the total number of all possible k-mers.

Although we illustrate our approach here using nucleotide sequences, the method is easily adapted for amino acids, requiring only a change of alphabet.

2.4.2 TF-IDF on texts

As mentioned above, TF-IDF was introduced to indicate the topic of a document, and distinguish that document from others in the same corpus for a specific query. The classical usage of TF-IDF is as a smart retrieval system and for automatic document categorisation[32-34]. A variant uses prototype vectors to calculate relevance between documents with a nearest-neighbor learning method[35]. PrTFIDF[36] is an improved version of TF-IDF founded on a probabilistic model for text categorization, and there are other variants for calculating TF-IDF[37]. In recent years, TF-IDF has also been applied in other areas including decision-making and sentiment analysis[24,38].

TF-IDF is widely used in text mining and information retrieval because it allows the identification of terms that are characteristic of (and hence important for) one text or a set of texts. It is not sufficient for a term to be frequent in a text (TF); it must also be rare in other texts in the corpus (IDF). Importantly, IDF depends only on the occurrence of terms, not on their numerical frequencies. Drawing on analysis of documents in three independent domains, Salton and Yang[39] identified five situations relevant to the performance of TF-IDF:

1. Terms that appear frequently across a corpus contribute little to performance because they do not discriminate between relevant and non-relevant documents;
2. Terms that appear in a moderate number of texts and show somewhat skewed distributions provide good retrieval performance;
3. Terms with sharply skewed distribution occurring in very few documents are important only for those documents;
4. Rare terms are important for the few queries and documents in which they occur; and
5. Terms of low or moderate frequency, but with a flat distribution across documents, are similarly useful for the documents in which they occur.

Classically, the frequencies of terms in a corpus follow a power law (Zipf law), in which case TF-IDF performs well. However, TF-IDF can perform adequately even when this is not the case: TF-IDF requires only that terms relevant to the query are distributed intensively in a subset of documents within the corpus[23,40]; this might include the query terms themselves (*e.g. happy*), or related terms in the corpus (*pleased*, *delighted*).

2.4.3 TF-IDF on sequences

Molecular sequences have long been analogised with natural language[41] or treated as texts[42]. Alternatively, both molecular sequences and texts have been subsumed within a broader class of objects[43]. The analogy is not precise: in sequences, “terms” must be recognized computationally, *e.g.* by extracting *k*-mers. Fast approaches exist for extracting *k*-mers[44,45], and *k*-mer distribution in empirical sequences has been studied at some length[46-48]. Like words in text, short *k*-mers (*k* between three and eight) in DNA sequences show Zipf-like scaling[49], although this is not sufficient to confirm DNA sequences as a natural language[50].

Although there is dispute whether DNA is a language or not, some methods in text mining have been successfully applied to DNA analysis. For example, the first (to our knowledge) software to identify lateral transfer in biological datasets[51] was repurposed from the analysis of textual contamination in manuscripts, which in turn was built on software for phylogenetic inference from DNA sequences[52] (PHYLIP).

Sequences (genomes, genes, proteins) do, however, differ from texts in some properties. For example, *k*-mer frequency distributions in sequences are usually much flatter than term frequencies in texts. Experience from text mining indicates that this is not critical, but this remains to be explored and is in fact a goal of the current work. In the specific application here, genomic regions of lateral origin are expected to have *k*-mers that appear frequently in genomes of the donor taxon, but rarely in the host. This is analogous to conditions 2 and/or 4 above[39].

Our algorithm works by comparing the frequencies of identified *k*-mers in a group of sequences (our TF) with their frequencies in other groups (our IDF). If a *k*-mer in one sequence is prevalent in a different group but not in its own, then it may have arisen by LGT from the group in which it is prevalent, and the direction of the transfer should be from that (donor) group to the recipient sequence. We compare these TF and IDF statistics to appropriate thresholds to optimize detection performance, *i.e.* to balance precision and recall.

Our algorithm consists of four steps: extracting all *k*-mers from genomes within one dataset, calculating inverse document frequencies, constructing potential LGT segments, and calculating term frequencies. For pseudocode, see the Supplementary file.

2.4.4 Extracting k-mers

To extract *k*-mers we scanned all the genomes, incrementing one nucleotide at one time. If the genome length is L, then L-k+1 *k*-mers are found. Unique *k*-mers were indexed in a red-black tree[53] for further searching.

2.4.5 Calculating IDF

To calculate the inverse document frequency, we construct an $n \times m$ relationship matrix R, denoting the frequency (number of occurrences) at which *k*-mers in each sequence appear in each group. Each row in R corresponds to a sequence, and each column corresponds to a group. Suppose sequence i consists of *k*-mers $w_{i,1}$, $w_{i,2}$,…, $w_{i,L\text{-}k+1}$. If the word w appears in group j with frequency $f_j(w)$, then the entries of the relationship matrix are

$$R_{ij} = \sum_{a=1}^{L-k+1} f_j(w_{i,a}). \qquad (1)$$

The entries in R are our IDF values. Eq. (1) is applied only on $j \neq \Theta(i)$, where $\Theta(i)$ is the group containing sequence i. The larger the R_{ij}, the more likely that sequence i contains a region transferred laterally from group j. Note that this is in contrast to the original definition of IDF, where a higher IDF indicates that the word appears less frequently in other documents.

To detect potential lateral-transfer events, we compare the IDF values against a threshold t. This threshold is the average value of all entries in R:

$$\boldsymbol{t = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}.} \qquad \mathbf{(2)}$$

IDF values that are above the average are used for further analysis.

2.4.6 Constructing potential LGT segments

We then mark potential lateral segments in each sequence. For each sequence i and group j with a sufficiently high IDF value, we examine each k-mer in sequence i to see if it appears in group j. Then we join all consecutive k-mers which do, forming potential lateral segments. Because mutations or other genomic events may disrupt the perfect matching, we allow gaps between blocks of k-mers of size up to a threshold G. Here we set $G = 2k$, a value at which the total number of detections is not greatly affected in real application[28]. We then assess the significance of these potential lateral segments using term frequency.

2.4.7 Calculating TF

For each potential lateral segment σ in a sequence, we calculate the frequency (number of occurrences) at which each of its component k-mers appears in sequences of its own group, say j. Our TF statistic for σ is the sum of these:

$$\boldsymbol{\delta_\sigma = \sum_{w\in\sigma} f_j(w).} \qquad \mathbf{(3)}$$

If δ_σ is higher than some threshold, then σ occurs frequently in its own group, and as such is considered not to be the consequence of a lateral event; otherwise it is considered to be of lateral origin.

To set the threshold, we calculate the average frequency of all unique *k*-mers in the recipient group j, denoted by τ_j. Then we compare δ_σ to $l\tau_j$, where *l* is the number of *k*-mers contained in the segment. If δ_σ is smaller, we consider σ to have been transferred laterally from the other group into this sequence. Other approaches to setting the threshold are possible, but we do not consider them here.

Note that our method considers lateral transfers only within the dataset; like most other LGT methods, it is silent on potential transfers from sources external to the dataset. In addition, it can detect transfers only between groups, not between sequences in the same group.

2.4.8 Runtime analysis

The computational complexity of the algorithm is dominated by extraction of the unique *k*-mers in the dataset. To find these, we scan each of the *n* sequences of length L. As each unique *k*-mer is found it is added to a library, which is stored in a red-black tree[53]. A red-black tree is an approximately balanced tree, which guarantees that searching and insertion are efficient. On average, this step takes $O(nL \, log \, U)$ time, where U is the number of unique *k*-mers stored in the tree. The frequency of each *k*-mer is also computed at this time. The remaining calculations are much quicker because most of the frequency (*f)* terms are zero. Thus for biological sequences of standard complexity, runtime increases about log-linearly with sequence length. Note that the *k*-mer profiles of each sequence could in principle be stored and retrieved for future use.

2.4.9 Implementation

We have implemented this algorithm in C++. The program can be compiled using GCC 4.8.2 and run on Unix, Unix-like and Windows platforms. We use the *map* template from STL (Standard Template Library) to index all distinct *k*-mers in a dataset. The inner implementation of *map* is a red-black tree[53].

2.4.10 Comparisons with ALFY

ALFY finds putative homology (shared DNA segments) between pairs of sequences by matching shustrings (shortest unique substrings). If a match is found with a region in an otherwise distant sequence, it will be judged as a potential lateral transfer. This method shows high efficiency and effectiveness for LGT detection[15,16], so we use it to benchmark our method.

The inputs to both TF-IDF and ALFY are sequences. For TF-IDF the group information is compulsory, while ALFY requires a query sequence and subject sequences. Both TF-IDF and ALFY can process DNA sequences; TF-IDF can also process amino-acid sequences, but ALFY does not currently implement evolutionary models of amino-acid change. Only *k*-mer frequencies will be taken into consideration for calculating the value of TF-IDF.

In TF-IDF, if a *k*-mer has low a frequency in its own group but high frequencies in other groups, then this *k*-mer will be judged atypical. A set of contiguous atypical *k*-mers will be inferred as lateral, with the direction of the transfer from the *k*-mer prevalent group. In contrast, ALFY computes the average shustring length between segments of only two sequences at a time. The longer the average shustring, the closer the two segments; and if the sequences themselves are otherwise distant in the reference tree, the segment in question will be inferred as lateral, without any implication of which sequence was donor or recipient.

If the sequences are grouped such that each group is compact and boundaries between groups are clear, then TF-IDF should find lateral segments easily. ALFY does not use group information, so grouping does not affect its performance.

The computational complexity of TF-IDF is $O(nL \log nL)$, where n is total number of sequences in a dataset, and L the average length of sequences in a dataset. The computational complexity of the ALFY algorithm is $O(nL)$. However, TF-IDF will process all sequences and infer all potential lateral regions over an entire dataset, whereas ALFY makes all pairwise comparisons between a single query sequence and the others. For fairness of comparison, all sequences in a dataset should be set as queries to find all LGTs in a dataset, in which case the complexity of ALFY increases to $O(n^2L)$, which in practice is much slower than TF-IDF.

2.4.11 Simulation of datasets

In order to test the performance of TF-IDF in different situations, and to compare with ALFY, we simulated datasets under the HYK85[26] and F84[27] evolutionary models. Our simulation process is as follows:

(1) We start with one random sequence, which will become the ancestor of all sequences in the dataset. We vary the length L of this sequence from 1000 to 300000 characters to simulate sequences from a single gene to a significant part of a genome (but our algorithm can be applied to sequences of any length).

(2) To establish phyletic groups (*i.e.* to simulate speciation), the ancestral sequence is allowed to evolve along a balanced binary tree with four levels of equal branch lengths, using the evolutionary model. The branch length varies from 0.01 to 0.20 (substitutions per site) in steps of 0.05. We refer to this parameter as *variation between groups*.

(3) To populate these groups with sequences, each descendant (leaf) in the initial tree (above) is allowed to evolve along another phylogenetic tree under the same evolution model. Again we use a balanced binary tree with four levels of equal branch length, which vary from 0.001 to 0.020 in steps of 0.005. We refer to this parameter as *variation within groups*.

(4) We then simulate LGT events between groups. For the sake of simplicity, here we make transfers only into sequences in Group 1. We fix the number of LGT events at 20, with lengths normally distributed around mean 0.1 of the sequence length, and standard deviation half that amount. For each simulated event the recipient sequence (in Group 1) is selected at random, with (typically) several sequences receiving multiple transfers and others receiving none. Transfer events overwrite the equivalent positions in the recipient sequence, but (to simplify our simulation) cannot themselves be subsequently overwritten. Five of the 20 LGT events are simulated to come from the group (of 16 sequences) arising from the most-recent common ancestor on the binary tree (from Step 2), five from descendants of the second-most recent ancestor (32 sequences), five from the third (64 sequences) and five from the deepest bifurcation (128 sequences). Thus the probability of transfer decreases with increasing distance (on the tree) between donor and target.

(5) In a final evolutionary process, we further evolve each of the 256 sequences along a balanced two-level tree, with branch lengths varying from 0 to 0.1 in steps of 0.025. We refer to this parameter as *variation post-LGT*.

(6) In some simulations, Step 5 also includes a stochastic process (implemented by using a shell script to call ALF[54], not to be confused with ALFY) which deletes from 0 to 10% of a sequence. The proportion was varied using the *deletion rate* setting in ALF, while keeping *deletion length distribution* at its default value. We refer to this parameter as *deletion*. We did not simulate duplications here because bacterial genomes contain very few repetitive components.

After the above steps, we select one descendant of each tree to yield our final dataset (256 sequences per simulation).

In addition to varying the parameters mentioned above for both TF-IDF and ALFY, for TF-IDF only we also varied the word length *k*, in steps of 10 from 20 to 50. As the number of possible parameter combinations above is very large, at Step 2 we varied only the *variation between groups* parameter while keeping all others fixed at minimal-impact settings. Similarly at Step 3 we varied only the *variation within groups* parameter. For each parameter combination we simulated 50 datasets under the F84 model of sequence change, and 50 under HYK85. This process is depicted in Figure 2.10, and is explained in greater detail in the Supplementary file. We also analysed smaller datasets omitting Step 4, to examine whether TF-IDF inferred LGT when none was present; no segments met the IDF (*k*-mers frequent in donor groups) and TF (*k*-mers infrequent in the recipient group) criteria simultaneously, so no LGT was inferred.

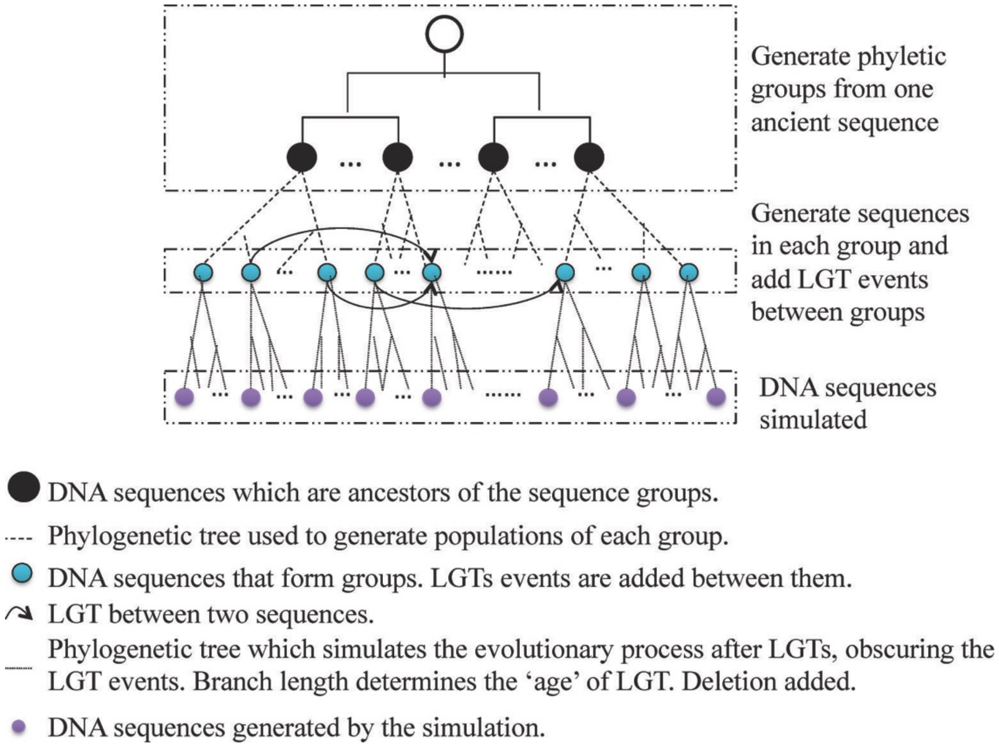

Generate phyletic groups from one ancient sequence
Generate sequences in each group and add LGT events between groups
DNA sequences simulated
● DNA sequences which are ancestors of the sequence groups.
---- Phylogenetic tree used to generate populations of each group.
● DNA sequences that form groups. LGTs events are added between them.
LGT between two sequences.
Phylogenetic tree which simulates the evolutionary process after LGTs, obscuring the LGT events. Branch length determines the 'age' of LGT. Deletion added.
● DNA sequences generated by the simulation.

Figure 2.10. Overview of data simulation. Flowchart of the simulation process. The simulation starts with a single ancestor and generates 16 sequences, which serve as ancestors for each group (variation between groups). Within each group we generate 16 descendants (variation within groups). Following the simulation of lateral genetic transfer, we simulate a final stage of evolution (variation post-LGT) which may include deletion. From each initial ancestor the simulation generates 256 sequences.

2.4.12 Performance measures

We assessed the performance of our algorithm on simulated data using two measures. Precision is the proportion of inferred LGT events which are real (*i.e.* were actually simulated):

$$Precision = tp \,/\, (tp + fp) \qquad (4)$$

where *tp* and *fp* are the total lengths of all true and false positives respectively. Recall is the proportion of true (simulated) LGTs which were inferred by the algorithm:

$$Recall = tp \,/\, (tp + fn) \qquad (5)$$

where *fn* is the total length of false negatives (simulated LGTs which were not found).

Figure 2.11 illustrates the output of TF-IDF analysis of a simulated dataset, showing the 20 regions of (simulated) lateral origin of which 19 were detected (wholly or in part) by TF-IDF. Positions 797-877 of Sequence 11 represent a false positive inference of LGT, and positions 58-117 of Sequence 2 a false negative. Overall for this dataset (*i.e.* LGT from Groups 2-16 into Group 1), precision was

0.82 and recall 0.95. Complete details (start and end coordinates) are presented in the Supplementary file.

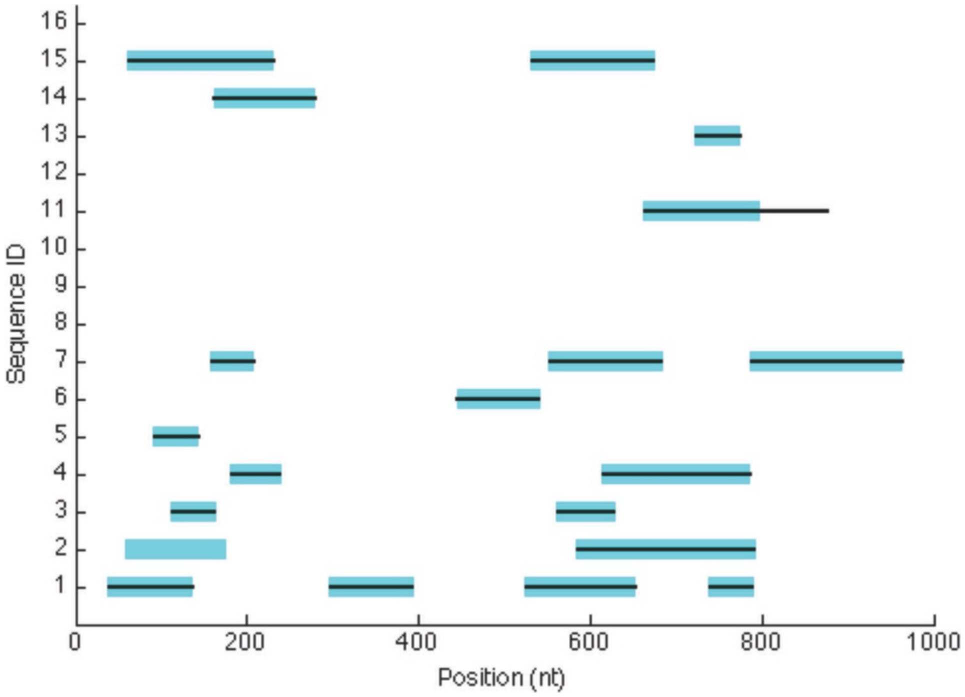


Figure 2.11. An example of simulated and inferred LGTs. The *x*-axis displays the nucleotide position, and the *y*-axis the sixteen sequences generated in our first (recipient) group. The wide bars show the lateral regions actually simulated, and the narrow black bars the regions inferred as lateral by TF-IDF. Here, variation between groups is 0.1, variation within groups is 0.001, variation post-LGT is 0.01, deletion is zero, *k* = 40 and sequence length is 1000 nt.

2.5 References

1 Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304, doi:10.1038/35012500 (2000).

2 Schmitt, R. M. Zur Variablilität der Enteritis-bakterien. *Zeitschr Infektionskrankh parasit Krankh Hyg Haustiere* **9**, 188 (1911).

3 Davies, J. Origins and evolution of antibiotic resistance. *Microbiologia* **12**, 9-16 (1996).

4 Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284**, 2124-2129 (1999).

5 Martin, W. Mosaic bacterial chromosomes: a challenge on route to a tree of genomes. *Bioessays* **21**, 99-104, doi:10.1002/(Sici)1521-1878(199902)21:2<99::Aid-Bies3>3.0.Co;2-B (1999).

6 Beiko, R. G., Harlow, T. J. & Ragan, M. A. Highways of gene sharing in prokaryotes. *Proc. Natl Acad. Sci. USA* **102**, 14332-14337, doi:10.1073/pnas.0504068102 (2005).

7 Raymond, J., Siefert, J. L., Staples, C. R. & Blankenship, R. E. The natural history of nitrogen fixation. *Mol Biol Evol* **21**, 541-554, doi:10.1093/molbev/msh047 (2004).

8 Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* **3**, 711-721 (2005).

9 Skippington, E. & Ragan, M. A. Lateral genetic transfer and the construction of genetic exchange communities. *FEMS Microbiol Rev.* **35**, 707-735, doi:10.1111/j.1574-6976.2010.00261.x (2011).

10 Chan, C. X., Darling, A. E., Beiko, R. G. & Ragan, M. A. Are protein domains modules of lateral genetic transfer? *PLoS ONE* **4**, e4524, doi:10.1371/journal.pone.0004524 (2009).

11 Ragan, M. A. & Beiko, R. G. Lateral genetic transfer: open issues. *Philos Trans R Soc Lond B Biol Sci* **364**, 2241-2251, doi:10.1098/rstb.2009.0031 (2009).

12 Lawrence, J. G. & Ochman, H. Amelioration of bacterial genomes: Rates of change and exchange. *J Mol Evol* **44**, 383-397, doi: 10.1007/Pl00006158 (1997).

13 Ragan, M. A. On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol Lett* **201**, 187-191, doi:10.1111/J.1574-6968.2001.Tb10755.X (2001).

14 Lawrence, J. G. & Ochman, H. Reconciling the many faces of lateral gene transfer. *Trends Microbiol* **10**, 1-4, doi:10.1016/S0966-842x(01)02282-X (2002).

15 Domazet-Lošo, M. & Haubold, B. Alignment-free detection of horizontal gene transfer between closely related bacterial genomes. *Mob Genet Elements* **1**, 230-235, doi:10.4161/mge.1.3.18065 (2011).

16 Domazet-Lošo, M. & Haubold, B. Alignment-free detection of local similarity among viral and bacterial genomes. *Bioinformatics* **27**, 1466-1472, doi:10.1093/bioinformatics/btr176 (2011).

17 Domazet-Lošo, M. & Haubold, B. Efficient estimation of pairwise distances between genomes. *Bioinformatics* **25**, 3221-3227, doi:10.1093/bioinformatics/btp590 (2009).

18 Saitou, N. & Nei, M. The neighbor-joining method - a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406-425 (1987).

19 Taniguchi, Y., Yamada, Y., Maruyama, O., Kuhara, S. & Ikeda, D. The purity measure for genomic regions leads to horizontally transferred genes. *J Bioinf Comput Biol* **11**, 1343002, doi:Artn 1343002 doi:10.1142/S0219720013430026 (2013).

20 Gusfield, D. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* (Cambridge University Press, 1997).

21 Luhn, H. P. The automatic creation of literature abstracts. *IBM J Res Dev* **2**, 159-165 (1958).

22 Spärck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *J Docum* **28**, 11-21 (1972).

23 Salton, G. & Buckley, C. Term-weighting approaches in automatic text retrieval. *Inform Process Manag* **24**, 513-523, doi: 10.1016/0306-4573(88)90021-0 (1988).

24 Wu, H. C., Luk, R. W. P., Wong, K. F. & Kwok, K. L. Interpreting TF-IDF term weights as making relevance decisions. *ACM Trans Inform Syst* **26**, doi:Artn 10.1145/1361684.1361686 (2008).

25 Holden, M. T., Lindsay, J. A., Corton, C., Quail, M.A., Cockfield, J.D., Pathak, S., Batra, R., Parkhill, J., Bentley, S. D. & Edgeworth, J. D. Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant *Staphylococcus aureus*, sequence type 239 (TW). *J Bacteriol* **192**, 888-892, doi:10.1128/JB.01255-09 (2010).

26 Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**, 160-174 (1985).

27 Felsenstein, J. & Churchill, G. A. A hidden Markov model approach to variation among sites in rate of evolution. *Mol Biol Evol* **13**, 93-104 (1996).

28 Cong, Y., Chan, Y.-b. & Ragan, M. A. Exploring lateral genetic transfer among microbial genomes using TF-IDF. *Sci Rep* **6**, 29319 (2016).

29 Popa, O., Hazkani-Covo, E., Landan, G., Martin, W. & Dagan, T. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res* **21**, 599-609, doi:10.1101/gr.115592.110 (2011).

30 Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: The complexity hypothesis. *Proc Natl Acad Sci U S A* **96**, 3801-3806, doi: 10.1073/Pnas.96.7.3801 (1999).

31 Robinson, D. A. & Enright, M. C. Evolution of *Staphylococcus aureus* by large chromosomal replacements. *J Bacteriol* **186**, 1060-1064, doi: 10.1128/Jb.186.4.1060-1064.2004 (2004).

32 Salton, G. *The SMART Retrieval System; Experiments in Automatic Document Processing* (Prentice-Hall, 1971).

33 Salton, G. & McGill, M. J. *Introduction to Modern Information Retrieval* (McGraw-Hill, 1983).

34 Salton, G. & McGill, M. J. The SMART and SIRE experimental retrieval systems in *Readings in information retrieval* (eds Karen Spärck, J. & Willett. P.), 381-399 (Morgan Kaufmann Publishers Inc., 1997).

35 Salton, G. Developments in automatic text retrieval. *Science* **253**, 974-980, doi: 10.1126/Science.253.5023.974 (1991).

36 Joachims, T. A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization. In *Proc Fourteenth Int Conf Machine Learning*, 143-151 (1996).

37 Justin, Z. M., Alistair. Exploring the similarity space. *SIGIR Forum* **32**, 18-34 (1998).

38 Paltoglou, G. & Thelwall, M. A study of information retrieval weighting schemes for sentiment analysis. *Proc 48th Ann Mtg Assoc Comput Linguistics* 1386-1395, (Association for Computational Linguistics, Uppsala, Sweden, 2010).

39 Salton, G. & Yang, C.-S. On the specification of term values in automatic indexing. *J Docum* **29**, 351-372 (1973).

40 Salton, G., Yang, C.-S. & Yu, C. T. A theory of term importance in automatic text analysis. *J Am Soc Inf Sci* **26**, 33-44 (1975).

41 Nussinov, R. Some rules in the ordering of nucleotides in the DNA. *Nucl Acids Res* **8**, 4545-4562 (1980).

42 Koonin, E. V. & Galperin, M. Y. *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics* (Kluwer Academic, 2003).

43 Kruskal, J. B. An overview of sequence comparison - time warps, string edits, and macromolecules. *SIAM Rev* **25**, 201-237, doi:10.1137/1025045 (1983).

44 Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764-770 (2011).

45 Greenfield, P., Duesing, K., Papanicolaou, A. & Bauer, D. C. Blue: correcting sequencing errors using consensus and context. *Bioinformatics* **30**, 2723-2732 (2014).

46 Chor, B., Horn, D., Goldman, N., Levy, Y. & Massingham, T. Genomic DNA *k*-mer spectra: models and modalities. *Genome Biol* **10**, R108, doi:10.1186/gb-2009-10-10-r108 (2009).

47 Burden, C. J., Leopardi, P. & Forêt, S. The distribution of word matches between Markovian sequences with periodic boundary conditions. *J. Comput Biol* **21**, 41-63, doi: 10.1089/Cmb.2012.0277 (2014).

48 Kurtz, S., Narechania, A., Stein, J. C. & Ware, D. A new method to compute *k*-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* **9**, 517 (2008).

49 Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, S. K., Simons, M. & Stanley, H. E. Linguistic features of noncoding DNA sequences. *Phys Rev Lett* **73**, 3169-3172, doi:Doi 10.1103/Physrevlett.73.3169 (1994).

50 Tsonis, A. A., Elsner, J. B. & Tsonis, P. A. Is DNA a language? *J Theor Biol* **184**, 25-29, doi: 10.1006/Jtbi.1996.0239 (1997).

51 Ragan, M. A. & Lee, A. R. III Making phylogenetic sense of biochemical and morphological diversity among the protists, in *The Unity of Evolutionary Biology: 4th International Congress of Systematic and Evolutionary Biology* (ed Dudley, T. R.) **2**, 432-441 (Dioscorides Press, Portland, Oregon, 1991).

52 Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, Universityy of Washington, Seattle.* (2005).

53 Guibas, L. J. & Sedgewick, R. A dichromatic framework for balanced trees. *Proceedings of the 19th Annual Symposium on Foundations of Computer Science*, 8-21 (Institute of Electrical and Electronics Engineers, 1995).

54 Dalquen, D. A., Anisimova, M., Gonnet, G. H. & Dessimoz, C. ALF -- a simulation framework for genome evolution. *Mol Biol Evol* **29**, 1115-1123, doi:10.1093/molbev/msr268 (2012).

2.6 Acknowledgements

Simulations requiring large-memory hardware were performed on QRIScloud (Queensland Cyber Infrastructure Foundation). *Funding*: YC acknowledges the China Scholarship Council and The University of Queensland for stipend support. The project was funded in part by a grant from the James S. McDonnell Foundation to MAR.

Supplementary Material

A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF

Yingnan Cong
Yao-ban Chan
Mark A. Ragan

1 Pseudocode for the TF-IDF algorithm

```
Begin
  Recognise all the different $k$-mers of size $k$ and label as $1, 2, \dots, U$.
  // Compute Matrix M
  $\mathbf{M} \leftarrow Zeros(n \times U)$      // $Zeros$ returns an all-zero matrix of the specified dimensions.
  For each sequence $i$ do
    For each $k$-mer $\kappa$ in sequence $i$ do
      $\mathbf{M}(i, L(\kappa)) \leftarrow \mathbf{M}(i, L(\kappa)) + 1$    // $L(\kappa)$ returns the label of $k$-mer $\kappa$.
    End For
  End For
  // Compute Matrix R
  $\mathbf{R} \leftarrow Zeros(n \times m)$
  For each sequence $i$ do
    For each group $j$ except $\Theta(i)$ do          // $\Theta(i)$ returns the group of sequence i.
      For each sequence $i'$ in group $j$ do
        $\mathbf{R}(i, j) \leftarrow \mathbf{R}(i, j) + \mathbf{\Omega}(i, i')$    // $\mathbf{\Omega}(i, i')$ returns the number of
                                                      // common elements between sequences i and i'.
      End For
    End For
  End For
  $t \leftarrow \mathbf{avg}(\mathbf{R})$          // $\mathbf{avg}(\mathbf{R})$ returns the average value of elements of R.
  // Compute $\tau$
  $\boldsymbol{\tau} \leftarrow Zeros(m)$
  For each group $j$ do
    K ← $\mathbf{M}(j, \circ)$          // $\mathbf{M}(j, \circ)$ returns the rows that represent the sequences in $j$
    $\tau(j) \leftarrow numel(\mathrm{K})/numel(unique(\mathrm{K}))$
  End For
  // Detect LGTs
  $(i, j, v) \leftarrow \mathbf{F}\boldsymbol{max}(\mathbf{R})$   // $\mathbf{F}\boldsymbol{max}(\mathbf{R})$ returns the maximum value of R as v, with
                                 // corresponding sequence i and group j.
  While $v > t$
    // Cut sequence $i$
    $\boldsymbol{\omega} \leftarrow Zeros(m)$
    For each $k$-mer $\kappa$ in sequence $i$ do
      $\boldsymbol{\omega}(i) \leftarrow \boldsymbol{ismember}(\kappa, j)$          // $\boldsymbol{ismember}(\kappa, j)$ returns 1 if $\kappa$ exists in at
                                               // least one sequence of species j, 0 otherwise.
    End For
    $TagS \leftarrow 0$
    $TagE \leftarrow 0$
    $Intrpt \leftarrow 0$
    For each element $\zeta$ ($p$-th) in $\boldsymbol{\omega}$ do
      If $\zeta$=1
```

```
                **If** $TagS = 0$
                    $TagS \leftarrow p$
                    $TagE \leftarrow p$
                **Else**
                    $TagE \leftarrow p$
                **End If**
                $Intrpt \leftarrow 0$
            **Else**
                **If** $Tag \neq 0$
                    $Intrpt \leftarrow Intrpt + 1$
                    **If** $Intrpt > 2 \times k$
                        Add $(TagS, TagE)$ to $\boldsymbol{f}$          // $\boldsymbol{f}$ denotes segments of interest
                        $TagS \leftarrow 0$
                    **End If**
                **End If**
            **End If**
        **End For**
        **For** each segment $\phi$ in $\boldsymbol{f}$ **do**
            $\varepsilon \leftarrow 0$
            **For** each sequence $i'$ in group $\vartheta(i)$ **do**
                **For** each $k$-mer $\kappa$ in $\phi$ **do**
                        $\varepsilon \leftarrow \varepsilon + \mathbf{M}(i', L(\kappa))$
                **End For**
            **End For**
            **If** $\varepsilon < \boldsymbol{\Gamma}(i) \times l$       // l denotes the length of the fragment, i.e., $TagE - TagS$
                Add $\boldsymbol{f}$ as a LGT.
                // Update Matrix **R**
                $\mathbf{R}(i, j) \leftarrow 0$
            **End If**
        **End For**
        $(i, j, v) \leftarrow \mathbf{F}\boldsymbol{max}(\mathbf{R})$
    **End While**
**End Begin**
```

2 Details of our simulation

The simulation process is as follows:

Step 1: Generate groups

- Generate one random sequence as ancestor of all sequences.
- Set a phylogenetic tree to generate different groups. Here, we use a 4-level full binary tree, thus generating 16 sequences, each of which will become the ancestor of a group. The branch lengths of the tree are identical and control the variation between groups.

Step 2: Generate individuals in each group

- Pick one sequence generated in last step as the ancestor of a group.
- Set a phylogenetic tree to generate individuals. As with the last step, we use a 4-level full binary tree to generate 16 individuals. The branch lengths of the tree are identical and control the variation within groups.
- Repeat the previous two lines until all sequences generated in Step 1 have been used.

256 sequences are generated in this step.

Step 3: Add LGT events

- Set the total number of LGT events. We use 20 in our experiments.
- Determine the distribution of LGT events. We take LGTs only to group 1 from other groups. The LGT donors are distributed evenly (5 each) among the following four sets: group 2, groups 3 and 4, groups 5 to 8, and groups 9 to 16.

Step 4: Evolve post-LGT

- Set a 2-level full binary tree. The branch lengths are identical and control the variation post-LGT.
- Let every sequence evolve following this tree, and add deletion simultaneously to get 4 descendants.
- Randomly pick one descendant from each sequence.

This generates a final simulation dataset with 256 sequences.

Parameters

Variation between = 0.01, 0.05, 0.1, 0.15, 0.2, within = 0.01, post_LGT = 0, deletion = 0, *k*=40

Variation between = 0.1, within = 0.001, 0.005, 0.01, 0.015, post_LGT = 0, deletion = 0, *k*=40

Variation between = 0.1, within = 0.01, post_LGT = 0, 0.02, 0.04, 0.06, 0.08, 0.1, deletion = 0, 0.025, 0.05, 0.075, 0.1, 0.125, k = 40

$\pi_T = \pi_c = \pi_A = \pi_G = 0.25, \kappa = 2$ under HYK85 model.

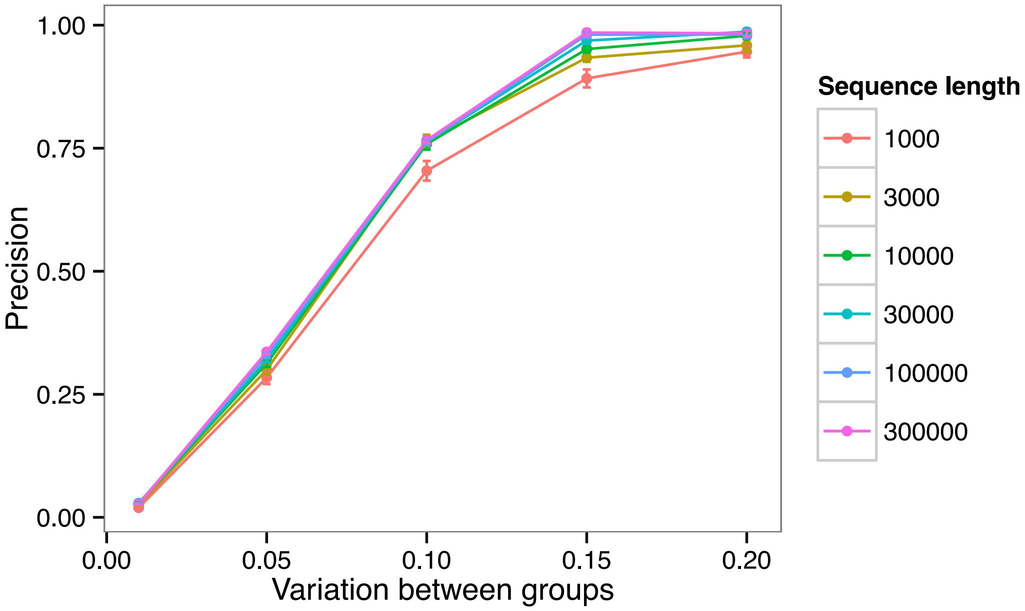
$\pi_T = 0.291, \pi_c = 0.275, \pi_A = 0.304, \pi_G = 0.130, \kappa = 2$ under F84 model.

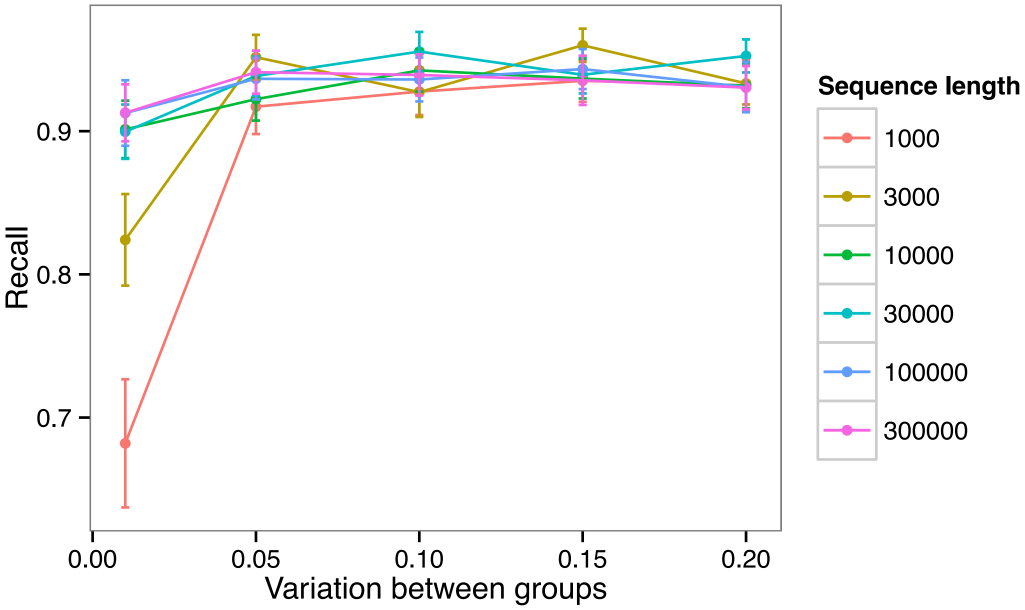
3 Coordinates of simulated and inferred LGT regions in Group 1 for Figure 2.2

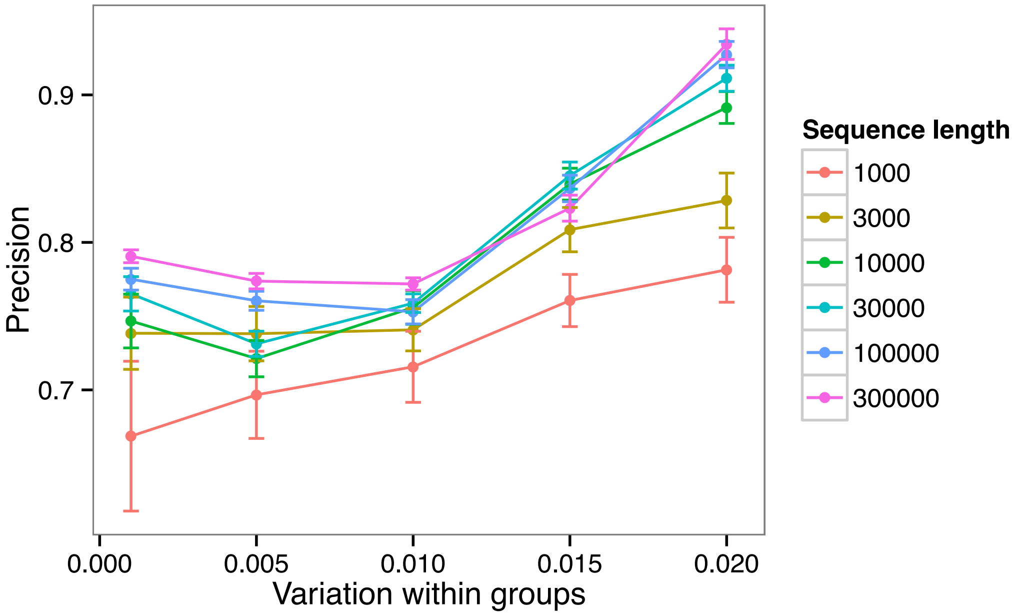
Simulated LGT		**Inferred LGT**		**Recipient (Sequence)**	**Donor (Group)**
Start	**Length**	**Start**	**Length**		
297	95	297	97	1	16
182	56	182	56	4	16
786	177	786	177	7	15
614	170	614	172	4	15
532	142	532	143	15	11
552	131	552	131	7	6
157	50	157	50	7	8
739	50	739	50	1	5
722	51	722	53	13	6
92	50	92	53	5	7
445	95	444	96	6	3
112	50	111	52	3	4
163	115	161	118	14	4
62	167	62	169	15	3
585	206	585	206	2	3
662	134	662	215	11	2
562	66	562	66	3	2
525	127	525	127	1	2
39	96	38	98	1	2
58	117	--	--	2	2

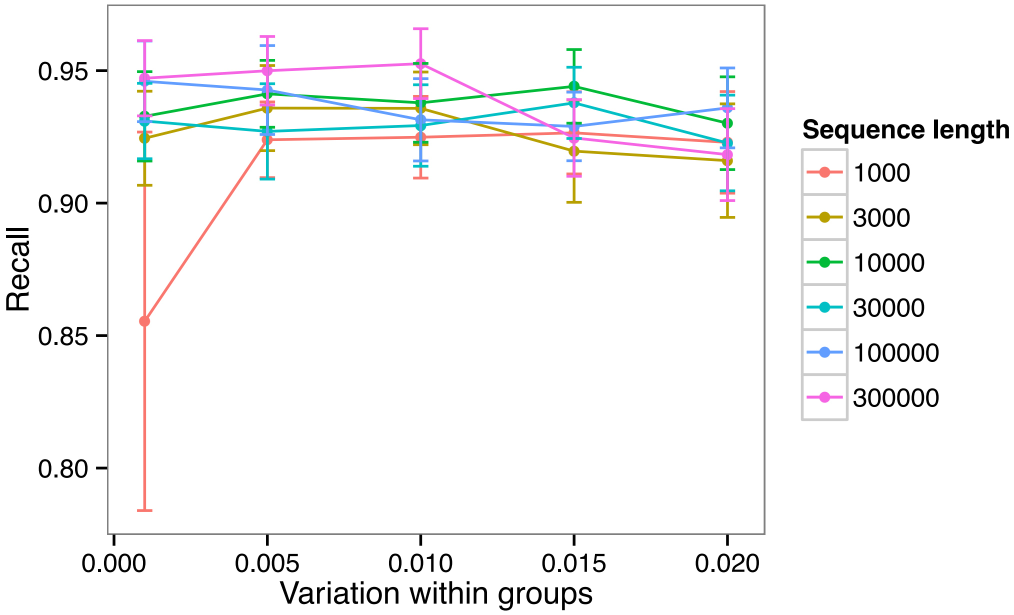
4 Performance of TF-IDF under the F84 model (variation between and within groups)

Here we replicate the TF-IDF analyses shown in Figures 2.3 and 2.4, under the F84 evolution model.

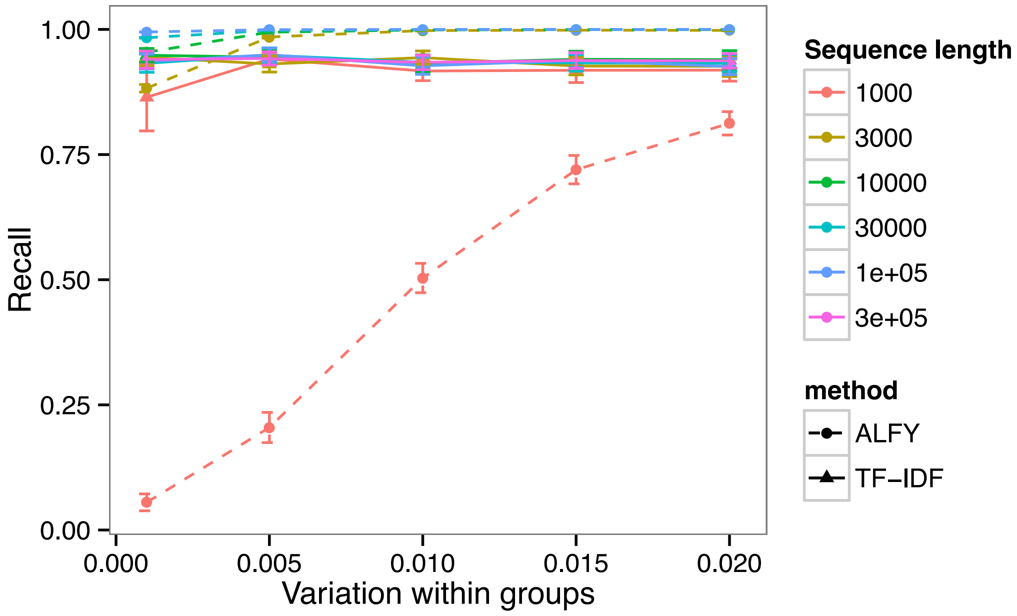





Precision
0.9
0.8
0.7
0.000
0.005
0.010
0.015
0.020
Variation within groups
Sequence length
1000
3000
10000
30000
100000
300000

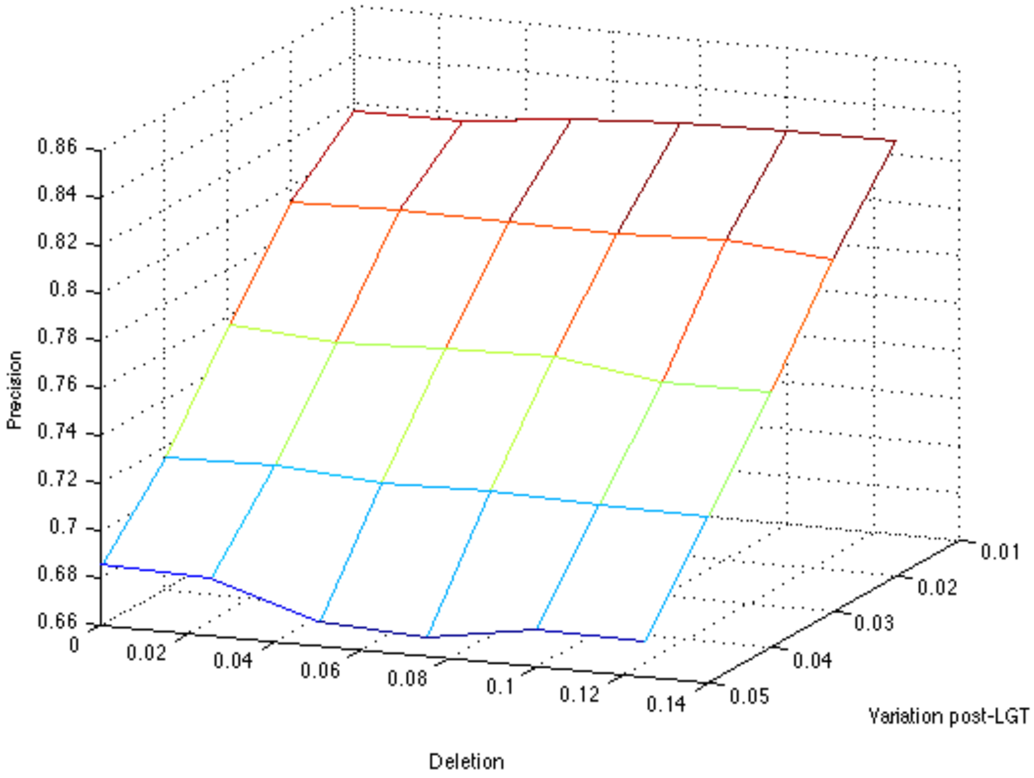

Recall
0.95
0.90
0.85
0.80
0.000
0.005
0.010
0.015
0.020
Variation within groups
Sequence length
1000
3000
10000
30000
100000
300000

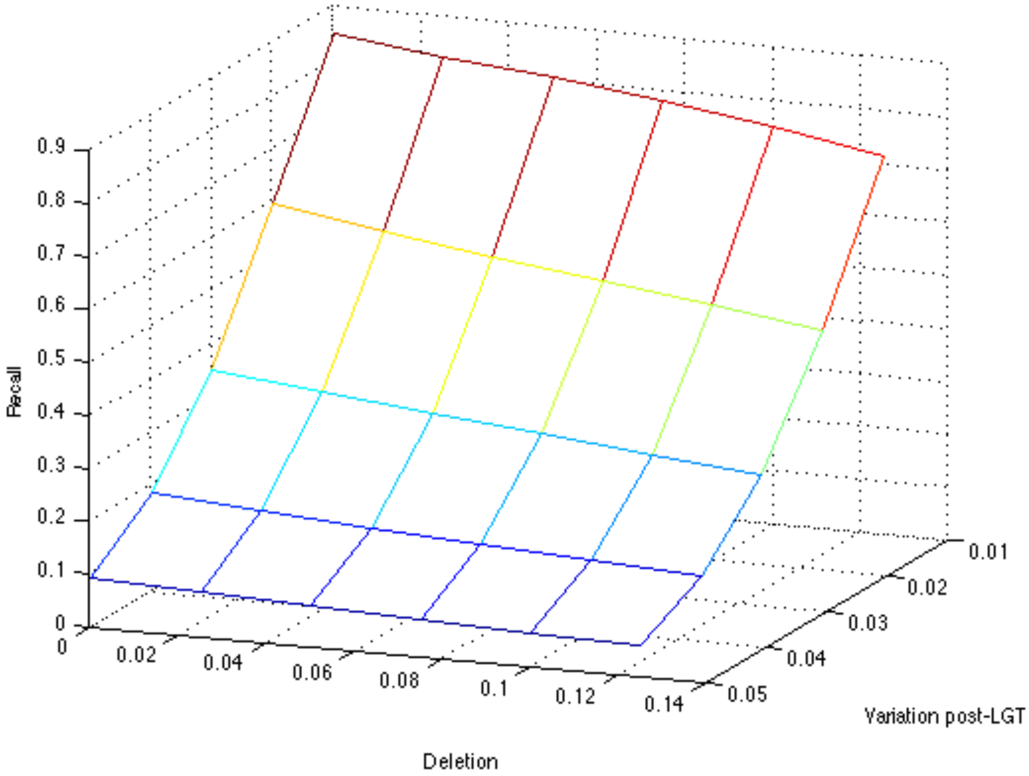
5 Full comparison of recall of TF-IDF and ALFY for variation within groups



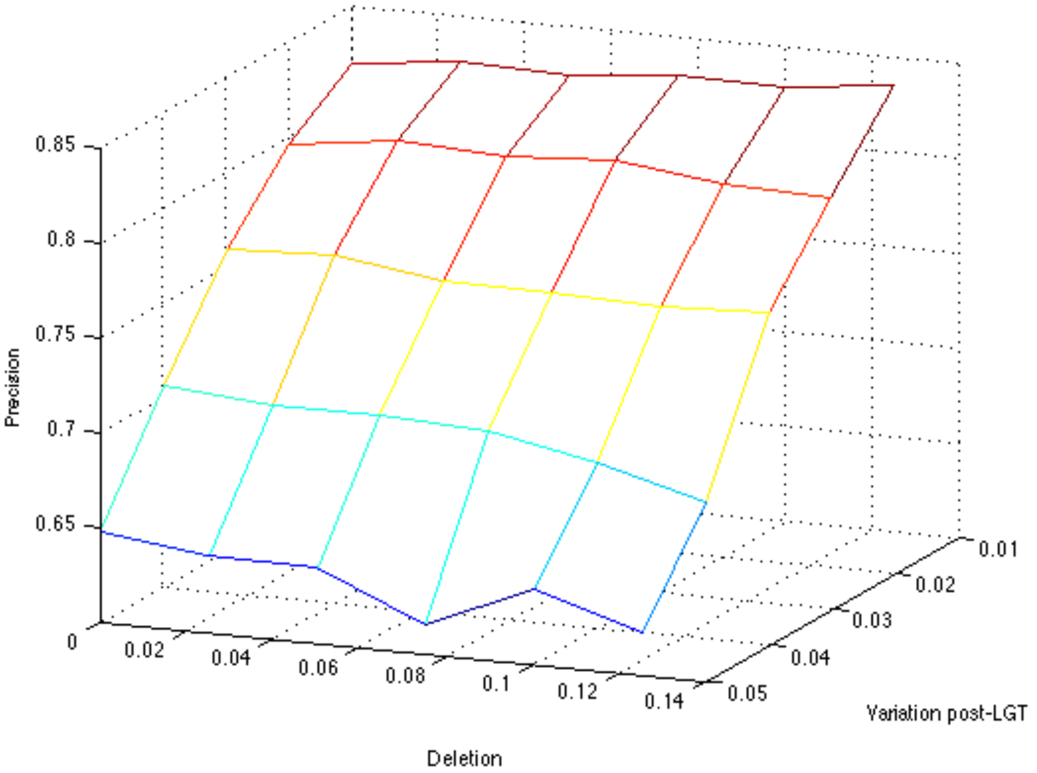
6 Performance of TF-IDF with variation post-LGT and deletion with different sequence lengths.

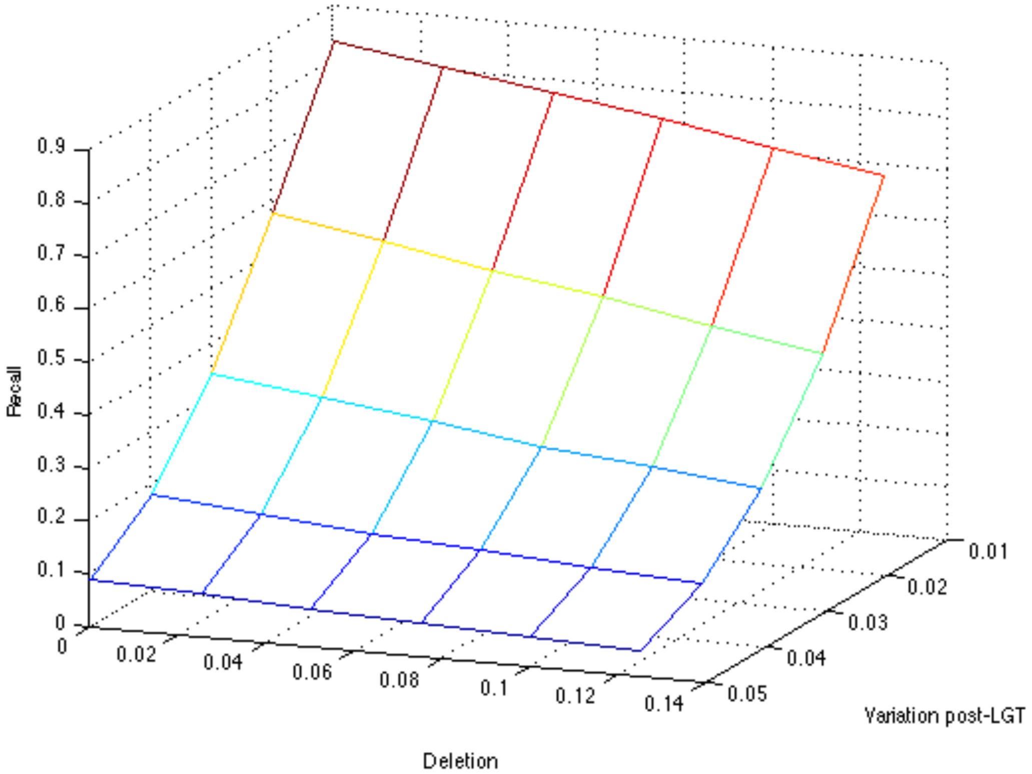
Performance on variation post-LGT and deletion (sequence length = 100,000 nt)



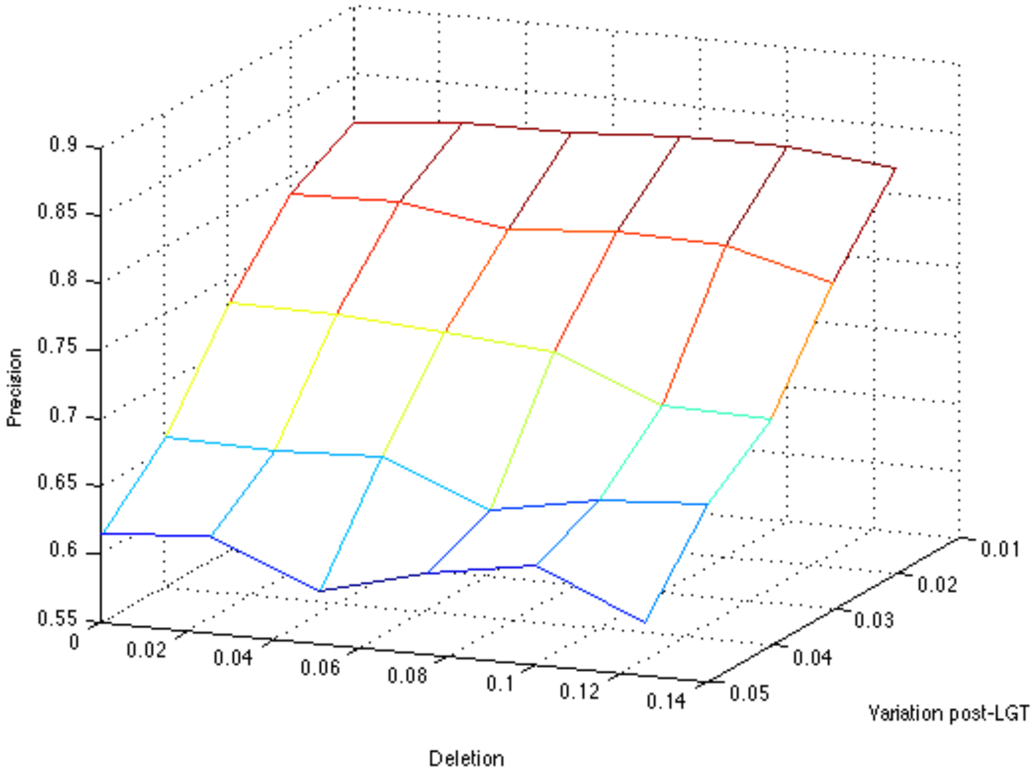


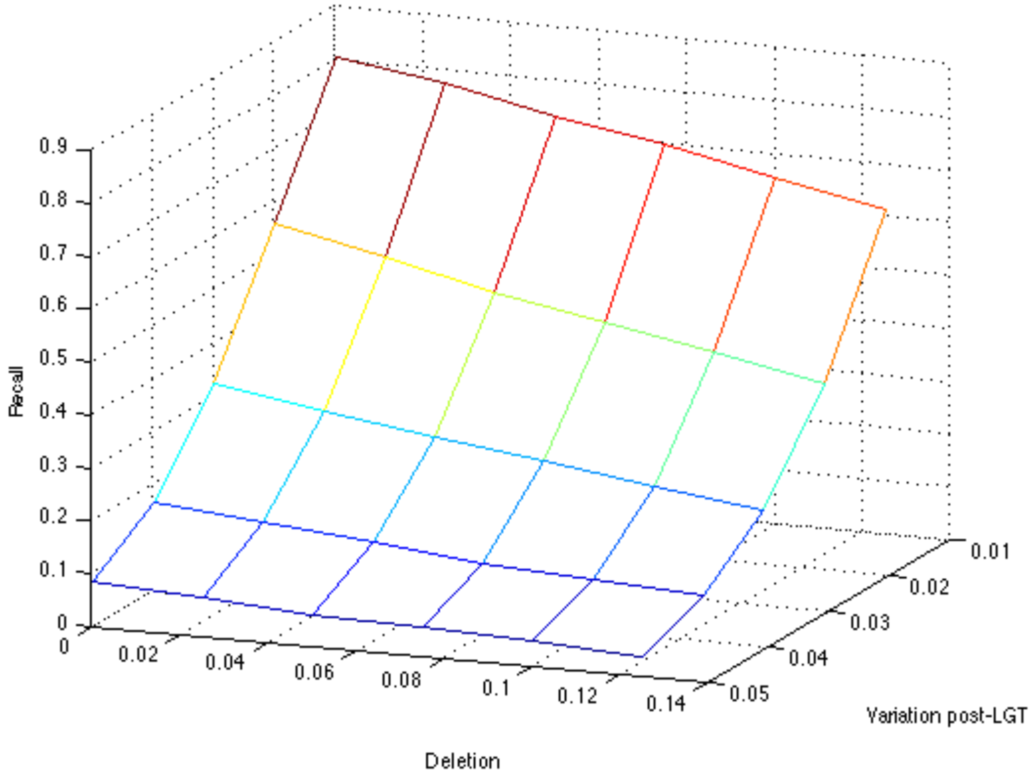
Performance on variation post-LGT and deletion (sequence length = 30,000 nt)



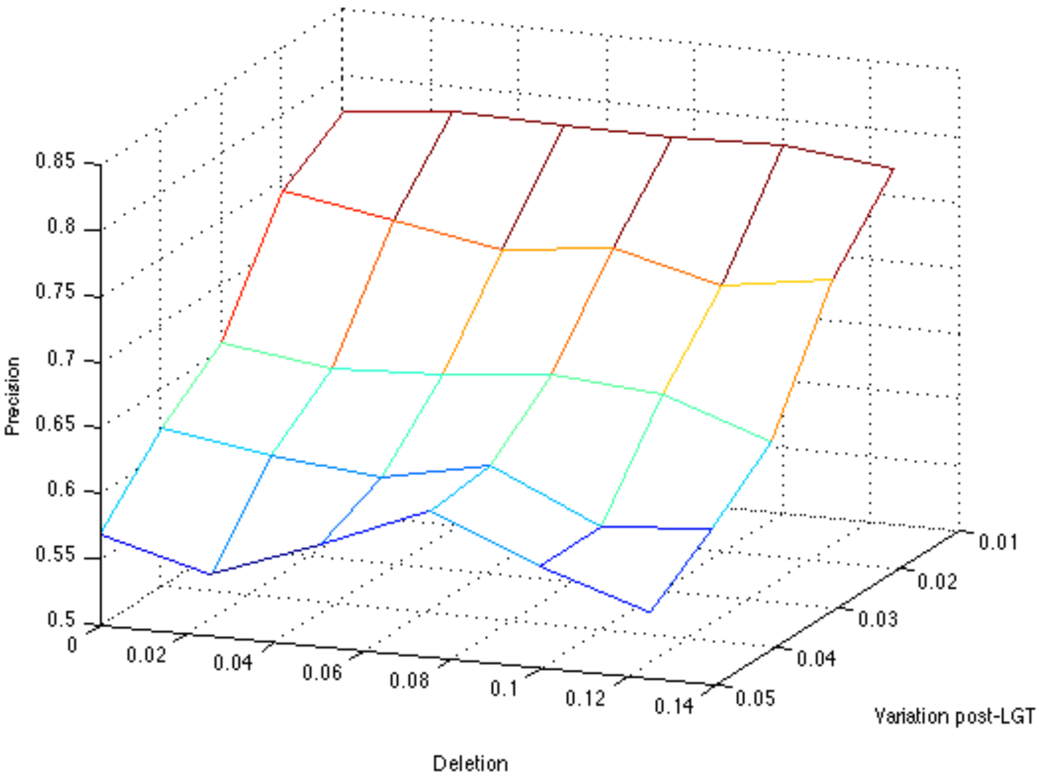


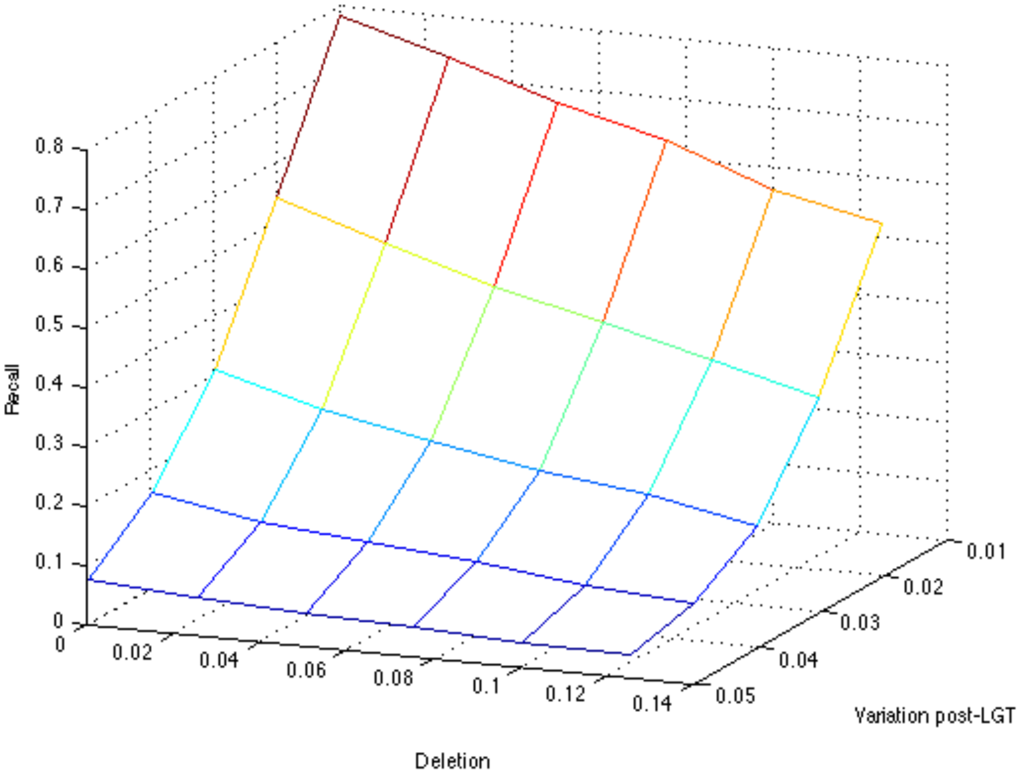
Performance on variation post-LGT and deletion (sequence length = 10,000 nt)



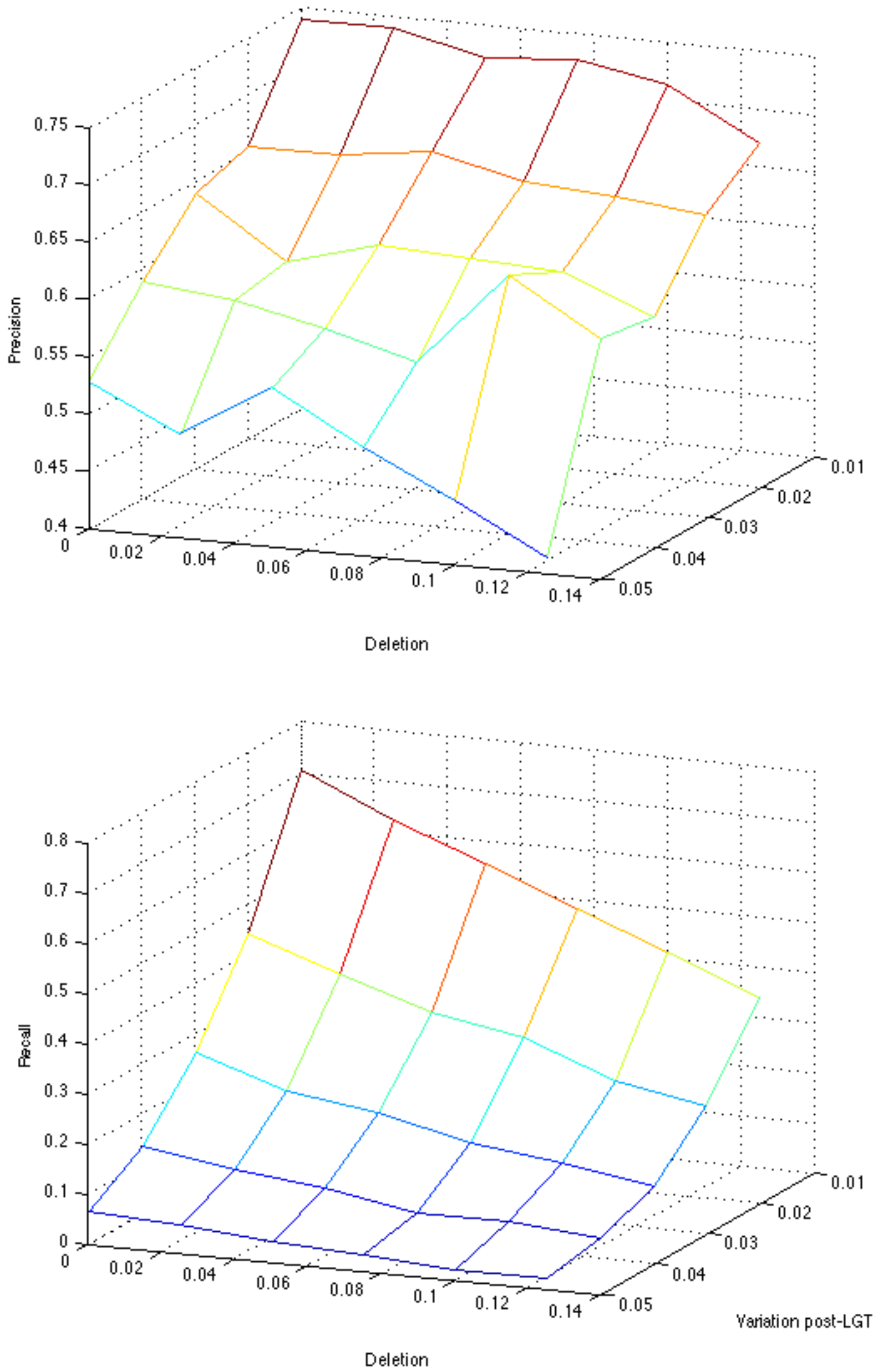


Performance on variation post-LGT and deletion (sequence length = 3,000 nt)





Performance on variation post-LGT and deletion (sequence length = 1,000 nt)



7 Supplementary Table S1. Detection of lateral regions in *Staphylococcus aureus* TW20 by TF-IDF, at *k* = 30 and *k* = 40

Annotated functions of proteins fully or partially contained within an LGT region of *Staphylococcus aureus* TW20, as discovered in this dataset by TF-IDF (*k* = 40). The first row is the length range of LGT segments selected for analysis.

		2000-3999 nt	4000-5999nt	6000+ nt	2000+ nt	2000+ nt
Annotated function[1]	Annotated in genome	Number	Number	Number	Number	%
adhesion / adhesion	4	0	0	2	2	50
antiporter	10	1	0	1	2	20
capsular polysaccharide	16	0	0	16	16	100
capsule	3	0	0	3	3	100
coagulase	1	0	0	1	1	100
efflux	4	1	0	1	2	50
integrase	9	0	1	0	1	11
lactamase	8	0	0	5	5	62
lysine	13	0	0	1	1	8
metalloproteinase /metallopeptidase	3	1	0	1	2	66
penicillin	6	1	0	3	4	66
permease	45	8	5	18	31	69
phage	249	12	14	0	26	11
recombinase	6	0	0	0	0	0
resistance protein	9	0	1	1	2	22
restriction	9	0	0	3	3	33
siderophore	5	0	0	5	5	100
surface protein	5	1	1	4	6	100

toxin	19	1	0	0	1	5
transport protein	28	3	1	7	11	39
transporter	98	14	3	28	45	46
transposase	30	0	2	2	4	13
uptake	4	1	1	0	2	50
ribosomal protein	60	2	1	5	8	13
polymerase (DNA/RNA)	15	5	2	1	8	53

Annotated functions of proteins fully or partially contained within an LGT region of *Staphylococcus aureus* TW20, as discovered in this dataset by TF-IDF (*k* = 30). The first row is the length range of LGT segments selected for analysis.

		2000-3999 nt	4000-6499 nt	6500+ nt	2000+ nt	2000+ nt
Annotated function[1]	Annotated in genome	Number	Number	Number	Number	%
adhesion / adhesion	4	0	0	4	4	100
antiporter	10	7	0	1	8	80
capsular polysaccharide	16	0	0	16	16	100
capsule	3	0	0	3	3	100
coagulase	1	0	0	1	1	100
efflux	4	1	1	1	3	75
integrase	9	2	1	0	3	33
lactamase	8	1	0	4	5	62
lysine	13	2	0	3	5	38
metalloproteinase /metallopeptidase	3	2	0	1	3	100
penicillin	6	3	2	2	(7)[2]	100
permease	45	11	7	15	33	73
phage	249	13	17	0	30	12
recombinase	6	2	0	0	2	33
resistance protein	9	1	0	1	2	22
restriction	9	0	0	3	3	33
siderophore	5	0	0	5	5	100
surface protein	5	0	1	2	3	60
toxin	19	1	0	0	1	5
transport protein	28	8	1	11	20	71

transporter	98	26	11	22	59	60
transposase	30	4	0	2	6	20
uptake	4	1	1	1	3	75
ribosomal protein	60	13	9	32	54	90
polymerase (DNA/RNA)	15	5	2	5	12	80

Notes:

1. As annotated in GenBank NC_017331.1

2. The 5′ and 3′ ends of the penicillin binding protein 2B gene fall into different inferred LGT regions.

Chapter 3 Exploring lateral genetic transfer among microbial genomes using TF-IDF

Lateral genetic transfer is an important source of phenotypic innovation in microbes. Studying this phenomenon is thus a key way to gain insight into microbial evolution. The traditional way to depict the evolutionary history of organisms is on a phylogenetic tree, the leaves of which represent contemporary organisms, the internal nodes show a hierarchy of common ancestors, and the branches depict the vertical (parent to offspring) transmission of genetic material. LGT might be added as lateral branches, although often there is not enough knowledge of donor and recipient lineages, or the ancestry of transfer events. More recently, patterns of genetic relatedness among microbes have been depicted as a network.

Here, I applied the TF-IDF method to three empirical multi-genome datasets of different size (numbers of genomes), evolutionary relatedness and group structure to detect LGT, and I generated networks in which the edges represent LGT events. These networks are not uniform, and their connectivity is affected by evolutionary distance between groups. The genes affected by these LGT events are described by a range of Gene Ontology terms, particularly those of metabolic, regulatory, and intracellular and trans-membrane transport processes. I carried out the first systematic analysis of genes inferred to have received LGT from more than one donor group, and showed that some of them might equally well be explained by a single ancient transfer. Initial analysis indicated the presence of potential community structures within the LGT networks; this question forms the basis of the following chapter.

Results presented as a publication

The work reported in this chapter has been published as a research article in the Nature Publishing Group journal *Scientific Reports*, volume 6, article 29319 (publication date 25 July 2016). Supplementary material for this publication follows the main text. Supporting datasets are assigned DOI numbers http://dx.doi.org/10.14264/uql.2016.483-487 and are available for download from UQ eSpace.

Exploring lateral genetic transfer among microbial genomes using TF-IDF

Yingnan Cong[1], Yao-ban Chan[2] and Mark A. Ragan[1*]

[1] Institute for Molecular Bioscience and ARC Centre of Excellence in Bioinformatics, The University of Queensland, St Lucia, Brisbane, QLD 4072, Australia

[2] School of Mathematics and Statistics, The University of Melbourne, Parkville, Melbourne, VIC 3010, Australia

Email addresses:

YC: y.cong@uq.edu.au

YBC: yaoban@unimelb.edu.au

MAR[*]: m.ragan@uq.edu.au

Abstract

Many microbes can acquire genetic material from their environment and incorporate it into their genome, a process known as lateral genetic transfer (LGT). Computational approaches have been developed to detect genomic regions of lateral origin, but typically lack sensitivity, ability to distinguish donor from recipient, and scalability to very large datasets. To address these issues we have introduced an alignment-free method based on ideas from document analysis, term frequency-inverse document frequency (TF-IDF).

Here we examine the performance of TF-IDF on three empirical datasets: 27 genomes of *Escherichia coli* and *Shigella*, 110 genomes of enteric bacteria, and 143 genomes across 12 bacterial and three archaeal phyla. We investigate the effect of *k*-mer size, gap size and delineation of groups on the inference of genomic regions of lateral origin, finding an interplay among these parameters and sequence divergence. Because TF-IDF identifies donor groups and delineates regions of lateral origin within recipient genomes, aggregating these regions by gene enables us to explore, for the first time, the mosaic nature of lateral genes including the multiplicity of biological sources, ancestry of transfer and over-writing by subsequent transfers. We carry out Gene Ontology enrichment tests to investigate which biological processes are potentially affected by LGT.

3.1 Introduction

Many microbes can acquire DNA from an exogenous source (other microbes, or the environment) and maintain it for transmission to subsequent generations, either incorporated into the new host genome or stabilised on a plasmid or other extra-chromosomal element. This process, lateral genetic transfer (LGT; also known as horizontal genetic transfer), generates size and gene-content diversity among microbial genomes, and is a major driver of metabolic innovation[1-3] including resistance to antibiotics[4,5].

Computational approaches have been applied to detect regions of lateral origin in microbial genomes since the 1990s[6,7]. In the accompanying article[8] and elsewhere[9] we review the main biological and computational factors that make LGT detection so challenging. Briefly, there is great diversity (and little predictability) with regard to the length, source or features of the introgressed DNA. LGT events can overwrite an existing sequence, including other lateral regions, rendering the new host genome an evolutionary pastiche or mosaic. Over time, features (*e.g.* G+C content or codon usage) indicative of lateral origin will be "ameliorated" to become indistinguishable from those of the new host genome[9-11]. For these reasons, accurately identifying regions of lateral origin can be very challenging.

Next-generation sequencing technologies are increasingly making it possible for researchers to address large-scale questions in the biological sciences, including open questions regarding the mechanisms and impact of LGT[2,9,12]. Several computational approaches are available to detect regions of probable exogenous origin in a genome, among which tree-based methods are considered to be the gold standard[13]. Taking genes (gene families) as the units of analysis, these approaches delineate orthogroups, multiply align sets of sequences, infer gene trees and compare their topologies against that of a reference "species" tree; well-supported instances of topological incongruence are taken as *prima facie* instances of LGT[14-16]. Such workflows are computationally demanding, yet cannot identify recombination breakpoints in individual genomes, and often fail to resolve the direction of transfer. They can be accelerated by use of approximate methods, better matching of computational tasks to hardware, and parallelisation, but nonetheless remain slow with large datasets[17].

For these reasons there is much interest in approaches that avoid altogether the potentially NP-hard steps of multiple sequence alignment, tree inference and tree reconciliation, while keeping track of regions of each individual genome in a manner that is agnostic to the number, size and nature of units of transfer. Alignment-free approaches have much to offer in this context. Among the main families of alignment-free approaches, those based on word counts or on substring match lengths have received the most attention[18,19]. The former compute a measure of similarity between two sequences based on the number or frequency distribution of matching "words" of length *k*, whereas the latter assess the length of the longest word that occurs in two sequences, or the shortest word unique to one of the sequences. In either case the match may be required to be perfect, or a defined number of mismatches may be permitted. In the simplest case, each pairwise measure can be transformed into a distance, and a matrix of such distances used as input for computing a distance tree, *e.g.* by neighbour-joining[20-22]. Evidence is accumulating that in phylogenetic inference *per se*, these alignment-free methods can offer acceptable performance – in certain cases better than approaches based on multiple sequence alignment – at much greater computational speed and scalability[19]. Other approaches to alignment-free sequence comparison, including methods based on compressibility[20,21], nucleotide correlations[23,24], gene order or recombination breakpoints[25,26], have seen more-limited application. There has, however, been little exploration of how any of these alignment-free methods might be extended to other steps in an LGT workflow.

In the accompanying article[8] we introduce TF-IDF as a scalable alignment-free approach to identify directional LGT in large molecular-sequence datasets. Variants of TF-IDF are widely used in text mining and information retrieval, for example to find important words, group and classify documents by topic, or retrieve documents that match a user query[27,28]. Using synthetic nucleotide-sequence data, we showed that by using TF-IDF we can detect LGT events with high precision and recall under a range of biologically realistic scenarios including different rates of deletion and nucleotide substitution[8]. We reported that TF-IDF performs well with a small empirical dataset (seven genomes of *Staphylococcus aureus*[29]) even though our target group consisted of a single sequence, presumably reducing the influence of the IDF term. The regions identified by TF-IDF as lateral matched closely with those inferred by a well-regarded method, ALFY[30], while in addition we identified two regions not found using ALFY that include genes encoding transporters and regulators of multidrug resistance and pathogenicity[8].

Here we explore the strengths and limitations of TF-IDF as applied to the discovery of regions of lateral origin among different-sized sets of empirical microbial genome sequences. Specifically, we investigate the effects of key parameter-value settings (*k*, and gap size *G*), and strategies for delineating, including or excluding, and subdividing groups. We consider how to interpret multiple inferred transfers into the same genome sequence, and look for evidence for overwriting. Furthermore, we determine the biological process annotations over- or under-represented among the genes we infer to have been affected by LGT, and report new LGT networks. Three empirical datasets (and variants) have been selected to illustrate a diversity of potential use cases, and address the above issues.

3.2 Results

3.2.1 Parameter values for TF-IDF analysis

To investigate the performance of TF-IDF on empirical data, we compare the number of regions identified as of potential lateral origin (Figure 3.1a,c,e) and the total length of these regions (Figure 3.1b,d,f) as a function of *k* and *G* in our three datasets. We examine the results in more detail for each dataset separately, and then discuss how to select suitable parameters in different situations.

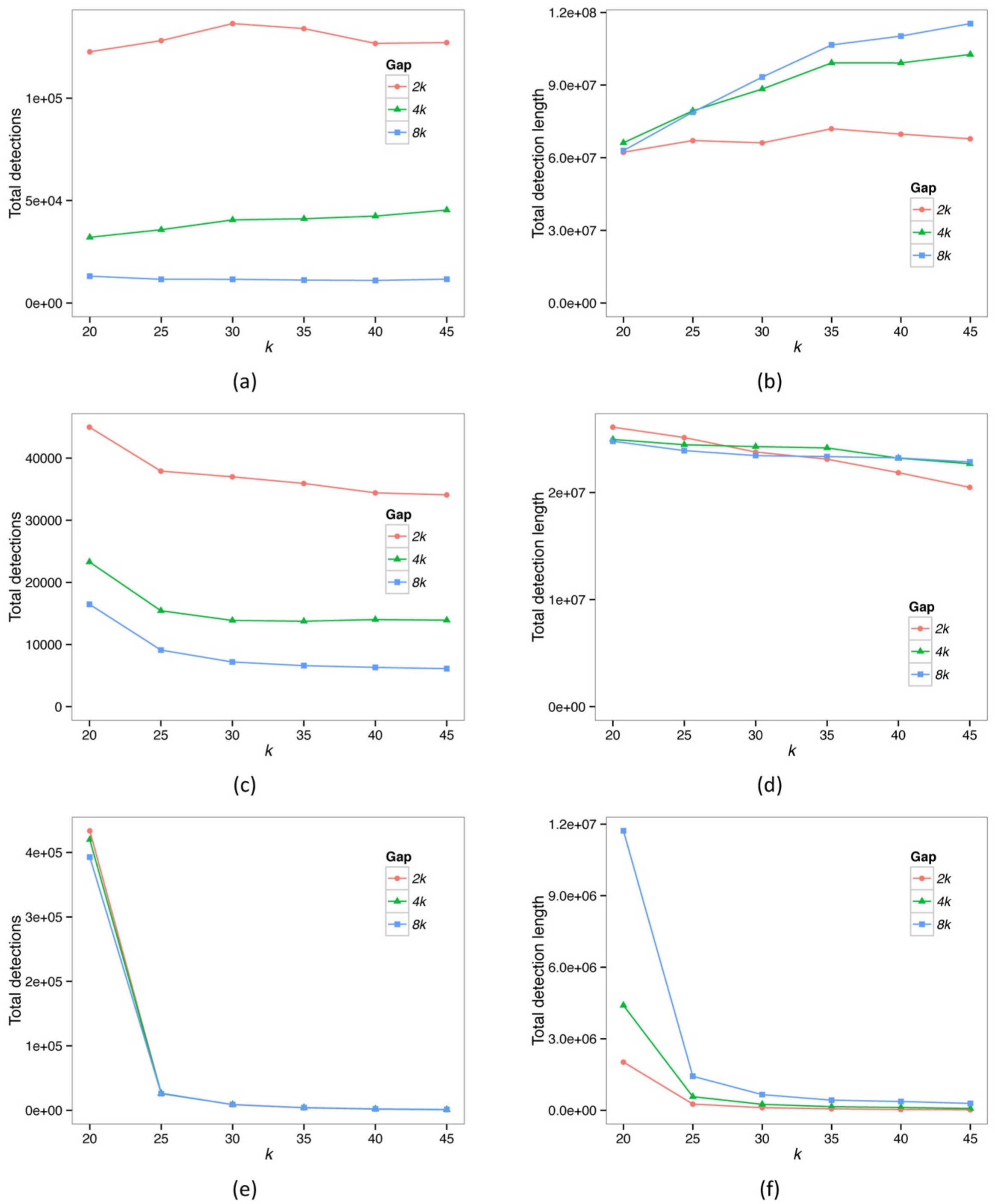


Figure 3.1. Number of regions detected as lateral, as a function of *k* and *G*. The panels on the left show total numbers of LGT detections in the ECS (panel a), EB (panel c) and BA (panel e) datasets. The panels on the right show the total length (in nucleotides) of all LGT detections in the same datasets.

Dataset 1 (E. coli and Shigella: ECS)

Here we use the six groups suggested by Skippington and Ragan[31]. TF-IDF presents the 27 ECS genomes as having sustained very extensive LGT from within the ECS clade itself. Inferred lateral segments hundreds of nucleotides in length are common (Supplementary Table S1), and the gaps between these segments tend to be small.

We show the dependence of the total number of lateral regions detected and the total length of all detections on *k* and *G* in Figures 3.1a and 3.1b respectively. As gap size *G* increases, the total number of detections decreases sharply, indicating that many potential LGT segments are being merged together. When *k* is large, we see a corresponding rise in the total length. However, when *k* is small, the total length is relatively stable with respect to *G*, in part because the gaps between segments contain *k*-mers that are also frequent in the recipient genome's own group, causing further proposed mergers to fail the TF hurdle. These *k*-mers can be false positives. With these considerations in mind, we set G to 2*k.* At this value, we see that both number and length of detections are relatively stable with respect to *k*. We choose the value *k* = 40, which we have shown to work well in simulations[8].

Dataset 2 (enteric bacteria: EB)

The EB clade is biologically more ancient than ECS, and accordingly their genomes show smaller similarity values (Table 3.1). Delineating groups within the EB dataset by genus, we find fewer and shorter LGT detections than in the ECS dataset (Supplementary Table S1). As before, we see a dramatic decrease in total number of detections as *G* increases (Figure 3.1c); however, the total detection length (Figure 3.1d) remains relatively stable with respect to *G* at all values of *k*. This again indicates that a large number of false positive segments are being merged with increasing *G*, and thus we again set *G = 2k.* Here there is a substantial decrease in the total number of detections as *k* increases from 20 to 25, suggesting that there are too many common *k*-mers at this value. We again choose a large value of *k = 40* to avoid this problem.

Dataset 3 (bacteria and archaea: BA)

The 143 BA genomes are much less closely related among themselves, with their common biological ancestor dating nearly to origin of cellular life[32]. These genomes share many fewer identical *k*-mers than do ECS or EB (Table 3.1), and *k* plays a much more important role than does *G.* Because regions of inferred lateral origin in this dataset present a much weaker signal than in the previous datasets, we should set *k* to a small value in order to detect these signals. We observe (Figures 3.1e and 3.1f) a precipitous drop in the both the number of detections and detection length from *k* = 20 to 25, again indicating the presence of too many common *k*-mers at the former value to make any useful detections. However, the detections are more stable for *k* ≥ 25, so we set *k* = 25. The value of *G* appears to make relatively little difference, so we again select *G = 2k* for consistency.

We note that TF-IDF is not biased toward detecting more LGT events in larger datasets. With suitable settings of *k* and *G* (as discussed above), fewer regions of within-dataset lateral origin, totaling fewer nucleotides, are detected in EB and BA than in ECS even though they contain many more sequences. For the subsequent analyses, we fix *k* and *G* at the optimal values we have found above.

Table 3.1. Summary of genome similarity (percentage of pairwise shared 12-mers) for the three datasets. For abbreviations see text.

Dataset	Minimum %	Maximum %	Mean %	Standard deviation %
ECS	37.15	94.5	64.06	10.58
EB	29.08	99.99	45.43	15.65
BA	0.4243	99.42	15.94	10.44

3.2.2 LGT networks and effect of grouping

Next we investigate the networks of inferred LGT among the genomes in each of our datasets. TF-IDF requires that we recognise or delineate groups of sequences in the dataset; an inferred LGT event represents transfer into a genome from a donor group (other than that containing the recipient genome). Using Dataset 1, we explore the effect of different ways of delineating groups. With Datasets 2 and 3 we ask whether adding further potential donor groups affects the inference. As our results will form the basis of functional analysis (see next section), here we aggregate inferred LGT events by gene. Although genes are not units of LGT[33,34], they are our link to functional annotation, notably in the GO database[35]. This mapping moreover allows us to explore, for the first time, multiple and overlapping transfers in a functional context. As intergenic regions account for only minor proportions of these genomes, we anticipate that results aggregated by gene will be substantially applicable to whole genomes as well.

Dataset 1 (E. coli and Shigella)

We apply two strategies for delineating groups in the ECS dataset. One uses six established phyletic groups[36], thereby reflecting the diverse biological and physiological features that underlie the recognition of taxa in ECS. The other is explicitly phylogenetic: we cut the MRP supertree of Skippington and Ragan[31] at basal branches to yield four groups. For details of group membership, please consult Supplementary Figure S1. As a control, we also generate 50 pseudo-replicate groupings based on the latter.

We begin with the biological (phyletic) delineation of groups. Of the 124717 genes annotated in the 20 *E. coli* and 7 *Shigella* genomes, we infer 45412 (36.4%) to have received LGT from at least one source group. Figure 3.2 shows the directed LGT network connecting the groups of *E. coli* (A, B1, B2, D and E) and *Shigella* (S). Of these 27 genomes, we infer 24 to have accepted an LGT event (Table 3.2). Group E (*E. coli* O157:H7 EDL933 and O157:H7) has been the most-active donor group, supplying genetic material to a total of 18059 genes across all the other groups. Group E has also been the most-active recipient of LGT on a per-gene basis, with 7177 of its 10490 genes (68.4%) showing evidence of LGT, all donated from Group B1. The genome *E. coli* O157:H7 in group E is known to have acquired substantial genetic material by LGT[37]; this was notably not found by using a classical LGT detection method[31]. Group B1 (*E. coli* E24377A, 55989, SE11 and IA11) has been second most-active both as donor and recipient, donating to 16237 genes across all the other groups, and accepting LGT into 12131 of 18751 genes (64.7%), with all other groups as donors (Table 3.2 and Figure 3.2).

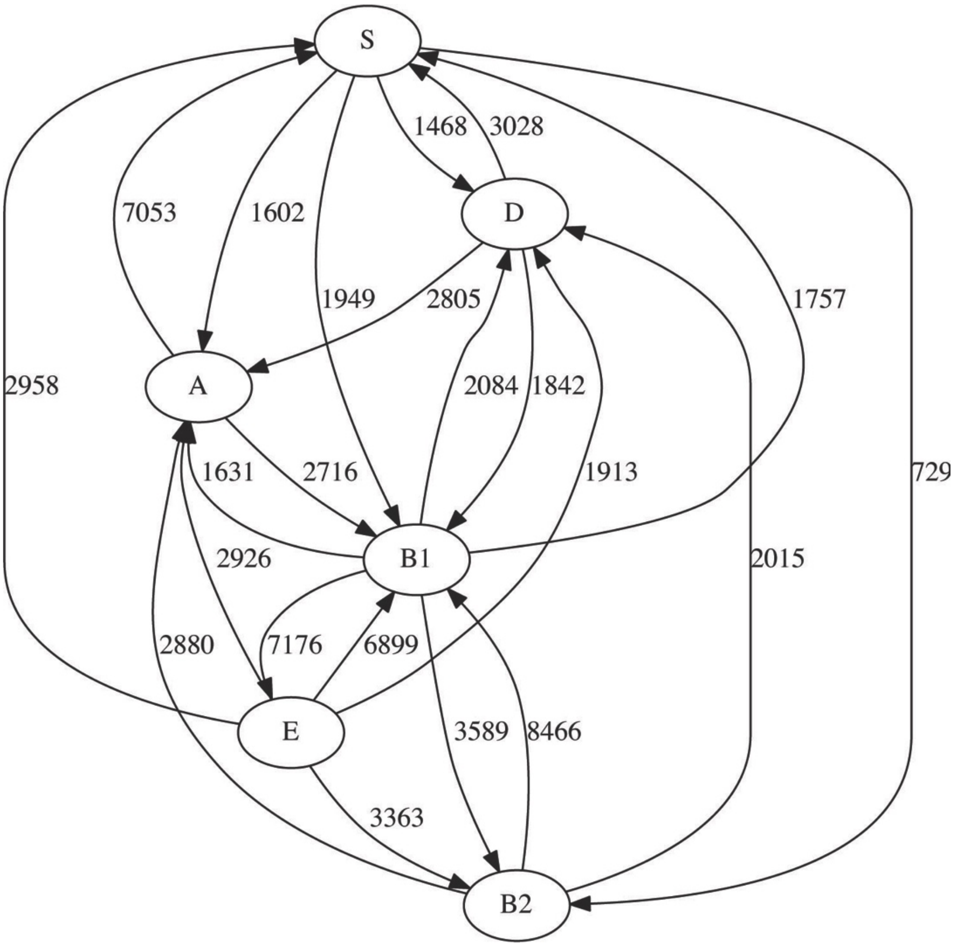


Figure 3.2. Inferred network of LGT within the ECS dataset with six biologically based groups. The numbers on each edge show the total number of genes involved in LGT events from one group into genomes in the other group.

Table 3.2. Numbers of lateral genes with single or multiple donors in each genome within the ECS dataset. The genomes are grouped into six groups following a phyletic criterion (see text).

Group	Organism	Number of genes	Number of lateral genes	Donor groups	Number of lateral genes with 1 donor	Number of lateral genes with 2 donors	Number of lateral genes with ≥ 3 donors
D	*E. coli* SMS 3 5	4744	0	-	-	-	-
D	*E. coli* IAI39	4725	1469	S	1469	-	-
D	*E. coli* UMN026	4878	2930	S, B1, B2	1004	769	1157
S	*Shigella flexneri* 5 8401	4336	1030	E, A, B1	855	120	55
S	*Shigella flexneri* 2a	4053	1142	D, E, A	650	288	204
S	*Shigella flexneri* 2a 2457T	4385	1188	E, A, B1	880	158	150
S	*Shigella sonnei* Ss046	4563	1842	D, E, A, B1	638	626	578
S	*Shigella boydii* Sb227	4391	1859	D, E, A, B1	974	538	347
S	*Shigella boydii* CDC 3083 94	4532	1280	A, B1	1093	187	-
S	*Shigella dysenteriae*	4063	1300	A	1300	-	-
E	*E. coli* O157 : H7	5204	3685	B1	3685	-	-
E	*E. coli* O157 : H7 EDL933	5286	3492	B1	3492	-	-
A	*E. coli* K12 substr W3110	4213	1211	S, E, B1	1050	117	44
A	*E. coli* K12 substr MG1655	4140	2045	D, E, B1, B2	927	465	653
A	*E. coli* HS	4366	2585	D, S, E, B1, B2	891	667	1027
A	*E. coli* C ATCC 8739	4434	0	-	-	-	-
B1	*E. coli* E24377A	4729	2945	S, E, A, B2	1469	1068	412
B1	*E. coli* 55989	4953	3108	S, E, A, B2	1526	1081	501
B1	*E. coli* SE11	4684	2918	S, E, A, B2	1448	1049	421

B1	*E. coli* IAI1	4385	3157	D, S, E,A, B2	1161	991	1005
B2	*E. coli* 0127 H6 E2348 69	4809	1561	E, B1	1120	448	-
B2	*E. coli* 536	4542	1183	E, B1	855	328	-
B2	*E. coli* CFT073	4897	0	-	-	-	-
B2	*E. coli* ED1a	5012	1230	S, E	941	305	-
B2	*E. coli* UTI89	4827	563	B1	563	-	-
B2	*E. coli* S88	4688	1088	E, B1	801	287	-
B2	*E. coli* APECO1	4878	592	B1	592	-	-

To evaluate how grouping affects the inference of LGT using TF-IDF, we delineated a different number of groups (four) using a phylogenetic criterion (see above). As physiology is not entirely orthogonal to phylogeny, the two groupings are not unrelated. Phylogenetic Group 1 includes B2 and two members of D; Group 2 encompasses E, one member of D and one *Shigella*; Group 3 has the same membership as A; and B1 and the remainder of S are merged into Group 4 (for complete lists see Supplementary Table S2). We now infer 18200 genes to have received LGT from at least one source group (Table 3.3), only 40.1% of the number detected in the previous grouping. The directed network is shown in Figure 3.3.

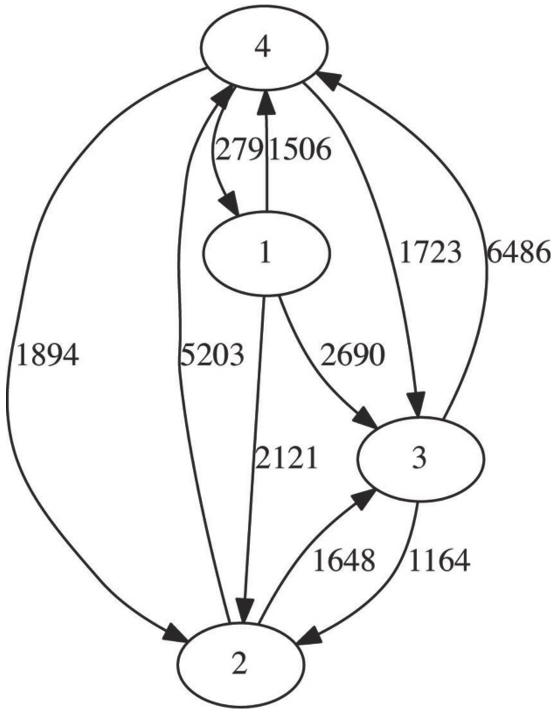


Figure 3.3. Inferred network of LGT within the ECS dataset with four phylogenetically based groups. The numbers on each edge show the total number of genes involved in LGT events from one group into genomes in the other group.

Table 3.3. Numbers of lateral genes with single or multiple donors in each genome within the ECS dataset. The genomes are grouped into four groups by cutting the MRP supertree (see text). LGT events were detected in only 17 of these 27 genomes.

Group	Organism	Number of genes	Number of lateral genes	Donor groups	Number of lateral genes with 1 donor	Number of lateral genes with 2 donors	Number of lateral genes with ≥ 3 donors
2	*E. coli* UMN026	4878	2122	1	2122		
4	*Shigella flexneri* 5 8401	4336	923	2, 3	680	243	
4	*Shigella flexneri* 2a	4053	590	2, 3	445	145	
4	*Shigella flexneri* 2a 2457T	4385	999	2, 3	674	325	
4	*Shigella sonnei* Ss046	4563	988	2, 3	688	300	
4	*Shigella boydii* Sb227	4391	1314	2, 3	848	466	
4	*Shigella boydii* CDC 3083 94	4532	1244	2, 3	879	365	
2	*Shigella dysenteriae*	4063	1363	3, 4	498	865	
2	*E. coli* O157 : H7 EDL933	5286	832	4	832		
3	*E. coli* K12 substr W3110	4213	875	2, 4	596	279	
3	*E. coli* K12 substr MG1655	4140	1255	1, 4	860	395	
3	*E. coli* HS	4366	1901	1, 2, 4	948	587	366
4	*E. coli* E24377A	4729	694	2, 3	495	199	
4	*E. coli* 55989	4953	1193	1, 2, 3	670	361	162
4	*E. coli* SE11	4684	683	2, 3	457	226	
4	*E. coli* IAI1	4385	944	1, 2, 3	507	300	137
1	*E. coli* ED1a	5012	280	4	280		

The fortuitous stability of Group A or 3 (see above) allows us to make a meaningful comparison of the results between the two groupings. We infer the Group A genomes to have accepted LGT into 5841 (34.1%) of their genes, and the very same Group 3 to have accepted LGT into 4031 genes (23.5%). This difference should reflect compositional changes within the donor groups, which affects the IDF step. Conversely, group A is inferred to have affected 39571 other genes by transfer, but only 14169 genes are affected by group 3 in the second grouping. This reflects the increased *k*-mer diversity of each recipient group, which decreases the TF threshold (that the frequency must fall below) and thus results in fewer detections. The four Group A/3 genomes show grossly similar trends between the two groupings, although details differ *e.g.* in number of genes with ≥ 3 LGT donor groups.

Finally, using results from TF-IDF analysis of the ECS dataset with the biological grouping, we compared the mean G+C content of the lateral genes (not just their inferred lateral regions) with that of their host genome, using a paired *t*-test. We find that G+C content is significantly higher in the inferred lateral genes; the P-value is 0.0017. Anomalous G+C content has often been used to detect transferred genes[10,38].

To explore the level of transfer signal in the dataset, we generate 50 randomised groupings based on the four phylogenetic groups as described above. In Figure 3.4, we compare the total detection length for the randomised groups against the real grouping. The total detection length in the real grouping is much greater than for any randomised grouping; indeed it is 6.2 standard deviations above the mean. Thus we are very confident that there exists strong lateral signal in this dataset, and that the grouping we have selected is effective in showing it. In Figure 3.5 we show the LGT network for the randomised groups. For seven of the twelve directional edges, we detect more genes in the actual grouping than in most (44-50) of the replicates. No LGT is found from group 2 or 3 to group 1 using the actual grouping information, while from group 4 to group 1 we inferred fewer genes than in 48 of the randomised groupings.

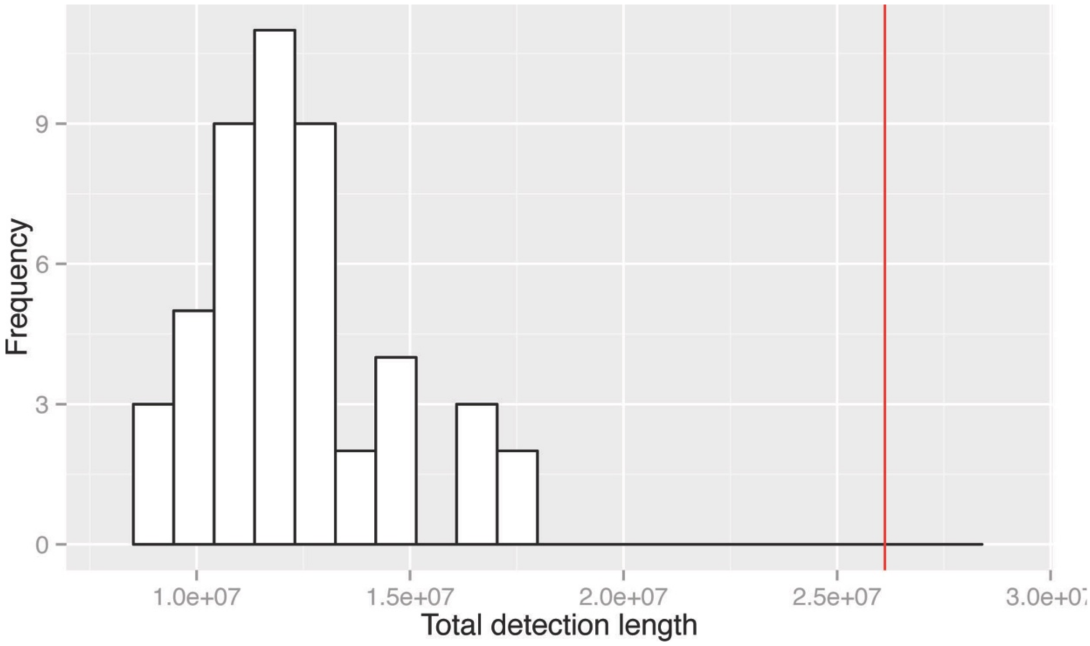


Figure 3.4. Histogram of the distribution of total length of all detections of the 50 randomly assigned replicates. The total detection length in the actual grouping (based on cutting the MRP supertree) is shown as a red line.

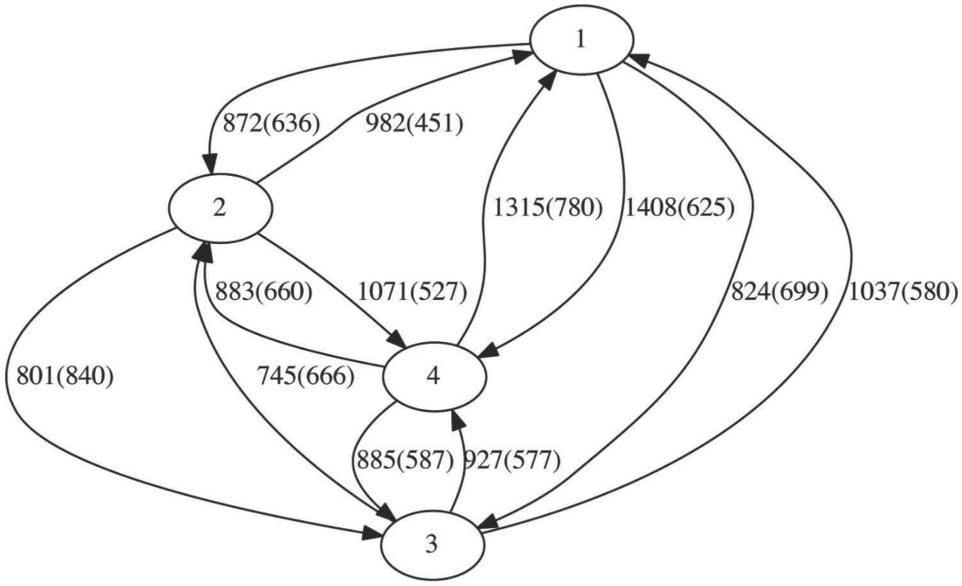


Figure 3.5. Summary of the LGT networks inferred using TF-IDF of the 50 random replicates of the ECS data. The numbers on each edge show the total number of genes involved in LGT events from one group into genomes in the other group, averaged over the 50 groupings. Number in parentheses are the standard deviations.

Dataset 2 (enteric bacteria)

Dataset 2a is a superset of ECS, containing additional genomes of *E. coli* and *Shigella*, plus genomes from *Klebsiella*, *Salmonella* and *Yersinia*. Naïvely taking the five genera as groups, we infer LGT only between *E. coli* and *Shigella* (Figure 3.6a). This happens because the lateral signal is dominated by the 62 *E. coli* and *Shigella* genomes, which are far more similar to each other (and thus share many more identical *k*-mers) than with the remaining genera. Since by default we set the IDF threshold to the average frequency of shared *k*-mers between a sequence and a group, only the *E. coli* and *Shigella* transfers are strong enough to overcome this threshold. There are potentially several non-exclusive ways to circumvent this situation, *e.g.* by manually overriding the default use of the mean value, or reducing the number of ECS genomes or groups. Here we reduce the number of groups by alternatively merging *E. coli* and *Shigella* into a single group, keeping only one or the other, or deleting both (Table 3.4). Figure 3.6 shows the LGT networks inferred in each case. Although the actual numbers of inferred transfers (even outside *E. coli* and *Shigella*) depend strongly on how we deal with *E. coli* and *Shigella*, common trends are nonetheless apparent, *e.g.* that *Salmonella* genomes are always inferred to have accepted more LGT from *Klebsiella* than *vice-versa*, and that *Yersinia* is only weakly connected.

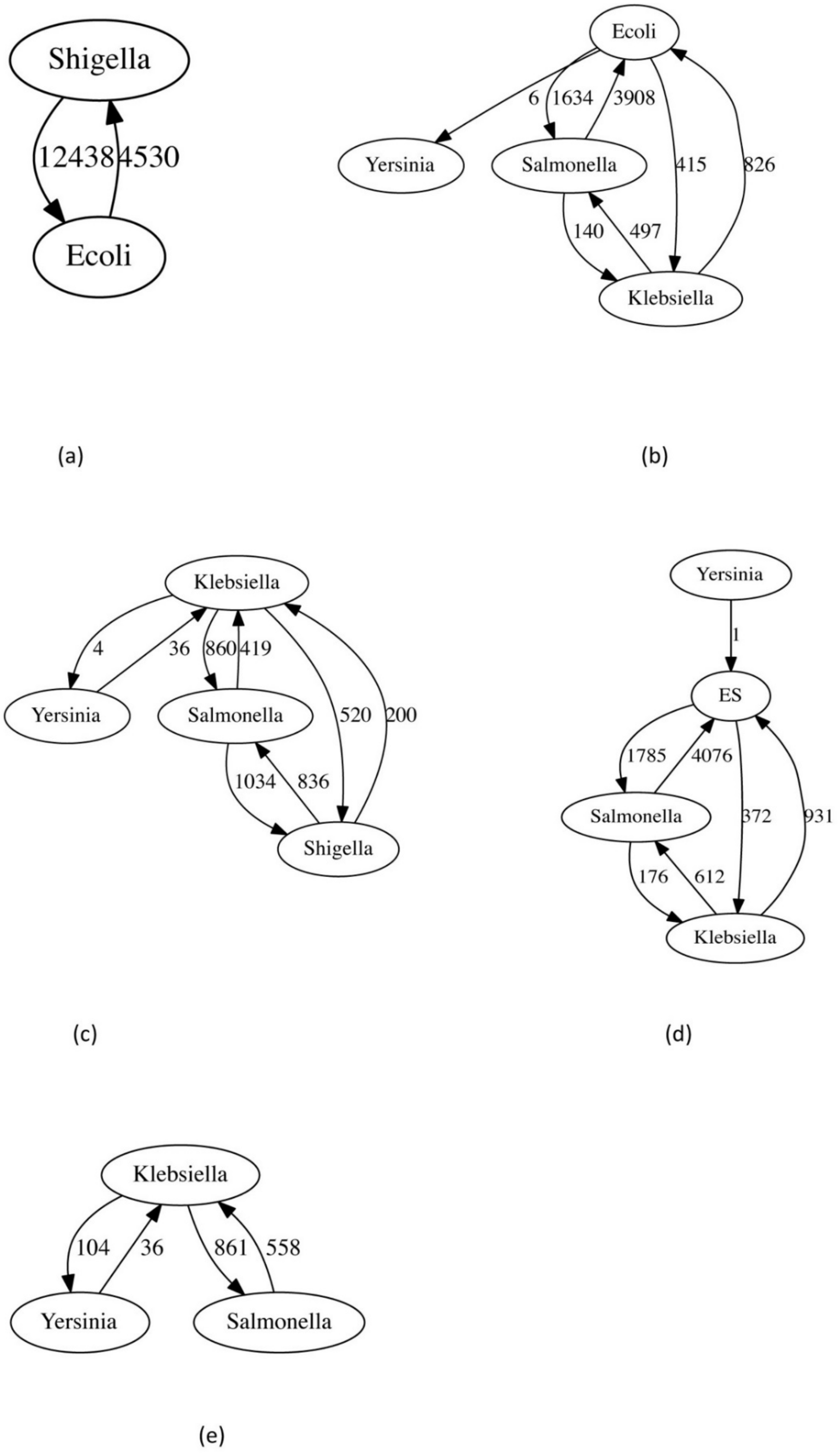


Figure 3.6. LGT networks of the EB dataset and its variants. We treat the *E. coli* and *Shigella* genomes in different ways: (a) assigned to separate groups, (b) with *Shigella* removed, (c) with *E. coli* removed, (d) combined into a single group, and (e) with both groups removed from the analysis.

Table 3.4. Variants of the EB dataset based on treatment of the *E. coli* and *Shigella* groups.

Ways of grouping	**Group information**	**Number of sequences**
Combine *E. coli* and *Shigella*	*Yersinia*, *E. coli* + *Shigella*, *Salmonella*, *Klebsiella*	110
Keep only *E. coli*	*Yersinia*, *E. coli*, *Salmonella*, *Klebsiella*	92
Keep only *Shigella*	*Yersinia*, *Shigella*, *Salmonella, Klebsiella*	57
No *E. coli* or *Shigella*	*Yersinia*, *Salmonella*, *Klebsiella*	48

Dataset 2 also allows us to investigate the effect of external groups (here *Klebsiella*, *Salmonella* and *Yersinia*) on inference within a clade (ECS). We generate Dataset 2b by replacing the 62 *E. coli* and *Shigella* genomes in Dataset 2a with the 27 ECS genomes as above. Using the phyletic (biological) grouping introduced above for the ECS dataset and retaining the default criterion for the IDF threshold (see previous paragraph), TF-IDF infers many more transfers within ECS (Figure 3.2, Figure 3.7 and Supplementary Table S3). Inclusion of the more-distantly related genomes has lowered the mean value of elements in the relationship matrix, thereby allowing many more regions within the ECS genomes to exceed the IDF threshold. Even with the additional TF filter, which remains unchanged, this results in a great increase in the number of transfers. As might be expected, all transfers detected in the ECS dataset by itself are still detected. Interestingly, this increase is non-uniform across the ECS subgraph: in every case where we inferred no LGT from one group into another when only the ECS dataset was examined, we found abundant LGT after adding the three additional genera.

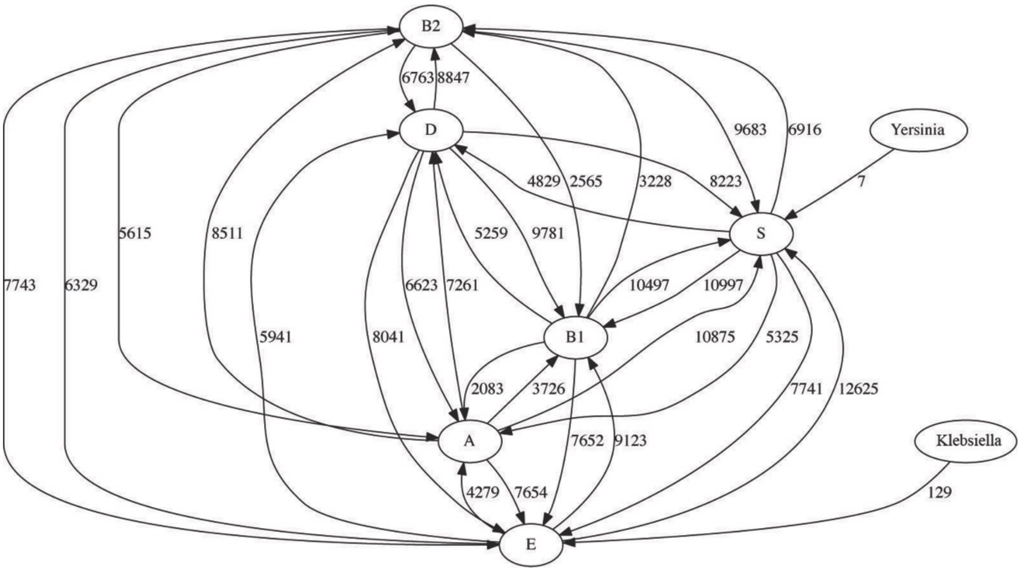


Figure 3.7. The LGT network inferred using TF-IDF from Dataset 2b, using six phyletic groups for *E. coli* and *Shigella* and grouping the remaining genomes according to genus. The numbers on each edge show the total number of genes involved in LGT events from one group into genomes in the other group.

As mentioned above, another option is to manually override the default use of the mean *k*-mer frequency value as the IDF threshold. When we set the threshold from the ECS dataset as the IDF threshold in Dataset 2b, we infer exactly the same genes in the ECS genomes to have accepted LGT from an ECS donor group. That is, presence or absence of external groups does not affect the performance of TF-IDF beyond their effect on the IDF threshold. We consider this further in the Discussion.

Dataset 3 (bacteria and archaea)

The 143-genome Bacteria and Archaea (BA) dataset allows us to examine the effect of within-group heterogeneity on inference using TF-IDF. Here we delineate groups taxonomically by phylum (15 phyla) or alternatively by class (31 classes). Grouping the genomes by phylum, we infer 686 genes as affected by LGT, many fewer than in the smaller but less-divergent previous datasets. Indeed, we infer no inter-phylum LGT involving the archaeal phyla (Crenarchaeota, Euryarchaeota, Nanoarchaeota) or three of the bacterial phyla (Aquificales, Planctomycetes, Thermotogales), presumably for the reason indicated above for Datasets 2a and 2b: potential matches fail to pass the IDF threshold. Of the nine remaining bacterial phyla, eight are inferred to have been both donors and recipients, while one (Chlamydiales) has been a recipient only (Figure 3.8). The highest-activity pathways ("highways"[14]) lie between Proteobacteria and High-G+C Firmicutes (378 genes affected), followed by those between Proteobacteria and Low-G+C Firmicutes (101 genes). Two phyla represented by one genome each, Thermus/Deinococcus and Chlorobi, contribute 13.9% and 6.3%

of total inter-phylum LGT; if more sequences had been included, these groups might be recognised as even more-active in inter-phylum LGT.

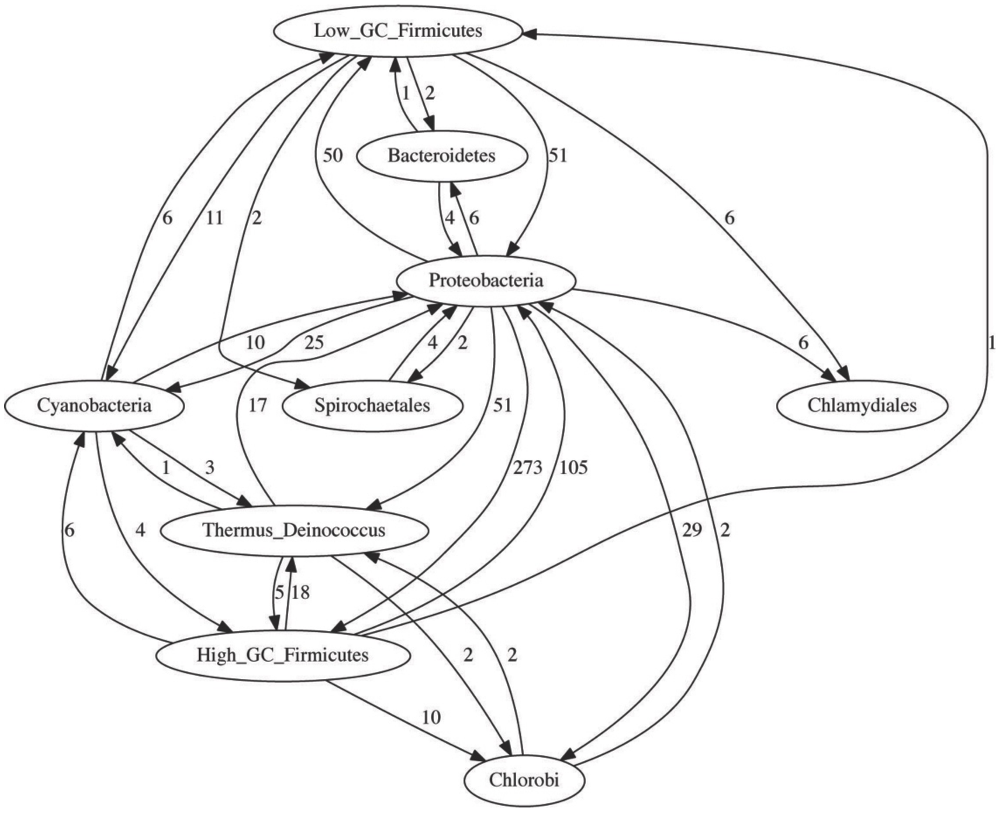


Figure 3.8. The LGT network inferred for the BA dataset grouped by phylum. The numbers on each edge show the total number of genes involved in LGT events from one group into genomes in the other group.

When we alternatively group the 143 genomes into 31 classes (Figure 3.9), the number of genes inferred to have accepted inter-class LGT increases nearly five-fold to 3043. We infer 24 lateral genes among eight archaeal classes with a ninth class, Archaeoglobales, silent to inter-class LGT. As above, no LGT is detected between archaea and bacteria. This grouping divides Proteobacteria into four subdivisions (α, β, γ, ε) at class level; genomes of the former three are rich in inferred lateral genes, whereas the ε subdivision is relatively silent. In accordance with our phylum-level analysis, the *Bacillus*/*Clostridium* class and Actinomycetales (from High-G+C Firmicutes) are inferred to have engaged in LGT with genomes across the subdivisions of Proteobacteria. By contrast, and in contradiction to earlier reports, we infer no LGT involving the *Thermotoga*[39] or *Aquifex*[40] genomes. This may be due to features of our dataset *e.g.* the number, size, balance, composition and cohesion of groups, and/or the phylogenetic distinctiveness of these genomes (see Discussion and conclusions).

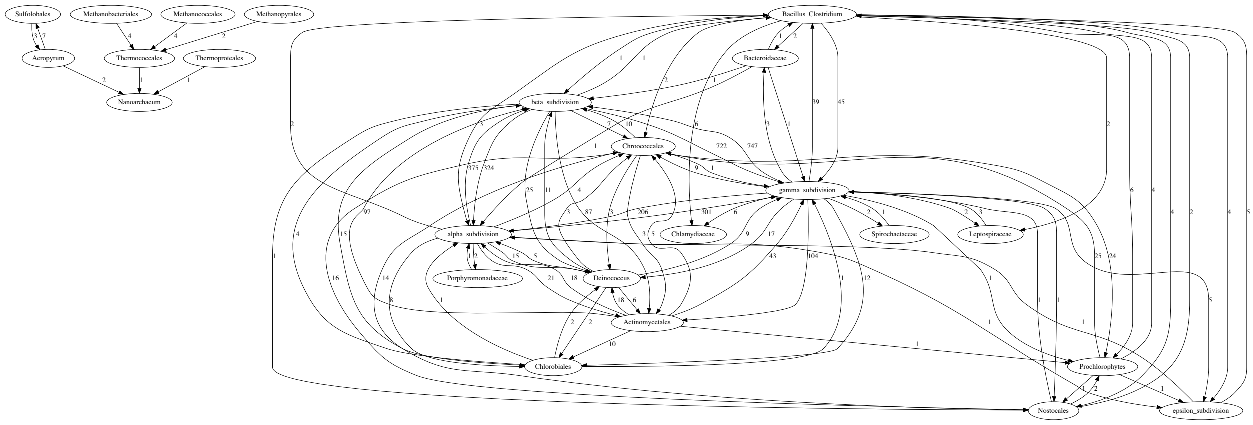


Figure 3.9. The LGT network inferred for the BA dataset grouped by class. The numbers on each edge show the total number of genes involved in LGT events from one group into genomes in the other group.

Using a dataset of 657 bacterial and archaeal genomes and a multi-step LGT inference approach based on anomalous G+C content and phylogenetic discordance, Popa *et al.*[41] identified 4700 genes of inferred lateral origin. Nine of these were also identified as lateral in our TF-IDF analysis of the 143-genome dataset, although with different inferred donors. Our comparison (Supplementary Section 2 and Tables S4-S7) indicates that at least at this phyletic scale, TF-IDF provides access to LGT events spanning broader phyletic distances than does the approach of Popa *et al.*[41].

3.2.3 Multiple donor groups and superimposed transfers

In the ECS dataset we observe a large number of transfers; correspondingly, we find many instances in which a gene is inferred to have accepted genetic material laterally from more than one donor group. This is especially prevalent in the phyletic grouping of six groups, whereas the phylogenetic grouping (four groups) contains fewer transfers as observed above. Here we look more closely at genomes which contain genes with multiple donors, to determine if we can untangle the sources of multiple transfer.

There are two possible explanations for such instances. One is that the gene is truly a mosaic, having accepted multiple transfers in the past. However, an alternative explanation is that there was only one transfer, but it was more ancient. For example, if a sequence is inferred to have accepted genetic material from groups G1 and G2, then it is possible that instead there was a single ancestral transfer from an ancestor of G1 and G2. This explanation is parsimonious only if G1 and G2 are closely related (*i.e.* monophyletic, or adjacent on the phylogenetic tree) and the events are inferred to affect overlapping regions on the genome. If either of these conditions is not met, it is more likely that more than one transfer event has occurred.

In Table 3.5, we examine the relative frequencies of possible ancestral transfers in eighteen ECS genomes. For ease of analysis, we consider only genes which are inferred to have accepted material from exactly two donor groups. Of the 2240 such genes in *E. coli* K12 W3110, K12 MG1655, HS and IAI1, which contain monophyletic relationships in their donor-group pairs, we observe 18446 events into these genes, forming 6549 overlapping regions. Of these overlapping regions, the donor groups are monophyletic in 1869 cases. Thus there is considerable evidence for both ancient transfers and mosaicism. However, ancient transfers (identified in this way) represent only 22-30% of overlapped events in these four genomes. For the other genomes no monophyletic overlapping regions are found, *i.e.* most overlaps may be the result of multiple lateral events. This is the first time a computational method has given us broad accessibility to data that can indicate the presence of these phenomena.

Table 3.5. Numbers of inferred lateral genes with two donors in the ECS dataset, and the phyletic relationship of the donors.

Genome name	Lateral genes with two donors	Number of overlaps (pairwise)	Number monophyletic	Number non-monophyletic	Proportion monophyletic
E. coli K12 substr W3110	117	223	50	173	22.4%
E. coli K12 substr MG1655	465	1514	423	1091	27.9%
Shigella flexneri 2a	288	709	0	709	0
Shigella flexneri 2a 2457T	158	361	0	0	0
Shigella sonnei Ss046	626	1564	0	1564	0
Shigella boydii Sb227	538	1247	0	1247	0
E. coli 536	328	522	0	522	0
Shigella flexneri 5 8401	120	259	0	259	0
E. coli HS	667	1885	564	1321	29.9%
E. coli E24377A	1068	2272	0	2272	0

Shigella boydii CDC 3083 94	187	277	0	277	0
E. coli SE11 (#20)	1049	2250	0	2250	0
E. coli 0127 H6 E2348 69	448	670	0	670	0
E. coli IAI1	991	2927	832	2095	28.4%
E. coli S88	287	413	0	413	0
E. coli ED1a	305	431	0	431	0
E. coli 55989	1081	2344	0	2344	0
E. coli UMN026	769	2445	0	2445	0

3.2.4 Biological process enrichment

To determine the frequencies at which different sorts of proteins are implicated in our LGT detections, we extracted protein-name annotations from the corresponding GenBank files (Table 3.6). The most-frequent name annotation is in LGT events is *hypothetical protein*, followed by *membrane protein*, *transcriptional regulator* and *transporter*, protein types known to be exchanged among bacteria[42,43]. To further investigate the biological processes affected by LGT, we carried out functional enrichment tests (see Methods), selecting a false discovery rate of 0.05 as significance threshold. Here we present a general discussion of biological processes over- or under-represented in the datasets; full lists of terms are given in Supplementary Tables S8-S10.

Table 3.6. Top 25 protein names (as extracted from the GenBank files) inferred to be affected by an LGT event in the ECS dataset.

Counts	**Protein name**
11944	hypothetical protein
1426	membrane protein
1056	transcriptional regulator
615	transporter
454	oxidoreductase
265	transposase
251	LysR family transcriptional regulator
232	tail protein
192	two-component system response regulator
190	lipoprotein
182	two-component system sensor histidine kinase
162	diguanylate cyclase
158	hydrolase
144	AraC family transcriptional regulator
141	MFS transporter
135	ABC transporter ATP-binding protein

125	porin
124	peptide ABC transporter permease
121	putative DNA-binding transcriptional regulator
115	protease
105	fimbrial protein
104	sensor histidine kinase
102	multidrug ABC transporter ATP-binding protein
100	glycosyl transferase
95	ABC transporter permease

Among the ECS genomes, enrichment analysis identifies metabolite and trans-membrane transport, carbohydrate metabolic processes, and small-molecule biosynthesis and catabolism as particularly over-represented as inferred targets of LGT; at least 42 of the 50 most over-represented terms refer specifically to such processes. By contrast, the 21 most under-represented terms refer to transposition, genetic recombination, translation, or metabolism of peptides or nitrogenous compounds.

Within the Enteric Bacteria (EB) dataset, enrichment of terms can depend on how we group the *E. coli* and *Shigella* genomes. When 62 *E. coli* and *Shigella* genomes are combined into a single group, biological processes related to translation, nitrogen-compound and RNA biosynthesis, and viruses dominate the most over-represented functions, while trans-membrane transport and polysaccharide metabolism are under-represented. Removing all *E. coli* and *Shigella* genomes, only the *E. coli* genomes, or only the *Shigella* genomes, does not greatly affect this picture. When the *E. coli* and *Shigella* genomes are retained but grouped separately, the TF-IDF analysis is dominated by LGT between these groups (Figure 3.6a); viral processes including entry into and release from host cells, and extracellular (lipo)polysaccharide biosynthesis, come to the fore among over-represented processes, while translation, transposition, and purine and ribose metabolism are now under-represented. These results illustrate how grouping can affect the functional interpretation of LGT in bacterial genomes.

With the BA dataset grouped by phylum, relatively few genes are inferred to have accepted LGT (above). Thirty-five processes are found to be over-represented (see Supplementary section 4), with *translational elongation* (GO:0006414) being by far the most significant. No under-represented process passes our FDR threshold. Grouping instead by class, diverse metabolic processes appear as over-represented, while only two processes appear as (slightly) under-represented.

3.3 Discussion and conclusions

TF-IDF is an alignment-free method for the detection of regions of exogenous origin in molecular sequences. Based on the content of *k*-mers in a specific dataset, the method can identify regions of exogenous origin in a sequence, and their inferred donor groups within the dataset, with high efficiency and effectiveness[8]. Here we apply TF-IDF on three empirical microbial-genome datasets of different sizes and sequence diversity to explore the advantages and limitations of this method. We systematically varied two key parameter-value settings (for word length k and gap size G), and investigated how the delineation of groups affects the performance of TF-IDF.

Our results indicate that it may not be possible to identify a value of k optimal for all datasets. However, for these microbial genome datasets, the distribution of shared 12-mers helps us to select k. Within the ECS genomes, the vertical components of which share a relatively recent common ancestor, the proportion of identical 12-mers is relatively high (median >60%) and a longer k ($35 \leq k \leq 50$) supports high-confidence detections while not missing too many real LGTs. By contrast, in the highly divergent BA dataset in which most genomes share <30% identical 12-mers pairwise, almost no LGT is detected at $k \geq 40$. To ensure adequate LGT signal in such a dataset, k must be set smaller ($20 \leq k \leq 30$). However, at $k \leq 15$, k-mers are too frequently matched pairwise at random, leading to an unacceptable level of false positives. In general, larger values of k are appropriate for high-similarity datasets, and shorter k for low-similarity data.

G determines how aggressively nearby lateral k-mers are consolidated into a single region. Given a sufficient density of such k-mers, a larger G causes intervening non-lateral regions to be merged into the consolidated region. This can cause some false positive regions to be detected by TF-IDF. At shorter G, the total number of detections increases without greatly affecting total detection length. Thus shorter G is preferred for precise delineation of lateral segments. In most cases, $G = 2k$ is a satisfactory option.

Apart from k and G, TF-IDF is also sensitive to how groups are recognised within the dataset. Many more transfers were inferred within the ECS dataset when six, rather than four, groups were recognised. It is difficult to disentangle the effects of group number, size, composition and phylogenetic cohesion, but we use the fortuitous stability of Group A/3 to argue that both TF and IDF terms can contribute to this sensitivity. We further demonstrate that the presence or absence of external groups does not affect the performance of TF-IDF beyond their effect (*via* the constituent sequences) on the IDF threshold. These results emphasise that as implemented here, TF-IDF is deterministic and is self-tuning to the dataset.

We suspect that these effects manifest so strongly in ECS Groups E and B1 because these lineages have been particularly active in LGT. Analysing a 64-genome superset of our ECS dataset, Lukjancenko *et al.*[44] find that members of Group E contribute almost half of the new gene families, while our four Group B1 strains contribute aggressively to the rise in pan-genome size. Additional effects arising from variation in gene content may contribute further.

TF-IDF delivers the most power when applied to a sequence dataset with high within-group similarity but uniformly low between-group similarity. However, group structure can be arbitrary in real-life cases. Our results with the ECS subset of Dataset 2 illustrate strategies for dealing with uneven or unbalanced data. The substantial loss of LGT signal when we assign sequences randomly to groups strongly indicates that groups should be delineated so as to capture the underlying phylogenetic structure where possible. This may not always be possible, as in our BA dataset, where *Thermotoga* and *Aquifex* are represented by single genomes of uncertain but relatively distant phylogenetic relationship. Other strategies for delineating groups can be imagined, but lie outside the scope of the present study.

It has long been considered that lateral transfers from different donor groups can be superimposed in the recipient genome, yielding mosaic or pastiche genes[45,46]. We have now demonstrated this in the ECS dataset. Most genes that have accepted LGT have done so in multiple events, often from different donor groups. Where group structure reflects evolutionary history and neighbouring genomic regions are inferred to have been donated by groups adjacent on the tree, the transfer may have been ancestral. Taking a gene-centric approach restricted (for simplicity of analysis) to genes with only two inferred lateral origins, we find that a modest proportion (22-30%) might best be explained by ancestral transfer. To our knowledge, this is the first systematic computational study of multiple or overlapping origins in empirical genome-scale data.

We mapped genes containing the inferred lateral regions to Gene Ontology (GO) terms using BLAST2GO then applied enrichment tests, identifying a wide range of biological processes as preferentially affected by LGT. Many processes known to be shared laterally are indeed over-represented, although others (including e.g. transposition) are under-represented, whether as a consequence of their actual distribution in the dataset, or their presumed origin from a donor group not represented in the dataset.

Our inference that genes annotated as involved in *translational elongation* (GO:0006414) in the BA dataset, and in *translation* (GO:0006412) in the EB dataset, are overrepresented among the LGT sets bears comment, as “informational” functions are considered less-susceptible to LGT than “operational” genes *e.g.* those involved in cellular transport or metabolism[42]. Closer examination reveals that (1) substantial subsets of our LGT-enrichment sets annotated with *translational elongation* (in BA) or *translation* (in EB) are not core informational genes, or indeed informational genes at all, but appear in our lists *via* secondary annotations *e.g.* involving specialised regulatory relationships; (2) many informational genes[42] are well-known to be susceptible to LGT; (3) even "core" informational genes are sometimes transferred laterally; and (4) in a few cases, core informational genes that we infer as lateral have features or properties (*e.g.* constraints, domains, paralogs, phyletic distributions) that could indicate a lateral history, or help explain why a lateral history has gone unrecognised by classical methods (for details see Supplementary material). Further, *translation* is over-represented when the 62 *E. coli* and *Shigella* genomes are combined into a single group; when they all are removed from the analysis; or when only *E. coli*, or only *Shigella*, are removed. However, when we include these *E. coli* and *Shigella* genomes but group them separately, *translation* becomes under-represented. That is, the LGT "translation" signal is being driven from parts of the dataset other than the *E. coli-Shigella* axis, and is completely overshadowed (indeed driven to under-representation) by the (much stronger) signal from the mostly non-translational transfers between *E. coli* and *Shigella*.

For each dataset, the groups (nodes) and inferred transfers (edges) constitute the LGT network. Each of these three networks exhibits one or more densely connected regions (subgraphs), as well as nodes that are more-weakly connected or unconnected. The lack of connection between archaea and bacteria in the BA dataset is a case in point: far fewer transfers are inferred between archaea and bacteria than internally among archaea, or internally among bacteria. In the TF-IDF analysis of Dataset 2a, *Yersinia* remains almost unconnected to other genera; this illustrates that even among LGT-active groups, some genera can remain inactive.

In summary, our results demonstrate that TF-IDF can be applied on diverse empirical genome-scale datasets, resulting in the inference of inter-group directional LGT and providing first steps toward the systematic reconstruction of multiple and superimposed transfer events. These inferred transfers affect a broad range of biological processes, including many already known or suspected to be affected by LGT. Future work will explore whether and how the settings of *k* and *G* affect topological features of the inferred LGT networks, hence our interpretation of lateral biology in microbial communities and the biosphere.

3.4 Methods

3.4.1 Datasets

From our earlier simulation study[8] we know that the performance of TF-IDF can be affected by how groups are delineated within a dataset, and by the divergence of sequences within a group. If sequences within groups are similar to one another (expected mutations up to 0.16/nucleotide) and the groups are dissimilar from one another (expected mutations between neighbouring groups above

0.2/nucleotide), the boundaries between groups are clear and TF-IDF can achieve high precision (>80%) and recall (>90%)[8]. Here, we select three empirical datasets that differ in number of sequences and divergence among sequences, to explore the performance of TF-IDF under a range of biologically realistic situations. Table 3.7 shows general information (number of sequences, sequence lengths and G+C content) on these datasets, while further information is presented in the following paragraphs.

Table 3.7. General description of the datasets investigated in this research.

Name	**Number of sequences**	**Mean length**	**Range of lengths**	**Mean G+C content (%)**	**Range of G+C content (%)**
ECS	27	4906162	4369232-5528445	50.76	50.39-51.33
EB	110	4920079	3976195-6097032	51.03	47.00-57.68
BA	143	3011345	490885-9105828	45.67	22.48-72.12

Dataset 1: *Escherichia coli* and *Shigella* (abbreviated hereafter as ECS), represented by 20 and 7 genomes respectively. Here and elsewhere[47] the *Shigella* genomes are resolved as one or more lineages within the genus *Escherichia*. Some genomes within ECS are known to be rich in regions of inferred lateral origin[47]. Using alignment-based methods, we have previously shown that lateral transfer of protein-coding regions within ECS is biased by phylogeny (*i.e.* genetic relatedness and/or sequence similarity) more than by environment[31], whereas the distribution of small RNAs has been affected more by gene loss than by LGT[48]. For the present work we recognise groups within ECS in two alternative ways: (1) by cutting the MRP supertree[31] at certain levels (see Supplementary Figure S1), or (2) by using recognized phyletic groups[36]. These approaches yield four and six groups respectively.

Dataset 2a: 110 genomes from the Enterobacteriaceae (53 *Escherichia*, 9 *Shigella,* 9 *Klebsiella*, 22 *Salmonella* and 17 *Yersinia*), here abbreviated EB. Among these *Escherichia*, *Shigella, Klebsiella* and *Salmonella* are considered relatively susceptible to LGT. Strains of *Yersinia* harbour plasmids that encode genes of probable lateral origin[49,50] but our datasets exclude plasmid sequences. *Yersinia* appears not to be naturally competent[51] and although its main chromosome shows evidence of pathogenicity islands, their genes match sequences outside the Enterobacteriaceae[52] and thus would not be recognised as lateral in our analyses of Dataset 2. We recognise each genus as a separate group except for *E. coli* and *Shigella*, which we treat in different ways (see Results).

Dataset 2b: These 75 genomes constitute a subset of Dataset 2a (pruned to 20 *E. coli* and 7 *Shigella*) and a superset of Dataset 1 (addition of 58 genomes from the other genera). We expect to see the same LGT detections within *E. coli* and *Shigella* as in the ECS dataset when the threshold is the same. Together, Datasets 2a and 2b allow us to explore the effects of group inclusion/exclusion (of groups other than ECS) and subdivision (ECS).

Dataset 3: 143 genomes across 12 bacterial and 3 archaeal phyla, abbreviated here as BA. This dataset allows us to explore the effects of phyletic breadth, degree of sequence divergence, unbalanced group size and disruptive genomes on LGT inference. This dataset has been well-explored in our group using classical alignment-based (and some novel) methods for more than ten years[14,33,34]; MRP[53] and 16S rRNA reference trees are available. This dataset moreover offers a more-general (less-biased) selection of Gene Ontology (GO) Biological Process (BP)[35] annotations than do specialist datasets dominated by human and animal pathogens (our Datasets 1 and 2).

These datasets span a variety of evolutionary divergences. Information on the divergence among a dataset is important for setting the parameters of TF-IDF; however, typical approaches based on alignments are time-consuming and do not scale well with increasing number of sequences. To quantify this variation, we thus compute a rough measure of sequence similarity by calculating the percentage of identical 12-mers shared between each pair of sequences. Summary information is presented in Table 3.1. The distribution of similarities is shown in Figure 3.10; here we see that as expected, the ECS genomes are most similar pairwise, the EB genomes are more divergent (with a small bimodality consistent with the ECS subset) and the BA genomes the most divergent, with most sequence pairs sharing fewer than 30% of their 12-mers.

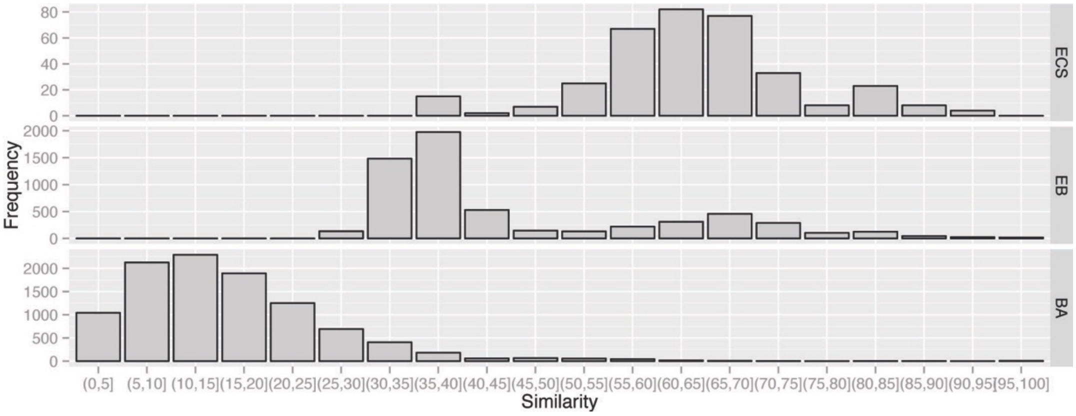


Figure 3.10. Distribution of 12-mer frequencies in the ECS, EB and BA datasets. The *x*-axis shows binned proportions of 12-mers shared pairwise over all genomes in each dataset; the left *y*-axis shows frequencies (counts) of 12-mers, and the right *y*-axis gives the name of the dataset.

3.4.2 TF-IDF and parameterisation

In this study we apply the TF-IDF method we devised in previous research[8]. TF-IDF is an alignment-free method that detects LGT by the relative frequencies of *k*-mers in pre-determined groups. The method proceeds in four steps:

1. Extract all unique *k*-mers in a dataset and build a *k*-mer dictionary of the dataset.
2. IDF: we count the identical *k*-mers between each sequence and each group other than its own. A relationship matrix *R* is built in which rows are genomes, columns are groups, and individual elements count the number of identical *k*-mers shared between a sequence and a group. For consistency across group sizes and genome lengths, we normalise these counts by dividing by the number of genomes in the group (column), and by the number of nucleotides in the genome (row). We then compute the mean over all elements in *R*. If the value of an element exceeds the mean, the corresponding genome potentially contains lateral events (segments) donated by that group.
3. For each genome with potential transfers from a donor group, we construct potential LGT segments by amalgamating all neighbouring *k*-mers in the genome which also appear in that group. These segments are further merged by joining all segments which are separated by an amount less than a threshold, which we refer to as gap size (*G*).
4. TF: if the average frequency of all lateral *k*-mers in a candidate LGT segment is lower than the average frequency of all *k*-mers in the group containing that genome, then that segment is considered to have arisen by LGT.

In this work we vary word length *k* and gap size *G* (see Table 3.8). Based on the results of our previous study[8], we limit *k* to the range 20-45; when $k < 20$ many detected events are false positives, while at $k \geq 50$ common *k*-mers become too rare, resulting in decreased performance. Values of *G* were selected to cover a biologically reasonable range of granularity consistent with computational feasibility.

Table 3.8. Range of parameter values investigated with the TF-IDF method. *k* is the size of *k*-mers.

Parameter	**Parameter values**
Word length *k*	20-45 in steps of 5
Gap size *G*	2*k*, 4*k*, 8*k*

For the ECS dataset, we also vary group composition in order to study its effect on inference using TF-IDF. We recognise groups in two ways as described above; in addition, we also generate 50 randomised groupings patterned on the first grouping (into four groups by phylogeny) by allocating each sequence to a group chosen at random, while preserving the number of sequences in each group. By doing this we generate a control set in which vertical inheritance signal is greatly attenuated, and against which we can compare our actual grouping. The total detection length based on actual groups (generated by cutting the MRP tree) is significantly higher than from the random replicates.

3.4.3 Gene Ontology mapping and enrichment tests

For each recipient genome, our TF-IDF analysis outputs a list of regions (coordinates) inferred to be of lateral origin, and the inferred donor group of each. To identify the biological functions affected by these regions, we map these coordinates to genes annotated in the host genome (as given in the NCBI .ffn and .gbk files). For both biological and statistical reasons, we examine only relatively long regions: biologically we are interested only in LGT events with potential to have functional consequence, while statistically we seek to minimise false positives and noise. Thus, a gene is considered only if it contains at least one segment which is longer than a given threshold. These thresholds (given in Table 3.9) are selected on different datasets to be close to the average length of all LGT detections in that dataset. This accounts for the variation in sequence diversity among the datasets.

Table 3.9. Thresholds for mapping LGT segments to genes. The thresholds are selected by means of all detections in three empirical datasets.

Dataset	**Thresholds for laterally transferred genes**
ECS	> 500 contiguous *k*-mers (500 + *k*-1 nt)
EB	> 100 contiguous *k*-mers (100 + *k*-1 nt)
BA	> 10 contiguous *k*-mers (10 + *k*-1 nt)

For each dataset we used blastp[54] at $E \leq 10^{-5}$ to match protein-coding regions annotated in all genomes to the Swiss-Prot database[55]. Genes were distinguished by GI number and position. Gene Ontology (GO) terms associated with the matches were retrieved using BLAST2GO[56,57] version 3.3.1 (mapping and annotation functions) from GO database version b2g_may15, yielding the background database for enrichment testing. We then submit a list of genes implicated as recipients of LGT, querying this list against the entire database. Regardless of the number of inferred lateral regions or donor groups involved, each gene is counted only once. We use a two-tailed Fisher's exact test with a false discovery rate (FDR) of 0.05. This yields a list of GO Biological Process (BP) annotations which are over- and under-represented in the test set[58].

3.5 References

1 Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* **36**, 760-766, doi:10.1038/ng1381 (2004).

2 Gogarten, J. P. & Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* **3**, 679-687, doi:10.1038/nrmicro1204 (2005).

3 Boto, L. Horizontal gene transfer in evolution: facts and challenges. *Proc R Soc B* **277**, 819-827, doi:10.1098/rspb.2009.1679 (2010).

4 Barlow, M. What antimicrobial resistance has taught us about horizontal gene transfer. In *Horizontal Gene Transfer: Genomes in Flux* (ed. Gogarten, M.B., Gogarten, J. P. & Olendzenski, L. C.) 397-411 (Springer, 2009).

5 Gyles, C. & Boerlin, P. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Vet Pathol* **51**, 328-340, doi:10.1177/0300985813511131 (2014).

6 Médigue, C., Rouxel, T., Vigier, P., Henaut, A. & Danchin, A. Evidence for horizontal gene transfer in Escherichia coli speciation. *J Mol Biol* **222**, 851-856 (1991).

7 Mrázek, J. & Karlin, S. Detecting alien genes in bacterial genomes. *Ann N Y Acad Sci* **870**, 314-329, doi: 10.1111/J.1749-6632.1999.Tb08893.X (1999).

8 Cong, Y., Chan, Y.-b. & Ragan, M. A. A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF. *Sci Rep* **6**, 30308.

9 Ragan, M. A. & Beiko, R. G. Lateral genetic transfer: open issues. *Philos Trans R Soc Lond B Biol Sci* **364**, 2241-2251, doi:10.1098/rstb.2009.0031 (2009).

10 Lawrence, J. G. & Ochman, H. Amelioration of bacterial genomes: Rates of change and exchange. *J Mol Evol* **44**, 383-397, doi: 10.1007/Pl00006158 (1997).

11 Harrison, E. & Brockhurst, M. A. Plasmid-mediated horizontal gene transfer is a coevolutionary process. *Trends Microbiol* **20**, 262-267, doi:10.1016/j.tim.2012.04.003 (2012).

12 Chan, C. X. & Ragan, M. A. Next-generation phylogenomics. *Biol. Direct* **8**, 3. doi:10.1186/1745-6150-8-3 (2013).

13 Keeling, P. J. & Palmer, J. D. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet* **9**, 605-618, doi:10.1038/nrg2386 (2008).

14 Beiko, R. G., Harlow, T. J. & Ragan, M. A. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A* **102**, 14332-14337, doi:10.1073/pnas.0504068102 (2005).

15 Galtier, N. A model of horizontal gene transfer and the bacterial phylogeny problem. *Syst Biol* **56**, 633-642, doi:10.1080/10635150701546231 (2007).

16 Ané, C., Larget, B., Baum, D. A., Smith, S. D. & Rokas, A. Bayesian estimation of concordance among gene trees. *Mol Biol Evol* **24**, 412-426 (2007).

17 Zhang, G. J., Li, C., Li, Q., *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311-1320, doi:10.1126/science.1251385 (2014).

18 Haubold, B. Alignment-free phylogenetics and population genetics. *Brief Bioinform* **15**, 407-418, doi:10.1093/bib/bbt083 (2014).

19 Chan, C. X., Bernard, G., Poirion, O., Hogan, J. M. & Ragan, M. A. Inferring phylogenies of evolving sequences without multiple sequence alignment. *Sci Rep* **4**, 6504. doi:Artn 6504 10.1038/Srep06504 (2014).

20 Höhl, M., Rigoutsos, I. & Ragan, M. A. Pattern-based phylogenetic distance estimation and tree reconstruction. *Evol Bioinform Online* **2**, 359-375 (2006).

21 Höhl, M. & Ragan, M. A. Is multiple-sequence alignment required for accurate inference of phylogeny? *Syst Biol* **56**, 206-221, doi:10.1080/10635150701294741 (2007).

22 Allman, E. S., Rhodes, J. A. & Sullivant, S. Statistically-consistent *k*-mer methods for phylogenetic tree reconstruction. Preprint, *arXiv:*1511.01956 (2015).

23 Liu, Z., Meng, J. & Sun, X. A novel feature-based method for whole genome phylogenetic analysis without alignment: application to HEV genotyping and subtyping. *Biochem Biophys Res Commun* **368**, 223-230, doi:10.1016/j.bbrc.2008.01.070 (2008).

24 Gao, Y. & Luo, L. Genome-based phylogeny of dsDNA viruses by a novel alignment-free method. *Gene* **492**, 309-314, doi:10.1016/j.gene.2011.11.004 (2012).

25 Wang, L. S., Warnow, T., Moret, B. M., Jansen, R. K. & Raubeson, L. A. Distance-based genome rearrangement phylogeny. *J Mol Evol* **63**, 473-483, doi:10.1007/s00239-005-0216-y (2006).

26 Blanchette, M., Kunisawa, T. & Sankoff, D. Gene order breakpoint evidence in animal mitochondrial phylogeny. *J Mol Evol* **49**, 193-203 (1999).

27 Zhang, W., Yoshida, T. & Tang, X. J. A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Syst Appl* **38**, 2758-2765 (2011).

28 Srividhya, V. & Anitha, R. Evaluating preprocessing techniques in text categorization. *Int J Comput Sci Applic* **47** (2010).

29 Domazet-Lošo, M. & Haubold, B. Alignment-free detection of horizontal gene transfer between closely related bacterial genomes. *Mob Genet Elements* **1**, 230-235, doi:10.4161/mge.1.3.18065 (2011).

30 Domazet-Lošo, M. & Haubold, B. Alignment-free detection of local similarity among viral and bacterial genomes. *Bioinformatics* **27**, 1466-1472, doi:10.1093/bioinformatics/btr176 (2011).

31 Skippington, E. & Ragan, M. A. Phylogeny rather than ecology or lifestyle biases the construction of *Escherichia coli-Shigella* genetic exchange communities. *Open Biol* **2**, doi:Artn 120112 10.1098/Rsob.120112 (2012).

32 Ragan, M. A., McInerney, J. O. & Lake, J. A. The network of life: genome beginnings and evolution. *Phil Trans R Soc Lond B Biol Sci* **364**, 2169-2175, doi:10.1098/rstb.2009.0046 (2009).

33 Chan, C. X., Beiko, R. G., Darling, A. E. & Ragan, M. A. Lateral transfer of genes and gene fragments in prokaryotes. *Genome Biol Evol* **1**, 429-438, doi:10.1093/gbe/evp044 (2009).

34 Chan, C. X., Darling, A. E., Beiko, R. G. & Ragan, M. A. Are protein domains modules of lateral genetic transfer? *PLoS ONE* **4**, e4524, doi:10.1371/journal.pone.0004524 (2009).

35 Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res* **38**, D331-335, doi:10.1093/nar/gkp1018 (2010).

36 Gordon, D. M., Clermont, O., Tolley, H. & Denamur, E. Assigning Escherichia coli strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environ Microbiol* **10**, 2484-2496, doi:10.1111/j.1462-2920.2008.01669.x (2008).

37 de la Cruz, F. & Davies, J. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol* **8**, 128-133 (2000).

38 Garcia-Vallvé, S., Romeu, A. & Palau, J. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* **10**, 1719-1725 (2000).

39 Nelson, K. E., Clayton, R. A., Gill, S. R., *et al.* Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323-329, doi:10.1038/20601 (1999).

40 Deckert, G., Warren, P. V., Gaasterland, T., *et al.* The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**, 353-358, doi:10.1038/32831 (1998).

41 Popa, O., Hazkani-Covo, E., Landan, G., Martin, W. & Dagan, T. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res* **21**, 599-609, doi:10.1101/gr.115592.110 (2011).

42 Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: The complexity hypothesis. *Proc Natl Acad Sci U S A* **96**, 3801-3806, doi: 10.1073/Pnas.96.7.3801 (1999).

43 Skippington, E. & Ragan, M. A. Lateral genetic transfer and the construction of genetic exchange communities. *FEMS Microbiol Rev* **35**, 707-735, doi:10.1111/j.1574-6976.2010.00261.x (2011).

44 Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 sequenced Escherichia coli genomes. *Microb Ecol* **60**, 708-720, doi:10.1007/s00248-010-9717-3 (2010).

45 Dowson, C. G., Coffey, T. J., Kell, C. & Whiley, R. A. Evolution of penicillin resistance in Streptococcus pneumoniae; the role of *Streptococcus mitis* in the formation of a low affinity PBP2B in *S. pneumoniae*. *Mol Microbiol* **9**, 635-643 (1993).

46 Dowson, C. G., Hutchison, A., Brannigan, J. A., George, R. C., Hansman, D., Liñares, J., Tomasz, A., Smith, J. M., & Spratt, B. G. Horizontal transfer of penicillin-binding protein genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*. *Proc Natl Acad Sci U S A* **86**, 8842-8846 (1989).

47 Skippington, E. & Ragan, M. A. Within-species lateral genetic transfer and the evolution of transcriptional regulation in *Escherichia coli* and *Shigella*. *BMC Genomics* **12**, 532, doi:10.1186/1471-2164-12-532 (2011).

48 Skippington, E. & Ragan, M. A. Evolutionary dynamics of small RNAs in 27 *Escherichia coli* and *Shigella* genomes. *Genome Biol Evol* **4**, 330-345, doi:10.1093/gbe/evs001 (2012).

49 Hinnebusch, B. J., Rosso, M. L., Schwan, T. G. & Carniel, E. High-frequency conjugative transfer of antibiotic resistance genes to *Yersinia pestis* in the flea midgut. *Mol Microbiol* **46**, 349-354 (2002).

50 Hu, P., Elliott, J., McCready, P., Skowronski, E., Garnes, J., Kobayashi, A., Brubaker, R. R. & Garcia, E. Structural organization of virulence-associated plasmids of *Yersinia pestis*. *J Bacteriol* **180**, 5192-5202 (1998).

51 Lewin, A., Hertwig, S., Strauch, E. & Appel, B. Is natural genetic transformation a mechanism of horizontal gene transfer in *Yersinia*? *J. Basic Microb.* **38**, 17-26, doi:10.1002/(Sici)1521-4028(199803)38:1<17::Aid-Jobm17>3.0.Co;2-D (1998).

52 Parkhill, J., Wren, B. W., Thomson, N. R. *et al*. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**, 523-527, doi:10.1038/35097083 (2001).

53 Ragan, M. A. Phylogenetic inference based on matrix representation of trees. *Mol Phylogenet Evol* **1**, 53-58 (1992).

54 Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**, 203-214, doi: 0.1089/10665270050081478 (2000).

55 Boeckmann, B., Blatter, M.C., Famiqlietti, L., Hinz, U., Lane, L., Roechart, B. & Bairoch, A.. Protein variety and functional diversity: Swiss-Prot annotation in its biological context. *C R Biol* **328**, 882-899, doi:10.1016/j.crvi.2005.06.001 (2005).

56 Gotz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Roles, M., Talón, M. D., Dopazo, J. & Conesa, A. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucl Acids Res* **36**, 3420-3435, doi:10.1093/nar/gkn176 (2008).

57 Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M. & Robles, M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676, doi:10.1093/ bioinformatics/bti610 (2005).

58 Primmer, C. R., Papakostas, S., Leder, E. H., Davis, M. J. & Ragan, M. A. Annotated genes and nonannotated genomes: cross-species use of Gene Ontology in ecology and evolution research. *Mol Ecol* **22**, 3216-3241, doi:10.1111/mec.12309 (2013).

3.6 Acknowledgements

This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government. *Funding*: YC acknowledges the China Scholarship Council and The University of Queensland for stipend and tuition fee support. The project was funded in part by a grant from the James S. McDonnell

Supplementary Material

Exploring lateral genetic transfer among microbial genomes using TF-IDF

Yingnan Cong

Yao-ban Chan

Mark A. Ragan

Overview of Supplementary Material:

Section 1. Supplementary material for ECS (*E. coli* and *Shigella* genomes) dataset, and for comparison among our datasets

Supplementary Tables S1-S3 and Supplementary Figure S1

Section 2. Comparison between TF-IDF analysis of our BA (bacteria and archaea) dataset with grouping by class, and results reported by Popa *et al*. (2011) on a 657-genome bacteria and archaea dataset.

Supplementary Tables S4-S7

Section 3. Index of sheet names and contents for Excel spreadsheets

These spreadsheets (Supplementary Tables S8-S10), reporting Gene Ontology enrichment results for our three main datasets, are too large to be displayed here. They are being assigned DOI numbers and will be available for download from UQ eSpace.

Section 4. Supplementary material for Gene Ontology enrichment tests

Supplementary Tables S11-S12

Section 1. Supplementary material for ECS (*E. coli* and *Shigella* genomes) dataset, and for comparison among our datasets

Table S1. Summary of lengths of the inferred lateral segments, showing the mean, median first quartile, and median third quartile lengths of all inferred segments.

Dataset	Mean	Median	First quartile	Third quartile
ECS	550.70	269	113	622
EB	116.22	56	8	104
BA (phylum level)	7.188	3	2	6
BA (class level)	9.84	4	2	8

Table S2. Group assignments for ECS genomes into six or four groups (see text and Figure S1).

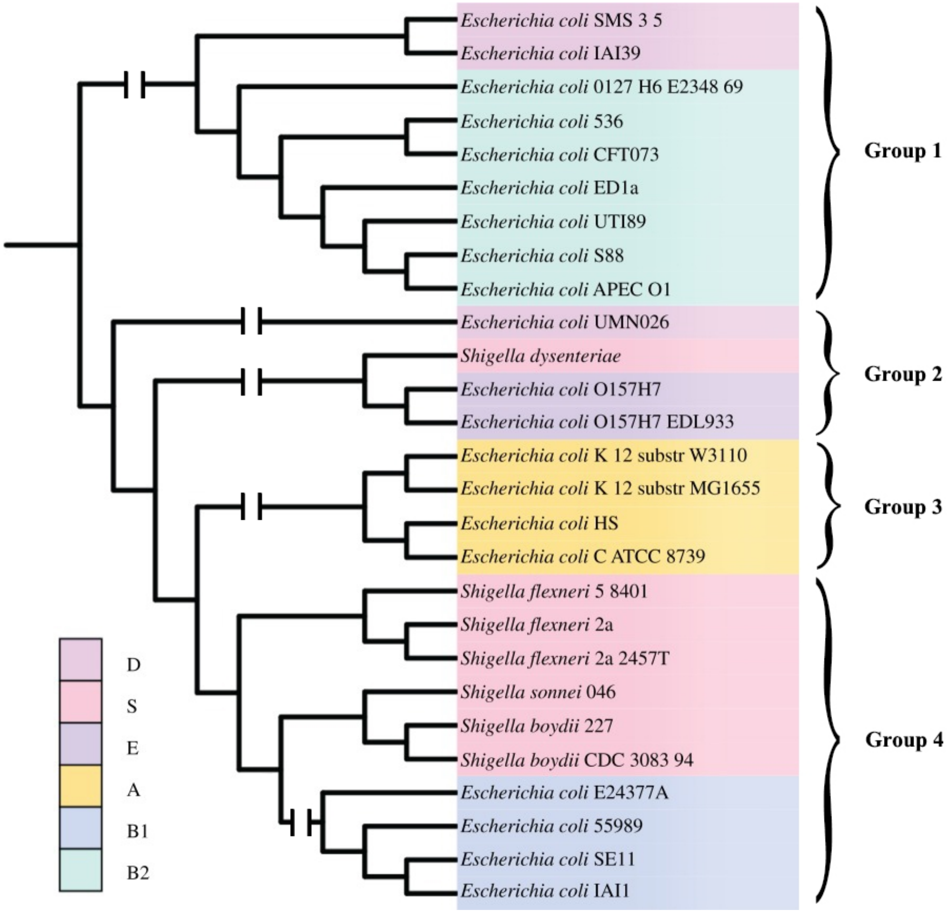
Six groups	Four groups	Organism
D	1	*E. coli* SMS 3 5
D	1	*E. coli* IAI39
D	2	*E. coli* UMN026
S	2	*Shigella dysenteriae*
S	4	*Shigella flexneri* 5 8401
S	4	*Shigella flexneri* 2a
S	4	*Shigella flexneri* 2a 2457T
S	4	*Shigella sonnei* Ss046
S	4	*Shigella boydii* Sb227
S	4	*Shigella boydii* CDC 3083 94
E	2	*E. coli* O157:H7
E	2	*E. coli* O157:H7 EDL933
A	3	*E. coli* K12 substr W3110
A	3	*E. coli* K12 substr MG1655
A	3	*E. coli* HS
A	3	*E. coli* C ATCC 8739
B1	4	*E. coli* E24377A
B1	4	*E. coli* 55989
B1	4	*E. coli* SE11
B1	4	*E. coli* IAI1
B2	1	*E. coli* 0127 H6 E2348 69
B2	1	*E. coli* 536
B2	1	*E. coli* CFT073
B2	1	*E. coli* ED1a
B2	1	*E. coli* UTI89
B2	1	*E. coli* S88
B2	1	*E. coli* APECO1

Table S3. Summary of lateral regions inferred among the 27 ECS genomes in Dataset 2b (*i.e.* when these 27 ECS genomes replace the 52 *E. coli* and *Shigella* genomes in the EB dataset, and *Salmonella*, *Klebsiella* and *Yersinia* genomes are present).

Group	Organism	Number of genes	Number of lateral genes	Donor groups
D	*E. coli* SMS 3 5	4744	3346	-

D	*E. coli* IAI39	4725	3212	S
D	*E. coli* UMN026	4878	3615	S, B1, B2
S	*Shigella flexneri* 5 8401	4336	1989	E, A, B1
S	*Shigella flexneri* 2a	4053	1850	D, E, A
S	*Shigella flexneri* 2a 2457T	4385	2091	E, A, B1
S	*Shigella sonnei* Ss046	4563	2436	D, E, A, B1
S	*Shigella boydii* Sb227	4391	2388	D, E, A, B1
S	*Shigella boydii* CDC 3083 94	4532	2347	A, B1
S	*Shigella dysenteriae*	4063	2236	A
E	*E. coli* O157:H7	5204	4489	B1
E	*E. coli* O157:H7 EDL933	5286	4570	B1
A	*E. coli* K12 substr W3110	4213	2534	S, E, B1
A	*E. coli* K12 substr MG1655	4140	2580	D, E, B1, B2
A	*E. coli* HS	4366	2983	D, S, E, B1, B2
A	*E. coli* C ATCC 8739	4434	3183	-
B1	*E. coli* E24377A	4729	3628	S, E, A, B2
B1	*E. coli* 55989	4953	3836	S, E, A, B2
B1	*E. coli* SE11	4684	3616	S, E, A, B2
B1	*E. coli* IAI1	4385	3512	D, S, E,A, B2
B2	*E. coli* 0127 H6 E2348 69	4809	3021	E, B1
B2	*E. coli* 536	4542	2702	E, B1
B2	*E. coli* CFT073	4897	2844	-
B2	*E. coli* ED1a	5012	2925	S, E
B2	*E. coli* UTI89	4827	2681	B1
B2	*E. coli* S88	4688	2709	E, B1
B2	*E. coli* APECO1	4878	2781	B1

Figure S1. MRP supertree of the ECS dataset [1], further annotated to show the two ways we group the 27 ECS genomes: by a biological criterion (into Groups D, S, E, A, B1 and B2) or by cutting the supertree on deep branches (into Groups 1, 2, 3 and 4).


Escherichia coli SMS 3 5
Escherichia coli IAI39
Escherichia coli 0127 H6 E2348 69
Escherichia coli 536
Escherichia coli CFT073
Escherichia coli ED1a
Escherichia coli UTI89
Escherichia coli S88
Escherichia coli APEC O1
Escherichia coli UMN026
Shigella dysenteriae
Escherichia coli O157H7
Escherichia coli O157H7 EDL933
Escherichia coli K 12 substr W3110
Escherichia coli K 12 substr MG1655
Escherichia coli HS
Escherichia coli C ATCC 8739
Shigella flexneri 5 8401
Shigella flexneri 2a
Shigella flexneri 2a 2457T
Shigella sonnei 046
Shigella boydii 227
Shigella boydii CDC 3083 94
Escherichia coli E24377A
Escherichia coli 55989
Escherichia coli SE11
Escherichia coli IAI1
Group 1
Group 2
Group 3
Group 4
D
S
E
A
B1
B2

[1] Skippington, E. & Ragan, M. A. Phylogeny rather than ecology or lifestyle biases the construction of *Escherichia coli-Shigella* genetic exchange communities. *Open Biology* **2**, 120112 (2012).

Section 2. Comparison between TF-IDF analysis of our BA (bacteria and archaea) dataset with grouping by class, with results reported by Popa *et al*. (2011) on a 657-genome bacteria and archaea dataset.

Popa and colleagues [1] analysed a dataset of 657 bacterial and archaeal genomes (sets of individual genes) using an approach that allows genes to be identified as lateral. In some cases (about 7% of the total) they can infer the direction of transfer. Lists of their genomes (Genome_list_dLGT.txt) and inferred lateral genes (dLGT-data.txt) are available as Supplemental Material [1]. Here we compare these lists with the corresponding results from our TF-IDF analysis of Dataset 3 (143 bacterial and archaeal genomes) analysed at class level, and report on overlaps.

POPA [1]:
Groups: 17 groups (mostly classes & phyla)
Genomes: 657
Genes: 2,129,548
Lateral genes: 52,621 (from dLGT-data.txt)
Unique lateral GI numbers: 41,392 (from dLGT-data.txt)
Unique lateral inter-generic GI numbers (donors and recipients): 5819
Genomes contributing these 5819 inter-generic lateral GIs: 317
Unique lateral inter-generic GI numbers (recipients only): 4700
Genomes contributing these 4700 recipient GIs: 277

OUR DATASET 3 (BACTERIAL & ARCHAEAL GENOMES):
Level-3 groups: 31 classes
Genomes: 143
Genes: 390,801
Unique GI numbers: 375,468
Lateral events: 3623 ($k = 25$, $G = 2k = 50$)
TF-IDF between-class lateral recipient GIs: 3043 (coverage threshold: see Main text, Table 9)
Genomes contributing these 3043 recipient GIs: 100

OVERLAP DATA:
Genomes in both POPA and BA: 40 (genus/species/strain descriptors identical or nearly identical)*
Group overlap: substantial overlap or similarity (see Tables S5 and S6)
GIs common to POPA 41,392 and BA 375,468: 4513 (POPA laterals in BA)
GIs common to POPA 41,392 and BA 3403: 81 (POPA laterals in BA laterals)
GIs common to POPA 5819 and BA 375,468: 800 (POPA inter-generic laterals in BA)
GIs common to POPA 5819 and BA 3043: 9 (POPA inter-generic laterals in BA laterals)

*Because Popa *et al*. do not report genome versions, precise (version) identity would have to be reverse-engineered from GI lists (which themselves are incomplete). However, the numbers of matched GI numbers suggest that at least half of the 40 genomes appearing in both the POPA and BA lists are represented by the same, or nearly identical, versions.

Thus when the two methods are applied to large, similarly diverse (bacteria plus archaea) datasets, they detect about the same density of unique "long-distance" lateral recipient GIs:

POPA: 4700 inter-generic recipients / 317 contributing genomes
TF-IDF: 3623 inter-class recipients / 100 contributing genomes

However, the POPA "long-distance" transfers are almost always within-class or even closer (*Escherichia* and *Shigella* are almost certainly the same genus). This is apparent from Figure 2A of Popa *et al*. [1], where very few clusters (connected components) encompass nodes of different colours. By contrast, all the "long-distance" transfers inferred using TF-IDF are necessarily (given the way we delineate groups in this case) between-class.

As illustrated by the nine inferred recipient genes above, the appearance of a gene in LGT lists from these two approaches does not imply that the same (or a compatible) lateral event has been inferred. A decade ago one of us [2] pointed out that different "surrogate" (non-phylogenetic) methods may agree less often than expected under a purely stochastic model. G+C-based methods preferentially identify relatively recent transfer events [3]. The method employed by Popa *et al*. [1] is a hybrid (G+C plus phylogenetic) method, but its initial screening step is based on G+C content.

We conclude that TF-IDF provides access to LGT events spanning broader phyletic distances than does the approach of Popa *et al*. [1].

REFERENCES FOR SECTION 2

[1] Popa O, Hazkani-Covo E, Landan G, Martin W & Dagan T. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res*. 21: 599-601 (2011).

[2] Ragan MA. On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Lett*. 201: 187-191 (2001).

[3] Ragan MA, Harlow TJ & Beiko RG (2006) Do different surrogate methods detect lateral genetic transfer events of different relative ages? *Trends Microbiol*. 14: 4-8 (2006).

Table S4. Nine GIs common to POPA 5819 and BA 3043 datasets (see above).

1
RECIPIENT 16272575 Haemophilus influenzae Rd KW20 (Gamma-proteobacteria)
GENE NP_438792.1 Elongation factor Tu
DONOR POPA 875 Actinobacillus_pleuropneumoniae_L20 (Gamma-proteobacteria)
DONOR TF-IDF Bacillus/clostridium

2
RECIPIENT 16762890 Salmonella enterica subsp. enterica serovar Typhi str. CT18 (Gamma-proteobacteria)
GENE NP_458507.1 B12-dep homocysteine-N5-methyltetrahydrofolate transmethylase
DONOR POPA 822 Klebsiella_pneumoniae_MGH_78578 (Gamma-proteobacteria)
DONOR TF-IDF Deinococcus

3
RECIPIENT 16763372 Salmonella enterica subsp. enterica serovar Typhi str. CT18 (Gamma-proteobacteria)
GENE NP_458989.1 ABC transporter ATP-binding protein
DONOR POPA 818 Citrobacter_koseri_ATCC_BAA-895 (Gamma-proteobacteria)
DONOR TF-IDF Alpha-proteobacteria

4
RECIPIENT 16767822 Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 (Gamma-proteobacteria)
GENE NP_463437.1 ABC transporter ATP-binding protein
DONOR POPA 818 Citrobacter_koseri_ATCC_BAA-895 (Gamma-proteobacteria)
DONOR TF-IDF Alpha-proteobacteria

5
RECIPIENT 27467230 Staphylococcus epidermidis ATCC 12228 (Low-GC Firmicutes)
GENE NP_763867.1 Elongation factor Tu
DONOR POPA 324 Enterococcus_faecalis_V583 (Low-GC Firmicutes)
DONOR TF-IDF Gamma-proteobacteria

6
RECIPIENT 30018378 Bacillus cereus ATCC 14579 (Low-GC Firmicutes)
GENE NP_830009.1 Elongation factor Tu
DONOR POPA 281 Lysinibacillus_sphaericus_C3_41 (Low-GC Firmicutes)
DONOR TF-IDF Gamma-proteobacteria

7
RECIPIENT 30260299 Bacillus anthracis str. Ames (Low-GC Firmicutes)
GENE NP_842676.1 Translation elongation factor Tu
DONOR POPA 281 Lysinibacillus_sphaericus_C3_41 (Low-GC Firmicutes)
DONOR TF-IDF Gamma-proteobacteria

8
RECIPIENT 33593471 Bordetella pertussis Tohama I (Beta-proteobacteria)
GENE NP_881115.1 Isocitrate dehydrogenase

DONOR POPA 57 Ralstonia_eutropha_H16 (Beta-proteobacteria)
DONOR TF-IDF Gamma-proteobacteria

9
RECIPIENT 56480411 Shigella flexneri 2a str. 301 (Gamma-proteobacteria)
GENE NP_709501.2 DNA gyrase subunit B
DONOR POPA 801 Escherichia_coli_O157H7 (Gamma-proteobacteria)
DONOR TF-IDF Beta-proteobacteria

(In addition, 56480411 appears twice as a donor in POPA, in each case donating to an Escherichia coli O157:H7 strain)

TABLE S5. Genomes in our BA data, grouped at Level 3 (by class).

Crenarchaeota:		
	1	Aeropyrum (1 genome)
	2	Sulfolobales (2 genomes)
	3	Thermoproteales (1 genome)
Euryarchaeota:		
	4	Archaeoglobales (1 genome)
	5	Halobacteriales (1 genome)
	6	Methanobacteriales (1 genome)
	7	Methanococcales (1 genome)
	8	Methanopyrales (1 genome)
	9	Methanosarcinales (2 genomes)
	10	Thermococcales (3 genomes)
	11	Thermoplasmales (2 genomes)
Nanoarchaeota:		
	12	Nanoarchaeum (1 genome)
Aquificales:		
	13	Aquificaceae (1 genome)
Bacteroidetes:		
	14	Bacteroidaceae (1 genome)
	15	Porphyromonadaceae (1 genome)
Chlamydiales:		
	16	Chlamydiaceae (7 genomes)
Chlorobi:		
	17	Chlorobiales (1 genome)
Cyanobacteria:		
	18	Chroococcales (4 genomes)
	19	Nostocales (1 genome)
	20	Prochlorophytes (3 genomes)
High G+C Firmicutes:		
	21	Actinomycetales (12 genomes)
Low G+C Firmicutes:		
	22	Bacillus/Clostridium group (34 genomes)
Planctomycetes:		
	23	Planctomycetales (1 genome)
Proteobacteria:		
	24	alpha subdivision (9 genomes)
	25	beta subdivision (8 genomes)
	26	epsilon subdivision (5 genomes)
	27	gamma subdivision (33 genomes)
Spirochaetales:		
	28	Leptospiraceae (1 genome)
	29	Spirochaetaceae (2 genomes)
Thermotogales:		
	30	Thermotoga (1 genome)
Thermus/Deinococcus group:		
	31	Deinococcus (1 genome)

TABLE S6. Genome numbers in groups from Popa *et al.* (2011), file “dLGT-data.txt”

59	Actinobacteria

89	Alphaproteobacteria
125	Bacilli
8	Bacteroidetes
71	Betaproteobacteria
6	Chlamydiae
4	Chlorobi
7	Chloroflexi
31	Clostridia
7	Crenarchaeota
29	Cyanobacteria
4	Deinococcus-Thermus
16	Deltaproteobacteria
15	Epsilonproteobacteria
16	Euryarchaeota
203	Gammaproteobacteria
12	Mollicutes
9	Spirochaetes
4	Thermotogae

TABLE S7. Strain names identical, or nearly identical, between the Popa *et al.* 657-genome list and ours. Close similarity of names does not guarantee identical assemblies or annotation versions (hence identical GI lists); conversely, in a few cases strain designations were changed or abbreviated, disguising potentially similar records. Our comparison of GIs was not pre-filtered through this name list, so full disambiguation of strain designators is not necessary for the purposes of this Supplementary analysis.

Bacillus_anthracis_Ames
Bacillus_cereus_ATCC_14579
Bacteroides_thetaiotaomicron_VPI-5482
Bordetella_bronchiseptica_RB50
Bordetella_parapertussis_12822
Bordetella_pertussis_TohamaI
Bradyrhizobium_japonicum_USDA110
Brucella_melitensis_16M
Brucella_suis_1330
Chlorobium_tepidum_TLS
Chromobacterium_violaceum_12472
Enterococcus_faecalis_V583
Escherichia_coli_CFT073
Escherichia_coli_O157:H7
Escherichia_coli_O157:H7_EDL933
Haemophilus_ducreyi_35000HP
Lactococcus_lactis_lactis
Mesorhizobium_loti_MAFF303099
Neisseria_meningitidis_MC58
Neisseria_meningitidis_Z2491
Nitrosomonas_europaea_ATCC_19718
Nostoc_sp._PCC_7120
Oceanobacillus_iheyensis_HTE831
Pasteurella_multocida_Pm70
Salmonella_typhimurium_LT2
Salmonella_enterica_Typhi_Ty2
Shigella_flexneri_2a_2457T
Sinorhizobium_meliloti_Rm1021
Staphylococcus_aureus_MW2
Staphylococcus_aureus_Mu50
Staphylococcus_aureus_N315
Streptococcus_agalactiae_2603V/R
Streptococcus_agalactiae_NEM316
Streptococcus_pneumoniae_R6
Streptococcus_pneumoniae_TIGR4
Vibrio_vulnificus_CMCP6
Vibrio_vulnificus_YJ016
Wigglesworthia_brevipalpis_Str.
Yersinia_pestis_CO92
Yersinia_pestis_KIM

Section 3. Index of sheet names and contents for Excel spreadsheets

Table S8. GO enrichment results on the ECS dataset (Supplementary Table S8: two sheets).

Sheet name	Description of results
Enrich_27_OVER	GO terms over-represented in six phyletic groups
Enrich_27_UNDER	GO terms under-represented in six phyletic groups

Table S9. GO enrichment results on the BA dataset (Supplementary Table S9: ten sheets).

Sheet name	Description of results
E_S_o	GO terms over-represented, *E. coli* and *Shigella* grouped separately
E_S_u	GO terms under-represented, *E. coli* and *Shigella* grouped separately
Ecoli_o	GO terms over-represented, *Shigella* genomes removed
Ecoli_u	GO terms under-represented, *Shigella* genomes removed
Shigella_o	GO terms over-represented, *E. coli* genomes removed
Shigella_u	GO terms under-represented, *E. coli* genomes removed
No_E_S_o	GO terms over-represented, *E. coli* and *Shigella* genomes removed
No_E_S_u	GO terms under-represented, *E. coli* and *Shigella* genomes removed
ES_combined_o	GO terms over-represented, *E. coli* and *Shigella* combined into one group
ES_combined_u	GO terms under-represented, *E. coli* and *Shigella* combined into one group

Table S10. Enrichment test results of BA dataset (Supplementary Table S10: three sheets).

Sheet name	Description of results
Enrich_143_L2_OVER	GO terms over-represented, genomes grouped by phylum
Enrich_143_L3_OVER	GO terms over-represented, genomes grouped by class
Enrich_143_L3_UNDER	GO terms under-represented, genomes grouped by class

Section 4. Supplementary material for Gene Ontology enrichment tests

Section 4.1 GO:0006414 translational elongation: 35 genes affected by LGT in BA dataset

2	ABC transporter ATP-binding protein
1	alanyl-tRNA ligase (synthetase)
2	leucyl-tRNA ligase (synthetase)
2	valyl-tRNA ligase (synthetase)
4	elongation factor G
21	elongation factor Tu
3	GTP-binding protein LepA

Rivera *et al*. [1] identified as “informational” genes as those functioning in translation (including tRNA synthetases), transcription and (DNA) replication, as well as homologs of vacuolar ATPases and GTPases. The following year the same group [2] included almost the same categories (omitting *replication*), and pointed to the translational and transcriptional complexes as examples. In the former they identified initiation, elongation (EF-Tu, EF-Ts, EF-G) and termination factors, ribosomal proteins, rRNAs, tRNAs and mRNAs, as well as “nongene products such as ions, small molecules such as GTP, GDP, etc., and membranes”.

These informational genes were considered less-susceptible to LGT [2]. This idea has persisted, although one subsequent study found no LGT bias between informational and operational genes [3], and another study found the bias limited to the “translation, ribosomal structure, and biogenesis” category once correction was made for connectivity bias [4]. Another study found translational genes to be the functional category for which within-bacteria LGT is the MOST frequent [5]. Functional category J (translation, ribosomal structure and biogenesis) shows strong net-like relationships among some although not all bacteria [6, Figure 4J].

Not all informational genes are “resistant” to LGT. Genes encoding aminoacyl-tRNA ligases (synthetases) are well-known to be susceptible to LGT [5,7,8], specifically including the three types we found: alanyl [9], leucyl [5,10,11] and valyl [5,7,12,13] ligases.

ABC transporter subunits, including the ATP-binding protein, are likewise well-known to be susceptible to LGT [14-17].

By contrast, the “core” elongation factors EF-G and EF-Tu are often considered resistant to LGT because the deeper branches of their phylogenetic trees agree with the corresponding 16S rRNA tree. However, recent studies have added quite a lot of nuance to this generalisation. Two large systematic studies of evolutionary dynamics in the EF-G protein family [18,19] identify four [19] or five [18] classes of EF-G paralogs, several of which have been affected by gene duplication, LGT and loss. Using somewhat different data, these authors identified about 14 instances of ancestral or more-recent LGT of EF-G paralogs involving α-, β- and γ-proteobacteria, actinomycetes, cyanobacteria and spirochaetes; other LGT events were considered possible. Some LGT events turned up in both studies [18,19]; others were supported by presence/absence of indels [19]. Using TF-IDF, we infer that EF-G genes in four genomes (of 143 in our BA dataset) have accepted LGT: one *Staphylococcus aureus*, one *Streptococcus*, one chlamydia (all transferred from Proteobacteria) and *Deinococcus* (transferred from high-G+C Firmicutes); see Table S4 below. The latter genome has been particularly accepting of LGT [5,20,21].

Using TF-IDF, we infer that 21 EF-Tu genes in 18 of these 143 genomes have been affected by LGT. In two γ-proteobacterial genomes (*Haemophilus influenzae* and *Vibrio parahaemolyticus*), both copies of EF-Tu are inferred to have accepted LGT from the Low-G+C Firmicutes, while both copies in *Deinococcus* have accepted LGT from Proteobacteria. In all, our 21 inferred LGT events involve transfer from the low-G+C Firmicutes into γ- (four) or ε-proteobacterial genomes (one), or from Proteobacteria into a low- (seven) or a high-G+C Firmicute, or a member of Chlamydia, Cyanobacteria, *Deinococcus* or Fusobacteria (Table S11). The involvement of Firmicutes in all but four of these 21 inferred events is notable, and recalls the results of Ke *et al*. [22] with 17 species of the low-G+C firmicute *Enterococcus*: in all 11 species with two copies of EF-Tu, the *tufB* copy had arisen laterally, likely in a single event, whereas there was no evidence of LGT in the six species in which EF-Tu is single-copy. Ke *et al*. [22] mention potentially similar situations in the low-G+C firmicute *Clostridium*, and in the high-G+C firmicute *Steptomyces*. We also note evidence for homologous recombination affecting the EF-Tu ortholog EF-1α in archaea [23], and for LGT being responsible for the distribution of the EF-1α-like factor EFL among eukaryotes [24].

Finally, our list also includes the highly conserved "fourth EF", leader peptidase A (LepA), which is present in bacteria and almost all eukaryotes, but absent from archaea. We are aware of only a single report examining its phylogeny [25]. No evidence was found for inter-kingdom LGT; taxon sampling within Bacteria was limited and bootstrap support modest (and/or not shown), but the topology of the LepA branch gives no reason to suspect the involvement of LGT.

References for Section 4.1

[1] Rivera, M.C., Jain, R., Moore, J.E. & Lake, J.A. Genomic evidence for two functionally distinct gene classes. *Proc. Natl Acad. Sci. USA* **95**, 6239-6244 (1998).

[2] Jain, R., Rivera, M.C. & Lake, J.A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl Acad. Sci. USA* **96**, 3801-3806 (1999).

[3] Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat. Genet*. **36**, 760-766 (2004).

[4] Cohen, O., Gophna, U. & Pupko, T. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol. Biol. Evol.* **28**, 1481-1489 (2011).

[5] Kanhere, A. & Vingron, M. Horizontal gene transfers in prokaryotes show differential preferences for metabolic and translational genes. *BMC Evol. Biol*. **9**, Art. 9 (2009).

[6] Puigbò, P., Wolf, Y.I. & Koonin, E.V. The tree and net components of prokaryote evolution. *Genome Biol. Evol*. **2**, 745-756 (2010).

[7] Wolf, Y.I., Aravind, L., Grishin, N.V. & Koonin, E.V. Evolution of aminoacyl-tRNA synthetases – analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* **9**, 689-710 (1999).

[8] Woese, C.R., Olsen, G.J., Ibba, M. & Söll, D. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev*. **64**, 202-236 (2000).

[9] Boussau, B., Guéguen, L. & Gouy, M. Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of Aquificales in the phylogeny of Bacteria. *BMC Evol. Biol*. **8**, Art. 272 (2008).

[10] Dohm, J.C., Vingron, M. & Staub, E. Horizontal gene transfer in aminoacyl-tRNA synthetases including leucine-specific subtypes. *J. Mol. Evol*. **63**, 437-447 (2006).

[11] Andam, C.P., Harlow, T.J., Papke, R.T. & Gogarten, J.P. Ancient origin of the divergent forms of leucyl-tRNA synthetases in the Halobacteriales. *BMC Evol. Biol*. **12**, Art. 85 (2012).

[12] Brown, J.R. Ancient horizontal gene transfer. *Nat. Rev. Genet*. **4**, 121-132 (2003).

[13] Adato, O., Ninyo, N., Gophna, U & Snit, S. Detecting horizontal gene transfer between closely related taxa. *PLoS Comp. Biol.* **11**, Art. e1004408 (2015).

[14] Nelson, K.E., *et al*. Evidence for horizontal gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323-329 (1999).

[15] De Ruggiero, J. *et al*. Evidence of recent lateral gene transfer among hyperthermophilic archaea. *Mol. Microbiol.* **38**, 684-693 (2000).

[16] Noll, K.M., Lapierre, P., Gogarten, J.P. & Nanavati, D.M. Evolution of *mal* ABC transporter operons in the Thermococcales and Thermotogales. *BMC Evol. Biol.* **8**, Art. 7 (2008).

[17] Meehan, C.J. & Beiko, R.G. Lateral gene transfer of an ABC transporter complex between major constituents of the human gut microbiome. *BMC Microbiol.* **12**, Art. 248 (2012).

[18] Atkinson, G.C. & Baldauf, S.L. Evolution of elongation factor G and the origins of mitochondrial and chloroplast forms. *Mol. Biol. Evol.* **28**, 1281-1292 (2011).

[19] Margus, T., Remm, M. & Tenson, T. A computational study of elongation factor G (EFG) duplicated genes: diverged nature underlying the innovation on the same structural template. *PLoS ONE* **6**, Art. e22789 (2011).

[20] Makarova, K.S. *et al*. Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. *Microbiol. Mol. Biol. Rev.* **65**, 44-79 (2001).

[21] Yuan, M. *et al*. Genome sequence and transcriptome analysis of the radioresistant bacterium *Deinococcus gobiensis*: insights into the extreme environmental adaptations. *PLoS ONE* **7**, Art. e34458 (2012).

[22] Ke *et al*. Evidence for horizontal gene transfer in evolution of elongation factor Tu in Enterococci. *J. Bacteriol.* **182**, 6913-6920 (2000).

[23] Inagaki, Y., Doolittle, W.F., Baldauf, S.L. & Roger, A.J. Lateral transfer of an EF-1α gene: origin and evolution of the large subunit of ATP sulfurylase in Eubacteria. *Proc. Natl Acad. Sci. USA* **103**, 4528-4533 (2006).

[24] Kamikawa, R., Inagako, Y. & Sako, Y. Direct phylogenetic evidence for lateral transfer of elongation factor-like gene. *Proc. Natl Acad. Sci. USA* **105**, 6965-6969 (2008).

[25] Qin, Y. *et al*. The highly conserved LepA is a ribosomal elongation factor that back-translocates the ribosome. *Cell* **127**, 721-733 (2006).

Table S11. Recipient genomes and donor groups for elongation factors inferred for our BA dataset.

EF-G
Staphylococcus aureus [Low-G+C Firmicutes] from Proteobacteria
Streptococcus [Low-G+C Firmicutes] from Proteobacteria
Chlamydophila [Chlamydiae] from Proteobacteria
Deinococcus from High-G+C Firmicutes

EF-Tu
Haemophilus [Gamma-proteobacteria] (two species) from Low-G+C Firmicutes
Pseudomonas [Gamma-proteobacteria] (two copies in one genome) from Low-G+C Firmicutes
Vibrio [Gamma-proteobacteria] (two copies in one genome) from Low-G+C Firmicutes
Yersinia [Gamma-proteobacteria] from Low-G+C Firmicutes
Campylobacter [Epsilon-proteobacteria] from Low-G+C Firmicutes
Bacillus [Low-G+C Firmicutes] (two species) from Proteobacteria
Listeria [Low-G+C Firmicutes] from Proteobacteria
Oceanobacillus [Low-G+C Firmicutes] from Proteobacteria
Staphylococcus [Low-G+C Firmicutes] (three strains in two species) from Proteobacteria
Bifidobacterium [High-G+C Firmicutes] from Proteobacteria
Prochlorococcus [Cyanobacteria] from Proteobacteria
Fusobacterium [Fusobacteria] from Proteobacteria
Chlamydia [Chlamydiae] from Proteobacteria
Deinococcus (two copies in one genome) from Proteobacteria

LepA
Bifidobacterium [High-G+C Firmicutes] from Thermus-Deinococcus
Lactococcus [Low-G+C Firmicutes] from Proteobacteria
Deinococcus from High-G+C Firmicutes

Section 4.2 GO:0006412 translation: 964 genes in Dataset 2a (EB dataset)

Ribosomal proteins = 583
Translation elongation factors = 40
Translation initiation factors = 40
Translation miscellaneous (probably non-core) = 6
tRNA synthetases / ligases = 130
Metabolic/operational & other non-core-translational = 165

See Section 4.1 (above) for an introduction to the "complexity hypothesis". Here we consider the over-enrichment of genes annotated with GO:0006412 in the EB dataset under three different groupings: 64 *E. coli* (EC) and *Shigella* (S) genomes combined into a single group (Dataset 2a), all ECS genomes removed, or only these S genomes removed. When these EC and S genomes are included but grouped separately, genes annotated with GO:0006412 translation become under-represented (see text).

The numbers below refer to TF-IDF inference on Dataset 2a (all ECS combined in a single group) at the default (mean value) IDF threshold. Of the 964 "translational" genes inferred as potentially lateral (Table S12), 130 encode tRNA synthetases/ligases, *i.e.* are well-known to be susceptible to LGT [1-3]. A further 165 encode metabolic/operational or other biological processes, and as such are not suspected of being LGT-resistant. Manual examination of a subset confirmed that the annotation is not in error, although the connection with translation can be indirect, *e.g.* as part of a specialised regulatory mechanism.

This leaves genes encoding 583 ribosomal proteins, 40 elongation factors, 40 initiation factors and six miscellaneous proteins.

Elongation factors have been discussed in Part 3.1 above. For the 40 EFs we identify as having accepted LGT, the breakdown is: EF-2 (6 instances), EF-G (22), EF-P and related (10), EF-Ts (1) and EF-Tu (5).

Jain *et al.* [4] included initiation factors 1, 2 and 3 among the translational apparatus components whose genes should be less-susceptible to LGT. For the 40 IFs we identify as having accepted LGT, the breakdown is: IF-1 (13), IF-2 (18) and IF-3 (9). We do not know of prior reports of LGT involving these genes.

Ribosomal proteins (r-proteins) comprise by far the greatest single component of this set. Collectively these sequences are considered to provide a conservative vertical central signal [4], and a subset of 16 r-proteins has been used to infer a three-Domain tree [5]. Individually, however, r-proteins are short, compositionally biased within and across Domains [4], and difficult to align [6]. Moreover, topologies of the inferred trees depend strongly on how the poorly alignable sites are treated [6]. For many although not all r-protein families, further complication is provided by multiple gene losses, restricted phyletic distributions and/or the presence of potentially subfunctionalised paralogs [7]. Yutin *et al.* [4] identified paralogs in 536 of 995 analysed bacterial genomes. All divergent paralogs contain a zinc-binding motif (zinc ribbon) [4,8], opening a path for (true or false) detection of LGT based on presence or absence of this conservative motif.

Keeping the above provisos in mind, LGT has previously been inferred for r-proteins S4 [8], S14 [9,10], S18 [7], L16 [10], L22 [10], L23 [6], L27 [11], L28 [7], L31 [7], L32 [7], L33 [7] and L36 [7]. Our list of r-proteins inferred by TF-IDF as affected by LGT includes most of these, and many others (S1 through S14, S16 through S21, L1 through L7, L9 through L11, L13 through L25, L25p, L27, L29, L30, L32 and L36).

Finally, in many bacteria and archaea, genes encoding many r-proteins occur adjacent to one another on the chromosome. Combined with the short length characteristic of r-protein genes, this means that an inferred lateral region that maps to the gene for one r-protein has an excellent chance of impinging on the gene for another. Given the overall weak gene-order conservation across and within many groups of Bacteria and Archaea, this "neighbour effect" is likely to affect r-proteins, hence the biological function *translation*, more than almost any other category.

References for Section 4.2

[1] Wolf, Y.I., Aravind, L., Grishin, N.V. & Koonin, E.V. Evolution of aminoacyl-tRNA synthetases – analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res* **9**, 689-710 (1999).

[2] Woese, C.R., Olsen, G.J., Ibba, M. & Söll, D. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev* **64**, 202-236 (2000).

[3] Kanhere, A. & Vingron, M. Horizontal gene transfers in prokaryotes show differential preferences for metabolic and translational genes. *BMC Evol Biol* **9**, Art. 9 (2009).

[4] Yutin, N., Puigbò, P., Koonin, E.V. & Wolf, Y.I. Phylogenomics of prokaryotic ribosomal proteins. *PLoS ONE* **7**, Art. e36972 (2012).

[5] Hug, L.A., *et al*. A new view of the tree of life. *Nat Microbiol* **1**, Art. 16048 (2016).

[6] Hansmann, S. & Martin, W. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. *In.t J Syst Evol Microbiol* **50**, 1655-1663 (2000).

[7] Makarova, K.S., Ponomarev, V.A. & Koonin, E.V. Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biol* **2**, Art. RESEARCH0033 (2001).

[8] Chen, K., Roberts, E. & Luthey-Schulten, Z. Horizontal gene transfer of zinc and non-zinc forms of bacterial ribosomal protein S4. *BMC Evol Biol* **9**, Art. 179 (2009).

[9] Brochier, C., Philippe, H. & Moreira, D. The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends Genet* **16**, 529-533 (2000).

[10] Coenye, T. & Vandamme, P. Organisation of the S10, spc and alpha ribosomal protein gene clusters in prokaryotic genomes. *FEMS Microbiol Lett* **242**, 117-126 (2005).

[11] Garcia-Vallvé *et al*. Simultaneous horizontal gene transfer of a gene coding for ribosomal protein L27 and operational genes in *Arthrobacter* sp. *J Mol Evol* **55**, 632-637 (2002).

Table S12. Details of gene types within GO:0006412 translation

Ribosomal proteins = 583
Ribosomal proteins (small subunit) = 289
Ribosomal proteins (large subunit) = 294

Metabolic/operational & other = 165
ABC transporter ATP-binding protein = 2
Acetyl-CoA carboxylase subunit beta = 6
Aconitase hydrases & hydratases = 32
Aspartate-ammonia ligase (asparagine synthetase A) = 1
ATP-binding component of transport system (putative) = 2
ATT-dependent RNA helicase DeaD = 13
Carbamoyl-phosphate synthase large chain / large subunit = 35
Cold-shock DEAD-box protein A = 2
DEAD/DEAH box helicase domain protein = 8
GTP-binding protein = 7
GTP-binding protein TypA (includes TypA/BipA) = 9
HF-I host factor = 1
Host factor I for bacteriophage Q beta replication = 2
Hypothetical proteins = 26
Integration host factor beta subunit = 1
N5-glutamine SAM-dependent methyltransferase = 1
Peptide deformylase = 1
Ribosome recycling factor = 1
RNA chaperone /binding protein Hfq = 9
RNA helicases = 4
Short-chain dehydrogenase/reductase SDR = 1
Sigma modulation protein (putative) = 1

Translation miscellaneous (probably non-core) = 6
Energy-dependent translational throttle protein EttA = 1
GTP-binding elongation factor family protein = 2
Peptide chain release factor 3 = 1
Peptidyl-tRNA hydrolase = 1
Release factor (putative) = 1

Translation elongation factors = 40
Translation elongation factor 2 = 2
Translation elongation factor G = 22
Translation elongation factor P = 9
Translation elongation factor P – (R) beta lysine ligase = 1
Translation elongation factor Ts = 1
Translation elongation factor Tu = 5

Translation initiation factors = 40
Translation initiation factor 1 = 13
Translation initiation factor 2 = 18
Translation initiation factor 3 = 9

tRNA synthetases / ligases = 130
Asparaginyl = 23
Glutaminyl = 9
Glycyl = 3
Isoleucyl = 1
Lysyl = 5
Methionyl = 4
Prolyl = 14
Threonyl = 2
Valyl = 69

Updates for Chapter 3

The results of LGT detection on a variant of the EB dataset (Dataset 2b, page 83) showed that lowering the IDF threshold leads to more lateral events being inferred. So the IDF threshold can be set as a tunable parameter of TF-IDF. However, a lower IDF threshold does not guarantee the quality of detections, *e.g.* false positives may be inferred.

Tuning the IDF threshold should be approached systematically, within an evaluation system in which the results generated at different threshold settings can be compared. In real applications, in which researchers need to detect LGT in empirical datasets, the quality of detections cannot be evaluated. So this much be approached by simulation. In addition, the TF component uses input from the IDF component, so the cooperation between the two must be taken into consideration. All these issues associated with performance tuning of the TF-IDF method will be discussed in future work.

In the meantime, the IDF threshold can adjusted by the user by modifying the function called in the code at a point corresponding to Line 20 of the pseudocode, replacing the current function $Avg(R)$.

Chapter 4 Finding genetic exchange communities

In the proceeding chapters, I showed that the total number of LGT detections using TF-IDF is affected by *k* and gap size *G*. Particularly for the BA (bacterial and archaeal genome) dataset, the inferred LGT networks are not uniform, as each contains densely and sparsely connected regions. The densely connected regions represent potential communities of organisms that, over time, have exchanged genetic material amongst themselves. In 2011 Skippington and Ragan proposed that such genetic exchange communities (GECs) could be recognised as structures in an LGT network, and mentioned paths, transitively closed sets, cliques and near-cliques (paracliques) as possible graphical structures.

In this chapter, I adopted *clique* (the most-densely connected of these structures) as the definition of a GEC, and inferred GECs using a data-driven approach. To do this I first built LGT networks from pairwise transfers inferred using TF-IDF; the nodes in these networks represent groups of sequences, and the edges represent inferred LGT. Recipient sequences were subsumed into the corresponding group, directionality of the edges was ignored, and incoming and outgoing edges connecting pairs of nodes were collapsed into a single edge. Then I used the software package GrAPPA (Graph Algorithms Pipeline for Pathway Analysis) to find maximum and maximal cliques in the network.

Analyses of these networks demonstrates that not all nodes within an LGT network are equally involved in GECs. Further, some nodes (which were always associated with high counts of LGT events) persisted when *k* was increased, and as such were considered *core nodes* in the corresponding GEC. I extracted the lateral genes underpinning these GECs, and carried out enrichment tests on their Gene Ontology terms. A wide range of biological processes are over-represented among these annotations, in particular those having to do with cellular metabolism, regulation and transport. This research extends our understanding of microbial communities that share genetic material via lateral processes, and of the biological processes affected by this sharing. The methodology itself exemplifies a highly scalable approach to the systematic study of GECs in diverse environments, including potentially the entire microbial biosphere.

This research chapter has been prepared as a manuscript. Supplementary material for this manuscript follows the main text.

Finding genetic exchange communities

Yingnan Cong[1], Yao-ban Chan[2] and Mark A. Ragan[1*]

[1] Institute for Molecular Bioscience and ARC Centre of Excellence in Bioinformatics, The University of Queensland, St Lucia, Brisbane, QLD 4072, Australia

[2] School of Mathematics and Statistics, The University of Melbourne, Parkville, Melbourne, VIC 3010, Australia

Email addresses:

YC: y.cong@uq.edu.au

YBC: yaoban@unimelb.edu.au

MAR[*]: m.ragan@uq.edu.au

Abstract

Bacteria and archaea can exchange genetic material across lineages through lateral genetic transfer (LGT). Collectively, these exchange relationships can be abstracted as a network and analysed using concepts from graph theory. In particular, densely connected regions within an LGT network have been defined as genetic exchange communities (GECs).

Here we apply the alignment-free method introduced in the previous chapters, TF-IDF, to study patterns of LGT among microbial genomes. We examine three empirical datasets (and selected variants of one of them) of different size (number of genomes) and phyletic breadth, varying *k* within limits identified earlier. We map the inferred lateral regions to genes in recipient genomes, and build networks in which the nodes are groups of genomes, and edges represent LGT. Finally we use the software package GrAPPA to find maximum and maximal cliques (*i.e.* GECs) in these graphs.

In a 27-genome dataset of *Escherichia coli* and *Shigella* genomes, we recognise one clique that encompasses all six groups based on MLST type. For a dataset of 110 genomes of enteric bacteria, the inferred LGT network is largely driven by exchange among *E. coli* and *Shigella* genomes. Removing genomes from either *E. coli* or *Shigella*, we find cliques that involve the remaining genus plus *Klebsiella* and *Salmonella*. *Yersinia* is mostly silent to LGT, and not involved in any GEC. For a dataset of 143 genomes across bacteria and archaea, network structure depends on *k*, and on whether we delineate groups by phylum or by class. Four phyla (or six classes) are core members, *i.e.* present in the cliques across all investigated *k*. No cliques join bacteria and archaea. GECs inferred for these three datasets contain most of the lateral genes, *i.e.* most surviving lateral transfer has happened within these exchange communities.

Gene Ontology enrichment tests demonstrate that biological processes associated with metabolism, regulation, and transport are often over-represented. This result is largely stable to change of *k*.

4.1 Introduction

Bacteria and archaea comprise much of the planet's biodiversity. Although individually inconspicuous, communities of these organisms are responsible for key biological and geochemical processes including nitrogen fixation, aerobic and anaerobic digestion of biomass, and oxidative dissolution of minerals. Bacteria also cause a range of diseases in plants, animals and humans. Since 1996, genome-sequencing technologies have been applied initially to study bacterial pathogenesis, and more-recently to understand environmental processes and explore biodiversity. Genome sequences are publicly available for more than 30,000 bacteria and archaea[1], and large international projects[2] are underway to sequence many thousands more.

Arguably, the two most-notable discoveries from the first two decades of microbial genomics have been the extent of strain-to-strain variation in gene content[3-5], and the prevalence of lateral genetic transfer (LGT). It has long been known that bacteria can take up genetic material from their surroundings, incorporate it into their main genome (or maintain it on extrachromosomal elements) and transmit it to subsequent generations. More than 35 years ago, unexpected patterns of gene presence among bacterial taxa and anomalous topologies of phylogenetic trees inferred for bacterial proteins were attributed, somewhat controversially, to LGT[6-9]. In the last 10-15 years, large-scale analysis has revealed the surprising extent of LGT among bacteria and archaea, with many estimates in the range 10-40% or greater[10]. Thus while all organisms transmit genetic information vertically from parent to offspring, bacteria and archaea simultaneously operate an orthogonal genetics that links (important components of) their genomes with viruses, phage, plasmids and free environmental DNA in a vast web[11-20].

We and others[13,15,16,21] have sought to abstract this web of genetic relationships as a network graph in which vertices represent observed entities that carry DNA (genomes, and in some applications also plasmids and phage), and edges the inferred transmission of genetic material between them. Precise definition of this mapping, however, turns out to be unexpectedly tricky. Two genomes that have descended only recently from a common ancestor are unlikely to differ greatly in sequence (through the accumulation of mutations) or gene content (through LGT and/or gene loss), and if they are to be accorded individual vertices, the edge between them will be dominated by vertical signal. To the extent that our network graph is intended to help us understand patterns of LGT, it makes sense to combine such genomes into a single vertex (node). As genomes diversify through time, it becomes increasingly useful to represent them as separate vertices, because doing so potentially increases the resolution at which LGT and gene loss can be inferred; but pairwise edges begin to represent a mixture of vertical and lateral signal. Moreover, older LGT (i.e. more-basal in the tree of vertical signal) becomes established in lineages and begins to be allocated among present-day genomes in hierarchical patterns that reinforce local vertical signal[22,23]. Thus by flattening the temporal (historical) dimension into the plane of the (present-day) network graph, we hide sequence diversity in the vertices and admix vertical and lateral signal in the edges. Although an optimal balance (or multiple locally optimal balances across the tree) can be sought, these issues remain.

Until now, the nature of the edges has received most attention. All-versus-all sequence comparison of genes[17,24-26] e.g. using BLAST[27] yields a similarity matrix based on all matching signals, both vertical and lateral. This can be skewed toward lateral signal by requiring best BLAST hits to be reciprocal[28,29], raising match thresholds[17] and/or subtracting the edges implied by a trusted reference tree[15,16]. A converse strategy was employed by Clarke et al.[30], who were interested only in the vertical component. A similar concept underlies phylogenetically based approaches: a test tree (typically inferred for a putatively orthologous gene or protein family) is compared with a reference (genome or organismal) tree, and instances of topological incongruence are considered *prima facie* cases of LGT[13,31-33]. However, reconstructing the pathway of inferred LGT as shortest edit paths is computationally hard, and may not yield a unique solution, or any solution at all[34]. Popa et al.[21] employ a hybrid approach in which only genes assessed as having regions of anomalous G+C content are input into phylogenetic discordance analysis.

Several objections have been raised to these approaches, both individually and collectively. We have repeatedly argued that as genes are not the actual units of LGT, gene families should not be the primary units of analysis[35,36]. Doolittle[37,38] has argued that by using a reference tree external to the analysis, we impose a higher standard of evidence on rejecting the reference topology (and thereby inferring LGT) than on accepting (or failing to reject) it, thereby according the vertical paradigm a methodologically unfair and theoretically unjustified advantage (but see[39]). A way is needed to infer LGT directly, positively and fairly in large genome-scale datasets.

In earlier chapters I introduced term frequency - inverse document frequency (TF-IDF) as an accurate, scalable approach to infer LGT among microbial genomes[40,41]. Using TF-IDF, edges represent lateral signal and can be inferred directly from whole genomes without first parsing them into individual genes. These edges are directional: transfers are inferred from a group of donor genomes to a single recipient genome. This may offer a partial solution to the issue of node heterogeneity; we return to this point below. No comparison with an external topology is required, although inference may be improved if the group structure reflects phylogeny (Chapter 3)[41].

Direct access to edges that represent only the lateral component of genetic relationships greatly simplifies interpretation of network graphs: these are natively LGT networks. In 2011, Skippington and Ragan[42] defined a genetic exchange community (GEC) as a densely connected region of an LGT network. Recognising the limitations of existing methods and data, these authors operationally defined a GEC "as a set of entities, each of which has over time both donated genetic material to, and received genetic material from, every other entity in that GEC, via a path of lateral transfer". These GECs do not exist *a priori* in nature, but rather are "actively fashioned (and continually refashioned) by the complex ongoing interplay among habitats, donors, vectors, recipients, mechanisms, sequences, population structures and selection"[42]. Biological problems that could be abstracted as involving dense edge sets in LGT graphs include the number, size, geospatial extent, taxonomic or habitat diversity of GECs in the microbial biosphere, and the role of vectors in mediating the exchange of pathogenicity, virulence or resistance factors among pathogens, primary hosts and secondary hosts[17,21,42,43].

Skippington and Ragan[42] further proposed that dense regions in LGT graphs might be described using concepts from graph theory, including cliques, paracliques (near-cliques), transitively closed sets, paths or walks, but were not in a position to recommend one of these over the others. My work in previous chapters makes it clear that edges, hence dense edge sets in LGT graphs and their biological interpretations, can be sensitive to the choice of TF-IDF parameters. Notably, precision and recall can be sensitive to the size of *k* (Chapter 2)[40], and edge number to the structure and delineation of groups (Chapter 3)[41]. The three empirical genome-scale datasets studied in Chapter 3 now provide a solid foundation for addressing these issues.

In this Chapter I examine (a) whether and how *k* affects these GECs; (b) whether there are core nodes (stable to variation of *k*) in different GECs; and (c) whether and how our biological process (functional) interpretation is consequently affected. More broadly, the approach pioneered in this Chapter will provide a framework for understanding the extent and biological significance of LGT in complex environments.

4.2 Methods

Since our TF-IDF method infers LGT from a donor group to a recipient sequence, the nodes (vertices) in our inferred networks are of two types: individual genomes when they are recipients of LGT, and groups of genomes when they are donors. Of course, members of a group may individually be (and often are) recipients. Edges are directional, so we depict them using an arrow from donor to recipient (Figure 4.1).

We aim to delineate GECs, introduced above as sets of nodes that have both donated genetic material to, and received genetic material from, each other. However, it is difficult to extract these relationships when nodes are of two types, individual sequences and groups. For this reason we take only groups as nodes in our network analysis. Member genomes are subsumed into their own-group identity, and all directed edges into those sequences are merged into a single directed edge from the donor group to the recipient group. The integer on each edge gives the total number of LGT events which are inferred in this way between donor and recipient groups.

As introduced above, GECs might best be described as *paths*, *transitively closed sets*, *paracliques* or *cliques*[42]. The first two structures fail to capture the density of connectivity, and many such structures of nearly equivalent size or value can often be found in relatively highly connected graphs such as the LGT networks we derive below. Paracliques differ from the corresponding cliques in lacking a defined number (or fraction) of edges; in the absence of theory or practice with respect to an application area (e.g. discovery of GECs), this number would have to be optimised for each dataset, requiring intense computation. Constrained by these considerations, here we use the strictest yet clearest definition of GEC, as *a clique containing groups (nodes) which share (donate or receive) genetic material from other nodes within this clique*. In this way we will focus on the most densely connected regions within an LGT network.

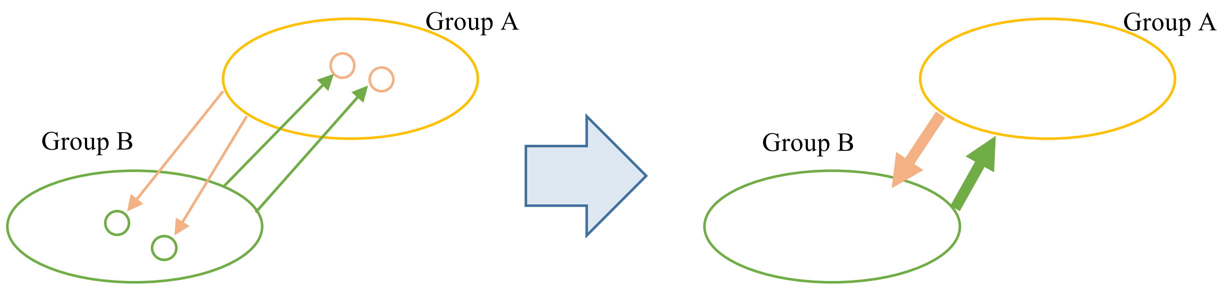


Figure 4.1 Merger of LGT relationships from group-to-sequence into group-to-group.

Discovery and analysis of GECs proceeds in four main steps:

(a) construct LGT networks based on the results of TF-IDF;

(b) collapse recipient genomes to recipient groups;

(c) find GECs from the LGT network; and

(d) carry out enrichment tests on biological processes underlying the GECs.

The datasets we used in this chapter are identical to those in Chapter 3: ECS (27 *Escherichia coli* and *Shigella* genomes), EB (110 enteric bacterial genomes) and BA (143 genomes from Bacteria and Archaea). From Chapter 2 and Chapter 3, we know that *k* can strongly affect the detection of LGT, hence potentially the topologies of LGT networks. For that reason, here we explore different values of *k* to test the stability of GEC topology. In step (a), we explore values of *k* from 20 to 40, with gap size *G* fixed at 2*k*. The step size is 10 for the ECS and EB datasets, while for the BA dataset (where LGT signal is much weaker than in the other two) we set step size as 5 for improved resolution against *k*. In step (b), we construct the LGT network based on groups which contain sequences involved lateral events.

To find GECs in a network (step (c)), we find cliques in the network. Finding cliques is known to be an NP-hard problem[44], although a parameterised complexity approach is available[45]. To approach this problem, we use the Graph Algorithms Pipeline for Pathway Analysis (GrAPPA) software package[46]. GrAPPA integrates multiple graph-theoretical tools including those designed to find maximum, maximal and para-cliques. The algorithm for maximal clique enumeration is an implementation of the Bron-Kerbosh algorithm[47]; here we set the minimum node threshold as three. For finding a maximum clique, GrAPPA employs a custom algorithm called Maximum Clique Finder (MCF)[48], a branch-and-bound algorithm of MCQ designed by Tomita and Kameda[49]. It implements tools to extract patterns (including cliques) efficiently from networks, but can deal only with undirected and unweighted networks. For this reason, we reformulate our directed networks as undirected networks (*i.e.* disregard the arrowhead), delete all weight annotations (number of inferred LGT events) on each edge, and merge edges between pairs of nodes which are both donors and recipients. Such reformulation does not make full use of the LGT information (*e.g.* directionality) provided by TF-IDF, but nonetheless preserves information sufficient for discovery of GECs as currently defined.

In Step (d), we find which biological processes are enriched in GECs. As we see below (Results), the GECs inferred for the ECS and EB datasets encompass almost all the respective lateral genes, so the biological process enrichments are essentially the same as described in Chapter 3. Thus here I report biological process enrichment only for cliques inferred, at different values of *k*, for the BA dataset. As before (Chapter 3), as a first step I map the inferred lateral regions to genes; thus if a gene is overlapped by at least one lateral segment of at least ten *k*-mers (with any intervening gaps up to $G = 2k$), the gene is considered lateral. These genes are extracted from the GenBank genome record using GI numbers and coordinates, and collected as a test set. All genes in each dataset form the reference set. The enrichment statistic is a Fisher's exact test, for which I set FDR = 0.05 as the threshold for selecting over- and under-represented Gene Ontology[50,51] terms. BA is the only dataset (among these three) for which LGT network topologies change greatly with *k*.

4.3 Results

Detailed results including LGT networks, maximum and maximal cliques, and gene lists are available as Supplementary Material (Supplementary Figures S1-S19, and Supplementary Tables S1-S31). Very large or detailed Supplementary Figures are also available for download in high resolution at http://bioinformatics.org.au/tools-data/ as "TF-IDF network diagrams.zip".

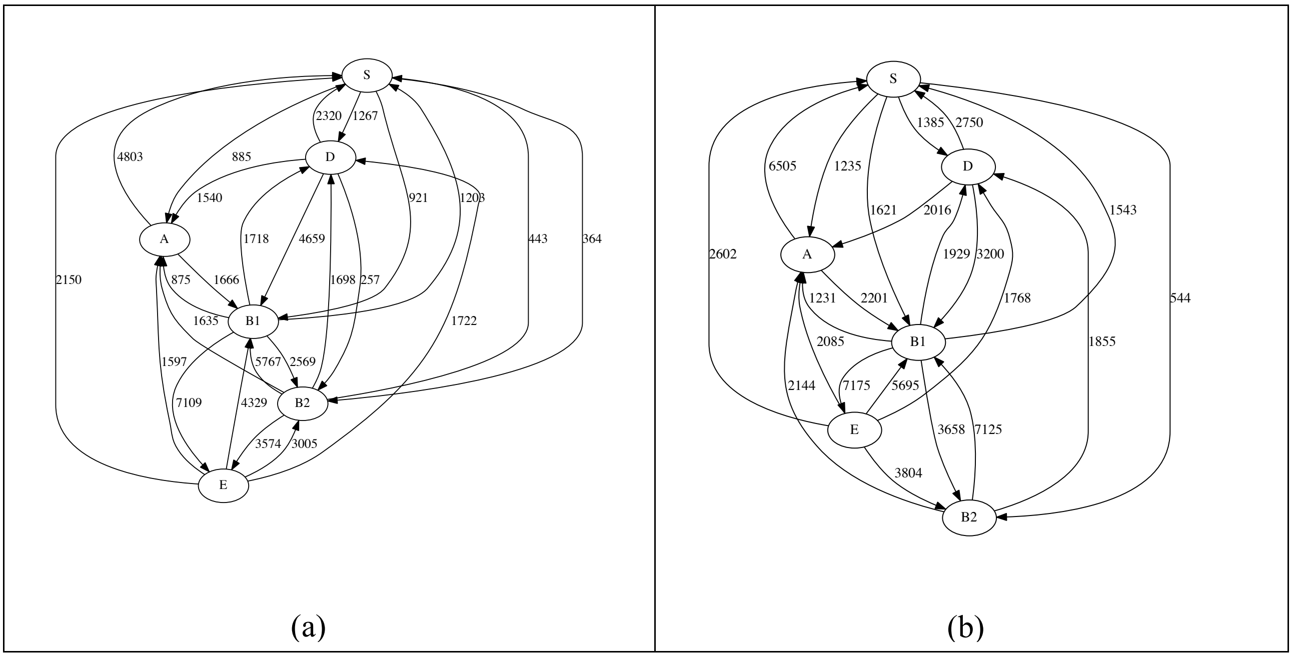
4.3.1 ECS dataset

We divide the ECS dataset (20 *Escherichia coli* and 7 *Shigella* genomes) into six groups according to multi-locus sequence type (MLST) [43,52]. In an earlier analysis of this dataset [43], lateral events identified by topological incongruence between trees inferred from putative orthogroups and an MRP [53] reference were shown to be biased more by phylogeny than by environment or lifestyle; concern was also expressed that defining GECs as cliques or paracliques might be too rigorous a standard. Here, we use our TF-IDF method to infer LGT networks (Figure 4.1). At $k = 30$ or 40, topologies of the two networks are identical (as before: [41]), although there are fewer detections on each edge at $k = 40$. At $k = 20$ we find three additional edges, from group D to group B2 (257 transfers), from B2 to S

(443 transfers) and from B2 to E (3574 transfers). For all *k* examined here, all six phyletic groups belong to a single clique. Summary information of the LGT networks can be found in Table 4.1.

Table 4.1 Numbers of lateral genes, LGT network edges and nodes inferred for the ECS dataset at *k* = 20, 30 or 40.

k	# of LGT events	# of edges	# of nodes in GECs
20	58076	25	6
30	64071	22	6
40	70849	22	6



(a)

(b)

Figure 4.2 LGT networks for ECS. (a) k = 20, and (b) k = 30. At k = 40, connectivity is the same as in (b), although values on the edges are larger (see Chapter 3, Figure 3.2).

Although the topology of the GECis stable for 20 ≤ *k* ≤ 40, the total number of lateral genes in each GEC increases with *k* (Figure 4.3). This increase might appear to contradict our earlier finding that when *k* increases, the total number of detections and detection length should remain the same or decrease (at *G* = 2*k*). However, when *k* is small, more short segments tend to be detected as lateral (Supplementary Table S1). For example, at *k* = 20, 26% of lateral segments are ≥ 500 nt in length, our threshold for selecting the segments for mapping to genes. This proportion increases to 31% at *k* = 40. Thus, we infer more lateral segments ≥ 500 nt at *k* = 40, which leads to more lateral genes being found.

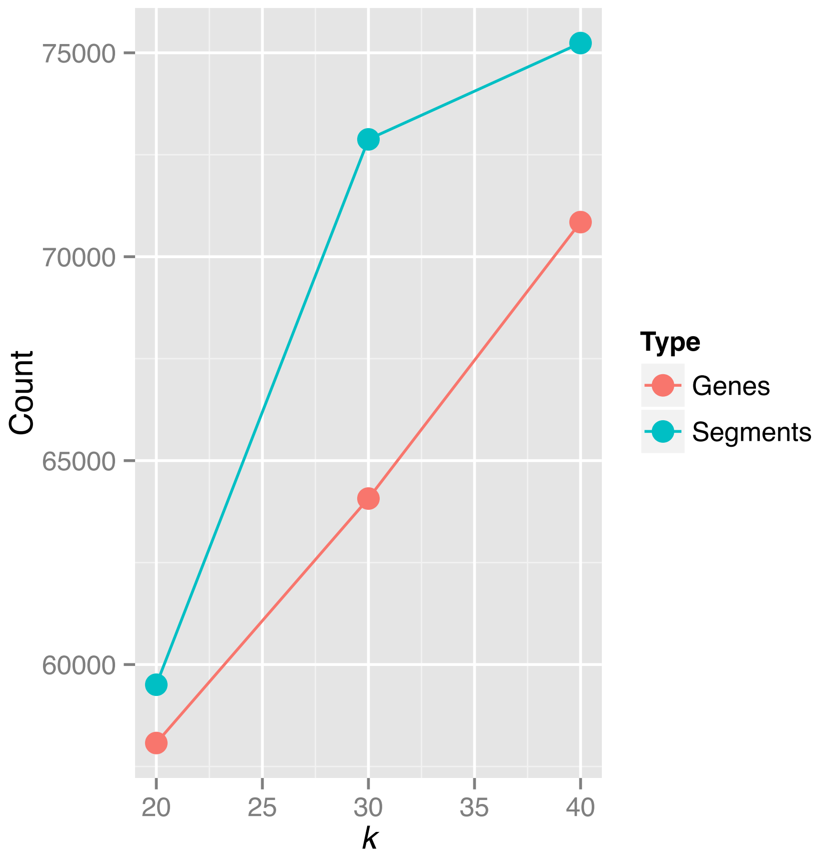


Figure 4.3 Total number of lateral segments of length ≥ 500 k-mers (i.e. not only those mapping to genes), and total number of lateral genes, within the (maximum) clique inferred for the ECS dataset, as a function of *k*.

4.3.2 EB dataset

The enteric bacteria dataset contains 110 genome sequences from five genera: *Escherichia*, *Shigella*, *Salmonella*, *Klebsiella* and *Yersinia*. Because the delineation of groups affects the detection results (Chapter 3), we infer LGT networks and extract GECs from five variants of this dataset: all genera present (referred to as EB-1); all genera except *Shigella* (EB-2) or alternatively, all except *Escherichia* (EB-3); with *Escherichia* and *Shigella* combined into a single group (EB-4); and with both *Escherichia* and *Shigella* removed (EB-5). These are five of the six variants examined earlier, in Chapter 3.

If we keep all 110 sequences and group them by genus (EB-1 dataset), the GEC topologies change as *k* steps from 20 to 30 to 40. At *k* = 20, *Escherichia*, *Shigella*, *Klebsiella* constitute a single GEC. At *k* = 30, we find two GECs, one consisting of *Escherichia* and *Shigella*, the other *Escherichia* and *Klebsiella*. At *k* = 40 only one GEC is found, consisting of *Escherichia* and *Shigella*. We infer many more LGT events between *Escherichia* and *Shigella* than between any other pair of genera. As *Escherichia* and *Shigella* are present in the GEC across the examined range of *k*, we can say that they are the *core nodes* of this GEC.

Because genomes from *Escherichia* and *Shigella* share many more identical *k*-mers than do other groups, the IDF values (elements of the *R* matrix) for these genomes are much higher than for the others (Chapter 2). This pushes up the IDF threshold, with the consequence that few lateral events are detected involving the other genera. To explore this effect, we removed the *Shigella* genomes from the dataset while retaining those from *Escherichia*, thereby eliminating the effect of *Shigella* (EB-2 dataset). We now infer additional LGT events between *Escherichia* and *Salmonella*, *Escherichia* and *Klebsiella*, and *Salmonella* and *Klebsiella.* All these node-pairs exchange genetic material bidirectionally, and we find a GEC composed of *Escherichia*, *Klebsiella* and *Salmonella* which remains stable as *k* increases from 20 to 40. We find the same when we instead retain *Shigella* sequences while removing those of *Escherichia* (EB-3 dataset). We also infer LGT events between *Klebsiella* and *Yersinia* in EB-3, but *Yersinia* cannot reach *Salmonella* and *Shigella*, so the GEC is restricted to *Shigella*, *Klebsiella* and *Salmonella*. In EB-4 we combine *Escherichia* and *Shigella* into a single group (ES) to test whether this grouping affects the topology of the GEC; more LGT events were inferred from *Salmonella* to ES, but membership of the GEC remained ES, *Salmonella* and *Klebsiella*. Lastly, to eliminate the effects of *Escherichia* and *Shigella* on LGT inference, we use only *Klebsiella*, *Salmonella* and *Yersinia* as input. At *k* = 20, the sole GEC contains all three genera, but at *k* = 30 or 40, the previous GEC is split into two GECs, one containing *Klebsiella* and *Salmonella* and the other *Klebsiella* and *Yersinia*. Details are provided in Table 4.2.

Table 4.2. Lateral genes and cliques inferred for variants of the EB dataset at *k* = 20, 30 or 40.

Datase t	*k* size	Nodes in clique	Number of lateral genes in cliques	Number of lateral genes in network	Proportion (%)
EB-1	20	*Escherichia, Shigella, Klebsiella*	29527	29527	100%
	30	*Escherichia, Shigella*	29258	29264	99.9%
	30	*Escherichia, Klebsiella*	6	29264	0.1%
	40	*Escherichia, Shigella*	16968	16968	100%

EB-2	20	*Escherichia, Klebsiella, Salmonella*	23964	23970	99.9%
	30	*Escherichia, Klebsiella, Salmonella*	10840	10840	100%
	40	*Escherichia, Klebsiella, Salmonella*	7420	7426	99.9%
EB-3	20	*Klebsiella, Salmonella, Shigella*	15290	15290	100%
	30	*Klebsiella, Salmonella, Shigella*	6473	6501	99.5%
	40	*Klebsiella, Salmonella, Shigella*	3869	3909	98.9%
EB-4	20	*ES, Klebsiella, Salmonella*	24806	24811	99.9%
	30	*ES, Klebsiella, Salmonella*	10762	10762	100%
	40	*ES, Klebsiella, Salmonella*	7951	7952	99.9%
EB-5	20	*Klebsiella, Salmonella, Yersinia*	6721	6721	100%
	30	*Klebsiella, Yersinia*	123	2586	4.8%
	30	*Klebsiella, Salmonella*	2463	2586	95.2%
	40	*Klebsiella, Yersinia*	140	1559	9%
	40	*Klebsiella, Salmonella*	1419	1559	91%

4.3.3 BA dataset

The BA dataset consists of 143 genome sequences across Bacteria and Archaea. We group these genomes into fifteen phyla or, alternatively, into 31 classes. With more nodes than in the two previous datasets, there is potential for inferred LGT networks to be more complex. On the other hand, sequences are more dissimilar across the BA dataset, so fewer *k*-mers are shared and fewer instances of LGT are inferred. These detections are also greatly affected by change of *k* (Chapter 3).

When groups are delineated by phylum, the number of total LGT detections decreases significantly as k increases. The decrease is sharp, which causes edges in the LGT network to vanish and the GECs to shrink. At $k = 20$ six GECs are found, each with five phyla. Five of these GECs contain the High G+C Firmicutes, Proteobacteria and Low G+C Firmicutes, which together contain 14797 lateral genes, 95.5% of the total inferred over the entire network. Thus these phyla form the core of inter-phylum GEC. We also observe a community of Nanoarchaeota, Euryarchaeota and Crenarchaeota; although only ten LGT events are inferred between these nodes, it is notable for showing potential GECs among Archaea. In addition, the Thermus/Deinococcus phylum contributes 244 lateral events, 1.5% of the total; as our dataset contains only one strain in this phylum, this particular genome appears to be more LGT-active than many other bacterial genomes.

The number of detections drops sharply at $k > 20$, the total detections continues to fall with increasing k, to 34 at $k = 40$. Recall that the simulation studies in Chapter 2 indicate potential false positives at $k \leq 20$, presumably due to identical k-mers shared between sequences and groups simply by coincidence. As LGT detections decrease in number, some edges in the LGT network vanish but the core component of the GECs – the High G+C Firmicutes, Low G+C Firmicutes and Proteobacteria – persists unchanged. Thermus/Deinococcus also remains active in sharing LGT with Proteobacteria for all investigated k.

When the BA genomes are alternatively grouped by class, the LGT networks are more complex. All but one of the 31 classes are involved in LGT at $k = 20$, with 30696 genes inferred to have received one or more lateral segments. These genes are distributed over 169 edges, mostly (77.7%) in the Actinomycetales (5377 genes), Bacillus/Clostridium (2277), Alpha-proteobacteria (5944), Beta-proteobacteria (7322) and Gamma-proteobacteria (8596). We observe 23 maximal cliques (≥ 3 nodes) in this network. The Actinomycetales, Bacillus/Clostridium, Alpha-, Beta- and Gamma-proteobacteria all form part of the core, each being present in 17 cliques (≥ 5 nodes) and in the maximum clique (largest GEC).

Since the sequences within BA are relatively dissimilar from each other, many fewer k-mers are shared between sequences than in other two datasets. The LGT detections are very sensitive to k size: when k is increased from 20 to 30 and 40, the numbers of LGT events drop sharply from 30696 to 1239 and 373 respectively. Gamma-proteobacteria, Beta-proteobacteria, Actinomycetales and Bacillus/Clostridium are hubs and play key roles in most GECs at $k = 30$; at $k = 40$, fewer LGT are inferred and only the Beta- and Gamma-proteobacteria remain as the core of GECs.

Deinococcus is active in exchanging genetic material with Beta- and Gamma-proteobacteria for $20 \leq k \leq 40$. Other lateral events are also inferred between Deinococcus and Actinomycetales, and between Deinococcus and Chlorococcales at $k < 40$.

4.3.4 Enrichment test on biological processes within GECs

In addition to the topologies, we are also interested in what biological processes are enriched in GECs. This can help us better understand how the biological processes are represented in genetic exchange, and contribute to construction of GECs.

From analyses of LGT networks in the ECS and EB datasets, we find that more than 90% of the inferred lateral genes in each dataset are represented in the corresponding GEC. In the ECS dataset, all vertices are in the GEC. There is no need to carry out enrichment tests on such GECs: biological processes contributing to GEC formation will be indistinguishable from those of the whole networks to which the GECs belong, hence from the total LGT edge sets (Chapter 3).

For the BA dataset, however, GEC topologies change significantly with k. Few LGT events are detected at $k > 30$, particularly when sequences are grouped by phylum (Table 4.3). For optimal comparison, enrichment tests were applied on lateral genes of maximum cliques in each LGT network at $20 \leq k \leq 30$ with genomes grouped by phylum and by class.

Table 4.3 Number of LGT genes in the BA dataset at $20 \leq k \leq 40$, analysed at the level of phylum or class.

k size	# LGT detected at phylum level	# LGT detected at class level
20	12880	28628
25	686	3043
30	167	1076
35	69	576
40	0	310

The enrichment tests identify biological processes related to metabolism, transport and regulation as over-represented when sequences are grouped by phylum at $20 \leq k \leq 30$. The term *translational elongation* (GO:0006414) ranks in first position at $k = 20$, and seventh at $k = 25$, among over-represented terms. The most significantly under-represented biological processes relate to transposition and to RNA modification at k = 20. At k = 25, the under-represented terms relate to RNA processing and biosynthetic process. The only term under-represented at k = 30 describes the modification of macromolecules.

When the genomes are grouped by class, the main categories of GO terms significantly over-represented remain those describing metabolism, transport and regulation. Terms most under-represent terms relate to transposition, RNA metabolisms and regulation at k = 20 and 25; at k = 30, processes of protein modification are found to be under-represented.

In general, the patterns of over-representation are similar between analyses at phylum and class levels. Interestingly, *translation elongation* is significantly over-represented at phylum level, but much less so at class level. *Transposition* (GO:0032196) is significantly under-represented in most cases.

4.4 Discussion and conclusion

Here I inferred LGT networks for three datasets of different phyletic breadth (hence evolutionary depth). For the ECS dataset, the entire LGT network is inferred to form a single GEC encompassing all nodes, consistent with previous research[43] There is, however, interplay with the IDF threshold, as seen in the EB dataset and its variants. For the full EB dataset (EB-1), the LGT signal between *Escherichia* and *Shigella* is much stronger than that of any other pairwise comparison and dominates the lateral signal, with the result that the only community that can be found is *Escherichia* and *Shigella*. If we remove *Escherichia* or (alternatively) *Shigella*, or combine them into a single group, we now detect LGT events from (and/or to) *Klebsiella* and *Salmonella*. This reveals a wider GEC containing either *Escherichia* or *Shigella*, plus *Klebsiella* and *Salmonella* (Supplementary Figures S2-5). By contrast, *Yersinia* is relatively silent to LGT, and contributes little to the community.

Particularly in the BA dataset, we see that some parts of the LGT network are sensitive to change of k. As discussed elsewhere, when k is small (here k = 20) many k-mers are shared by chance, resulting in many false positive inferences. Edges supported by large numbers of lateral events (e.g. with high weights) will tend to persist, whereas those representing smaller numbers of events may disappear as k increases. Even so, when the sequences are grouped by phylum, the High-G+C Firmicutes, Low-G+C Firmicutes and Proteobacteria are found in all GECs inferred across the investigated range of parameter values (Supplementary Figures S6-10, Table S2-11). For this reason we identify them as core nodes of the GEC for the BA phyla. Although it does not contribute many LGT events, Thermus/Deinococcus is also a member of most communities.

When the BA dataset is grouped into 31 classes, many more clique structures are found. The Alpha-, Beta-, Gamma-subdivisions of Proteobacteria, the Actinomycetales and Bacillus/Clostridium are always present in (at least one) GEC (Supplementary Figures S11-15, Tables S12-S21), i.e. are core nodes. By contrast, the Epsilon-proteobacteria appear relatively silent to LGT, with fewer inferred events per genome (Supplementary Table S22). In the class-level LGT network, the sole *Deinococcus* genome is also involved in many GECs, linked through a lateral edge with subdivisions from Proteobacteria. More LGT events might be expected if more sequences from Deinococci and its immediate relatives were included in this dataset.

Although many fewer instances of LGT are inferred involving archaea, we nonetheless recognise one GEC among them. The low frequency of inferred LGT events may arise because these genomes are relatively diverse in gene content and phylogenetically distant from each other, and/or because in reality these genomes have exchanged little genetic material[13,21]. In the former case TF-IDF should find instances of LGT but the pairwise values may not pass the IDF threshold, whereas in the latter case there would be little true-positive LGT to be found and lowering the IDF threshold would lead only to false-positive inferences. Comparing the results of TF-IDF with those of classical alignment-based methods may help distinguish between these alternative explanations.

Enrichment tests on the BA data reveal that a wide range of biological processes are over-represented in the LGT events that underpin the GECs identified. As expected[54,55], metabolic processes, gene regulation, and trans-membrane and intracellular transport are broadly represented. For example, with genomes grouped by class and k = 25, 30 of the 50 most over-represented processes describe metabolism. Terms associated with transposition or antibiotic resistance are not seen: these genes are usually transferred within-phylum or within-class (or indeed more narrowly), and often occur on plasmids, which are not represented in the genome files we used. As expected, few terms describing processes of transcription, translation or DNA replication[54,55] are overrepresented.

Fewer biological process terms are under-represented among the LGT events that underpin the BA GECs. The term *transposition* (GO:0032196) is very significantly under-represented. A similar result was also found in the ECS dataset (Chapter 4 Supplementary Table S4). From previous research, we know that genes annotated with this term are widespread in the ECS dataset, making it difficult for genes annotated with this term to pass the TF threshold for detection. In the BA dataset, genomes of *E. coli* and *Shigella* are a major source of genes associated with transposition; as these are members of the same group (Gamma-proteobacteria), they are not detected by TF-IDF. In the EB dataset, when *Escherichia* and *Shigella* are not treated as separate groups, *transposition* is not significantly under-represented (Chapter 4 Supplementary. Table S5). Thus TF-IDF is not blind to such mobile biological processes, but the way groups are delimited limits their discovery.

This work represents the first systematic exploration of the sensitivity of GECs to choice of parameter values in an alignment-free framework. The results confirm the promise of this approach, but important challenges remain.

As the network analysis software package GrAPPA[46] can process only undirected and unweighed networks, for the analyses reported in this Chapter all edges were rendered undirected (with merger of incoming and outgoing edges) and weights were ignored. These steps represent a classic pre-processing on a directed network[56]. Other strategies have been introduced to find cliques in directed network; all of them involve weakening he edges, and do not guarantee a better interpretation of properties of the original directed network[57,58]. Comparing these strategies is a problem in graph theory beyond the scope of this project. The current operation has made it difficult to extract biological information from the graphs, e.g. directionality and relative magnitudes of inferred lateral exchange across environments. Nonetheless, some features of the role played by LGT in the evolution of microbes are still accessible, e.g. the frequent exchange among *Escherichia* and *Shigella* contrasted with the relative isolation of *Yersinia*.

Here we have defined GECs as cliques, a rigorous graph structure that maps particularly well onto biological concepts of the sharing of genetic information via LGT. In introducing the mapping of GECs onto LGT graphs, Skippington & Ragan[42] expressed concern that missing data might make clique too rigorous a definition; we share their concern, but find no criterion by which to select (much less optimise) the number or proportion of "missing" edges (paraclique), while for the analyses presented here, maximal paths would be numerous yet not particularly informative biologically.

4.5 References

1 International Nucleotide Sequence Database Collaboration <http://www.insdc.org/> (accessed 30 June 2016).

2 Joint Genome Institute <http://jgi.doe.gov/our-science/science-programs/microbial-genomics/phylogenetic-diversity/> (accessed 29 June 2016).

3 Segerman, B. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. *Front Cell Infect Microbiol* **2**, 116, doi:10.3389/fcimb.2012.00116 (2012).

4 Croucher, N. J., Coupland, P. G., Stevenson, A. E., Callendrello, A., Bentley, S. D. & Hanage, W. P. Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat Commun* **5**, 5471, doi:10.1038/ncomms6471 (2014).

5 Tettelin, H., Masignani, V., Cieslewiz, M. J., *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102**, 13950-13955, doi:10.1073/pnas.0506758102 (2005).

6 Ambler, R. P., Meyer, T. E. & Kamen, M. D. Anomalies in amino acid sequences of small cytochromes *c* and cytochromes *c*' from two species of purple photosynthetic bacteria. *Nature* **278**, 661-662 (1979).

7 Ambler, R. P. *et al.* Cytochrome c2 sequence variation among the recognised species of purple nonsulphur photosynthetic bacteria. *Nature* **278**, 659-660 (1979).

8 Dickerson, R. E. Evolution and gene transfer in purple photosynthetic bacteria. *Nature* **283**, 210-212 (1980).

9 Woese, C. R., Gibson, J. & Fox, G. E. Do genealogical patterns in purple photosynthetic bacteria reflect interspecific gene transfer? *Nature* **283**, 212-214 (1980).

10 Ragan, M. A. & Beiko, R. G. Lateral genetic transfer: open issues. *Philos Trans R Soc Lond B Biol Sci* **364**, 2241-2251, doi:10.1098/rstb.2009.0031 (2009).

11 Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284**, 2124-2129 (1999).

12 Bryant, D. & Moulton, V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* **21**, 255-265, doi:10.1093/molbev/msh018 (2004).

13 Beiko, R. G., Harlow, T. J. & Ragan, M. A. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A* **102**, 14332-14337, doi:10.1073/pnas.0504068102 (2005).

14 Kunin, V., Goldovsky, L., Darzentas, N. & Ouzounis, C. A. The net of life: reconstructing the microbial phylogenetic network. *Genome Res* **15**, 954-959 (2005).

15 Dagan, T., Artzy-Randrup, Y. & Martin, W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci U S A* **105**, 10039-10044, doi:10.1073/pnas.0800679105 (2008).

16 Dagan, T. & Martin, W. Getting a better picture of microbial evolution en route to a network of genomes. *Philos Trans R Soc Lond Ser B Biol Sci* **364**, 2187-2196, doi:10.1098/ rstb.2009.0040 (2009).

17 Halary, S., Leigh, J. W., Cheaib, B., Lopez, P. & Bapteste, E. Network analyses structure genetic diversity in independent genetic worlds. *Proc Natl Acad Sci U S A* **107**, 127-132, doi:10.1073/pnas.0908978107 (2010).

18 Puigbò, P., Wolf, Y. I. & Koonin, E. V. The tree and net components of prokaryote evolution. *Genome Biol Evol* **2**, 745-756, doi:10.1093/gbe/evq062 (2010).

19 Bapteste, E., van Iersel, L., Janke, A., Kelchner, S., Kelk, S., McInerney, J.O., Morrison, D. A., Nakhleh, L, Steel, M., Stougie, L. & Whitfield, J. Networks: expanding evolutionary thinking. *Trends Genet* **29**, 439-441, doi:10.1016/j.tig.2013.05.007 (2013).

20 Koonin, E. V. The turbulent network dynamics of microbial evolution and the statistical Tree of Life. *J Mol Evol* **80**, 244-250, doi:10.1007/s00239-015-9679-7 (2015).

21 Popa, O., Hazkani-Covo, E., Landan, G., Martin, W. & Dagan, T. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res* **21**, 599-609, doi:10.1101/gr.115592.110 (2011).

22 Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**, 2226-2238 (2002).

23 Gogarten, J. P. & Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* **3**, 679-687, doi:10.1038/nrmicro1204 (2005).

24 Bansal, A. K., Bork, P. & Stuckey, P. J. Automated pair-wise comparisons of microbial genomes. *Math Model Sci Comput* **9**, 1-23 (1998).

25 Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol Biol Evol* **25**, 762-777, doi:10.1093/molbev/msn023 (2008).

26 Fondi, M. & Fani, R. The horizontal flow of the plasmid resistome: clues from inter-generic similarity networks. *Environ Microbiol* **12**, 3228-3242, doi:10.1111/j.1462-2920.2010.02295.x (2010).

27 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

28 Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).

29 Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. & Yuan, Y. Predicting function: from genes to genomes and back. *J Mol Biol* **283**, 707-725, doi:10.1006/jmbi.1998.2144 (1998).

30 Clarke, G. D., Beiko, R. G., Ragan, M. A. & Charlebois, R. L. 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**, 2072-2080 (2002).

31 Goldman, N., Anderson, J. P. & Rodrigo, A. G. Likelihood-based tests of topologies in phylogenetics. *Syst Biol* **49**, 652-670 (2000).

32 Beiko, R. G. & Ragan, M. A. Detecting lateral genetic transfer: a phylogenetic approach. *Meth Mol Biol* **452**, 457-469, doi:10.1007/978-1-60327-159-2_21 (2008).

33 Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F. & Papke, R. T. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res* **16**, 1099-1108, doi:10.1101/gr.5322306 (2006).

34 Beiko, R. G. & Hamilton, N. Phylogenetic identification of lateral genetic transfer events. *BMC Evol Biol* **6**, 15, doi:10.1186/1471-2148-6-15 (2006).

35 Chan, C. X., Darling, A. E., Beiko, R. G. & Ragan, M. A. Are protein domains modules of lateral genetic transfer? *PLoS ONE* **4**, e4524, doi:10.1371/journal.pone.0004524 (2009).

36 Chan, C. X., Beiko, R. G., Darling, A. E. & Ragan, M. A. Lateral transfer of genes and gene fragments in prokaryotes. *Genome Biol Evol* **1**, 429-438, doi:10.1093/gbe/evp044 (2009).

37 Doolittle, W. F. & Bapteste, E. Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl Acad. Sci. U S A* **104**, 2043-2049, doi:10.1073/pnas.0610699104 (2007).

38 Doolittle, W. F. The practice of classification and the theory of evolution, and what the demise of Charles Darwin's tree of life hypothesis means for both of them. *Philos Trans R Soc Lond. B Biol Sci* **364**, 2221-2228, doi:10.1098/rstb.2009.0032 (2009).

39 O'Malley, M. A. & Koonin, E. V. How stands the Tree of Life a century and a half after The Origin? *Biol Direct* **6**, 32, doi:10.1186/1745-6150-6-32 (2011).

40 Cong, Y., Chan, Y. & Ragan, M. A. A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF. *Sci Rep* **6**, 30308 (2016).

41 Cong, Y., Chan, Y. & Ragan, M. A. Exploring lateral genetic transfer among microbial genomes using TF-IDF. *Sci Rep* **6**, 29319 (2016).

42 Skippington, E. & Ragan, M. A. Lateral genetic transfer and the construction of genetic exchange communities. *FEMS Microbiol Rev* **35**, 707-735, doi:10.1111/j.1574-6976.2010.00261.x (2011).

43 Skippington, E. & Ragan, M. A. Phylogeny rather than ecology or lifestyle biases the construction of Escherichia coli-Shigella genetic exchange communities. *Open Biol* **2**, doi:Artn 120112 10.1098/Rsob.120112 (2012).

44 Knuth, D. E. Postscript about NP-hard problems. *ACM SIGACT News* **6**, 15-16 (1974).

45 Downey, R. G., Fellows, M. R. & Stege, U. in *Contemporary Trends in Discrete Mathematics: From DIMACS and DIMATIA to the Future,* 49-99.

46 Lu, Y., Phillips, C. & Langston, M. A. <https://grappa.eecs.utk.edu/> (accessed on 30 May 2016).

47 Bron, C. & Kerbosch, J. Algorithm 457: finding all cliques of an undirected graph. *Commun ACM* **16**, 575-577 (1973).

48 Eblen, J. D. The Maximum Clique Problem: Algorithms, Applications, and Implementations. (2010).

49 Tomita, E. & Kameda, T. An efficient branch-and-bound algorithm for finding a maximum clique with computational experiments. *J Global Optimiz* **37**, 95-111 (2007).

50 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).

51 Consortium, G. O. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* **32**, D258-D261 (2004).

52 Gordon, D. M., Clermont, O., Tolley, H. & Denamur, E. Assigning Escherichia coli strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environ Microbiol* **10**, 2484-2496, doi:10.1111/j.1462-2920.2008.01669.x (2008).

53 Ragan, M. A. Phylogenetic inference based on matrix representation of trees. *Mol Phylogenet Evol* **1**, 53-58 (1992).

54 Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: The complexity hypothesis. *Proc. Natl Acad. Sci. USA* **96**, 3801-3806, doi:Doi 10.1073/Pnas.96.7.3801 (1999).

55 Jain, R., Rivera, M. C., Moore, J. E. & Lake, J. A. Horizontal gene transfer accelerates genome innovation and evolution. *Mol Biol Evol* **20**, 1598-1602, doi:10.1093/molbev/msg154 (2003).

56 Seidman, S. B. & Foster, B. L. A graph-theoretic generalization of the clique concept. *Journal of Mathematical sociology* **6**, 139-154 (1978).

57 Palla, G., Farkas, I. J., Pollner, P., Derenyi, I. & Vicsek, T. Directed network modules. *New journal of physics* **9**, 186 (2007).

58 Seidman, S. B. Clique-like structures in directed networks. *Journal of Social and Biological Structures* **3**, 43-54 (1980).

4.6 Acknowledgements

This research utilised resources of QCIF (Queensland Cyber Infrastructure Foundation), which is supported by the Queensland and Australian Governments. We thank Mr Brett Dunsmore and staff of Information Technology Services, Institute for Molecular Bioscience for additional computational support.

YC acknowledges the China Scholarship Council and The University of Queensland for stipend and tuition fee support. The project was funded in part by a grant from the James S. McDonnell Foundation to MAR. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Supplementary Material

Finding genetic exchange communities

Yingnan Cong
Yao-ban Chan
Mark A. Ragan

Overview of Supplementary Material:

Section 1. Supplementary material for EB and BA datasets. LGT networks inferred for different datasets at different values of *k*.

Supplementary Figures S1-S15

Section 2. Distribution of lengths of lateral segments inferred for the ECS dataset, at different values of *k*.

Supplementary Table S1

Section 3. Cliques in the BA dataset and its variants, with different values of *k* and grouping.

Supplementary Tables S2-S21

Section 4. Connectivity of classes within Proteobacteria.

Supplementary Table S22

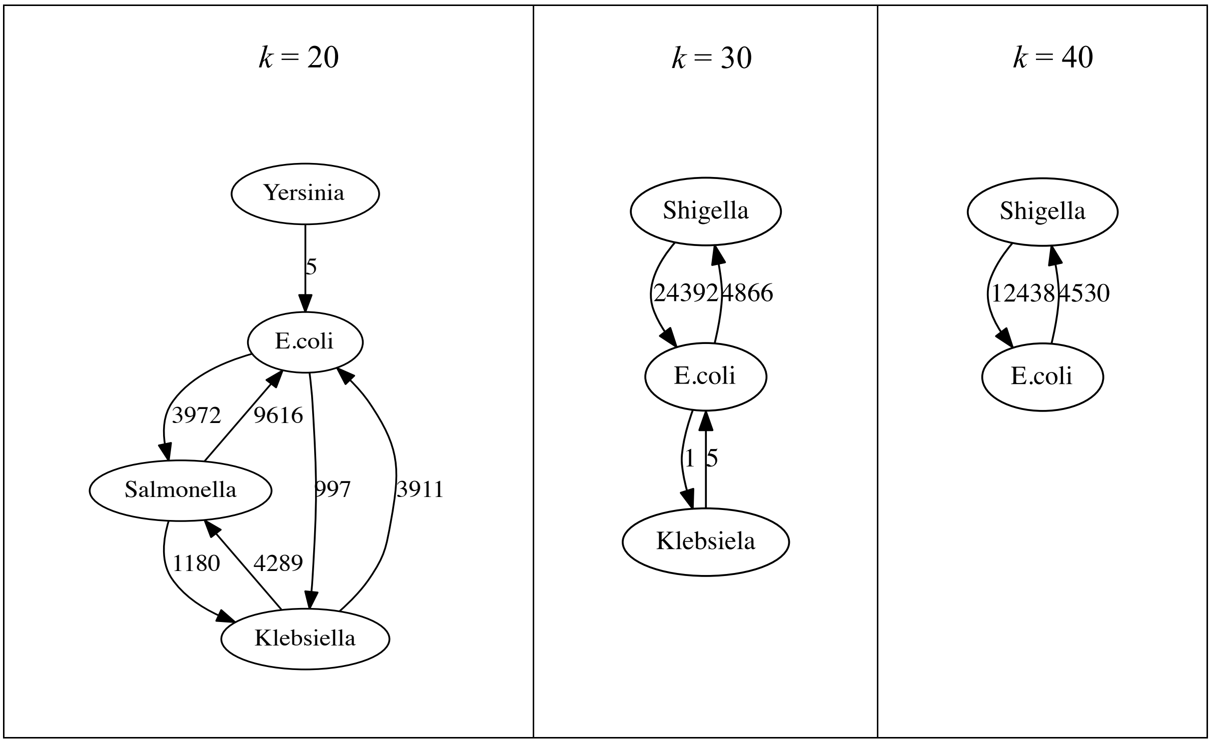


Figure S1. LGT networks for EB-1 dataset at different values of *k*.

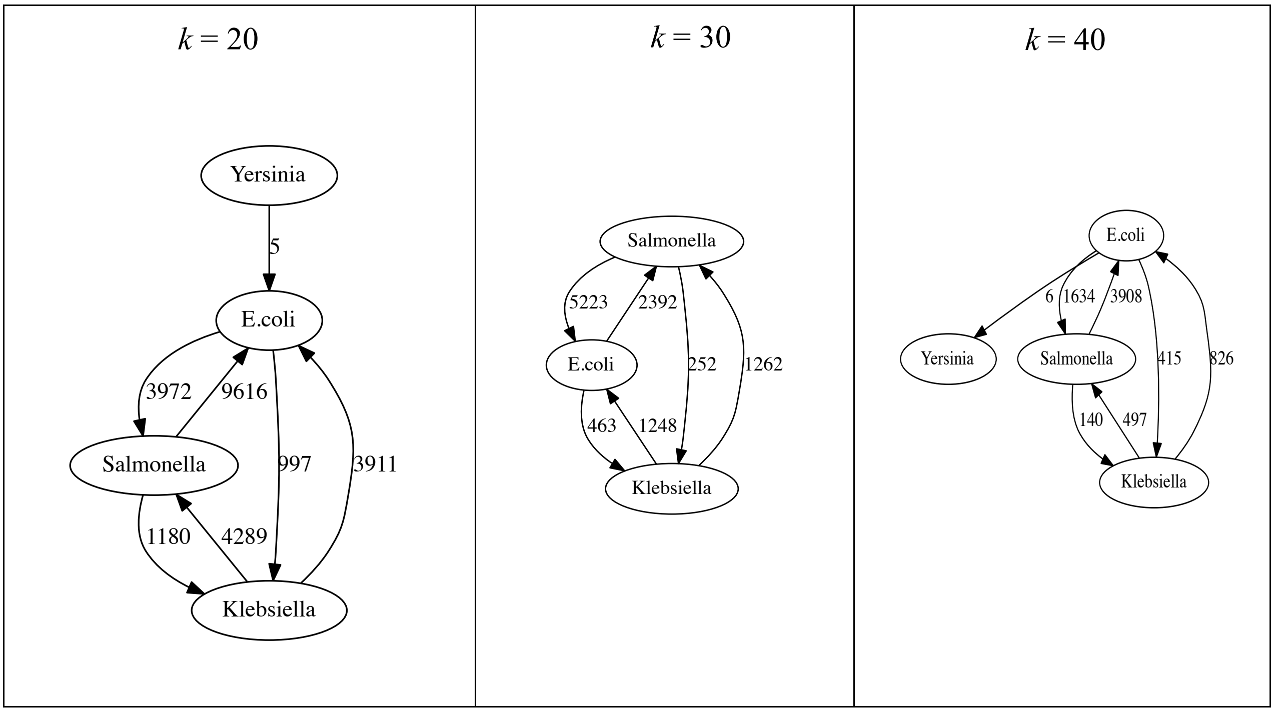


Figure S2. LGT networks for the EB-2 dataset at different values of *k*.

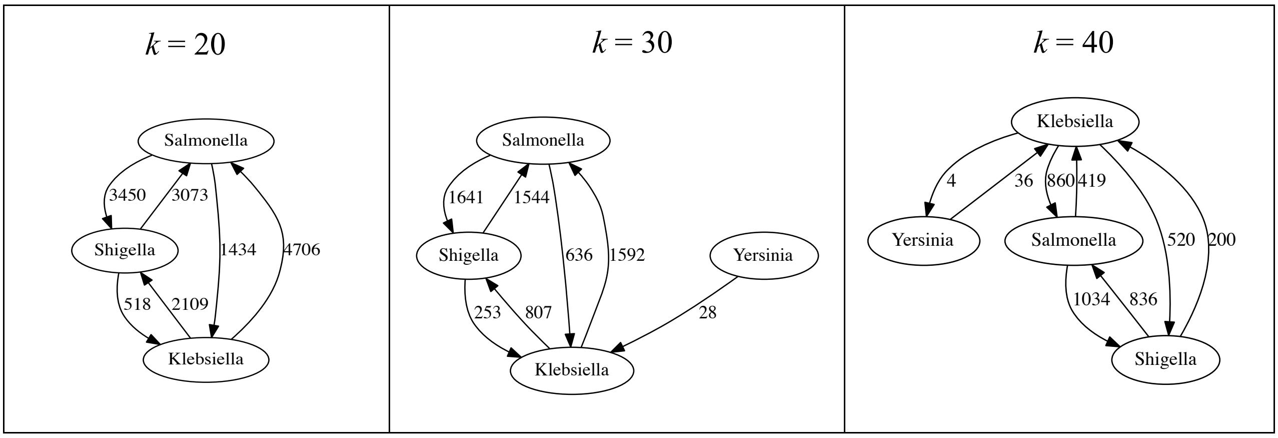


Figure S3. LGT networks for the EB-3 dataset at different values of *k*.

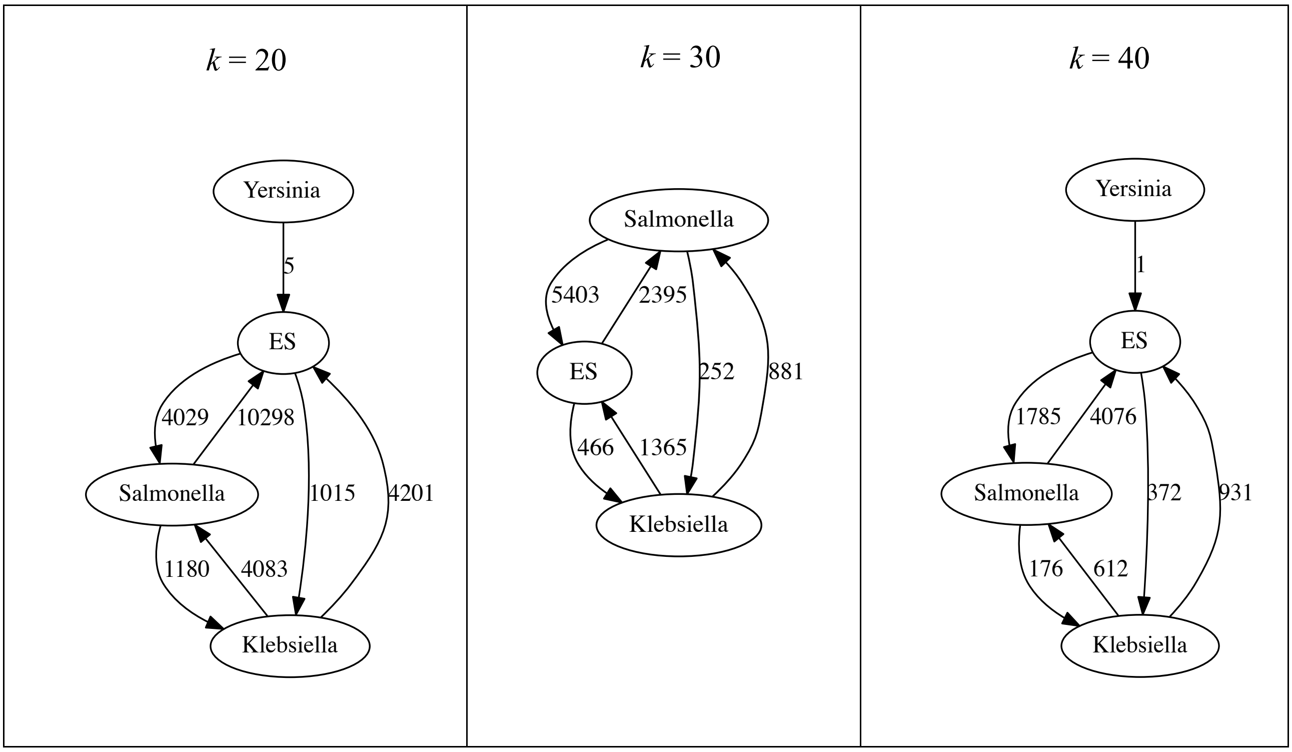


Figure S4. LGT networks for the EB-4 dataset at different values of *k*.

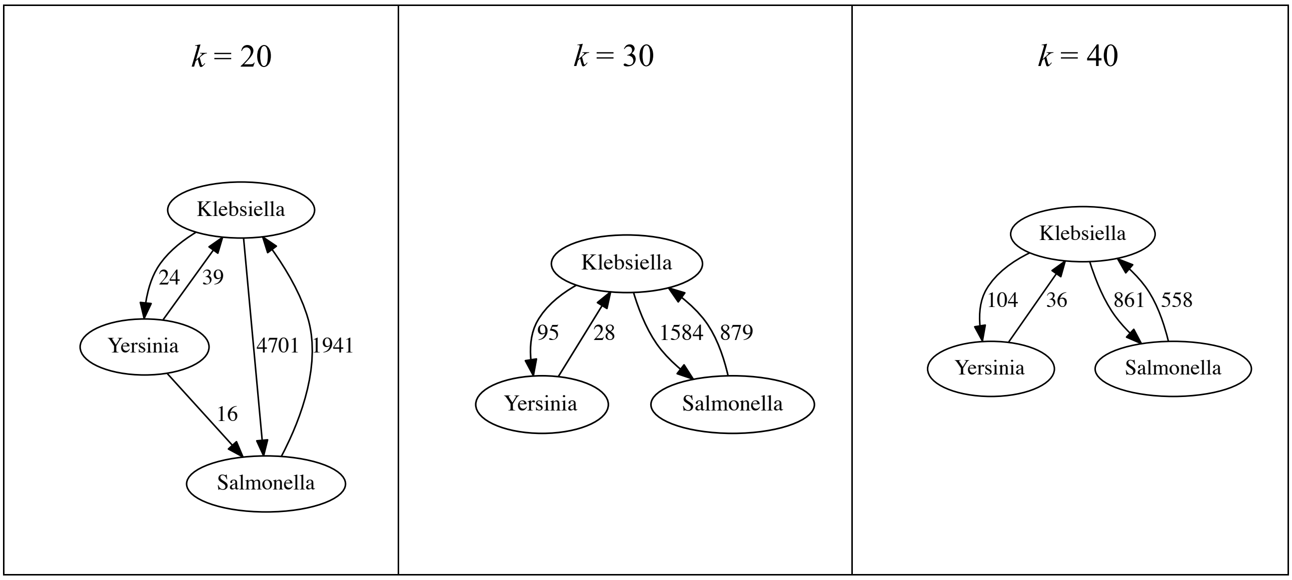


Figure S5. LGT networks for the EB-5 dataset at different values of *k*.

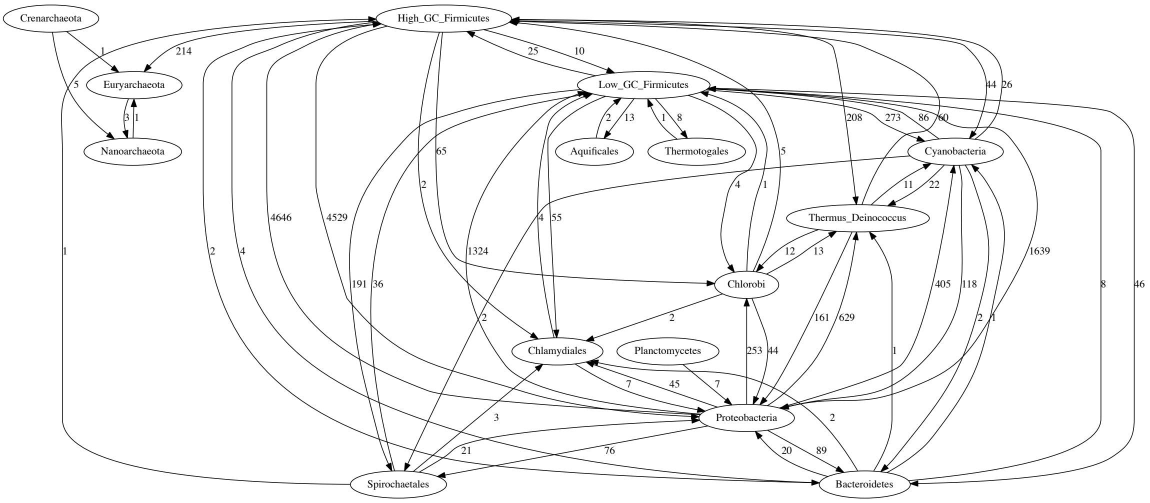


Figure S6. LGT network for the BA dataset: genomes grouped by phylum, *k* = 20.

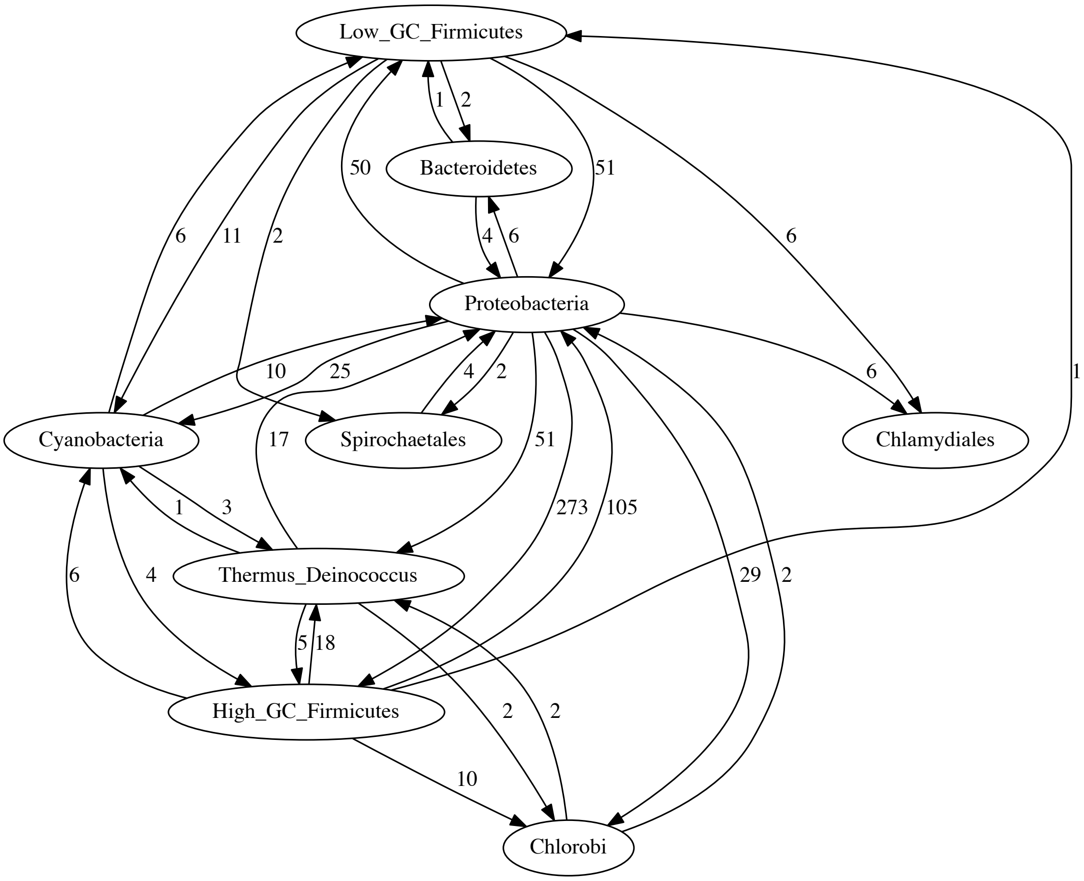


Figure S7. LGT network for the BA dataset: genomes grouped by phylum, *k* = 25.

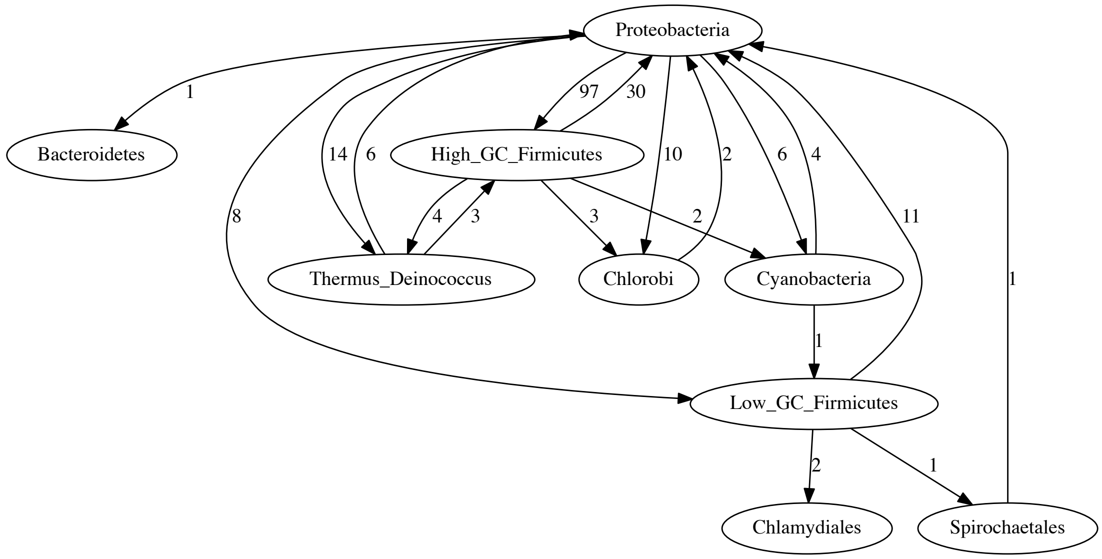


Figure S8. LGT network for the BA dataset: genomes grouped by phylum, *k* = 30.

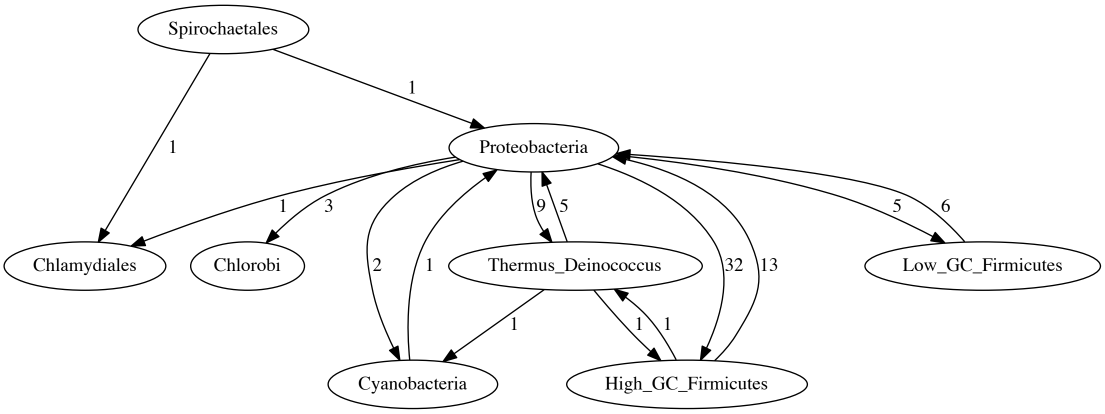


Figure S9. LGT network for the BA dataset: genomes grouped by phylum, *k* = 35.

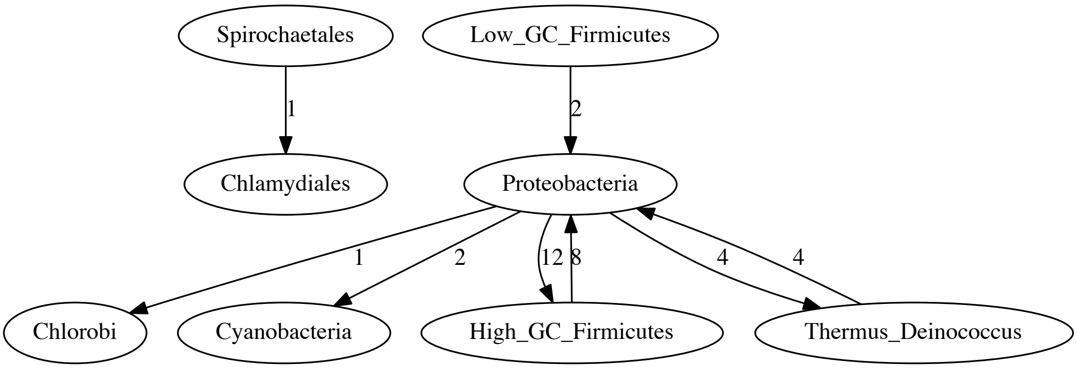


Figure S10. LGT network for the BA dataset: genomes grouped by phylum, *k* = 40.

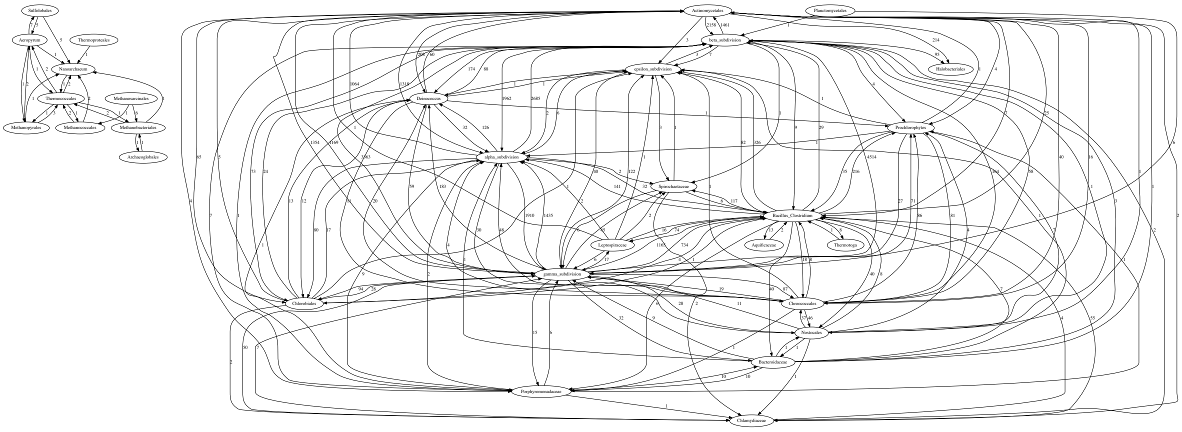


Figure S11. LGT network for the BA dataset: genomes grouped by class, *k* = 20.

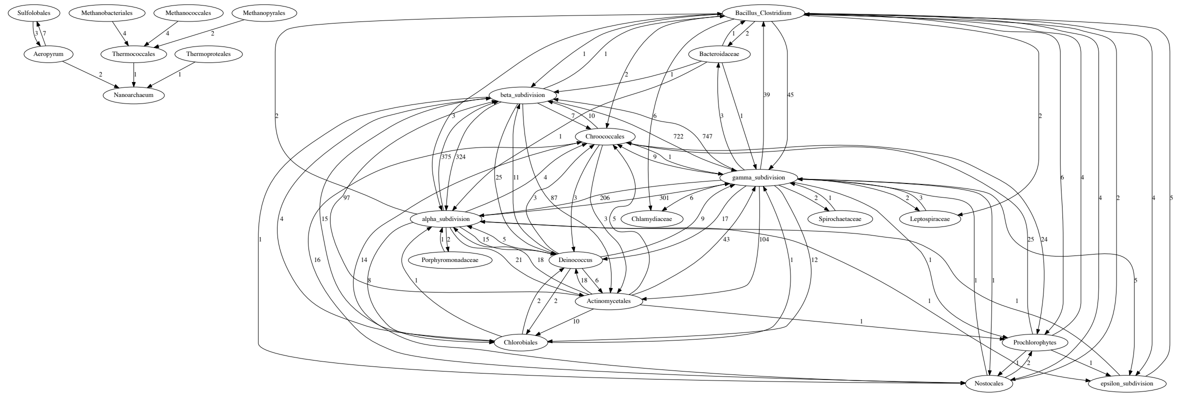


Figure S12. LGT network for the BA dataset: genomes grouped by class, *k* = 25.

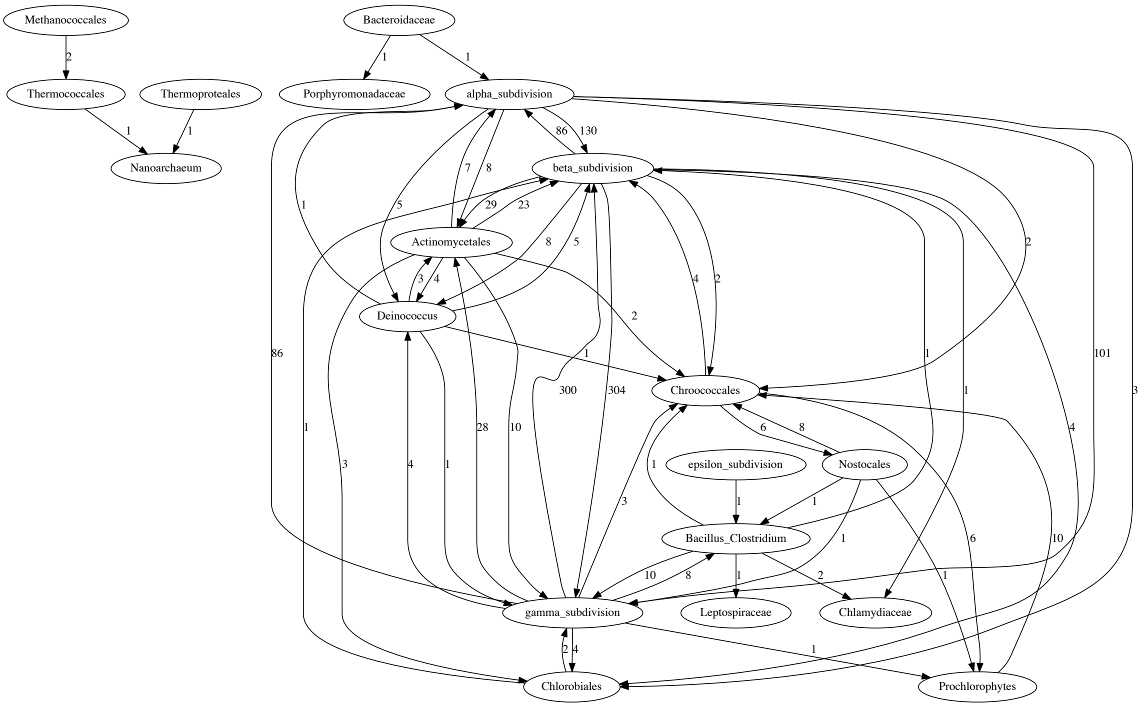


Figure S13. LGT network for the BA dataset: genomes grouped by class, *k* = 30.

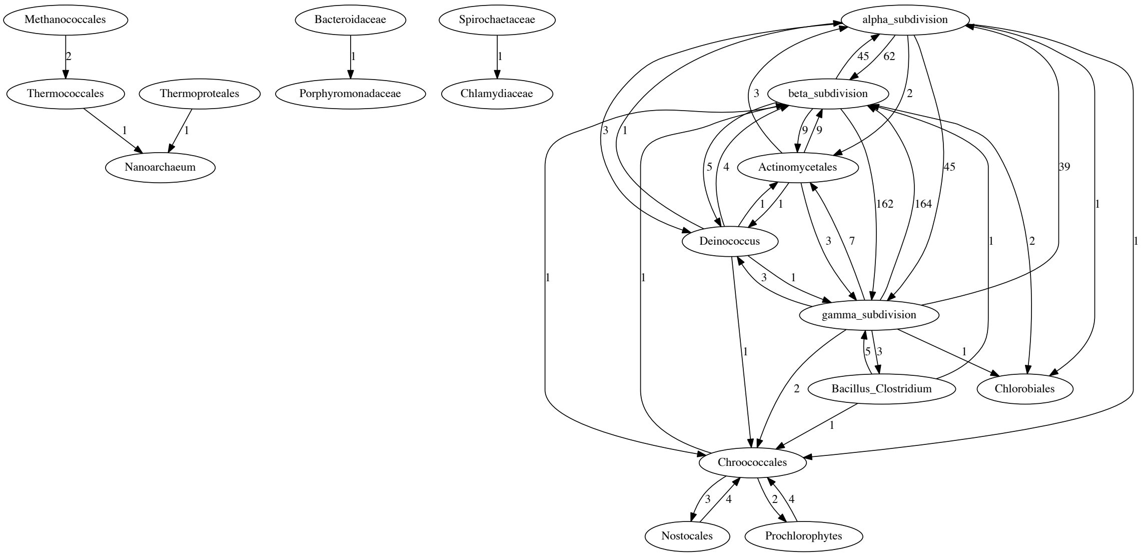


Figure S14. LGT network for the BA dataset: genomes grouped by class, *k* = 35.

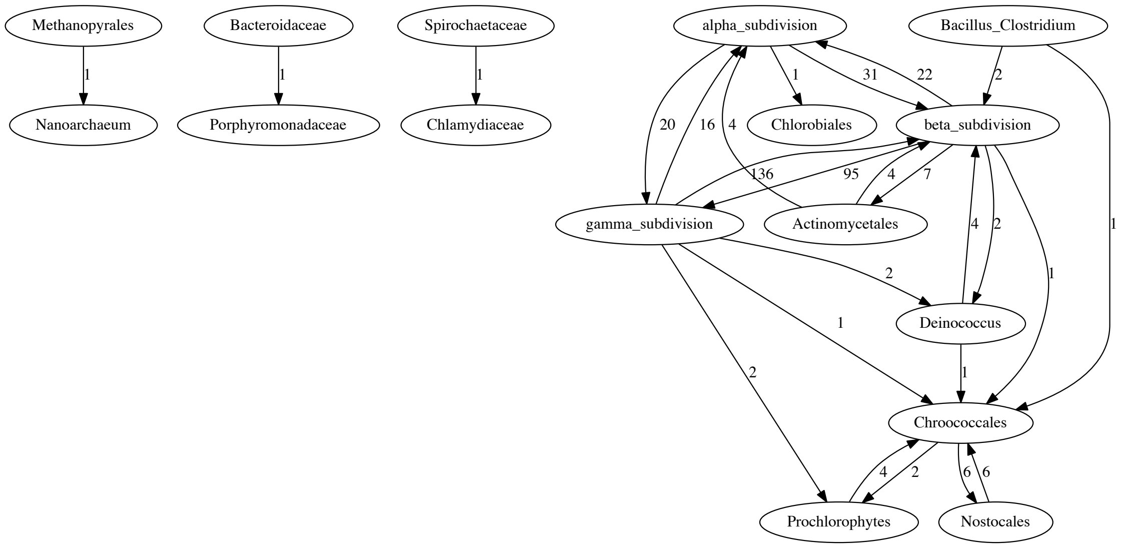


Figure S15. LGT network for the BA dataset: genomes grouped by class, *k* = 40.

Table S1. Distributions of lengths of inferred lateral segments in the ECS dataset at different *k*.

Length of lateral segment (nt)	***k* = 20**	***k* = 30**	***k* = 40**
0-100	41587	37297	27972
100-200	21620	26318	23783
200-300	13343	17161	16043
300-400	8890	11240	11333
400-500	6189	8394	8510
500-600	4656	6261	6235
600-700	3674	4814	5148
700-800	2971	3824	3926
800-900	2281	2894	3286
900-1000	1929	2449	2614
1000-1100	1634	2037	2187
1100-1200	1417	1644	1845
1200-1300	1264	1215	1539
1300-1400	970	1074	1354
1400-1500	917	1018	1106
1500-1600	736	824	991
1600-1700	659	701	844
1700-1800	625	789	722
1800-1900	581	488	563
1900-2000	452	449	598
2000-2100	399	432	504
2100-2200	413	414	460
2200-2300	359	335	424
2300-2400	320	358	379
2400-2500	314	283	342
2500-2600	255	272	267
2600-2700	260	216	275
2700-2800	218	159	236
2800-2900	224	225	214
2900-3000	190	171	213
3000-3100	210	166	182
3100-3200	182	143	153
3200-3300	157	142	157
3300-3400	153	121	138
3400-3500	179	136	165
3500-3600	123	126	119
3600-3700	125	85	104
3700-3800	121	87	98
3800-3900	91	87	109
3900-4000	86	82	81
4000-4100	68	73	78
4100-4200	80	44	66
4200-4300	52	44	82
4300-4400	77	67	72
4400-4500	72	65	45
4500-4600	63	65	56
4600-4700	71	40	51

4700-4800	65	50	51
4800-4900	50	58	51
4900-5000	44	34	31
5000+	1164	858	868

Table S2. BA dataset grouped by phylum, *k* = 20; maximum clique as the GEC.

1	2	3	4	5
Spirochaetales	Chlamydiales	High_GC_Firmicutes	Low_GC_Firmicutes	Proteobacteria

Table S3. BA dataset grouped by phylum, *k* = 20; maximal cliques as the GECs.

1	2	3	4	5
High_GC_Firmicutes	Proteobacteria	Low_GC_Firmicutes	Chlamydiales	Bacteroidetes
High_GC_Firmicutes	Proteobacteria	Low_GC_Firmicutes	Chlamydiales	Chlorobi
High_GC_Firmicutes	Proteobacteria	Low_GC_Firmicutes	Chlamydiales	Spirochaetales
High_GC_Firmicutes	Proteobacteria	Low_GC_Firmicutes	Cyanobacteria	Bacteroidetes
High_GC_Firmicutes	Proteobacteria	Low_GC_Firmicutes	Cyanobacteria	Spirochaetales
High_GC_Firmicutes	Proteobacteria	Thermus_Deinococcus	Bacteroidetes	Cyanobacteria
High_GC_Firmicutes	Proteobacteria	Thermus_Deinococcus	Chlorobi	
Nanoarchaeota	Euryarchaeota	Crenarchaeota		

Table S4. BA dataset grouped by phylum, *k* = 25; maximum clique as the GEC.

1	2	3	4
Thermus_Deinococcus	Cyanobacteria	High_GC_Firmicutes	Proteobacteria

Table S5. BA dataset grouped by phylum, *k* = 25; maximal cliques as the GECs.

1	2	3	4
Proteobacteria	Low_GC_Firmicutes	High_GC_Firmicutes	Cyanobacteria
Proteobacteria	Low_GC_Firmicutes	Bacteroidetes	
Proteobacteria	Low_GC_Firmicutes	Chlamydiales	
Proteobacteria	Low_GC_Firmicutes	Spirochaetales	
Proteobacteria	Chlorobi	Thermus_Deinococcus	High_GC_Firmicutes
Proteobacteria	Thermus_Deinococcus	Cyanobacteria	High_GC_Firmicutes

Table S6. BA dataset grouped by phylum, *k* = 30; maximum clique as the GEC.

1	2	3
Thermus_Deinococcus	High_GC_Firmicutes	Proteobacteria

Table S7. BA dataset grouped by phylum, *k* = 30; maximal cliques as the GECs.

1	2	3
Proteobacteria	High_GC_Firmicutes	Chlorobi
Proteobacteria	High_GC_Firmicutes	Cyanobacteria
Proteobacteria	High_GC_Firmicutes	Thermus_Deinococcus
Proteobacteria	Low_GC_Firmicutes	Cyanobacteria
Proteobacteria	Low_GC_Firmicutes	Spirochaetales

Table S8. BA dataset grouped by phylum, *k* = 35; maximum clique as the GEC.

1	2	3
High_GC_Firmicutes	Thermus_Deinococcus	Proteobacteria

Table S9. BA dataset grouped by phylum, *k* = 35; maximal cliques as the GECs.

1	2	3
Proteobacteria	Thermus_Deinococcus	Cyanobacteria
Proteobacteria	Thermus_Deinococcus	High_GC_Firmicutes
Proteobacteria	Spirochaetales	Chlamydiales

Table S10. BA dataset grouped by phylum, *k* = 40; maximum clique as the GEC.

1	2
Thermus_Deinococcus	Proteobacteria

Table S11. BA dataset grouped by phylum, *k* = 40; maximal cliques as the GECs.

1	2
Proteobacteria	Chlorobi
Proteobacteria	Cyanobacteria
Proteobacteria	High_GC_Firmicutes
Proteobacteria	Low_GC_Firmicutes
Proteobacteria	Thermus_Deinococcus
Chlamydiales	Spirochaetales

Table S12. BA dataset grouped by class, *k* = 20; maximum clique as the GEC.

1	2	3	4	5	6	7	8
Chroococcales	Prochlorophytes	epsilon_subdivision	alpha_subdivision	Actinomycetales	gamma_subdivision	Bacillus_Clostridium	beta_subdivision

Table S13. BA dataset grouped by class, *k* = 20; maximal cliques as the GECs.

1	2	3	4	5	6	7	8
beta_subdivision	gamma_subdivision	Bacillus_Clostridium	Actinomycetales	alpha_subdivision	Chroococcales	Prochlorophytes	Nostocales
beta_subdivision	gamma_subdivision	Bacillus_Clostridium	Actinomycetales	alpha_subdivision	Chroococcales	Prochlorophytes	Porphyromonadaceae
beta_subdivision	gamma_subdivision	Bacillus_Clostridium	Actinomycetales	alpha_subdivision	Chroococcales	Prochlorophytes	epsilon_subdivision
beta_subdivision	gamma_subdivision	Bacillus_Clostridium	Actinomycetales	alpha_subdivision	Bacteroidaceae	Nostocales	
beta_subdivision	gamma_subdivision	Bacillus_Clostridium	Actinomycetales	alpha_subdivision	Bacteroidaceae	Porphyromonadaceae	
beta_subdivision	gamma_subdivision	Bacillus_Clostridium	Actinomycetales	alpha_subdivision	Bacteroidaceae	epsilon_subdivision	
beta_subdivision	gamma_subdivision	Bacillus_Clostridium	Actinomycetales	alpha_subdivision	Chlorobiales	epsilon_subdivision	
beta_subdivision	gamma_subdivision	Bacillus_Clostridium	Actinomycetales	Chlamydiaceae	Porphyromonadaceae		
beta_subdivision	gamma_subdivision	Bacillus_Clostridium	Actinomycetales	Chlamydiaceae	Nostocales		
beta_subdivision	gamma_subdivision	Bacillus_Clostridium	Actinomycetales	Chlamydiaceae	epsilon_subdivision		
beta_subdivision	gamma_subdivision	Bacillus_Clostridium	Actinomycetales	Chlamydiaceae			
beta_subdivision	gamma_subdivision	Bacillus_Clostridium	Spirochaetaceae	Chlamydiaceae			
beta_subdivision	gamma_subdivision	Deinococcus	Actinomycetales	alpha_subdivision	Chroococcales	Prochlorophytes	Porphyromonadaceae
beta_subdivision	gamma_subdivision	Deinococcus	Actinomycetales	alpha_subdivision	Chroococcales	Prochlorophytes	epsilon_subdivision
beta_subdivision	gamma_subdivision	Deinococcus	Actinomycetales	alpha_subdivision	Chlorobiales	epsilon_subdivision	
beta_subdivision	gamma_subdivision	Planctomycetales					

beta_subdivision	Halobacteriales	Actinomycetales			
Aeropyrum	Nanoarchaeum	Thermococcales	Methanopyrales		
Aeropyrum	Nanoarchaeum	Sulfolobales			
Thermococcales	Nanoarchaeum	Methanobacteriales			
Thermococcales	Nanoarchaeum	Methanococcales			
Leptospiraceae	Bacillus_Clostridium	gamma_subdivision	alpha_subdivision	epsilon_subdivision	Actinomycetales
Leptospiraceae	Bacillus_Clostridium	gamma_subdivision	alpha_subdivision	epsilon_subdivision	Spirochaetaceae

Table S14. BA dataset grouped by class, *k* = 25; maximum clique as the GEC.

1	2	3	4	5	6
Deinococcus	Actinomycetales	Chroococcales	beta_subdivision	alpha_subdivision	gamma_subdivision

Table S15. BA dataset grouped by class, *k* = 25; maximal cliques as the GECs.

1	2	3	4	5	6
gamma_subdivision	Bacillus_Clostridium	alpha_subdivision	beta_subdivision	Bacteroidaceae	
gamma_subdivision	Bacillus_Clostridium	alpha_subdivision	beta_subdivision	Chroococcales	
gamma_subdivision	Bacillus_Clostridium	alpha_subdivision	epsilon_subdivision		
gamma_subdivision	Bacillus_Clostridium	Chlamydiaceae			
gamma_subdivision	Bacillus_Clostridium	Nostocales	Chroococcales	beta_subdivision	
gamma_subdivision	Bacillus_Clostridium	Nostocales	Chroococcales	Prochlorophytes	
gamma_subdivision	Bacillus_Clostridium	Prochlorophytes	epsilon_subdivision		
gamma_subdivision	Bacillus_Clostridium	Leptospiraceae			

gamma_subdivision	Chlorobiales	alpha_subdivision	Actinomycetales	beta_subdivision	Deinococcus
gamma_subdivision	Actinomycetales	Chroococcales	beta_subdivision	Deinococcus	alpha_subdivision
gamma_subdivision	Actinomycetales	Chroococcales	Prochlorophytes		

Table S16. BA dataset grouped by class, *k* = 30; maximum clique as the GEC.

1	2	3	4	5	6
Deinococcus	Actinomycetales	alpha_subdivision	beta_subdivision	Chroococcales	gamma_subdivision

Table S17. BA dataset grouped by class, *k* = 30; maximal cliques as the GECs.

1	2	3	4	5	6
gamma_subdivision	Chroococcales	beta_subdivision	Actinomycetales	alpha_subdivision	Deinococcus
gamma_subdivision	Chroococcales	beta_subdivision	Bacillus_Clostridium		
gamma_subdivision	Chroococcales	Nostocales	Bacillus_Clostridium		
gamma_subdivision	Chroococcales	Nostocales	Prochlorophytes		
gamma_subdivision	Chlorobiales	Actinomycetales	beta_subdivision	alpha_subdivision	
Chlamydiaceae	Bacillus_Clostridium	beta_subdivision			

Table S18. BA dataset grouped by class, *k* = 35; maximum clique as the GEC.

1	2	3	4	5
Actinomycetales	Deinococcus	alpha_subdivision	gamma_subdivision	beta_subdivision

Table S19. BA dataset grouped by class, *k* = 35; maximal cliques as the GECs.

1	2	3	4	5
beta_subdivision	gamma_subdivision	alpha_subdivision	Deinococcus	Chroococcales
beta_subdivision	gamma_subdivision	alpha_subdivision	Deinococcus	Actinomycetales
beta_subdivision	gamma_subdivision	alpha_subdivision	Chlorobiales	
beta_subdivision	gamma_subdivision	Bacillus_Clostridium	Chroococcales	

Table S20. BA dataset grouped by class, *k* = 40; maximum clique as the GEC.

1	2	3	4
Deinococcus	gamma_subdivision	Chroococcales	beta_subdivision

Table S21. BA dataset grouped by class, $k = 40$; maximal cliques as the GECs.

1	2	3	4
Chroococcales	beta_subdivision	gamma_subdivision	Deinococcus
Chroococcales	beta_subdivision	Bacillus_Clostridium	
Chroococcales	Prochlorophytes	gamma_subdivision	
alpha_subdivision	beta_subdivision	gamma_subdivision	
alpha_subdivision	beta_subdivision	Actinomycetales	

Table S22. Lateral events inferred involving Proteobacteria. In-degree, number of genes affected by lateral events; out-degree, number of LGT events originating from this node. Nodes in the network represent classes; $k = 25$, $G = 2k$.

Class name	In-degree	Out-degree	Total degree	Number of sequences	Number of LGT events per genome
Alpha-	560	729	1289	9	143
Beta-	1246	1182	2428	8	303
Gamma-	1128	1154	2282	33	69
Epsilon-	11	6	17	5	3

Chapter 5 General conclusions and future work

With the development and uptake of next-generation DNA-sequencing technologies over the past decade, more and more sequence data are becoming available not only to support the study of LGT in diverse natural and artificial environments, but also to begin to discern a global picture of evolutionary processes in microbes. How to process and analyse such large, dynamic data poses a significant challenge for bioinformaticians. My PhD project aimed to develop an approach to the inference of LGT in multi-genome datasets that does not depend on the computationally hard steps of multiple sequence alignment, inference of phylogenetic trees, and topological comparison of test and reference trees. To accomplish this I took an alignment-free approach.

In recent years, alignment-free methods have gradually been used in several areas of bioinformatics including phylogenetic inference. Alignment-free methods do not analyse a set of sequences *per se*, but instead decompose them into fixed-size segments (*k*-mers) and then analyse these segments. Typically, distributions of *k*-mers are compared to generate a matrix of pairwise distances among the sequences. This matrix then serves as input into a distance method, usually neighbour-joining. Other variants have been explored, but tend not to be as scalable[1,2]. Compared with alignment-based methods, alignment-free methods are much faster in processing large datasets, and can show comparable accuracy[3].

In contrast to phylogenetic inference *per se*, for which around ten alignment-free methods have been published[3,4], only two such methods are currently available for the computational inference of LGT: ALFY[5,6] and the so-called purity measure[7]. ALFY computes a phylogenetic tree of sequences based on K_r[8] and identifies subtrees topologically incongruent with the reference tree as lateral transfers. The purity measure approach finds the less-repeated segments in a sequence (those with "high purity") and identifies them as lateral transfers. Both methods infer the existence and locations of LGT events, but not their directionality; and ALFY retains the computationally hard topology-comparison step. To better address these problems, I designed and implemented a new alignment-free method, as discussed in Chapter 2, modifying and repurposing the concepts *term frequency* and *inverse document frequency* (TF-IDF).

5.1 TF-IDF for LGT detection

The TF-IDF method, as developed in Chapter 2, infers lateral segments by comparing matches of *k*-mers between donor groups and recipient sequences. If a segment contains *k*-mers infrequent in its own group, but frequent in other groups, then that segment is judged as lateral. The direction of the inferred LGT events is from the external (donor) group to the recipient sequence.

In Chapter 2, I demonstrated that the TF-IDF method can identify regions of lateral origin efficiently. Its time complexity is $O(nLlogU)$ where n is the number of sequences, L the average length of sequences, and U the number of unique k-mers within a dataset. Such time complexity guarantees that TF-IDF will be scalable on large genome-scale datasets. Based on extensive simulations, we understand that TF-IDF can infer LGT events with high accuracy and recall under a range of biologically relevant situations *e.g.* long evolutionary distances between groups (substitution rate *i.e.* probability of change per site ≥ 0.1), close similarity of sequences within each group (substitution rate ≤ 0.02), modest diversification following lateral transfer (substitution rate ≤ 0.01), and variable rates of deletion post-LGT (rate ≤ 0.1); details are given in Chapter 2.

Although TF-IDF offers significant advantages for the inference of LGT, it has certain limitations. Its performance depends on the size of k (and to a much lesser extent on G, the maximum spacing allowed between k-mers that are merged to form an inferred lateral region), and on how groups are delineated. From computational simulations and the study of empirical datasets, $k = 40$ was identified as a good default value in most cases, balancing precision and recall. However, this did not extend to cases where the sequences under investigation are short (≤ 1000 nt in the simulations) or very dissimilar (≥ 70% of the 12-mers are different). The former case is to some extent a consequence of the simulation strategy: the lengths of the added LGT segments are distributed around a mean value proportional with length of the target sequence. This was nonetheless valuable in showing that short segments (≤ 40 nt) can be ignored by TF-IDF. In the latter case, there can be too few identical 40-mers between sequences. If possible, the size of k should be matched to the specific dataset under investigation, and I presented an alignment-free way to carry out this matching (Chapter 4).

TF-IDF requires sequences to be arranged into groups, and the way this is done can play an important role in how the method performs with a particular dataset. If the sequences are grouped randomly, many fewer (ECS dataset) or no (simulated datasets) lateral events may be detected. Grouping the sequences by evolutionary relationship (*i.e.* on a phylogenetic tree) was, in general, found to yield good results. As a main aim of alignment-free approaches is to avoid computationally complex steps, this tree should be available in advance: indeed, comprehensive trees are available for bacteria and archaea[9,10], and groups can be formed simply by excising subtrees at the desired resolution (Chapter 3). Alternatively, an alignment-free approach could be used to generate this reference tree[3,11]. Where investigated, the use of known taxonomic labels (genus, class, phylum) was usually a good proxy for a phylogenetic grouping, although it may not be a coincidence that the two genera that so greatly elevated the IDF threshold, *Escherichia* and *Shigella*, are widely considered to be one and the same phylogenetic genus[12,13]. Within-genus, MLST (Multi Locus Sequence Typing) data can serve as a useful basis for grouping, if the feature genes have been defined in advance; but MLST types have been defined for relatively few bacteria and no archaea (or eukaryotes). Alternatively, features of lifestyle of ecology of the organisms might be used to delineate groups, depending on the specific research aim.

The current version of our TF-IDF method is implemented using in-memory computing for the core *k*-mer extraction and comparison. This makes the program run much faster than would be the case if data were swapped onto hard disk, but requires much more RAM. With the development of flash memory (1-TB RAM servers have been deployed on many computational clusters), such a memory requirement becomes affordable. In favourable cases (low overall dissimilarity among genomes), TF-IDF can process about 600 bacterial genomes using 1 TB of memory. Another option would be to use the MapReduce architecture[14] to store and process *k*-mers over a cluster, *i.e.* to take advantage of the memory and CPU resources of a number of computers rather than only one. This can be implemented on the Hadoop platform[15], an open-source framework for distributed storage and processing on big data that has been deployed (for instance) by Yahoo! to analyse and index bodies of text > 5 PB in size.

In the TF part of the current method, the mean value of the counts of all *k*-mers within a group is used to assess the "popularity" of lateral segments within that group, while in the IDF part, the mean of occurrences of *k*-mers shared between the recipient sequences and donor groups is taken as the threshold for existence of LGT events. In the applications reported in this thesis, we encountered no disadvantage arising from the use of the TF mean. In the IDF part, by contrast, the mean serves as a threshold on the number of *k*-mers that a sequence must share with a donor group before it is further processed by the TF part. Thus sequences containing only a few lateral events will be ignored. As shown by the simulations (Chapter 2), combining the two thresholds helps to ensure the performance (precision and recall). While the TF-IDF method as devised gives useful results, there is still some scope for improvement. We observed that the IDF and TF thresholds are selected to be self-tuning to the dataset, which is reasonable. However, they can be sensitive to divergence between and within groups, so better results might be obtained if we were to manually control these thresholds to some extent. This may have the effect of enabling the user to detect more-ancient transfers, or conversely to reduce the number of false positives. The effects of tuning the IDF or TF thresholds cannot be evaluated on empirical datasets, because there is no gold standard database with all true LGT events known. Rigorous simulations are required to understand how these thresholds affect the performance of TF-IDF in different phylogenetic situations, before they can be adopted and used safely as parameters of the method. This issue can be approached through computational simulation, and remains for future work.

Furthermore, the method detects LGT without ascribing a strength or confidence level to these detections. One way by which confidence could be assessed would be to construct a single statistic for each segment, combining the TF and IDF values in some way. It might then be possible to use resampling methods such as bootstrapping to gain an idea of the significance of each detection.

5.2 Exploring LGT among microbes using TF-IDF

In Chapter 3, I applied the TF-IDF method to three empirical datasets: ECS (twenty *E. coli* and seven *Shigella* genome sequences), EB (110 enteric bacterial genomes) and BA (143 genomes across Bacteria and Archaea). Two of these datasets were examined previously using alignment-based approaches[9,16], providing reference LGT sets for comparison. I examined the effect of two TF-IDF parameters, *k* and *G*, on LGT inference. *k* is the size (length) of *k*-mers, and *G* a threshold for merging two lateral segments into one. I showed that the number of detections decreases when *k* increases, whereas changing *G* does not greatly affect the total detection lengths when the lateral segments are not densely distributed (as in the EB and BA datasets). If the LGT segments are densely distributed on a genome, however, high values of *G* will cause many undetected regions to be merged into detected ones, subsequently affecting the performance of TF-IDF. In most cases, $k = 40$ and $G = 2k$ (*i.e.* 80) were the optimal parameter settings.

Details of the biological relationships play an important part in assessing the frequencies of LGT events. In Chapter 3 I found that closely related phyletic groups (*e.g.* those in ECS or EB compared with BA) share many more LGT events than distantly related ones. Similar results have been seen previously[17-19] and arise in part from increased ease of homologous recombination[20,21]. No LGT was inferred between bacteria and archaea using TF-IDF. This does not mean that there has been no LGT between bacteria and archaea; it could be the case that regions of genetic material transferred across domains are shorter, or were transferred a much longer time ago, than between lower-level phyletic groups, causing the former to fall beneath the IDF threshold and thereby be ignored by TF-IDF. As I demonstrate in Chapter 3, one possible solution is to manually lower the IDF threshold, but this may open the analysis to an unacceptable level of false positive detections.

It has long been considered that individual genes might have accepted LGT from two or more donors, and in Chapter 3 I examined this possibility. A gene (or other region) inferred to have multiple lateral origins may indeed have accepted LGT from different donors; but another explanation might be that there was a single ancient transfer, and subsequent divergence of the daughter lineages of the ancient donor have led to a situation where one part of the lateral region shares *k*-mers with one donor group (one daughter lineage) and a different part shares *k*-mers with another donor group. The latter scenario might be favoured if the regions in question are adjacent (or nearly so) in the recipient lineage, and the inferred donor groups are close relatives in a phylogenetic sense. In Chapter 3 I presented evidence for LGT from multiple donor groups into single genes; and in the first systematic analysis addressing these alternative scenarios, I was able to estimate that 28.5% of the genes with exactly two inferred donor groups can be interpreted as the result of a single ancient transfer. Instances of multiple (more than two) inferred donors could likewise be examined, ideally using datasets larger than these. TF-IDF is unique in enabling such innovative exploration of the vertical and lateral origins of genes and genomes.

Inferred lateral genes were also mapped to Gene Ontology (GO) terms for enrichment testing. These tests showed that LGT affects a wide range of biological processes in all three empirical datasets; the most-enriched terms describe metabolism, regulation, cellular and trans-membrane transport, and (often) energetics. These processes are "operational" in the sense of Lake and colleagues[22,23] and are often found to be affected by LGT[24-26] .

A direct implication of results in this Chapter is that the extent of genetic exchange is not uniform across genomes; some phyla, classes, genera and MLST types have exchanged genetic material frequently (and the exchanged material has persisted), whereas others have done so much less-often (and/or the material has not persisted in its new hosts). As a consequence, LGT networks will have both densely and sparsely connected regions. The former are potential genetic exchange communities (GECs). I explore these patterns in the next chapter.

5.3 Constructing GECs in LGT networks

In Chapter 4 I inferred LGT networks for the ECS, EB and BA datasets, and identified GECs within these networks. For this work, I defined GECs as cliques in LGT networks, although other definitions are possible[27].

All the ECS genomes and groups are connected by LGT and give rise to a single clique, which I interpret as a GEC. The corresponding organisms span a variety of environmental strains as well as human and animal commensals, and pathogens as well as non-pathogens. The TF-IDF analysis shows that genetic material has flowed among all groups, *i.e.* across environmental and lifestyle boundaries. In the EB dataset, *Escherichia* and *Shigella* have been strongly involved in LGT, contributing many lateral events to all inferred GECs. *Salmonella* and *Klebsiella* are seen as LGT-active within GECs from which either *Escherichia* or *Shigella* is excluded (or both are excluded), while *Yersinia* is relatively silent to LGT, contributing almost no lateral genetic material in any GEC. Thus even among the enteric bacteria and their immediate relatives (all of them Gamma-proteobacteria), not all genera have contributed to GECs. This presumably reflects multiple underlying mechanisms including natural competence, host ranges of plasmids and phage, efficacy of intracellular defence mechanisms, mechanisms of genetic integration and recombination, and subsequent population dynamics[27].

I grouped genomes in the BA dataset alternatively by phylum or class for discovery of cliques in the LGT networks. In the analysis at phylum level, it is clear that some microbial genomes and phyla have donated and/or accepted many regions, whereas others have been relatively inactive, at least when the IDF threshold is set by mean value (see above). The Proteobacteria, High-G+C Firmicutes and Low-G+C Firmicutes comprise the core of the GECs in this dataset. As the value of *k* is increased the number of inferred lateral genes decreases, but the core structure of the GECs remains stable.

When the sequences are instead grouped by class, *i.e.* at finer granularity, we see that genera within a phylum do not necessarily all exchange genetic material with equal frequency. For example, genomes in the Alpha-, Beta- and Gamma subdivisions of Proteobacteria exhibit many more lateral events than do those in the Epsilon subdivision, whether assessed by total number of lateral events or by average number of events per genome (Chapter 4, Supplementary file, Table S22). As the genomes are of similar mean size, normalising by genome size would scarcely change this conclusion. These classes have previously been identified as linked by "highways of gene sharing"[9].

Gene Ontology enrichment tests showed that the biological processes underpinning GECs inferred for the BA dataset strongly resemble those inferred earlier (Chapter 3) for LGT in this dataset more broadly, with processes of metabolism, transport and regulation strongly over-represented.

5.4 Future directions

The results of TF-IDF can be sensitive to properties of the data being analysed. In Chapter 3, sequences were simulated on a tree and we saw that TF-IDF can infer LGT with high precision and recall if the sequences are highly similar (but not identical) within-group while substantially dissimilar between-group. Although we did not extend the simulations to consider the balance or distribution of group sizes, there is the suspicion that TF-IDF will work best if groups are not immensely dissimilar in size. In real-life applications, however, we may not have control over these features: the empirical datasets in Chapter 4 ranged from a narrow phyletic range (ECS) up to nearly the entire span of prokaryotic diversity (BA). Groups were intrinsically unbalanced: even apart from sampling bias, Proteobacteria are numerous (and many have been sequenced), whereas only a few dozen species have been described within *Thermus/Deinococcus*. To ensure that parameter settings (especially *k*) are matched as well as possible to features of the data, an LGT detection pipeline based on alignment-free methods could incorporate efficient (perhaps alignment-free) methods for *e.g.* calculating the evolutionary distance among sequences, optimising group size, balancing the tree and removing outlier sequences.

As originally defined[27], GECs involve a relationship among sets of sequences. The TF-IDF method as developed here is best applied to sequences that are organised into groups, as this leads to better and more-stable values for the TF and IDF thresholds. LGT is inferred from donor groups to recipient sequences; the latter are of course also members of a group, but the inference is actually carried out from group to sequence. In the LGT networks in Chapter 4, recipient sequences were then collapsed into the respective recipient group, such that all nodes are groups rather than individual sequences. Information is lost in this process, and (as we saw in Chapter 4) the resulting simplification can make the resulting network less interesting from both biological and graph-theoretical points of view. There are, however, types of graphs in which nodes can be contained within other nodes (power graphs[28,29]) and it could be interesting to explore whether such frameworks might have advantages in representing GECs inferred using the TF-IDF method.

The edges inferred using TF-IDF are directed and weighted. In the process (described in the previous paragraph) of subsuming recipient sequences into groups, detail (of individual recipient sequences), edge weight and directionality are all lost, and with them much of the potential power of TF-IDF to inform about biology. Graph theory extends to densely connected groups where edges are directed and/or weighted[30,31] but to my knowledge, these methods have not so far been applied in the molecular biological sciences, and are not computationally scalable. The application of such methods to delineate GECs may nonetheless represent another possible direction for future research.

5.5 Thesis summary

Recognising genetic exchange communities within LGT networks is a way to systematically analyse LGT between organisms. To achieve this, detecting LGT in big sequence datasets is the cornerstone. Current methods for LGT detection, however, are not scalable to big datasets. In this thesis I describe the development, testing and application of a new alignment-free approach, based on TF-IDF, for detection of LGT in synthetic and empirical genome-scale datasets. This is the first time the TF-IDF idea has been used to detect LGT, and indeed (to our knowledge) its first application to any biological problem. Rigorous tests have shown that it can be efficient and effective. Applying this method to three empirical datasets, I inferred LGT, built LGT networks and found genetic exchange communities (as cliques) among them. This is the first time that the GEC concept has been applied on large empirical datasets. These GECs reveal how groups of microbes interact through LGT, including donor-recipient pairs and biological process enrichment. TF-IDF expands the bioinformatic toolkit of alignment-free methods and can be incorporated into LGT workflows to infer GECs in much larger genome datasets in the future.

5.6 References

1 Höhl, M., Rigoutsos, I. & Ragan, M. A. Pattern-based phylogenetic distance estimation and tree reconstruction. *Evol Bioinform Online* **2**, 359-375 (2006).

2 Höhl, M. & Ragan, M. A. Is multiple-sequence alignment required for accurate inference of phylogeny? *Syst Biol* **56**, 206-221, doi:10.1080/10635150701294741 (2007).

3 Chan, C. X., Bernard, G., Poirion, O., Hogan, J. M. & Ragan, M. A. Inferring phylogenies of evolving sequences without multiple sequence alignment. *Sci Rep* **4**, 6504. doi:Artn 6504 10.1038/Srep06504 (2014).

4 Haubold, B. Alignment-free phylogenetics and population genetics. *Brief Bioinform* **15**, 407-418, doi:10.1093/bib/bbt083 (2014).

5 Domazet-Lošo, M. & Haubold, B. Alignment-free detection of horizontal gene transfer between closely related bacterial genomes. *Mob Genet Elements* **1**, 230-235, doi:10.4161/mge.1.3.18065 (2011).

6 Domazet-Lošo, M. & Haubold, B. Alignment-free detection of local similarity among viral and bacterial genomes. *Bioinformatics* **27**, 1466-1472, doi:10.1093/bioinformatics/btr176 (2011).

7 Taniguchi, Y., Yamada, Y., Maruyama, O., Kuhara, S. & Ikeda, D. The purity measure for genomic regions leads to horizontally transferred genes. *J Bioinf Comput Biol* **11**, 1343002 (2013).

8 Domazet-Lošo, M. & Haubold, B. Efficient estimation of pairwise distances between genomes. *Bioinformatics* **25**, 3221-3227, doi:10.1093/bioinformatics/btp590 (2009).

9 Beiko, R. G., Harlow, T. J. & Ragan, M. A. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A* **102**, 14332-14337, doi:10.1073/pnas.0504068102 (2005).

10 Hug, L. A., Baker, B. J., Anantharaman, K. & Banfield, J.A. A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).

11 Bernard, G., Chan, C. X. & Ragan, M. A. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. *Sci Rep* **6**, 28970, doi:10.1038/srep28970 (2016).

12 Brenner, D. Introduction to the family Enterobacteriaceae, in Balows, A., Trüper, H. G., Dworkin, M., Harder, W. & Schleifer, K.-H. (eds) *The Prokaryotes: a Handbook on the Biology of Bacteria: Ecophysiology, Isolation, Identification, Applications* **3**, 2673-2695 (1981).

13 Pupo, G. M., Lan, R. & Reeves, P. R. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A* **97**, 10567-10572 (2000).

14 Dean, J. & Ghemawat, S. Mapreduce: Simplified data processing on large clusters. *Commun ACM* **51**, 107-113, doi: 10.1145/1327452.1327492 (2008).

15 Shvachko, K., Kuang, H. R., Radia, S. & Chansler, R. The Hadoop distributed file system. *IEEE S. Mass. Stor. Sys.* (2010).

16 Skippington, E. & Ragan, M. A. Phylogeny rather than ecology or lifestyle biases the construction of *Escherichia coli-Shigella* genetic exchange communities. *Open Biol* **2**, doi:Artn 120112 10.1098/Rsob.120112 (2012).

17 Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**, 2226-2238 (2002).

18 Andam, C. P. & Gogarten, J. P. Biased gene transfer and its implications for the concept of lineage. *Biol Direct* **6**, 47, doi:10.1186/1745-6150-6-47 (2011).

19 Andam, C. P., Williams, D. & Gogarten, J. P. Natural taxonomy in light of horizontal gene transfer. *Biol Philos* **25**, 589-602 (2010).

20 Gogarten, J. P. & Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* **3**, 679-687, doi:10.1038/nrmicro1204 (2005).

21 Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* **3**, 711-721, doi:10.1038/nrmicro1234 (2005).

22 Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: The complexity hypothesis. *Proc Natl Acad Sci U S A* **96**, 3801-3806, doi:10.1073/Pnas.96.7.3801 (1999).

23 Jain, R., Rivera, M. C., Moore, J. E. & Lake, J. A. Horizontal gene transfer accelerates genome innovation and evolution. *Mol Biol Evol* **20**, 1598-1602, doi:10.1093/molbev/msg154 (2003).

24 Ragan, M. A. & Beiko, R. G. Lateral genetic transfer: open issues. *Philos Trans R Soc Lond B Biol Sci* **364**, 2241-2251, doi:10.1098/rstb.2009.0031 (2009).

25 Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304, doi: 10.1038/35012500 (2000).

26 Garcia-Vallvé, S., Romeu, A. & Palau, J. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* **10**, 1719-1725 (2000).

27 Skippington, E. & Ragan, M. A. Lateral genetic transfer and the construction of genetic exchange communities. *FEMS Microbiol Rev* **35**, 707-735, doi:10.1111/j.1574-6976.2010.00261.x (2011).

28 Royer, L., Reimann, M., Andreopoulos, B. & Schroeder, M. Visualising protein interaction networks with power graphs. *BMC Syst Biol* **1**, 1 (2007).

29 Ahnert, S. E. Power graph compression reveals dominant relationships in genetic transcription networks. *Mol BioSyst* **9**, 2681-2685 (2013).

30 Vassilevska, V. Efficient algorithms for clique problems. *Inform Process Lett.* **109**, 254-257 (2009).

31 Seidman, S. B. Clique-like structures in directed networks. *J Soc Biol Struct* **3**, 43-54 (1980).