Freie Universität Berlin

Computational Methods for Integrative Structural Variant Analysis Across Species Boundaries

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

von

Kathrin Trappe

Berlin
August 2018

Betreuer:	PD Dr. Bernhard Renard Prof. Dr. Knut Reinert
Erstgutachter:	PD Dr. Bernhard Renard
Zweitgutachter:	Juniorprof. Dr. Tobias Marschall
Datum der Disputation:	01.10.2018

Abstract

Structural variations (SVs) are a phenomenon that have a tremendous impact on all species. SVs are the result of fundamental rearrangement mechanisms but can lead to severe human diseases like cancer. Rearrangement events also provide means that enable bacteria to adapt to environmental pressures where they can also happen across species boundaries in events called horizontal gene transfer (HGT). The incorporation of foreign genes from a donor into an acceptor genome can be investigated on the genomic level, the activity and protein expression changes, however, are better revealed on the proteomic level.

This thesis contributes four computational methods for the detection of complex SVs of various types and sizes including HGT events from genomic next-generation sequencing (NGS) data and proteomic shotgun mass-spectrometry (MS) data. Concerning HGT events, our methods address the questions of what organisms are involved in the transfer, what genes are exactly transferred and to what position, and what are the implications on proteomic level.

First, we present the generic SV detection tool Gustaf. Gustaf improves the size and type resolution compared to previous SV detection methods. A further specific advantage is the characterisation of translocations and dispersed duplications as a combination of simple, delocalised variants that have to be inferred from separate SV calls. With this basis for a more in-depth focus on HGT detection, we developed two mapping-based methods, Daisy and DaisyGPS. Daisy facilitates Gustaf and further SV detection strategies to precisely identify the transferred region within the donor and its insertion site in the acceptor genome. DaisyGPS uses metagenomic profiling strategies to identify suitable acceptor and donor references. In contrast to previous approaches based on sequence composition patterns or phylogenetic disagreements, our methods provide a detection based on sequence comparison and hence offer novel means of evidence. In the last project, we present a method for HGT detection, called Hortense, that is based on proteomic MS data. Hortense extends a standard database peptide search with a thorough cross-validation to ensure HGT properties, and is the first dedicated proteomics HGT detection method. Results from Hortense can also serve as supporting evidence and functional confirmation for HGT events proposed by our genomic-based methods. Taken together, the three HGT methods provide a full view of the transfer event that was not be possible before or with one of the methods alone.

Acknowledgements

First and foremost, I want to thank my supervisor Bernhard Renard for the opportunity to work on this thesis and being a true mentor during the past years. I appreciate that he was always available to give vital advice and encourage me.

I want to extend my warmest thanks to Tobias Marschall for the enjoyable and fruitful collaboration and his willingness to review my thesis.

I started my first PhD project during my masterthesis in the group of Knut Reinert and was lucky enough to have Anne-Katrin Emde as my supervisor back then. She got me hooked on the topic of structural variations and was an inspiration to work with. I also want to thank Knut Reinert for his advise during that time and the whole SeqAn team, especially Birte Kehr who shared an office with me, for that great working environment.

I am grateful to Bernhard who let me further dwell in the area of structural variations and who suggested to extend this to the astonishing phenomenon of horizontal gene transfer. I want to thank my bioinformatics team - past and present - at the RKI, with special thanks to my late office mates Martina, Christine, Tobias, and Thilo. We had sincere and helpful discussions, silly moments, and the spirit here is exceptional. With you, I know I could always do it again. A special thanks goes to my co-authors Anne-Katrin, Christian, Thilo, and especially Enrico and Ben. I had the pleasure of supervising both of you during the projects and you taught me how much it means to me to be able to give back, to share my experiences and knowledge, and watch you grow with your own effort, to then come back and surprise me.

Last but not least, I want to give my warmest thanks to my family, Erwin, my and his parents, my brother and his sister, for always being so encouraging and supportive. I appreciate how much I grew during that time, both personally and on the way to becoming somewhat of an expert in the field of my research, thanks to everyone who supported me and was with me on my journey.

Abbreviations

Abbreviation	Explanation
AA	Amino Acid
AMR	Antimicrobial Resistance
bp	Base Pairs
CNV	Copy Number Variant
DAG	Directed Acyclic Graph
dNTP	Deoxynucleotide Triphosphate
ddNTP	Dideoxynucleotide Triphosphate
DivIVA	A curvature sensitive membrane binding protein
DP	Dynamic Programming
EBI	European Bioinformatics Institute
EHEC	Enterohemorrhagic *Escherichia coli*
FM index	Ferragina-Manzini index
FP	False Positive
GCP	Genome Coverage depth Profile
HGT	Horizontal Gene Transfer
HTS	High-throughput Sequencing
LCA	Lowest Common Ancestor
LCB	Local Collinear Block
MALDI-TOF	Matrix-assisted Laser Desorption/Ionization Time-of-flight
MGE	Mobile Genetic Element
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MSA	Multiple Sequence Alignment
MSRA	Methicillin-resistant *Staphylococcus aureus*
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
PI	Pathogenicity Island
PSM	Peptide Spectrum Matches
PTM	Post-translational Modification
RefSeq	Reference Sequence database of the NCBI
SAM	Sequence Alignment/Map format
SC	Soft Clipped/Clipping
SNV	Small Nucleic Variant
SNP	Small Nucleic Polymorphism
SRA	Sequence Read Archive of the NCBI
STEC	Shigatoxigenic *Escherichia coli*
SV	Structural Variant
TP	True Positive
VCF	Variant Call Format
WGA	Whole-genome Alignment
WHO	World Health Organisation

Contents

1 Introduction

1.1 Integrating omics - From genomics to proteomics

The deduction of the three-dimensional double helix structure of DNA by Watson and Crick (Watson and Crick, 1953), based on the crystallographic work of Rosalind Franklin and Maurice Wilkins (Franklin and Gosling, 1953; Wilkins et al., 1953), lay the foundation of nowadays genomics research. The DNA of each cell in every DNA-based organism contains the hereditary information of the individual organism and the order of the nucleotides in the double helix determines the blueprint for all kinds of cell functions and biochemical properties. The size and composition of a genome varies largely between organisms: The human genome, e.g., has 3.2 billion bases and is structured in stretches of coding and non-coding segments, where the coding sequences - the genes - consist of introns and exons. Bacteria are way smaller, the *Escherichia coli (E. coli)* genome, e.g., has only about 4.5 million bases, without a distinct exon-intron structure.

The genome with its genes is important concerning inheritance and foundation for all processes within a cell. But it is the proteins that are the functional units in a cell to facilitate cell functions. Each protein is created in the cells' ribosome. The blueprint lies in the genomes' DNA sequence located in the nucleus. To get the blueprint from nucleus to ribosome, a DNA transcript made of RNA is created and send as a messenger to the ribosome, therefore called messenger RNA (mRNA). There the mRNA is translated into the corresponding protein code made of amino acids and a protein is synthesised. It was again Crick in 1958 who summarised the genetic information flow from DNA over RNA to protein, a concept later denoted as the "central dogma of molecular biology" (see also Figure 1.1). This dogma has been revolutionised since (Portin, 2014). RNA has manifold additional functionality, e.g. for gene regulation, and further important mechanisms are continuously discovered.

Many research areas and technologies spin around the three fields of genomics, transcriptomics and proteomics. But also the interplay of these fields, like e.g. proteogenomics, with integrative methods is important: Results from genomics methods unravel the tremendous potential but only proteomics tells us what genes are actually expressed at a particular point in time, cellular location or other circumstances.

All of these fields are relevant in some way or the other for public health. The most common definition that has also been taken up by the WHO, and that is also one of the primers at the Robert Koch Institute, is the one by Acheson in

1988: Public Health is ”the art and science of preventing disease, prolonging life and promoting health through the organized efforts of society”. The overall vision for public health stated by the WHO extends this aim to be sustainable and to also reduce inequalities. The WHO also emphasises that public health means not only treatment or eradication of diseases but strengthening all interdisciplinary efforts to improve the entire spectrum of health and disease.

The focus among these many facets of public health within the scope of this thesis is on the aspects of research of human health concerning inheritance, cancer and infectious diseases from genomic and proteomic data with the help of bioinformatics methods. A common important problem in both human health and infectious diseases from viruses and bacteria is variant analysis, especially structural variations (SVs). SVs can cause inheritable diseases and are a prominent phenomenon in cancer genomes. SVs are also due to the large variability in viruses and bacteria, where they facilitate also pathogenic potential in both bacteria and viruses, or are patterns of viral recombination. In bacteria, a special type of SV is caused by horizontal gene transfer (HGT), a concept where genes are transferred between different species. These genes can carry important functions like, e.g., antibiotic resistances.

The core task of bioinformatics is to derive actionable information from the tremendous amount of biological data. The goal within the scope of SVs relevant for public health is to develop a method for SV detection that is generic enough to be applicable to several types of SVs in human inheritance diseases but also applicable to SVs in bacteria or viruses. Here, the focus is on HGT events in bacteria as an example for a specific SV. To show the importance of integrative analysis of genomics and proteomics, we investigate HGT events further by providing methods that facilitate complement approaches and results for HGT events from next-generation DNA sequencing and mass spectrometry-based proteomics. The application of these SV and HGT detection methods will help to decipher the general SV landscape and in particular help to enlighten the origin and consequences of HGT events. The foundation for the addressed research questions is explained throughout this introduction.

1.2 Technologies and computational methods for genomic and proteomic research

The era of genome sequencing brought new technologies and opened up a completely new field of research and computational method development. The first complete bacterial genome was reported in 1995 (Fleischmann et al., 1995), shortly followed by the first eukaryotic genome in 1996 (Goffeau et al., 1996). The first report of a human genome draft in 2001 (Lander et al., 2001; Venter et al., 2001) was a turning point for research of human genomics. It took another three years before

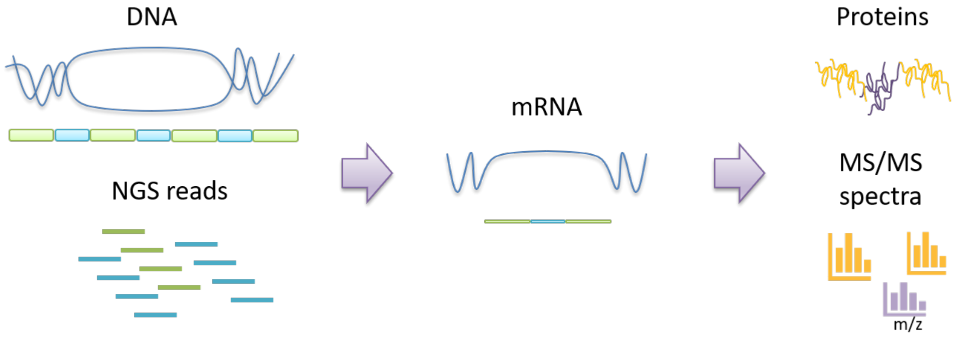


Figure 1.1. Central dogma of molecular biology and the corresponding technologies and data considered within the scope of this thesis. DNA is transcribed into mRNA and then translated into protein sequences. Next-generation sequencing (NGS) means produce billions of short DNA fragments, *reads*, that reveal potential features and gene of the sequenced organisms. Proteins can be captured by mass spectrometry (MS) technologies and the produced spectra of the peptides of each proteins can be analysed to reveal the expressed proteins of a cell under defined conditions.

the complete finished genome was reported in 2004 (International Human Genome Sequencing Consortium, 2004), further updates - published under the 1000 Genomes Project Consortium - of it are still used as the common reference genome for human sequencing data. Subgroups of the 1000 Genomes Project also aim to bring light into the structural variation landscape (Mills et al., 2011; The 1000 Genomes Project Consortium., 2012, 2015; Sudmant et al., 2015). In addition to the 1000 Genomes Project, several more consortia covering a diverse spectrum of fields and applications provide valuable resources for the research community, e.g. the Genome of the Netherlands project (Boomsma et al., 2013) (see also Table 2 in Reuter et al. (2015)). While the cost for the first sequenced human genomes ranges between a half and one billion dollars and was only done for research purposes, nowadays efforts go toward the vision of personalised genomes that offer precision medicine at low costs per genome. The National Human Genome Research Institute (NHGRI) tracks these costs and estimates the current cost at around 1000 dollars (Institute, 2018).

The clinical relevance and public health importance of genome sequencing is manifold. Starting from fundamental basic research to understand evolution, heredity and functionality of a genome, the list of possible applications is seemingly endless with, e.g., studies of alzheimers disease, diabetes, or disfunctionality in cancer in human, viral and bacterial transmission pathways and pathogenicity, infections from parasites and fungi, resistance characterisation and outbreaks of bacteria and so forth.

At the same time, genome characterisation can only reveal the potential of any organism whereas proteomic methods help to elucidate the functional impact from that potential (Tyers and Mann, 2003). In recent years, mass spectrometry (MS) has seen tremendous improvements both technology and computational analysis methods wise that allows for sensitive detection of disease relevant proteins (Radhouani et al., 2012; Lima et al., 2013; Van Oudenhove and Devreese, 2013). Matrix-assisted laser desorption ionisation–time of flight (MALDI-TOF) based proteomics is seen already in clinical applications for several years, e.g., in cancer diagnostics (Kriegsmann et al., 2014) or fingerprinting of pathogens for classification and identification (Wang et al., 2014; Dworzanski and Snyder, 2005). Increasing demands to understand and analyse antimicrobial resistance prompted a rising number of proteomics studies that investigate resistance properties and mechanisms (e.g. Pérez-Llarena and Bou (2016); Tomazella et al. (2012); dos Santos et al. (2010)). With tandem MS, one can further infer the amino acid sequence of proteins by a database search against *in silico* peptides (see also Figure 1.2).

Many technologies and computational methods exist to analyse genomics- and proteomics-based research questions. This thesis focusses on methods for the analysis of highthroughput sequencing and shotgun proteomics.

1.2.1 From Sanger to high-throughput sequencing technology

The first sequencing attempts were done on RNA molecules from, e.g., bacteriophages, as these were already single-stranded and comparatively shorter than genomic DNA. In 1965, 12 years after the deduction of the double helix structure, Holley et al. produced the first whole nucleic acid sequence from the alanine tRNA from Saccharomyces cerevisiae (Holley et al., 1965). In 1977, Frederick Sanger and colleagues developed a related technique now commonly referred to as *Sanger Sequencing* that is based on the natural chain elongation by a DNA polymerase during replication. In Sanger's "chain-termination" or dideoxy technique, dideoxynucleotides (ddNTPs) are used for chain reaction instead of dNTPs. Compared to dNTPs, ddNTPs lack a hydroxyl group that is necessary to perform a bond with the next dNTP to produce a chain. Hence, chain formation is terminated after a ddNTP (hence "chain-termination" technique). Via combination of multiple but separate reactions using radio-labelled - and later fluorescently labelled - versions of all four types of ddNTPs, the single chain reactions terminate at different positions according to the ddNTP used for that reaction and the inserted base can be inferred. If a wrong ddNTP has been inserted during sequencing, the variation from the reference due to the wrong insertion is called a sequencing error in order to distinguish them from naturally occurring point mutations (see section about structural variations below). Sanger sequencing usually produces high-quality reads with a low error rate.

Sanger Sequencing, especially after further technical improvements, became the hallmark of the first generation sequencing techniques and was also used to sequence the first human genome, which was estimated to cost 0.5–1 billion dollar (see, e.g., Heather and Chain (2016), Metzker (2010) or Morey et al. (2013) for a comprehensive review on the history of sequencing).

Sanger Sequencing, however, is limited in throughput, high cost, and read length. The upper length boundary for Sanger Sequencing is 1 kb, which is too short to span large scale rearrangements (see Chapter 1.3). The low throughput combined with high cost prevented Sanger Sequencing to be established for diagnostic purposes and personal medicine. The NHGRI's Advanced DNA Sequencing Technology program in 2008 (Schloss, 2008), that aimed to reduce the cost to $ 1,000 per human genome, set the starting point for emerging high-throughput sequencing (HTS), later also termed next-generation sequencing (NGS), technologies.

The general procedure of all technologies consists of a template preparation step followed by clonal amplification, and then cyclical rounds of massively parallel sequencing that provides the high sequencing depth (Reuter et al., 2015). Illumina is the mostly used sequencing technology throughout the NGS community. For a while, only 454 Life Science (later taken over by Roche Diagnostics) could compete with Illumina and hence, most methods and tools - including those developed within this thesis - were developed tailored to the sequencing properties of both technologies. Albeit the most prominent technology, Illumina did not produce the first sequencer on the market.

454 Life Science released the first NGS sequencer in 2005 (Margulies et al., 2005), i.e. even before the NHGRI program. The sequencing-by-synthesis reaction couples base insertion with luciferase-based light emission where one of the four dNTP types is added at a time to correlate the emission to the correct base. A camera records the emitted light and the inserted base is inferred with the signal strength being - theoretically - proportional to the number of bases. To start the next cycle, the unused dNTPs are washed away and a new dNTP mixture is added. This process shows disadvantages when synthesising the same base multiple times in a row, so called *homopolymers*, as the exact number of inserted bases is hard to determine from the light intensity profile when the homopolymer reaches a certain length (normally around 8 bp). This produces the typical error profile of small insertions and deletions of 454 reads (Metzker, 2010). Owing to the advantages of the longer, albeit single-end, reads of 400-500 bp, 454 was thought to dominate the future market. Unfortunately, Roche shut down the sequencing unit in 2013 deeming the technology not competitive.

Illumina (after acquiring Solexa) released the first sequencer in 2006 (Reuter et al., 2015). In Solexa/Illumina machines, the template strand is sequenced in a cyclic process of fluorescently labeled base insertion, imaging, and cleavage, based on a concept invented by Canard and Sarfati (1994). Compared to 454, only one nucleotide is inserted at a time using reversible terminator bases. After imaging, the dye and terminal blocker are removed and a new cycle starts. This process avoids homopolymer errors but produces typical error profiles of single base substitution errors if a wrong base is being inserted.

During DNA library preparation before the actual sequencing, DNA is first cut into many small pieces (hence also termed shot gun) that are rigorously amplified by PCR and then hybridised in a so-called bridge amplification step onto a flow cell, creating clusters of billions of fragments. Since the fragments in each cluster are the same, the fluorescent signal during sequencing is amplified according to the large fragment number, thereby multiplying the signal. Depending on the size of the flow cell and amplification rate during PCR and bridge amplification, this allows for billions of reads being sequenced at the same time - giving the technique the description of massive parallel sequencing.

Error rates of Illumina reads are still commonly below 0.1%. Reads produced by Illumina technologies were much shorter in comparison to Roche/454 pyrosequencing reads with ∼75 bp length in the first years and up to 300 bp today. Their smaller error rate made downstream analysis easier and more robust though. In later sequencers, introduced paired-end sequencing. Here the amplified fragment is sequenced from both sides in two sequential steps, creating a so called forward and reverse read. Because forward and reverse read are sequenced from the end towards the middle, both have opposite orientations. Both reads have distinct adapters such that they can be separated afterwards. Depending on the fragment and read sizes, there is a gap of defined size between forward and reverse read that can be used, together with the fixed orientation of forward and reverse read, by downstream analysis to determine the correct origin in case of ambiguities. Detected deviation from the expected insert size or read orientation can also be a sign for structural variations (see Section 1.3).

Other important sequencing technologies such as Ion Torrent, that also does sequencing-by-synthesis like 454 but where the base insertion is detected through change in pH level (Rothberg et al., 2011), or ABI's SOLiD (Valouev et al., 2008), are not the focus of this thesis and are well covered elsewhere (e.g., Metzker (2010); Morey et al. (2013)). Sanger is still a valuable alternative to NGS. Albeit the low throughput, Sanger reads have a high quality, much higher than NGS reads, and are up to 1 kb long. They are therefore often used for targeted validation on top of NGS results. For some applications where a high quality sequence of short fragments without high throughput is required, Sanger is still the preferred sequencing technology.

Despite the huge breakthroughs NGS has brought, there are caveats and limitations that hamper NGS as a one for all strategy especially in diagnostic settings where high quality results are necessary. During sequencing, PCR artefacts can result in chimeric reads. Short fragments and regions without GC bias result in a better amplification frequency, and hence an over- or underrepresentation of certain fragments. NGS technologies still provide rather short reads - even shorter than Sanger Sequencing. Reconstructing or assigning the short reads is difficult and error prone in complex or repetitive regions (Snyder et al. (2010), see also Chapter 1.3). Haplotype resolution is nearly impossible as long range connectivity information between variants is lost. Existing, highly computationally intensive methods are only feasible in smaller regions for now (Töpfer et al., 2014).

In recent years, so called third generation sequencing technologies are revolutionising the sequencing market again. Compared to the high amplification of many short fragments, third generation sequencing technologies amplify fewer, single but long molecules. Currently, the single-molecule real-time (SMRT) sequencing technology by Pacific Biosciences (PacBio) (Eid et al., 2009) and the nanopore sequencing technology by Oxford Nanopore Technologies (Clarke et al., 2009) dominate the market. Both produce reads up to several thousand base pairs, i.e. much longer also compared to Sanger sequencing. However, compared to the Illumina error profile, third generation error rates are even worse and are still around 11%. Similar to the 454 error profile, errors consist to a large amount of insertions and deletions. The long reads are able to span SVs or also complete virus genomes that are prone to recombination, but the high error profile deems them not suited for SNP and short indel discovery ($<$ 15 bp) (Chaisson et al., 2017). Hence, despite the advantage in read lengths, this high error rate together with high sequencing costs has prevented these technologies from taking the lead on the sequencing market so far. But with better - and also cheaper - developments these technologies are slowly coming of age (Ardui et al., 2018).

Independent from the technology of choice, sequencing reads can be either *de novo* assembled to reconstruct the original sequence, or they can be aligned - also called mapped or assigned - to a reference sequence. There are many possible fields and applications for NGS, including transcriptomics, genome structure characterisation, or population scale analysis, and many more that have been reviewed elsewhere. The focus within this introduction is on methods and applications relevant for research questions addressed in this thesis.

1.2.2 Methods for assembly of sequencing reads

Given a set of reads from a genome, the goal of a *de novo* assembly is to reconstruct the complete original genome. This is usually done if there is no reference sequence available or the reconstruction is in the interest of the research question. Denovo assembly is usually done graph-based to ensure feasibility. A graph consists of nodes and edges connecting the nodes. These edges can be either undirected, i.e., they can be passed in both directions, or directed, i.e., only one direction is allowed to pass. A *path* through a graph between two nodes is then the collection of all nodes and edges that are passed through on the way from the first to the second node. If there are two nodes marking the start and end of a genome, a complete genome assembly then corresponds to a path from start to end. To create the graph, either nodes or edges are labelled with sequences.

The most common assembly graph is the de Bruijn graph. Here nodes are labelled with sequences of length k, also called *k-mers*. Each unique k-mer has only one node, i.e. each sequence containing this k-mer is represented by and passing through this node. Nodes are connected by a directed edge if their sequences overlap by a length of $k-1$. Pevzner et al. (2001) proposed an algorithm to construct a de Bruijn graph from short read sequencing data, and all following assemblers build up on that. Velvet (Zerbino and Birney, 2008), SOAPdenovo (Li et al., 2009b) and EULER-SR (Chaisson and Pevzner, 2008) were among the first assemblers based on de Bruijn graphs. Velvet is still one of the mostly used short read assemblers today. ALLPATHS (Butler et al., 2008) and ALLPATHS-LG (Gnerre et al., 2010) use a variation of the de Bruijn graph notion, an unipath graph, for assembly representation. Consecutive, non-branching collections of edges (and nodes) are collapsed to produce so called unipaths (i.e. path without a branch). To facilitate assembly for large genomes, paired-end reads are joined in ALLPATHS-LG if they overlap to produce longer reads with a longer range connectivity. ABySS (Simpson et al., 2009) uses a distributed representation of de Bruijn graphs that allows parallel computation of billions of reads. One of the most recent and also still most widely-used assemblers is SPAdes (Bankevich et al., 2012).

It has been shown, however, that the quality of an assembly not only depends on the chosen assembler but also on read sequence quality and the sequenced genome (Salzberg et al., 2012). So there is no simple answer to the question of what is the best assembly tool or method. Tools like QUAST (Gurevich et al., 2013; Mikheenko et al., 2018) aim to provide comprehensive statistical assessments and quality metrics to help evaluate and compare assemblies. The Assemblathon competition (Earl et al., 2011) marks the first benchmarking efforts but, unfortunately, assembler parameters and settings were not made available to serve as a guide for the user. Salzberg et al. (2012) offered a more usable comparison by publishing all settings and preprocessed data sets. Still, this can only be a static snapshot and the comparison becomes outdated with recent software versions, new tools or advanced sequencing technologies.

Limitations of all assemblers are repetitive regions and heterozygosity for polyploid organisms, where the assembler cannot uniquely resolve read or segment/contig order. Most assemblers are tailored to the deep sequencing coverage and low error profile from the short NGS reads and are often not suited for long read sequencing data. Pacific Bioscience offers their own assembly tools as part of the SMRT Suite (see SMRT Analysis). Canu (Koren et al., 2017) is an open source alternative derived from the Celera Assembler (*wgs-assembler*, Myers (2000)). So called hybrid assembler such as SPAdes incorporate short read information from NGS data into the assembly to correct the long reads. In addition, tools such as LoRDEC (Salmela and Rivals, 2014) provide methods for long read correction using NGS reads.

Assembly is still not a solved task and draft genomes from automatic assembly have to be considered with great care as especially repeats can lead to - in the worst case undetected - misassemblies (Salzberg and Yorke, 2005).

1.2.3 Methods for read assignment and read mapping

Many research questions do not require the complete genome to be fully assembled if a meaningful reference genome is already available. Determining the origin of the read in the reference - called *mapping* - is hence one of the fundamental bioinformatics tasks. Mapping or aligning of reads to a reference genome also establishes similarity - or distance - information between the sequenced organism and a reference sequence.

Similarity information is usually used for identification (see also next subsection for metagenomic applications) or characterisation of gene content, strain level similarity and so on. Distance information is usually further analysed in follow-up variant detection methods that is used for all kinds of functional analysis and characterisation (see next section 1.3).

Read mapping usually requires an *alignment* of the read to the reference where every base of the read is matched to a part of the reference. In an error-free alignment, no mismatches or missing bases - called *gaps* - are allowed. However, errors in terms of wrong bases or indels (missing or additional bases) are likely introduced by the sequencer or present due to evolutionary distance between reference and sequenced sample. An error-free alignment would hence often result in a poor amount of reads being mapped. Alignments can be determined by dynamic programming (DP), where the best path, under a given a score model, of matching bases from read and reference or introduced gaps is traced in a matrix, e.g., using the Smith-Waterman (Smith and Waterman, 1981) or Needleman-Wunsch (Needleman and Wunsch, 1970) algorithm. These methods produce very exact alignments but are computationally very expensive, which deems them not suited for applications for large amounts of NGS data.

If the exact alignment is not needed for further analysis because one is only interested in the origin, i.e., position of the reads, one can fall back to read assignment methods, which are much faster. These are alignment free or rather do a pseudoalignment, e.g. based on k-mers (see below). Mapping can be seen as a trade off between alignment and assignment: common read mappers first identify candidate positions for the reads and then align them but without exploiting a full DP matrix. A plethora of mappers exists with different strategies: Some are based on a two-pass strategy of first indexing and then lookup. Some do a full alignment including exact positions to determine the best or all mappings of a read, e.g, Stampy (Lunter and Goodson, 2010), MOSAIK (Lee et al., 2014), BWA and BWA-MEM (Li and Durbin, 2009, 2010), Bowtie and Bowtie2 (Langmead et al., 2009; Langmead and Salzberg, 2012), Yara - the follower of MASAI (Siragusa et al., 2013), SOAP and SOAP2 (Li et al., 2008, 2009a). Read mappers can be grouped according to their indexing strategy: some use hash tables, or prefix or suffix trees, often using the Ferragina-Manzini (FM) index (Ferragina and Manzini, 2000) based on the Burrows-Wheeler transformation (Burrows and Wheeler, 1994).

To improve upon runtime, heuristics are often employed. Paired-end sequencing was developed to overcome the limitations from single reads: in regions with local similarities or regions characterised by structural variations, a single read can map ambiguously to multiple locations, or wrongly, or not at all if the read crosses the boundary of such SVs. Methods designed to exploit these signatures are able to detect these SVs and are explained in more detail in Section 1.3. The number of available read mappers is seemingly endless but still new mappers are being published that are either faster or more memory and space efficient, or that are tailored to specific applications, e.g., RNA-seq specific mappers that account for exon-intron structure in eukaryotic genomes, or organisms, e.g. viruses that have a much higher variability such that the mapper has to account for more possible errors but that are also smaller, which makes the utilisation of more expensive algorithms feasible.

1.2.4 Methods for metagenomic applications

In contrast to a sample with a single organism produced in the lab, metagenomics studies environmental or human samples with a complex ensemble of different communities. A prominent example is the human microbiome, its characterisation revealed startling insights into the influence the gut microbial composition has on human health and also disease traits. It is also highly depended on personal dietary characteristics, drug intake and personal life style, even to that extend that every individual has practically his or her own microbiome. Especially the gut microbiome is also a hotspot for horizontal gene transfer (Liu et al., 2012).

The number and quantity of organisms in the sample is often not known in advance, which presents additional challenges to the classical assembly and mapping problems. Hence, reads are generally mapped or compared to databases of various complexity instead of single references. Hence, exact read alignments are often not feasible. Depending on the sequence similarity of the organisms within the sample, reads could be mapped to more than one candidate reference. Also, despite a reasonable total sequencing coverage, the single organisms in the sample are usually only low and locally covered. First methods, when only short reads from 25-76 bp could be achieved, focussed on 16S rRNA regions in bacteria. Longer reads of several hundred base pairs make whole genome analysis possible. Methods to analyse metagenomics samples can be grouped into (i) identification or classification, and (ii) quantification or abundance estimation methods (Breitwieser et al., 2017). Identification approaches aim to classify the present organisms in the sample. Based on that taxon classification, abundance estimation methods aim to quantify the taxa in the sample in relation to each other. Tools for abundance estimation include, e.g. GASiC (Lindner and Renard, 2013), MetaPhlAn2 (Truong et al., 2015), Bracken (Lu et al., 2017), kallisto (Bray et al. (2016), Schaeffer et al. (2017)) or DiTASiC (Fischer et al., 2017).

Metagenomic classification methods can be taxonomy dependent or taxonomy independent (Lindgreen et al. (2016), Sedlar et al. (2017)). In both cases, reads are *binned* into groups where reads in one group belong to the same organism. Taxonomy independent approaches bin the reads by means of clustering methods based on sequence features. Taxonomic dependent binning approaches assign the reads to specific taxonomic groups, thereby inferring the occurrence of these taxa in the sample. Similar to general read mapping and assignment strategies, these methods use sequence composition patterns such as k-mers, e.g. Kraken (Wood and Salzberg, 2014), or they assign reads based on mapping similarities. MEGAN (Huson et al., 2007) is one of the first alignment-based taxonomic profilers and assigns reads to the ”lowest common (taxonomic) ancestor”. Compared to that, DUDes (Piro et al., 2016) aims for the ”deepest uncommon descendent”. Although the latter approach may induce ambiguous taxonomic assignments, the liberal approach allows for identifications in lower taxonomic levels. MicrobeGPS (Lindner and Renard, 2015) tries to account for species not present yet in a reference database. Instead of assigning reads to fixed taxa, MicrobeGPS tries to infer the distance of the reads to references in the database. If no reference fits - more or less - perfectly, the distances are used to describe the composition of the organism in the sample. In addition, it also provides an abundance estimation.

The tool Clinical Pathoscope (Byrd et al., 2014) focusses on identifying pathogens in clinical samples such as blood where a high amount of host contamination can be expected. Because the human genome is quite large compared to the smaller viral or bacterial genomes, the amount of reads from the human genome can dominate the sample after a sequencing run, which makes identification even more difficult. Hence, it is important to remove host contamination to then do a reliable pathogen detection. Like MicrobeGPS, Clinical Pathoscope also identifies and reports the nearest relative in case the sample contains novel or highly mutated strains. But since Clinical Pathoscope is tailored for pathogen identification, it also removes reads from non-target and non-pathogenic genomes.

Due to all the challenges mentioned above, metagenomic identification methods were for a long time not able to reliably resolve composition beyond species level to identify strains. Strains of the same species can have crucial differences when it comes to the presence of anti-microbial resistance genes or pathogenicity properties and is hence important for diagnostic applications. Ditasic and kallisto are a few methods to have shown to be able to identify on strain level.

1.2.5 Mass spectrometry-based proteomics

Proteomics comprises the study of all proteins - the proteome - and their isoforms together with their structures, modifications, functions and interactions (Tyers and Mann, 2003). Mass spectrometry (MS) is one of the fundamental methods in proteomics to identify proteins by deduction of their amino acid (AA) sequence.

The first step is usually the digestion of proteins to peptides using a protease such as Trypsin. The peptides are then separated, e.g. on a high-performance liquid chromatography column, and ionised to be analysed in the MS. The analysis process can be either one-step, where information is analysed on the so called MS1 level, or two-step, also termed tandem MS or MS/MS or MS2 (Steen and Mann, 2004).

A prominent method to process information on all levels is matrix-assisted laser desorption ionisation–time of flight (MALDI-TOF) (Karas and Hillenkamp, 1988). With the help of a matrix of crystallised molecules (matrix-assisted), peptides are first fixed onto a plate and then vaporised (desorption) with the help of a laser into a co-crystalised, ionised state. The ionised molecules are accelerated in the MS and their time of flight (TOF) is measured. From the TOF, the mass-to-charge ratio of the molecule can be induced, and from that the protein can be identified (Dworzanski and Snyder, 2005). A prominent application of MALDI-TOF is fingerprinting on the MS1 level. An alternative to MALDI is electrospray ionisation (Fenn et al., 1989). Here, the separated peptides from the column are electrostatically dispersed by a high electrical potential that also causes the ion containing droplets to solve over time. The analyte ions are again analysed in the MS by, e.g., TOF. MALDI-TOF is usually used for intact, single large molecules where the identification is sufficient and resolving the AA sequence not needed (Steen and Mann, 2004).

Tandem mass spectrometry (MS/MS) proteomics is capable of deciphering AA sequences on top of protein identification. This in turn allows the deduction of functional properties and taxonomical classification (Muth et al., 2016).

1.2.6 Tandem mass spectrometry

Methods for MS/MS have gained a growing popularity that also helped to improve protein identification (Tyers and Mann, 2003). After a peptide was analysed at the MS1 level, it can be further isolated and analysed in a second round of MS. The peptide ion is broken apart through collision with an inert gas, and the peptide sequence can be inferred from the mass spectrum of the resulting fragments, the MS/MS spectrum (see also Figure 1.2) (Steen and Mann, 2004). The isolated peptide ion before the collision is often called the *precursor ion*, the fragmented ions measured in the tandem MS are called *product ions*. The peptide spectra are searched in a database of *in silico* spectra from *in silico* peptides.

This general procedure of peptide digestion, separation and analysis in the MS

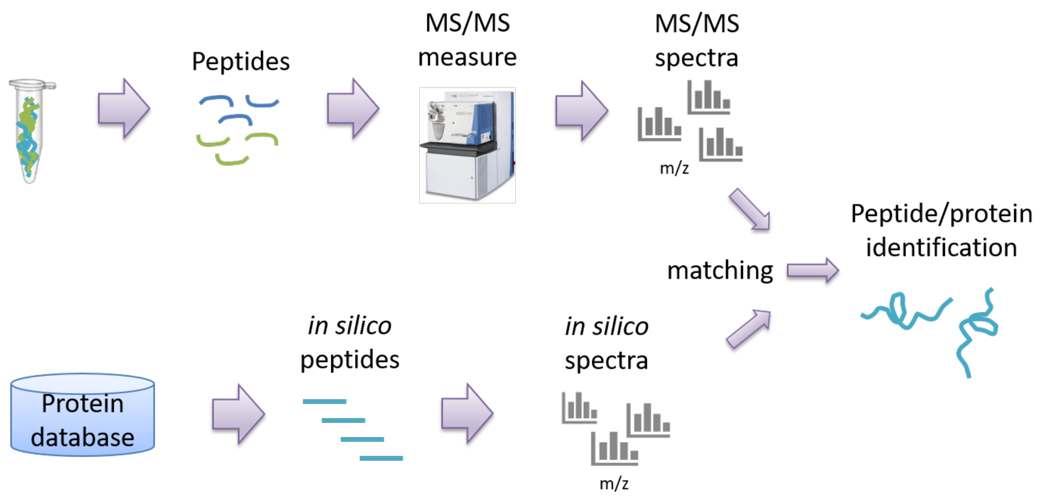


Figure 1.2. Tandem mass spectrometry peptide and protein identification. The spectra of peptides from a measured protein can be compared to a database of spectra from *in silico* peptides for peptide and follow-up protein identification.

is also called bottom-up MS (Karlsson et al., 2015). Disadvantages of the bottom-up approach are miscleavages during digestion, or undetected peptides that are lost during chromatography or not detected by the mass spectrometer. This can lead to a low total protein coverage, i.e. the protein may still be identified but possible post-translational modifications (PTMs) or isoforms may remain hidden. Alternative isoforms or PTMs render identification difficult as they change the mass of the protein and, hence, alter the proteins mass-to-charge ratio in the spectrum. On the one hand, this can be used to detect such modifications. Ambiguous spectra and mass-to-charge ratios, one the other hand, lead to false positive identifications.

One alternative to the bottom-up approach is the top-down approach (Chait, 2006). In the top-down approach, the protein is not cleaved before analysis, which preserves the complete structure including modifications. So application-wise and concerning a possible resolution, top-down is superior to the bottom-up strategy. On the downside, a complete protein is harder to analyse during chromatography and in the mass spectrometer since complete proteins differ much more in their properties such as solubility, size and detectability. Hence, there is no fit-all treatment for all proteins like it is possible for all peptides in the bottom-up approach. Due to the resulting low throughput, top-down analysis has been mainly used to analyse single proteins. A more practical alternative is targeted proteomics. Here, the goal is to focus on a few - targeted - proteins instead of aiming for a broad discovery. This is done by a target-specific assay that selects fragments according to the defined ratios in the MS (Doerr, 2013). Targeted proteomics is especially applied to detect bacteria (e.g. Peters et al. (2016); Ebhardt et al. (2015)).

1.2.7 Computational methods for tandem mass spectrometry data analysis

Much like what de novo assembly and mapping is for NGS reads, peptides and their sequences can be inferred de novo from the spectra, e.g., with PepNovo (Frank and Pevzner, 2005), or the spectra can be searched against databases of known proteins and peptides that allow inference of their presence given certain threshold criteria. De novo sequencing is dependent on high quality spectra and needs to be able to resolve modifications, which would be achieved best by having different fragmentation modes (Guthals et al., 2013). The mass difference between peaks in a MS/MS spectrum corresponds to a particular amino acid (AA). Except for leucin and isoleucin, all AAs have distinct masses and from all peaks and their mass differences, the AA sequence can be inferred. In a low quality spectrum, noisy peaks can lead to false positive inferences. Compared to de novo sequencing, database search approaches are still more reliable and widely applied (Muth and Renard, 2017).

A database contains a comprehensive list of theoretical spectra from *in silico* digested peptide sequences from all open-reading frames (Vaudel et al., 2012). That means, the contained sequences are way smaller than genome sequences, namely only few AAs. Hence, the number of entries is huge and the chance of a random match, especially considering modifications such as PTMs, quite high. In addition to that, many proteins share certain peptide sequences, e.g. from housekeeping genes, which requires rigorous quality assessment and scoring of the PSMs to infer which protein is actually present. Popular search engines for database search are, e.g., MS-GF+ (Kim and Pevzner, 2014) and X!Tandem (Craig and Beavis, 2004) (see Muth et al. (2016) for a holistic list of database search algorithms). PeptideShaker (Vaudel et al., 2015) is an extensive software suite that provides reanalysis methods by integrating various established tools like MS-GF+.

For downstream analysis such as taxonomic classification, a number of approaches and tools is available (McHugh and Arthur, 2008), e.g., BACid (Jabbour et al. (2010, 2011), based on work by Dworzanski et al. (2004)). TCUP (Boulund et al., 2017) aims to quickly determine the presence and characteristics of disease causing bacteria. To do that, TCUP facilitates a database comparison to then automatically detect peptides from which it is possible to characterise not only taxonomic composition but also expressed antimicrobial resistance genes.

Akin to the metagenomics application of NGS, there are methods for species identification from tandem mass spectrometry data of metaproteomic samples, e.g., MetaProteomeAnalyzer (MPA) (Muth et al., 2015, 2017) or Pipasic (Penzlin et al., 2014) that also aims to estimate the relative abundance of the species in the sample.

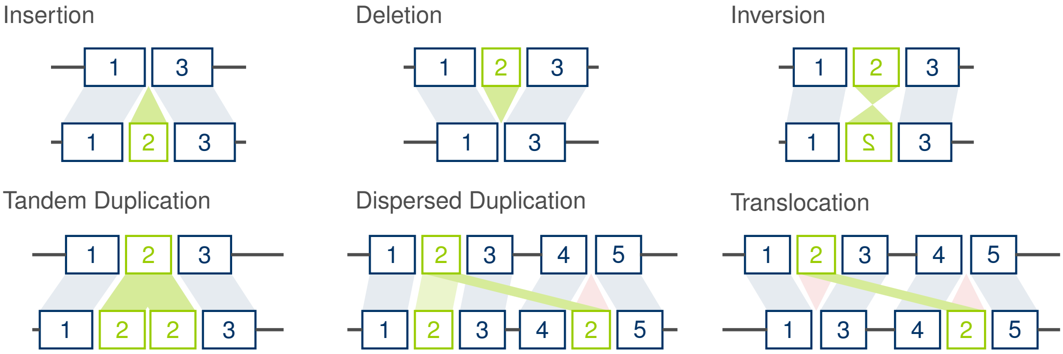


Figure 1.3. Structural variations: Shown are possible variations of two genome sequences with segments labelled 1-5. In case of an insertion, e.g., the first genome has segments 1 and 3 (blue), the second genome below an additional segment 2 (green). The breakpoint of the insertion is between segments 1 and 3. Insertions, deletions and inversions are classified as simple SVs as they have only one or two breakpoints indicating novel adjacencies. Complex SVs like tandem or dispersed duplications or translocation have more than two (reused) breakpoints. A translocation can also be defined as the combination of two simple SVs, namely a deletion signature of segment 2 between segments 1 and 3, combined with an insertion signature of segment 2 between segments 4 and 5. The dispersed duplication is missing the deletion signature of segment 2.

On top of information gained from genomics, metaproteomics allows insides into the functional role that the individual members of an economic community play (Wilmes and Bond, 2006). Muth et al. (2016) comprehensively review computational approaches and means for metaproteomic data analysis.

While there are studies that investigate the functional impact and expression of potentially acquired proteins, e.g., Tomazella et al. (2012), dos Santos et al. (2010), or Sirichoat et al. (2016), structural variations in general and HGT events in particular have not been directly addressed by methods or studies so far.

1.3 Structural variations in the human genome

Not every individual genome has exactly the same sequence, not even the genomes of twins. Alterations in sequence composition are hence a natural consequence of evolution and the basis for the large variety in shape, size, colour and more between individuals of the same species. But some alterations can lead to all kinds of diseases, intolerances, resistances and so forth. Hence sequence variations nurtured a whole field of research trying to decipher the meaning behind sequence alterations. Variations are usually defined in comparison to a certain reference sequence. The smallest variations are single nucleotide variations (SNVs), i.e. single bases are mutated between 'A', 'C', G' or 'T' compared to a reference. If a SNV is not only seen in a single individual but manifests in a part of the population, it is referred to as a single nucleotide polymorphism (SNPs) (usually the threshold is 1% of the population). Next to mutations, stretches of nucleotides can be missing between otherwise similar genomes. Depending on the declaration of the reference sequence, these stretches are defined as *deletions* or *insertions* or, more general, as *indels*. These indels comprise one up to few bases, with 'few' currently being defined as around 50 bp. To distinguish them from larger insertions and deletions over 50 bp in the context of structural variations (SVs), they are also often referred to as *small* indels. Compared to SNPs and small indels, SVs or rearrangements affect up to several thousands of base pairs and vary largely in shape and composition (see below for detailed definitions). Detection methods and discoveries around SNPs and small indels are reviewed elsewhere (e.g., Li et al. (2012b); Altmann et al. (2012); Kumar et al. (2014); Mielczarek and Szyda (2015)). At a local scale, SNPs are studied to characterise diversity in genes or regulatory regions that lead to a broad variety in functions. The study by Timmermann et al., e.g., is the first to analyse whole exome 454 sequencing data to identify somatic variants in colorectal cancer (Timmermann et al., 2010). At a global scale across the complete genome, SNPs and indels are exploit to reconstruct haplotypes, also called phasing, that is used in turn to characterise organisms with more than one genome copy (diploid or more) or also to study quasispecies in virus populations (Zagordi et al., 2011; Töpfer et al., 2014; Jayasundara et al., 2014; Barik et al., 2017). The possible resolution for global haplotyping, i.e., the maximal possible reconstructable length, highly depends on the variability of the genome, its ploidy and mostly on the available read lengths. Methods that combine long read information to connect local haplotypes are still to come.

According to results from the different genome sequencing projects (see, e.g., Table 2 in Reuter et al. (2015)), individuals harbour several million SNVs and several hundred thousand indels (The 1000 Genomes Project Consortium., 2010). High-throughput sequencing made it possible to study these variants on a genome-wide scale, identifying pathogenic mutations and linking them to Mendelian diseases (e.g., Ng et al. (2010); Mullaney et al. (2010)).

1.3.1 Relevance of structural variations for human health

With the broadened detection spectrum from modern technologies and methods, SVs were recognised to account for more differences in terms of the amount of bases between individual genomes than the well studied SNP variations (Redon et al., 2006; Alkan et al., 2011; Berger et al., 2011; Baker, 2012). Some efforts like the study by de Koning et al. (2011) estimate that over 50% if the human genome is repetitive, to some extend involving large duplications. Thanks to NGS, it was shown that SVs not only affect more bases than SNPs but also that they are involved in many disorders (Stankiewicz and Lupski, 2010). Lupski et al. (1991), e.g., discovered that a duplicated gene is the reason for the Charcot-Marie-Tooth disease, and Sebat et al. (2007) were the first to report the role that copy number variations (CNVs) play in a complex neuropsychiatric disease. Further associated conditions include autism (e.g. Sebat et al. (2007), Sanders et al. (2011), Levy et al. (2011), Pinto et al. (2014)), schizophrenia (Walsh et al., 2008), Hunter Syndrome (Bondeson et al., 1995), or haemophilia A (Lakich et al., 1993).

It has been also observed that DNA in cancer cells show significant larger amounts of SVs, which tends to also be heterogenic among different cells of the same tumor. SVs in cancer genomes are also referred to as somatic SVs. A direct contribution of SVs for carcinogenesis is given if, e.g., the SVs alter the expression of tumor suppressor or oncogenes (Liu et al., 2015), or create new oncogenes like, e.g., *the BCR-ABL* gene (MacConaill and Garraway, 2010).

Despite the growing recognition of the importance of SVs they are still less well studied than SNPs. This is mostly due to the complex nature SVs can inhabit. Nevertheless, understanding mechanisms of SVs and being able to detect them opens up new paths for early on detection of human health related SVs and therapeutic treatment possibilities.

1.3.2 Definition of structural variations

Due to their complexity, SVs are nowadays commonly defined via their *breakpoints*. Breakpoints define novel adjacencies between former unrelated sequence segments that are now in closer proximity due to the rearrangement events (see Chapter 2 for formal definitions of segments and breakpoints). Knowing the precise breakpoint is important to characterise the novel adjacency, e.g., novel fusion genes that bring novel functions - or disfunctions - with them. Breakpoints can also affect promoter or suppressor functionality, which has been observed, e.g., for some types of cancer. So called *simple* SVs have one or two breakpoints and define large insertions (one breakpoint), deletions, and inversions (both two breakpoints), i.e. segments that are inverted compared to a reference sequence. In contrast, *complex* SVs have three or more breakpoints. Figure 1.3 shows some types of complex SVs: a duplication is a segment that is copied and then inserted at a distinct position in the genome. A tandem duplication is a special type of duplication in the sense that the copy of the segment is located right next to the original segment. A translocation is similar to a duplication event except that the original segment is missing. Signature wise, a translocation can be seen as a combination of a deletion signature of the original segment (two breakpoints) and an insertion event (one breakpoint). Mobile elements - also mobile genetic elements (MGEs) - are elements that change their location within the genome actively and frequently. With respect to the *netto gain* of sequence content, SVs are also categorised as balanced (inversions, translocations) or unbalanced (insertions, deletions, duplications, see Figure 1.3). The extend of complexity is even not fully understood yet. Combinations of simple and complex events, such as inverted duplications or combined deletion and inversion events that share breakpoints, have already been observed. The most severe event observed so far is termed chromothripsis and describes a catastrophic event that results in the rearrangement of complete chromosomes (Stephens et al., 2011).

The variation in size and complexity in turn renders the detection of SVs difficult. The size definition also changed with the development of new technologies that offered a higher base pair resolution for sequence analysis methods. SV sizes starting from > 50 bp where first mentioned in Alkan et al. (2011), but at this point in time no tools were able to detect them (see below for more details). Kloosterman et al. (2015) describe mutation rates for SVs starting even from 20 bp.

1.3.3 Graph representations of rearrangements in whole-genome alignments

Whole-genome alignments (WGAs) are used to represent homologies between genomes and are a prerequisite for downstream analysis such as rearrangement studies. Hence, methods for WGA are designed to handle fully assembled reference genomes and not NGS data directly. Considering rearrangements or SVs, a WGA is not trivial, especially with more complex rearrangements. So called segments of a genome might align well to segments of another genome but they can be in a different order, duplicated or inverted.

Similar to de Bruijn graphs for genome assembly, graphs have proven to be an efficient way to handle the complexity in representing a WGA with rearrangements. A graph G has a set V of nodes and a set E of edges. The edges can be divided in groups of directed and undirected edges and may also contain weights. A path through the graph is then a series of nodes and edges through the graph where undirected edges can be passed in both directions but a directed edge in only one direction. Special graph types have defined properties. A directed acyclic graph (DAG), e.g., has only directed edges and must not contain cycles, i.e. each node may only be passed once on a path. Such properties allow to define special paths such as the shortest path. A shortest path from one node to another has the minimal total weight compared to other possible paths connecting these two vertices, and can be determined by established algorithms such as Dijkstra's algorithm.

Representing a WGA in a graph then works as follows. Matching segments or synteny blocks between two genomes are identified through local alignments as locally collinear blocks (LCBs) that have only SNPs and small indels, and are represented as nodes in a graph. The order and relationship of the segments is then captured by edges between segments within the graph, and each genome is a path through the graph. The edges therefore represent breakpoints from rearrangement events of segments. The amount and kind of constraints for connecting LCBs in the graph make up the most essential differences between the common methods and graph structures for WGA. In Kehr et al. (2014), we compared four common graph data structures, namely Alignment graphs (Kececioglu, 1993), A-Bruijn graphs (Pevzner et al., 2004), Enredo graphs (Paten et al., 2008) and Cactus graphs (Paten et al., 2011b,a). In that paper we could show that these graph structures can actually be converted into each other with regards to labels (Kehr et al., 2014).

It is important to note that a pairwise WGA can only reveal rearrangements apparent from the two compared genomes. In case gene loss events in a genome coincide with other rearrangements, some breakpoints only become detectable in a comparison of more than two genomes. These breakpoints are referred to as *hidden* breakpoints (Kehr et al., 2012) as they are not detectable by pairwise comparisons. Multiple WGAs are computationally expensive since the size of segments and blocks is not known beforehand and also varies depending on allowed errors for the synteny. They may also be in conflict between rearrangements of pairs of genomes. So there usually needs to be a tradeoff between breakpoint and SV type resolution, and computational expenses.

1.3.4 Detection from pre-NGS technologies

Early on, SVs were defined as variants larger than 1,000 bp because smaller variants could not be detected by pre-NGS methods. Before NGS, microarrays have been the state-of-the-art experimental methods to study CNVs, and have been used extensively before they were slowly replaced by NGS (Alkan et al., 2011). Comparative genomic hybridisation (CGH) was the first efficient technique for genome-wide location of CNVs between a sample and a reference. Sample and reference DNA are fluorescence labelled and competitively hybridised to an array (aCGH). The relative intensity signals of the sample and reference DNA for specific locations indicates the corresponding relative copy number for that location (Pinkel et al., 1998). Early aCGH technologies were able to reliably detect CNVs from about 1 kb to several 100 kb. So called high-density arrays extended this range to CNVs starting from 500 kb (Alkan et al., 2011).

Iafrate et al. (2004) were the first to report CNVs in the human genome using aCGH. This technique was used early on to associate copy number variations in cancer (Pinkel and Albertson, 2005) and other diseases. Unfortunately, it is not possible to determine exact breakpoints with microarray techniques, and hence accurate assessment of the altered sequence and gene content is impossible.

Compared to these early methods, NGS-based SV detection made it possible to detect also other types of SVs such as novel insertions and inversions with base pair resolution, resolving variants starting from 50 bp. Albeit this advantage in resolution, NGS was comparatively expensive and only slowly replaced microarray experiments. In 2007, Korbel et al. (2007) presented the first study that uses NGS for SV discovery. Also, the first NGS-based methods focussed on single types of SV like, e.g., indels (Emde et al., 2012) and only later evolved to include multiple SV types.

1.3.5 Methods for structural variant detection from NGS data

When it comes to SVs, especially complex rearrangements such as translocations or combinations of inversions and deletions, even modern read mappers reach their limit to accurately place reads spanning boundaries of such variations. The larger the variant and the more breakpoints it has, the less likely it gets that a single read spans the complete variant. Hence, the core objective of SV detection is to identify all breakpoints and segments belonging to a particular SV.

SNPs and small variants could be detected if they were smaller than the common read length, and hence, for some time there was a size gap between small variants (about <50 bp) and large variants (>1 kb). The most recent methods closed the gap and are able to detect SVs with 50 or more bases, and with this the definition adjusted accordingly. Today, a plethora of methods exists that exploit various signatures and that have hence also different advantages and limitations (Medvedev et al. (2009), Alkan et al. (2011), Abel and Duncavage (2013), Pabinger et al. (2014)). These signatures are discordant read pairs, read depth, and split-reads, and will be explained in detail in the following section.

The first methods focused on analysing discordant read pairs and later discordance in sequencing depth. The study by Tuzun et al. (2005) was the first to implement a paired-end sequencing approach to study SVs (Alkan et al., 2011). As described before, paired reads originate from different ends of the same DNA molecule and therefore have an expected distance and orientation. A discordant read pair is a pair that aligns with aberrant orientation or distance between the reads or to different chromosomes. Although discordant read pairs span at least one breakpoint (albeit not like split-reads, see below), methods exploring these signatures do not identify precise breakpoints. Tools based on discordant read pairs include, e.g., BreakDancer (Chen et al., 2009), PEMer (Korbel et al., 2009) VariationHunter (Hormozdiari et al., 2009, 2010), GASV (Sindi et al., 2009), SVDetect (Zeitouni et al., 2010), and inGAP-sv (Qi and Zhao, 2011).

Read depth refers to the mapping coverage of the reads versus a reference genome. Local pileups or sinks in coverage are indications for duplications or deletions, therefore these variants are also called CNVs. The most common methods exploiting read depth variations include, e.g., SegSeq (Chiang et al., 2008), EWT (Yoon et al., 2009a), CNVnator (Abyzov et al., 2011), CNV-seq (Xie and Tammi, 2009), ReadDepth (Miller et al., 2011), Genome STRiP (Handsaker et al., 2011), and GROM-RD (Smith et al., 2015). These methods heavily depend on the overall sequencing coverage and artificial fluctuations hamper their resolution. Also, balanced SVs such as inversions or translocations cannot be detected with these methods as they do not result in coverage profile changes. The precise breakpoint determination of deletions and duplications is also usually not possible with these methods.

Sequencing reads spanning borders of SVs allow the precise detection of SV breakpoints. A so-called split-read has one part of the read mapping to one site of the breakpoint and the other, with a potential long gap in between, to the other site of the breakpoint at a distant location in the genome. The tool Pindel (Ye et al., 2009) was among the first split-read aligner. It uses the normal mapping read of a read pair as an anchor to limit the search space for the split-read. Early methods arbitrarily split the read in two - sometimes three - parts, assuming that the breakpoint will roughly match the middle of the read, and then try to map the smaller parts of the read. Mapping can be further complicated by microindels around the breakpoint and the error-prone ends of a read. These mapping locations are then reported without breakpoint refinement. Prominent examples are the tools SVSeq (Zhang and Wu, 2011), PRISM (Jiang et al., 2012), CLEVER and MATE-CLEVER (Marschall et al., 2012; Marschall and Schönhuth, 2013; Marschall et al., 2013), and especially Delly (Rausch et al., 2012). There are also dedicated split-read aligner like AGE (Abyzov et al., 2011) or SplazerS (Emde et al., 2012) that solve a DP matrix around the breakpoint to really pin-point the most likely position according to the chosen scoring scheme also under consideration of gaps and mismatches (see Section 1.1 for details on read alignment). These (re-)alignments are, however, usually computationally expensive.

Split-read methods have the advantage, compared to, e.g, methods exploiting read depth fluctuations, to be able to precisely determine the breakpoint of an SV. Having only the exact mapping information from a fraction of the original read, however, made it a difficult task to really unambiguously map the read part to its origin, especially considering repetitive regions. Split-read approaches without realignment are also hampered by the presence of microindels around breakpoints. This was a particular issue in the first phase of NGS when read lengths from Illumina were still only 76 bp. Hence, split-read signatures need additional information from paired-end signatures to be able to resolve SVs in repetitive regions. There are also split-read based tools dedicated to RNA-seq data such as FusionHunter (Li et al., 2011), deFuse (McPherson et al., 2011) or Splitread Karakoc et al. (2012).

A similar principle to split-reads are soft-clipped reads or open end anchors (OEA), but here the aligned parts of the read are restricted to the 5' and 3' end of the read, i.e. they are always a prefix or suffix of the read. Soft clipping (SC) is often done by modern read mappers that use a seed and extend approach. Tools like ClipCrop (Suzuki et al., 2011), CREST (Wang et al., 2011), Socrates (Schröder et al., 2014), or NovelSeq (OEA) (Hajirasouliha et al., 2010) then build upon these SC or OEA reads to call SVs.

At early stages of SV detection, there appeared to be poor agreement between the called SVs based on different methods, often owing to the distinct strengths and weaknesses of each method or the restriction to certain SV types. So called hybrid

SV detection tools hence aimed to integrate two - sometimes three - approaches. Also future releases of previously mentioned tools aimed to incorporate additional information to improve their results. Examples for the combination of discordant read pairs and read-depth include, e.g. CNVer (first method to combine both, Medvedev et al. (2010)) and Genome STRiP. Most tools actually combine split-read with paired-end information, like e.g., later versions of Delly , SVMerge (Wong et al., 2010), SVSeq (Zhang and Wu, 2011), Meerkat (Yang et al., 2013), SMUFIN (Moncunill et al., 2014), or forestSV (Michaelson and Sebat, 2012). LUMPY offers another approach to combine the different strengths. As a meta caller, LUMPY combines the results of various tools (Layer et al., 2014) .

Apart from mapping, some tools also facilitate assembly strategies. Apart from complete genome assembly, this is usually done by a local assembly for larger, novel insertions, e.g., with MindTheGap (Rizk et al., 2014) or Anis and Basil (Holtgrewe et al., 2015). PopIns uses the assembler Velvet to facilitate a population scale insertion assembly that also implements a follow up genotyping (Kehr et al., 2015). To place the insertions during genotyping, PopIns uses split-reads and read pairs. Rearrangements between whole genomes can again be discovered with methods for whole genome alignments. Local assembly usually starts at SC positions. Unmapped reads are then assembled anchored to the identified 5' or 3' end of the breakpoint in an iterative fashion until the gap between two breakpoints is closed (targeted assembly Manta (Chen et al., 2015), SVMerge, TIGRA (Chen et al., 2013a), window-based assembly SOAPindel (Li et al., 2012a), DISCOVAR (Weisenfeld et al., 2014)). GRIDSS uses de Bruijn graphs for their targeted assembly (Cameron et al., 2017).

To date, plethora of papers exists presenting different SV detection methods and tools. The latest study performs haplotype-resolved SV detection in human genomes integrating multiple platforms (Chaisson et al., 2017). Fortunately, there are also joint community efforts toward a better understanding of the human SV landscape (Redon et al., 2006; Mills et al., 2011; The 1000 Genomes Project Consortium., 2012, 2015; Sudmant et al., 2015) that also aim to resolve complex SVs.

The list of tools mentioned in this section is by no means exhaustive owing to the huge number of available software that is sometimes tailored to specific SV types, sequencing technologies, read lengths or error profiles and so forth. While the aforementioned tools were mainly designed for human sequencing data, there are also tools designed for other organisms, e.g., *breseq* for bacteria (Barrick et al., 2014). While breseq accounts for traits specific for bacteria such as higher recombination rates and repeats, the tool focusses on variants within the genome and does not detect horizontal gene transfer events between bacteria.

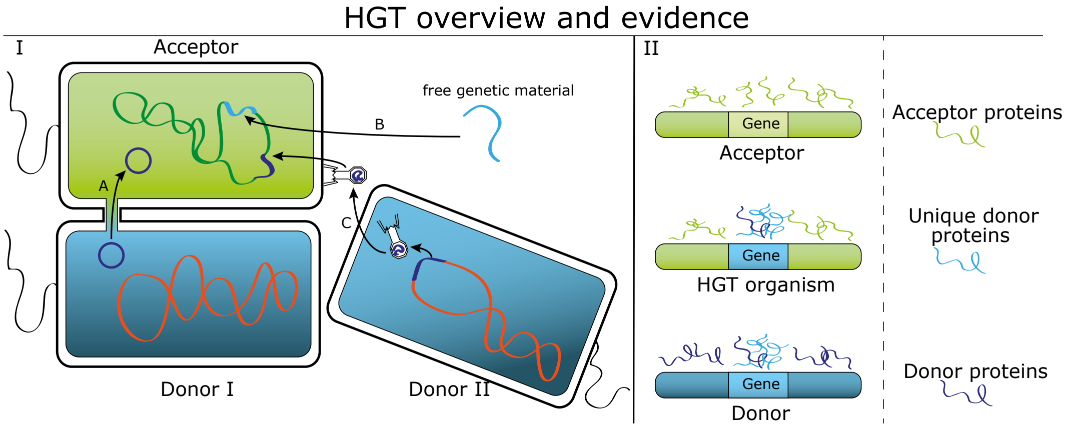


Figure 1.4. HGT overview and evidence. (I) Via horizontal gene transfer (HGT), genetic material is transferred from the donor cell to the acceptor cell by one of three possible ways. The genetic material can be part of a plasmid (A) that is exchanged directly between the acceptor and donor cell. Bacteria can also take up free genetic material from the environment (B). The gene(s) can also be part of a donor bacteriophage that transfers the gene(s) when it integrates into the genome after infection of the acceptor (C). (II) Regarding its gene - or protein - content, the HGT organism consists therefore mainly of the acceptor genome, and the transferred proteins should be unique within the acceptor and donor (light blue). Reprinted from Trappe et al. (2017).

1.4 Horizontal gene transfer in bacteria - variants across species boundaries

Traditionally, evolution is viewed as a bifurcation process: changes are only vertically inherited from parent to offspring and hence, only slow divergence leads to new species over time (Olendzenski and Gogarten, 2009). Along this line of thought, we view all species relations and their history as a ”tree of life”, depicted as bifurcating phylogenetic trees. Any genetic relation between distant lineages in the tree were regarded as a strange coincidence due to orthologous evolution.

The parent to offspring inheritance and the integrity of a phylogenetic tree was unquestioned until proof for a concept known as horizontal gene transfer (HGT) was established (Syvanen, 1985; Sprague, 1991; Lawrence, 2002). HGT is defined as the movement of genetic information between distantly related organisms (Crisp et al., 2015). Before genome sequencing shed light into the frequency and prominence of HGT, such events were considered to happen only rarely, and hence, the impact considered irrelevant (Koonin et al., 2001). Now, HGT is even thought to be the dominating factor in microbial evolution, at a gene gain and loss rate similar to point mutation rates (Koonin, 2016). Some extreme views even go that far to deny vertical tree-like evolution.

HGT is not limited to, but probably best studied in, bacteria. HGT happens wherever various bacteria live in close proximity like, e.g., in the human gut, or are brought in frequently from different sources like, e.g., in hospitals. Liu et al. (2012) claim to have identified over ten thousand high confidence HGT genes in human microbes.

Transferable genes are also referred to as mobile genetic elements (MGEs). These are loosely defined as DNA that is able to move within or between genomes (Siefert, 2009), and make up the unique mobilome of every genome. Bacterial MGEs are commonly classified into transposons, plasmids, and bacteriophages. Bacteriophages are viruses that infect bacteria. There are at least three known mechanisms for bacterial HGT through which the so called acceptor cell receives genetic material from the donor cell (see Figure 1.4). The acceptor cell can take up nascent DNA from the environment in a process called transformation (Mazodier and Davies, 1991). The acceptor can also gather packed DNA in form of plasmids or bacteriophages. For the direct transfer of a plasmid from donor to acceptor via conjugation, a cell-cell contact is established through a pilus. Bacteriophages are transferred from the donor cell through transduction.

The recognition of HGT has led to a postulation of a "net of life" with fusing branches compared to the traditional "tree of life" where fusions are only seen as inconsistencies (Hilario and Gogarten, 1993).

1.4.1 Relevance and impact of HGT for public health

In the past years, HGT in bacteria became a widely accepted mechanism but the impact concerning public health has only recently been fully acknowledged. Bacteria are prominent in basically every multi-cellular organism and various environments. Some are essential for our survival and hence harmless, e.g., considering the human gut microbiome (Eckburg, 2005; Backhed, 2005). Here they synthesise important vitamins or other nutrients, and protect us from invading pathogens. But they also pose maybe the greatest health risk when it comes to disease causing bacteria (Guarner and Malagelada, 2003), according to the WHO to some extent even more than viruses or cancer.

Upon the discovery of the first antibiotic penicillin in 1929 by Alexander Fleming, a new era began where humans were actually able to efficiently combat bacterial infections. For a long time, modern medicine thought to have even defeated death due to bacterial infections. But only few decades later, bacteria started to overcome antibiotics. And those who did would sometimes be even more aggressive, posing a greater risk and demand for stronger antibiotic alternatives. The problem of resistance was more prominent in hospitals where the use of antibiotics is more frequent. The essential question was, how can bacteria adapt to these antibiotics so quickly? Bacteria have a high evolutionary rate, but this could not explain the high rate of adaptation and resistance. Observations from acquired resistances of former susceptible bacterial strains that lived in close proximity to strains with known antimicrobial resistances (AMRs) led to suspicions about underlying, unknown transfer mechanisms. In 1960, Akiba et al. (1960) demonstrated such a transfer of resistances from *Shigella* to *E. coli* in a mixed cultivation. The underlying mechanism, called horizontal gene transfer (HGT), has been acknowledged over the years. AMR can also arise through point mutations as in the beta-lactamase resistance gene, in fact beta-lactamase can undergo both, HGT and evolution through point mutation.

AMR genes are also only one example of the importance of HGT events in bacteria for public health. Either through a natural HGT event or an artificially induced event - i.e. a forced genetic modification - virulence factors that increase pathogenicity can be introduced into former harmless organisms. Pathogenicity islands (PIs) are conserved, mobile sequence stretches that can be transferred by conjugation or transduction. Despite conservation, they are often modular such that new genes can be incorporated into the island. *Salmonella* Pathogenicity Island-7 (SPI-7), e.g., is the largest island identified so far with a size of 134 kb. SPI-7 is specific to *Salmonella enterica* subsp. enterica serovar Typhi, a human pathogen (Seth-Smith, 2008). *Yersinia pestis* that causes the disease plague that resulted in several epidemics throughout history including the *Black Death*, e.g., hosts two plasmids that are not carried by other *Yersinia* species but that contain the virulence factors responsible for the high pathogenicity. Transduction of these two plasmids could cause a pathogenicity increase in other *Yersinia* species, which also poses a potential thread in the context of biological warfare (Parkhill et al., 2001). Yu et al. identified a gene cluster in *Burkholderia pseudomallei*, a human pathogen causing melioidosis, that is similar to a gene cluster from *Y. pestis* and that appears to be horizontally transferred (Yu et al., 2006). This gene cluster is absent in the avirulent *B. thailandensis* strain.

These are just a few examples for the great potential that comes with the ability of HGT, and more will certainly be discovered with improved methods and technologies. Important for method developments in the context of public health is also the distinction of a single infection by a modified organism in contrast to double - or multi-infections - with distinct organisms or pathogens that may have conflicting resistance patterns and cause additional symptoms when treated inappropriately.

1.4.2 Computational methods for HGT detection and analysis

Pre-NGS genomic HGT methods can be roughly categorised into three approaches, namely sequence composition-based, phylogenetic based or BLAST based (Ravenhall et al., 2015). All of them work on fully or at least partially assembled sequences. In any case, the common assumption to identify a probably horizontally transferred gene is that the foreign gene has measurable features that distinguishes the gene from the remaining genes and sequence content. In a way, with the sequence composition and BLAST based methods, one can identify candidate regions or genes that can -

or rather should - then be further analysed and validated by phylogenetic methods or other means.

Each genome has unique patterns of sequence signatures such as, e.g., the GC content. A foreign gene thus has a theoretically detectable different signature compared to its new host genome (Karlin, 2001; Putonti et al., 2006; van Passel et al., 2005). Different studies or methods use one or more of these sequence signatures, e.g., the nucleotide composition (Daubin et al., 2003), *k*-mer frequency (Lawrence and Ochman, 1998), codon usage (Lawrence and Ochman, 2002) or also other structural features (Worning et al., 2000). Over time through the natural evolutionary process, the transferred genes loose their previous host specific signatures, a process called amelioration (Lawrence and Ochman, 1997), and become unrecognised by composition based methods. Cortez et al. (2009) compared different composition-based approaches on an *in silico* HGT organism to show the need for control methodologies concerning the assessment of false positive prediction, which they claim has been missing in the studies employing these approaches before. Methods that rely on the distinct sequence pattern between the host genome and introduced, foreign sequence content can only reliably identify newer HGT events. These signatures are also less distinct between closely related species. A drawback of composition-based methods is also that pattern changes can also have other reasons than HGT events (Ravenhall et al., 2015). For example, GC content has been shown to be higher in highly expressed genes (Wuitschick and Karrer, 1999).

The goal of BLAST-based methods is to identify homologous genes in a distantly related organism. This can be done either with a candidate gene or in an all versus all search. The hits are usually ranked by blast identity scores such as the bit score, where the highest ranked candidate gene, i.e. the best match, is identified for every query gene. If the best match is then in a distantly related organism it is categorised as potentially horizontally transferred. Tools implementing BLAST strategies are, e.g., Pyphy (Sicheritz-Ponten, 2001) PhyloGenie (Frickey, 2004), or DarkHorse (Podell and Gaasterland, 2007). BLAST-based approaches also have limitations. The best hit does not necessarily have to be the nearest neighbour and could be caused also by other means than HGT, e.g., unfinished genomes in databases with missing sequence information, gene loss events, or other database errors (Zhu et al., 2014). These methods are also challenged by HGTs in closely related species, or in case there are orthologs or paralogs that can be mistaken to have horizontal origin. HGTector (Zhu et al., 2014) facilitates a standard all-vs.-all BLASTP search with an additional, phylogenetically informed grouping that aims to normalise BLAST hit scores according to the phylogenetic category, thereby trying to circumvent some of these limitations. A downstream phylogenetic validation is integrated into the pipeline as well.

Both BLAST and sequence composition methods can only identify pattern abnormalities and possible candidates for horizontally transferred genes. Evidence, however, can only be established in the phylogenetic context (Eisen, 2000).

Phylogenetic HGT detection or validation procedures are based on the reconciliation between phylogenetic trees from different genes in a fixed set of genomes. If the genomes have an established relationship that confers with most of their genes, a foreign gene would fall out of line and suggest a different relationship. Explicit phylogenetic methods directly compare the phylogenetic trees to infer HGTs whereas implicit methods make use of other metrics that correlate with the gene tree history (Ravenhall et al., 2015). In the implicit tool DLIGHT, e.g, the authors use a likelihood ratio based on pairwise evolutionary distances (Dessimoz et al., 2008). Ravenhall et al. (2015) discuss both kinds of methods extensively.

One of the downsides is that phylogenetic analysis itself is a task with still unsolved challenges and pitfalls. Also, HGT is not the only mechanism or cause that can result in phylogenetic conflict patterns. Some causes for false positive predictions include poor data quality, ambiguous alignments that in addition might also lead to misinterpretation of genes with paralogy and orthology relation, or simply misapplication of (unsuitable) phylogenetic methods (Eisen, 2000). Hence, there is need for HGT detection methods based on other kind of evidence.

The mentioned approaches often detect HGT events at different phylogenetic distances and age. Similar to the phenomenon of early SV detection approaches, there is often a low overlap or agreement between the reported HGT events of different HGT detection methods on the same dataset (Lawrence and Ochman, 2002; Ravenhall et al., 2015). At the same time, HGT rates estimated upon different methods vary sometimes largely, causing again controversies about the frequency of HGT. Large scale HGT analysis without follow-up, supporting investigation is still viewed critically as many of such reported genes have proven to be false positives due to manifold reasons (Koonin et al., 2001).

NGS facilitates new means for HGT detection methods that can provide other evidence, e.g., in a mapping-based fashion. An example, although limited to the detection of AMR genes, is KmerResistance (Clausen et al., 2016) that uses k-mer profiles directly from sequencing reads to infer AMR genes. Large sequencing efforts are also stocking up and hence also improving the reference content in databases.

1.5 Terminology

I here define and clarify some important terms and phrases used throughout my thesis. Some of them have ambiguous meanings, others are not commonly used yet in the literature.

Coverage The term coverage has at least two important meanings when it comes to genomic sequencing technologies and data. Sequencing coverage, also called sequencing depth, refers to the theoretical, average number of reads generated for each base of the sequenced genome. So for a 100x coverage, each base is represented by an average of 100 reads. Note that one read covers more than one base, depending on the read length. The desired sequencing coverage can be defined before sequencing to obtain the number of fragments and cycles needed for the experiment.

When the sequenced reads are mapped to a reference genome, we can define the mapping coverage as the number of reads mapping to the reference at one position or also as the percentage of the reference covered by the reads. I refer to a genomic region as covered if there is one or more reads mapping to this region, thereby "covering" the base pairs at those positions. Depending on the sequence similarity of the reference and the sequenced genome, the average sequencing depth and the mapping coverage of a position in the genome can vary enormously.

When speaking of coverage within the scope of this thesis, I generally refer to the mapping coverage.

HGT organism During an HGT event, a piece of DNA is transferred from a donor organism to an acceptor organism. Strictly speaking, there is no biological HGT organism. It is a term we defined to address the resulting organism after an HGT event has occurred, i.e., after an acceptor organism has acquired a novel gene. Also, an HGT event effects both the genome and, if the gene is expressed, the proteome of an organism. If not distinguished further, I refer to both the genome and the proteome by the term HGT organism.

1.6 Open research questions

In this section, I want to highlight the open research questions opened up and motivated by the previous introduction that are addressed within the scope of this thesis. The first topic addresses the motivation to develop a SV detection tool for (human) NGS data. Its methods are then adapted to a more specific variation in bacteria, namely the detection of HGT events from NGS data. The last topic addresses a proteomic approach for HGT detection that integrates with and further characterises results from genomic-based HGT detection.

1.6.1 Development of a computational method for SV detection from NGS data

Structural variations have been shown to play a major role in complex disorders and diseases such as cancer, and methods are required that are able to detect and characterise even complex types of SVs from NGS data. The project to develop a structural variation detection method started back in 2011 when the field of SV detection tools was still scarce and Delly - the coming state of the art - not published yet. Existing methods could detect only simpler SV types such as indels and inversions, or CNVs but without breakpoint precision. The resolution for SV size of > 50 bp was first mentioned in Alkan et al. (2011), but no tools were developed so far that addressed this size span from 50 to 1 kb. Technologically wise, Illumina reads were still short (mostly 76 bp), and the common assumption was that the long 454 reads would be the future and dominate the sequencing market. Long reads of several hundred or thousand base pairs are likely to span one or even multiple SV breakpoints, with the size of the matching segments in between breakpoints unknown beforehand and likely to vary in size. Both points were still not addressed by existing methods. Inspired by methods and graph-based data structures for whole genome alignment under consideration of rearrangements, we wanted to build a method that finds those matching segments and is able to represent and detect the rearrangements connecting these segments. Representing the segments in a generic way, we aimed to also expand the types of SVs that can be detected to include inversions, duplications and translocations next to long indels. Especially duplications and translocations were not detectable with base pair resolution before. These complex SVs involve distinct regions within the genome and actually consist of a combination of simpler SV patterns. The insertion event is taking place at a distinct location from the copy (duplication) or deletion (translocation) event, creating distantly related events that are usually not spanned by the same read.

In short, we aimed at filling the size gap from 50-1000 bp - which we termed the NGS Twilight zone - being generic in variant type and size, and able to span multiple breakpoints by considering long read lengths. Our approach was not designed as a read mapper with SV detection on top but focusses on fewer reads spanning SVs but therefore trying a more sensitive but computational expensive approach.

To do that, we wanted to use the local aligner Stellar (Kehr et al., 2011) to identify the segments between breakpoints, and then represent these in a graph data structure. We also wanted to incorporate a dedicated split-read mapping as proposed in AGE (Abyzov et al., 2011) and used in SplazerS (Emde et al., 2012) to provide breakpoint precision in the presence of sequencing error or microhomologies.

1.6.2 Development of computational methods for HGT detection from NGS data

Horizontal gene transfer (HGT) has changed the way we regard evolution. Instead of waiting for the next generation to establish new traits, especially bacteria are able to take a shortcut via HGT that enables them to pass on genes from one individual to another, even across species boundaries. Existing HGT detection methods usually exploit phylogenetic discrepancies of a particular gene tree compared to a species tree. To be able to build up a gene tree, one has to identify potential HGT genes, which can be done using composition-based HGT detection methods. This approach answers the question of what particular gene is of foreign nature in a certain organism. What the approach does not answer is, where did the gene come from, and where exactly did it go? Once candidate genes are identified, phylogenetic approaches could offer clues about the whereabouts of their origin but do not give evidence about the transfer. Moreover, all methods depend on fully assembled genomes of the HGT organism rather than exploiting NGS data directly, and are also not consistent in their results. In the context of public health, it is also important to characterise HGT events directly from NGS data that enables the distinction between single infections of novel HGT organisms and double infections from the potential acceptor and donor organisms.

Given NGS data from an organism with a potential but unknown HGT event, two research questions can be derived: (i) what are the involved acceptor and donor organism, and (ii) what sequence - or gene - content from the donor with what characteristics has been transferred to what position in the recipient. Answering the second question allows to deduce the functional impact of the transfer and also, in a more general term, to characterise specific insertion sites. Answering the identification question of acceptor and donor organism is a prerequisite to define the search space for potential HGT regions and sites.

In a first proof of principle, we aimed to develop an approach for HGT site detection and characterisation based on read mapping that provides complementary evidence compared to existing methods at the cost of relying on the acceptor and donor references of the HGT organism being known. To achieve this, we adapted methods established for SV detection in human genomes to the more specific HGT detection problem. In an abstract point of view, an HGT event has the same SV pattern as an inter-chromosomal translocation, with acceptor and donor genome being considered as chromosomes. The problem of identifying organisms in a sample based on sequencing reads is addressed by metagenomic profiling tools. However, acceptor and donor references have certain properties such that these methods can not be directly applied. We wanted to develop a mapping-based pipeline that is able to identify acceptor and donor candidates of an HGT organism based on sequencing reads. To do that, we aimed to leverage properties of the metagenomic profiling tool MicrobeGPS, and to refine them for HGT candidate identification.

1.6.3 Development of a computational method for HGT detection from MS/MS data

The detection of HGT events on the genomic level can only reveal the potential functional gain but the expression also depends on where the gene has been inserted and further means of expression like, e.g., promotors. Proteomics is also an orthogonal type of evidence supporting genomic based clues. However, despite studies addressing impact and functionality of potentially horizontally transferred genes such as specific antibiotic resistances, there is no MS-based method or study investigating HGT events as such in an independent and detailed manner. Akin to the approach on the genomic level, one can investigate the two research questions of (i) involved acceptor and donor organisms identification and (ii) characterisation of the transferred gene separately in a two-stage process. Building on our genomic approach, we first wanted to further investigate the characterisation of transferred proteins from shotgun MS data.

To do that, we aimed to develop a computational workflow that can automatically detect a unique donor protein by an extended, standard database search. A rigorous cross-validation ensures that the protein conforms to the characteristics of a HGT protein, namely that it is not present among the acceptor database and that no other donor proteins are detected. In preparation for future integration with our genomic approach, we also aimed to determine the genomic origin in a proteogenomic fashion.

1.7 Thesis outline

In this thesis, I present my contributions to the research questions described in the previous section. Chapter 2 addresses the problem of SV detection from NGS data. Anne-Katrin Emde, Knut Reinert and I developed the concept of using local alignments and a graph data structure to present generic breakpoints of SVs, and to use these to detect SVs from NGS data. I further developed a combinatorial method to combine the patterns of simple SVs to infer complex translocations and duplications, and implemented that with the concept in the software *Gustaf*. Anne-Katrin Emde and I designed the evaluation, and Christian Ehrlich helped implementing the evaluation in an automised benchmark. I analysed the data and wrote the paper together with Anne-Katrin. Chapter 2 is based on the following publication:

> *Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone.* K. Trappe, A. K. Emde, H. C. Ehrlich, and K. Reinert. *Bioinformatics, Oxford University Press (OUP)*, 30 (24): 3484–90, 2014.
>
> Presented at the Conference for High Throughput Sequencing (Hit-Seq), ISMB 2014, Boston, USA.

While Gustaf is a tool for generic SV detection, the concept can be adapted to tailored variant detection. A HGT event is in a way an SV event involving genomes across species boundaries. The HGT detection problem involves two questions, (i) what are the involved acceptor and donor species, and (ii) what part of the donor genome has been transferred to what position in the acceptor genome.

Chapter 3 describes how we applied SV detection methods to the second problem, the detection and characterisation of the HGT site. Tobias Marschall, Bernhard Renard and I developed the concept of integrating these established methods to a distinct across-species event, and together designed a benchmark to evaluate our approach. I implemented the concept in the software *Daisy* and analysed the data. I wrote the paper together with Tobias Marschall, and Bernhard Renard helped in writing the manuscript and gave valuable advice during data analysis. Chapter 3 is based on this publication:

> *Structural Variant Detection across Species Boundaries: Mapping-Based Horizontal Gene Transfer Detection from Sequencing Data.* K. Trappe, T. Marschall, and B. Y. Renard. *Bioinformatics, Oxford University Press (OUP)*, 32 (17): i595–i604, 2016.
>
> Presented at the European Conference for Computational Biology (ECCB) 2016, The Hague, Netherlands, and the NGS'17 conference, Barcelona, Spain.

The problem of acceptor and donor identification was further investigated and a method to solve this problem implemented in *DaisyGPS*. These investigations are described in Chapter 4. I designed and conceived the general approach together with Bernhard Renard, and I had the lead in evaluation and in the data analysis and interpretation process. Enrico Seiler developed the scoring and ranking for acceptor and donor candidates under my supervision, and implemented it in the DaisyGPS software. Enrico and I did the evaluation and wrote the paper together. Chapter 4 is based on this preprint:

> *Where did you come from, where did you go: Refining Metagenomic Analysis Tools for HGT characterisation.* E. Seiler[1], K. Trappe[1], and B. Y. Renard. *bioRxiv*, preprint, 2018.

In Chapter 5, the concept behind Daisy is partly transferred to the field of proteomics and implemented in the tool *Hortense*. I designed the overall workflow with Thilo Muth and Bernhard Renard, where Thilo and Bernhard contributed to adapting the concept to the characteristics of proteomic MS/MS data. Ben Wulf developed and implemented the database search procedure and ranking in Hortense under Thilos and my guidance and critical supervision. Joerg Doellinger (ZBS 6, Robert Koch Institute) and Sven Halbedel (FG 11, Robert Koch Institute) provided biological data for the validation, and gave valuable proteomic and bacteriological insights for evaluation and interpretation of Hortense. Thilo Muth and I wrote, and Bernhard Renard contributed to writing, the manuscript, Ben Wulf designed the figures. The chapter is based on the following publication that was reviewed through the submission process for the German Conference on Bioinformatics, 2017:

[1] Joint first authors

Hortense: Horizontal gene transfer detection directly from proteomic MS/MS data. K. Trappe, B. Wulf, J. Doellinger, S. Halbedel, T. Muth, and B. Y. Renard. *PeerJ*, preprint, 2017.

Presented at the German Conference on Bioinformatics (GCB) 2017, Tübingen, Germany.

1.8 Further contributions

Besides the research I conducted as part of my PhD that is presented in the following chapters, I also contributed to the following research project. In this project, Birte Kehr and I conducted a theoretical comparison of available graph data structures for whole genome alignments in terms of their ability to represent alignment information. I contributed to establishing analogous structures and properties between the chosen graphs, and helped to define rules by which one graph can be transformed into another without loss of alignment information. This work greatly inspired the concept behind the SV detection tool Gustaf presented in Chapter 2.

Genome alignment with graph data structures: a comparison. B. Kehr, **K. Trappe**, M. Holtgrewe, K. Reinert. *BMC Bioinformatics*, 9: 15–99, 2014.

2 Detecting complex structural variants with Gustaf

Variation in the human genome remains only partially characterized. Theoretically, whole genome sequencing (WGS) data carries the potential to identify all genomic variation, including small variants such as single nucleotide variants (SNVs) and small indels, as well as larger structural variants (SVs, typically defined as >50 bp), such as large deletions, inversions, duplications and translocations. However, the difficulty lies in computationally analyzing the large-scale data obtained from WGS and reliably identifying the whole range of genomic variation.

With current algorithms and methods, only small variants, up to 30 bp, can be identified rather confidently, while larger variation still poses a major challenge (Alkan et al., 2011). Large and complex rearrangements are often accompanied by micro-homologies or microindels (Onishi-Seebacher and Korbel, 2011)) around their breakpoints which makes them even harder to detect, particularly with base pair resolution. Many bioinformatics tools have been developed in recent years that address SV detection through identification of certain signatures in the sequencing data (Alkan et al., 2011). Mainly, they rely on one or multiple of the following four approaches: 1) identifying discordant read pairs that span SV breakpoints (Chen et al., 2009; Marschall et al., 2012; Hormozdiari et al., 2009; Tuzun et al., 2005), 2) detecting regions of unexpectedly low or high read depth (Abyzov et al., 2011; Xi et al., 2011; Yoon et al., 2009b), 3) identifying split reads that span SV breakpoints (Ye et al., 2009; Emde et al., 2012) or 4) local reassembly of SV candidate regions (Chen et al., 2013b). Typically, overlap between different tools is low (Alkan et al., 2011). This is partially due to the fact that most tools are geared towards certain types of data, e.g. for specific read lengths, or towards certain SV size ranges and types, e.g. mid-size indels. Also, the different strategies suffer from various biases and sources of errors. Read depth methods are vulnerable to read depth fluctuations leading to non-uniform coverage, e.g. due to GC content and mappability issues. Even after successfully normalizing for coverage biases, the spectrum of SVs that can be identified through read depth signatures is limited to copy number variable regions, i.e. deletion and amplification. Read pair methods are vulnerable to mismappings caused by repetitive regions in the genome and are furthermore susceptible to chimeric read pairs (Maher et al., 2009).

Both read depth and read pair methods have problems identifying small SVs, i.e.

less than a few hundred base pairs, which we here term the NGS twilight zone of SVs. For read depth methods, it is hard to discern coverage changes in small regions from natural read depth fluctuations. For read pair methods, such relatively small deletions are difficult to identify since paired read spacing may still lie within the variance of the insert size distribution.

Of the four strategies, only split-reads and assembly-based methods yield single nucleotide resolution, by identifying reads or reconstructing contigs that directly span SV breakpoints. When correctly mapped onto the reference genome, the read- (or contig-)to-reference alignment of an SV-spanning read (contig) will be split into partial alignments at the breakpoint positions, hence yielding base pair resolution.

The split-read approach has been primarily used in conjunction with read pair methods that identify potential SV-spanning reads through abnormal paired read distance or orientation (Ye et al., 2009; Rausch et al., 2012; Marschall et al., 2013). This significantly reduces the search space for split mapping, which is computationally expensive to apply to the whole genome.

However, with increasing read lengths (and improved local reassembly approaches that generate contigs), the split mapping approach becomes more and more powerful, since it yields highest confidence and base pair resolution.

Therefore, we aimed to develop a method that can generically split-map contigs or reads of arbitrary length. Our tool Gustaf (Generic mUlti-SpliT Alignment Finder) allows for multiple splits at arbitrary locations in the read, is independent of read length and sequencing platform, and supports both single-end and paired-end reads.

Gustaf is based on finding local alignments of a read, and then essentially chaining local alignments into a semi-global read-to-reference alignment. Similar approaches are used in the context of whole-genome alignment, where large rearrangements between locally collinear blocks are maintained in graph or graph like structures like, e.g., the alignment graph (Kececioglu, 1993), the A-Bruijn graph (Pevzner et al., 2004) or the Enredo graph (Paten et al., 2008), or in the SHUFFLE-LAGAN glocal alignment algorithm (Brudno et al., 2003).

Local alignments are identified with Stellar (Kehr et al. (2011), `www.seqan.de/projects/stellar`), an edit distance local aligner, which guarantees to find all local alignments of a given minimal length, maximal error rate and maximal X-drop. In theory, however, our approach can take local alignments from any aligner as input, making it versatile and adaptable. Since local alignments are allowed to be anywhere in the reference genome, it allows for non-collinearity of chained local alignments, and hence has the power to identify all types of structural variation.

In contrast to other SV callers, Gustaf furthermore attempts to correctly classify SV events leading to multiple breakpoints such as translocations and dispersed duplications including the actual length of the duplication (see also Figure 2.3). These complex SV patterns incorporate pseudodeletions that make them harder to classify and that are often wrongly reported by other methods.

In the following, we introduce our method Gustaf, and compare it with two other popular methods that are able to report SVs with base pair resolution, Pindel (Ye et al., 2009) and Delly (Rausch et al., 2012).

2.1 Determining breakpoints and shortest paths in a breakpoint graph

Gustaf is an SV detection tool that uses a split read approach to detect exact breakpoints of SVs including indels, inversions, duplications and translocations. Gustaf uses the local aligner Stellar to detect partial alignments of a read and stores compatibility information of these partial alignments in a graph data structure.

We will first give a brief overview of Gustaf's general approach, and then define all important steps such as local alignments, adjacency of local alignments, breakpoints of different types and our split-read graph together with the necessary methods such as the split-alignment in the following subsections.

2.1.1 Gustaf's workflow

Gustaf takes as input a set of reads and optionally a set of local alignments for these reads. If no local alignments are supplied, they will be computed using Stellar. Our approach is based on maintaining all local alignments of a read within a graph structure so that we can use standard graph algorithms to evaluate relationships of the alignments. We call this graph *split-read graph* and will completely define it below (see also Figures 2.1 and 2.2). In this graph, each local alignment is represented by one vertex.

Breakpoints (like positions v, w, z in Figure 2.3) create new sequence adjacencies in a genome, e.g. block 4 and 5 in Figure 2.3 are adjacent in the reference genome but formed new adjacencies with the duplicated (or translocated) block 2 in the donor genome. Vertices of alignments that belong to the same variant, i.e. that span a breakpoint, are connected by an edge so that the edge represents the breakpoint. Each alignment has an edit distance. Roughly, each edge carries as weight the edit distance of its target vertex (see also Figure 2.2, more details will follow below).

All valid combinations of alignments are represented by their corresponding paths of vertices in the graph. The sum of the edge weights of each path will correspond to the edit distance of the individual split alignments, with additional penalties incurred for each split depending on the type and size of the SV it indicates (see Figure2.2 for an inversion example $inv(w, z)$ for a_1, a_2 with edit distance $d^{a_2} = 2$, inversion penalty 5 and gained split alignment score $d' = -4$). The path with the lowest total weight represents the most likely combination of local alignments for this read, and the edges give the breakpoints causing the split in the read-to-reference

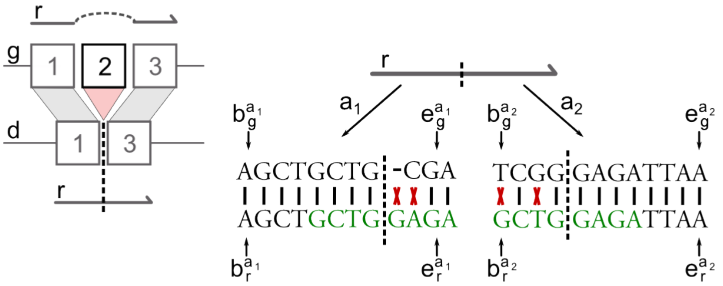


Figure 2.1. Local alignments a_1 and a_2 of a read r spanning blocks 1 and 3 in a donor genome d where block 2 is deleted compared to the upper reference genome g. Alignment a_1 denotes the local alignment of the read prefix with block 1, a_2 of the read suffix with block 3. Denoted are begin (b) and end (e) positions of the alignment a_1 in the reference (b_g^{a1}, e_g^{a1}) and the read (b_r^{a1}, e_r^{a1}), and positions of a_2 analogous to a_1. Alignments a_1 and a_2 are adjacent in the read and overlap at their sloppy end within the read (green bases GCTGGAGA). The correct split position of the deletion is indicated by the dotted line. Both a_1, a_2 have edit distance 2. The gained score d' of this split-alignment is -4, as we get rid of all errors within the sloppy ends when cutting the alignments at the dotted line.

alignment.

2.1.2 Local alignments

Gustaf uses SeqAn's local aligner Stellar (Kehr et al., 2011) to compute the set of local alignments $\mathcal{A}(r)$ of each read r. Stellar implements a seed and extend approach based on the SWIFT filter algorithm (Rasmussen et al., 2006), and guarantees to find all local alignments between two sequences given a minimal match length and maximal error rate for these alignments. During the extension phase, the seeds are extended as long as the final alignment is still valid with respect to the allowed maximal error rate. This produces sloppy alignment ends. In case a read spans an SV, the alignment is thus likely to extend past the SV breakpoint. This is an important feature that we use to our advantage in the subsequent split-graph construction (see also Figure 2.2).

Throughout the paper, we use the following definitions: Let $s = (b, e)$ be a segment of a sequence starting in b and ending in e, i.e. $b < e$. An alignment $a = \{s_g, s_r, o\}$ between a reference g and a read r aligns the segment $s_g = (b_g, e_g)$ of g to the segment $s_r = (b_r, e_r)$ of r. The orientation $o \in \{+, -\}$ indicates whether the read mapped to the forward (+) or reverse (−) strand of the reference. We denote the read segment positions b_r and e_r of an alignment a with b_r^a and e_r^a, and the reference positions b_g^a and e_g^a (see Figure 2.1). Every alignment a has an edit distance d^a.

For two alignments a_1, a_2, we say $a_1 < a_2$ if $b_r^{a_1} < b_r^{a_2}$, i.e. we can sort all alignments of one read according to their start position in the read. Analogously, we can impose an ordering of the alignments according to their reference sequence positions.

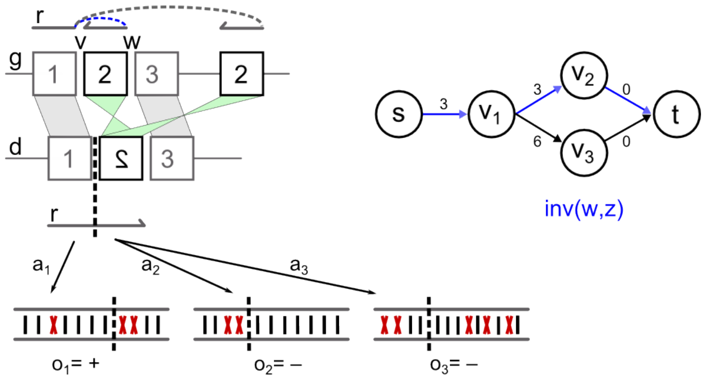


Figure 2.2. Gustaf's workflow detecting an inversion of block 2 in the donor genome d where another similar block 2 is present in the reference g: A read spanning blocks 1 and 2 results in alignments a_1, a_2 and a_3 (with $o_1 = +$ and $o_2 = o_3 = -$) where a_3 is an alignment to the region similar to block 2 upstream in the reference. The corresponding split-read graph (top-right) has artificial start (s) and end (t) vertices, representing start and end of the read, and vertices v_1-v_3 representing alignments a_1-a_3. Ingoing edge labels are edit distances of corresponding alignments adjusted by inversion penalty 5 and the gained split-alignment score (4 for both v_2 and v_3). The shortest path highlighted in blue from s over v_1 and v_2 to t represents the most likely alignment combination (a_1,a_2) of the inversion.

A read that spans a breakpoint is split up in alignments and we identify the alignments belonging to the same variant according to their adjacency, i.e. we say that two alignments span a breakpoint if they fulfill the following criteria of adjacency for alignments. Two alignments can be adjacent regarding their read or their reference sequence (or both). For read adjacency, two alignments a_1, a_2, $a_1 < a_2$, are adjacent if their read positions overlap such that $b_r^{a_2} < e_r^{a_1}$, or if the gap between the read segments is smaller than a predefined threshold T_g, i.e. $b_r^{a_2} - e_r^{a_1} < T_g$. The gap definition of adjacency accounts for possible microindels around a breakpoint and is per default 5 bp but can be adjusted by the user. For reference adjacency, a_1, a_2 are adjacent if their reference positions overlap such that either $b_g^{a_2} < e_g^{a_1}$ (and $b_g^{a_1} < b_g^{a_2}$) or $b_g^{a_1} < e_g^{a_2}$ (and $b_g^{a_2} < b_g^{a_1}$), or if the gap between the reference segments is smaller T_g.

The adjacency and relation of two alignments a_1, a_2, indicate a specific type of SV (see section SV classification). The positions of the SVs are determined by the begin and end positions of the reference segments in a_1 and a_2, refined by the split-alignment method described in the next subsection.

2.1.3 Split-alignment

Two adjacent alignments a_1, a_2 with $a_1 < a_2$ that overlap at their sloppy ends are realigned in their overlapping region to determine the exact breakpoint. This realignment is in principle similar to the split alignment approach in AGE (Abyzov et al., 2011) and is implemented as an alignment algorithm in SeqAn (Döring et al., 2008; Emde et al., 2012). Similar to AGE, two alignment matrices for the overlapping parts of two adjacent alignments a_1, a_2 of a read r are computed simultaneously. That is, we have a segment $s_o = (b_r^{a_2}, e_r^{a_1})$ denoting the overlapping region in the read, and two alignments $a'_1 = \{s'_1, s_o, o_1\}$ and $a'_2 = \{s'_2, s_o, o_2\}$ where s'_i is the subsequence of s_g in a_i that aligns to s_o. The matrices of a'_1, a'_2 determine the best split point p as the position in the read with the best total edit distance score for both alignments a'_1 and a'_2. The segments s_r, s_g of the alignments a_1 and a_2 are trimmed according to p, i.e. s_r^t of a_1^t is $s_r^t = (b_r^{a_1}, p)$ and s_r^t of a_2^t is $s_r^t = (p, e_r^{a_2})$, and s_g^t of a_1^t and a_2^t is the subsequence that aligns to s'_r, respectively. The new edit distances of a_1^t and a_2^t are lower, than the old ones d^{a_1} and d^{a_2} of the untrimmed alignments a_1, a_2, i.e. we have a gain d' of edit score when we split-align a_1 and a_2.

All adjacency and split-alignment information is represented in a graph that we call split-read graph and that is defined in the following subsection.

2.1.4 Split-read graph

All local alignments $\mathcal{A}(r)$ of a read r and their adjacency relation to each other are represented in a split-read graph $G = (V, E, s, t)$ (see Figure 2.2). When building the graph, we start with an almost empty graph containing only two artificial vertices representing *start (s)* and *end (t)* of a read, and then add a vertex $v \in V$ for each local alignment $a \in \mathcal{A}$ of the read. We add directed edges $e_s = (s, v)$ and $e_e = (v, t)$ for every a (represented by v) that misses $t_i^a < T_i$ base pairs to either start or end of the read, and weight them with $w(e_s) = d^a + t_i^a$ and $w(e_e) = t_i^a$.

We then add directed edges $e = (v_1, v_2)$ between all pairs of alignments a_1, a_2 that fulfill the criteria of adjacency defined above and weight the edge with $d^{a_2} - d'$, the edit score of a_2 adjusted by the gained score of the split-alignment of a_1 and a_2. Based on the type of split that an edge supports, an additional penalty is added to the weight of the edge. Splits that agree with a collinear alignment of adjacent local alignments, i.e. are suggesting a simple insertion or deletion event, receive a penalty of 0. All other splits, i.e. suggesting translocation, inversion or duplication, receive a higher penalty. We set all penalties to 5 here, i.e. for a 100 bp read, a non-collinear split will be weighted equivalently to 5 mismatches in a read-to-reference alignment, but these penalties can be adapted by the user depending on the application. Edges reflect adjacency of the alignments either within the read or within a reference genome. The direction of the edge depends only on the alignment order in the read, such that if a_1 and a_2 with $a_1 < a_2$ are adjacent, then there is an edge $e = (v_1, v_2)$, independent of whether the adjacency is in the read or reference. This definition guarantees that we have a directed acyclic graph (DAG) for which we use common graph algorithms like the DAG shortest path algorithm.

The adjacency edges create zero to multiple paths through the graph from *start* to *end* where the sum of the edge weights of each path correspond to the total edit distance of the alignments on the path plus penalties incurred for the indicated SV types (see Figure 2.2 for an inversion example with two alternative paths). Adjusting thresholds T_g and T_i for gaps allowed at beginning or within reads influences sensitivity and specificity when adding edges and therefore also the number of paths. We identify the most likely path using a DAG shortest path algorithm.

2.1.5 Paired-end data

We described the workflow for single-end reads so far. The paired-end version is an extension to the described approach. By joining both mates and treating them as a single read, we obtain a graph with an expected split at the joining position of the mates. Edges in the graph that stem from this artificial split and are in agreement with the library insert size, receive an edge weight bonus that makes any path through this edge more likely to be the best path, i.e. the shortest path. The benefit from this version lies in the higher probability of choosing the correct alignments of a split read and thereby increasing the specificity of the SV calling.

2.2 Inference of complex variants

2.2.1 SV classification

We define the different SV types according to the positions and sequence content affected by the variation. A deletion $del(b, e)$ is then a segment $s = (b, e)$ in the reference g that is absent in the donor genome (see absent block 2 in Figure 2.2). An inversion $inv(b, e)$ is a segment s that is inverted in g (see inverted block 2 in Figure 2.2). A dispersed duplication $dup(b, e, t)$ is a segment $s = (b, e)$ that appears again at position t in g (see duplicated block 2 at position z in Figure 2.3). If position e of s could not be inferred by the classification described below, we refer to the duplication as imprecise, denoted by $dup\_impr(b, t)$ (see imprecise duplication indicated by read 1 in Figure 2.3). A translocation is usually annotated by the new adjacencies formed by the translocation process. We denote a translocation $tra(b, m, t)$ by the three positions b, m, t involved in the new adjacencies, i.e. there is either a segment $s_1 = (b, m)$ translocated to positions t or a segment $s_2 = (m, t)$ translocated to position b (see translocated block 2 at position z in Figure 2.3).

The adjacency and relation of two alignments a_1 and a_2 indicate a specific type of SV. If g_1, g_2 of a_1, a_2 are different, both alignments are from different chromosomes indicating an inter-chromosomal translocation. When the orientation of both alignments is different, i.e. $o_1 \neq o_2$, a_1 and a_2 are spanning an inversion breakpoint (like a_1 and a_2, or a_1 and a_3, in Figure 2.2). If a_1, a_2 are adjacent in the read and there is a gap $b_g^{a_2} - e_g^{a_1} > T_g$ between the reference positions then the donor genome is missing sequence content at this breakpoint caused by a deletion event

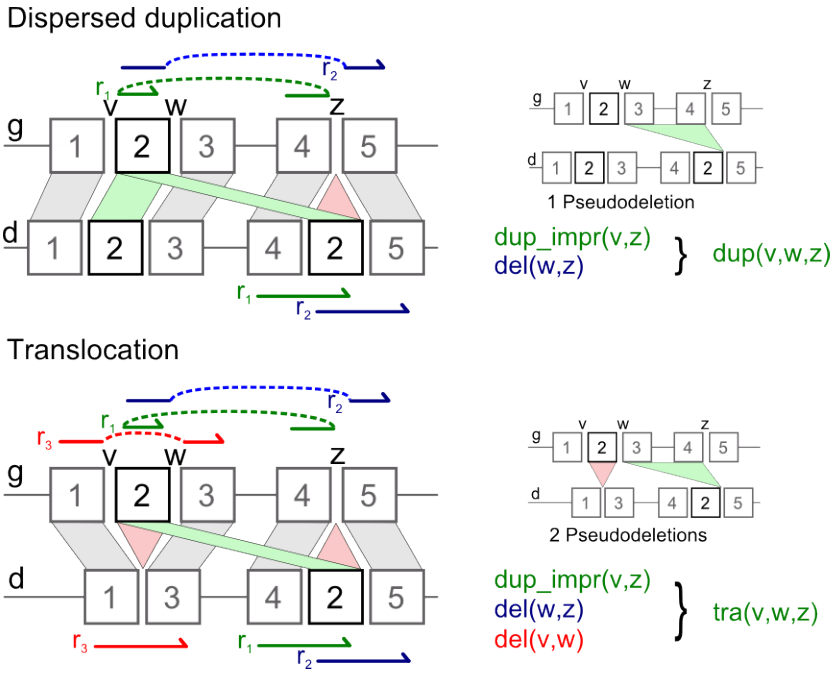


Figure 2.3. Duplication and translocation alignment patterns: On the left hand side, we show alignment patterns of a reference g (upper sequence) with a donor genome d where sequence block 2 is either duplicated (upper figure) or translocated within the donor genome. In a read-to-reference alignment, read r_1 (green) indicates the duplication or translocation event ($dup\_impr(v,z)$) through the different order of the read parts within the reference. We also observe pseudodeletions for both variants (highlighted on right hand side) through read 2 (blue) and 3 (red): An upstream duplication of block 2 creates a pseudodeletion $del(w,z)$, an upstream translocation pseudodeletions $del(w,z)$ and $del(v,w)$. From observing both $dup\_impr(v,z)$ and $del(w,z)$, we infer the duplication $dup(v,w,z)$. If we also observe $del(v,w)$, we infer the translocation $tra(v,w,z)$.

$del(e_g^{a_1}, b_g^{a_2})$. After an insertion event, a_1, a_2 are adjacent in the reference but not the read. Adjacency in both reference and read can indicate small indels or tandem duplications.

Tandem duplications, dispersed duplications, and intra-chromosomal translocations cause a change in the order of the alignments between read and reference, i.e. $b_r^{a_1} < b_r^{a_2}$ and $b_g^{a_2} < b_g^{a_1}$ (like alignments for read 1 in Figure 2.3). From this observation, we can only infer an imprecise duplication $dup\_impr(b_g^{a_2}, e_g^{a_1})$. We therefore cross validate observed deletions and possible duplications to infer the missing middle coordinate of the duplication or to distinguish a duplication from an intra-chromosomal translocation. The difference between a duplication and an intra-chromosomal translocation event is an additional deletion pattern of the translocated region (see pseudodeletions and read 2 and 3 in Figure 2.3), upstream or downstream of the read containing the duplication pattern which can usually not be observed by a single read. So given a $dup\_impr(b_g^{a_2}, e_g^{a_1})$, if we observe a $del_1(b_g^{a_2}, m)$ we infer a $dup(m, e_g^{a_1}, b_g^{a_2})$. If we observe a $del_2(m, e_g^{a_1})$, we infer $dup(b_g^{a_2}, m, e_g^{a_1})$. Only if we see both del_1 and del_2, we know one of them is the upstream or downstream deletion marking the event as a translocation $tra(b_g^{a_2}, m, e_g^{a_1})$ (see also Figure 2.3).

Sometimes a read reaches over a breakpoint, with one end being mappable whereas the other is not, e.g. if the non-mappable part belongs to a large insertion or is simply too short for a mapper to be found. We refer to these breakpoints which can only be observed from one end as *breakends*. Breakends that do not support already observed and classified SVs are reported as unrefined (due to the missing second alignment) single breakends.

2.3 Experimental setup

The analysis of the benchmark results consists of the following three parts. In a first stage, we want to compare all tools on even ground independent of their SV classification. Therefore, we compare all predicted single breakpoints, i.e. all novel adjacencies in the donor genome, of all three tools with the simulated variants from Mason without considering the called SV type. A single breakpoint here is a single position of an SV, e.g., for a called deletion $del(b, e)$ which starts in position b and ends in position e, we would compare the positions b and e separately to the set of all single positions simulated by Mason and for now also ignore that both positions belong to a deletion. We count a predicted single breakpoint as true positive (TP) if the difference to the simulated single breakpoint is at most 10 bp.

Next, we want to evaluate the SV classification of all tools and therefore compare the called SVs according to their SV type including a classification for deletions, inversions and duplications, i.e. we compare a called deletion $del_1(b_1, e_1)$ to a simulated deletion $del_2(b_2, e_2)$ considering both positions b_i and e_i. Here, we consider all predicted duplications as imprecise denoted as $dup\_impr(b,t)$. This way, we compare the begin and target positions b,t predicted by all tools to the begin and target position of the dispersed duplications $dup(b,m,t)$ simulated by Mason but disregarding the known length so that we can directly compare Gustaf with Pindel and Delly.

In contrast to Delly and Pindel, Gustaf classifies and fully annotates dispersed duplications and intra-chromosomal translocations. In the last part, we therefore compare the dispersed duplications $dup(b,m,t)$ and translocations $tra(b,m,t)$ predicted by Gustaf considering also the predicted length of the duplication.

For the second and third analysis part, we require a reciprocal overlap of 80% of the simulated SV length and the predicted SV length to count the predicted SV as a TP. All predicted SVs not in the simulated set are counted as false positive (FP), all not recovered simulated SVs as false negative (FN). We compare the results using sensitivity values $S = TP/(TP + FN)$ and positive-predictive-values (or precision) $PPV = TP/(TP+FP)$. The results for coverages 5, 10, 15 (default run), 30, insert sizes for paired-end data $(\mu,\sigma) = (600,60)$ and $(\mu,\sigma) = (1000,100)$, read length 150 bp (with $(\mu,\sigma) = (600,60)$ to avoid read overlaps) and different SV size range from 500 to 5000 bp are reported in Table 2.1.

2.4 Results and discussion of Gustaf

In general, we want a well performing tool to have both a high sensitivity and precision for a given set of parameters. We will first evaluate the results of all methods on the set of small SVs with the tested parameters settings for coverages, insert size and read length, and then evaluate the set of large SVs below.

Gustaf has the highest sensitivity for the small SVs usually over 95% over all tested parameter settings usually with a PPV over 90%. Note that Gustaf requires a support of three reads for coverage 10, and only two for support 5. Therefore, Gustaf's sensitivity for coverage 5 is higher (97.9% to 70.8%) but at the cost of a lower PPV (26.3% to 35.4%).

For high enough coverage, Gustaf yields a perfect sensitivity with a PPV of about 92.1%, and Pindel and Delly improve towards their highest single breakpoints sensitivities of 92.8% and 71.8%, respectively. Otherwise, Delly's and Pindel's sensitivity values vary with coverage, insert size and read length. We observe a slightly higher precision for Pindel for the single breakpoints (>99.2%) and inversions (>97.6%) compared to Gustaf (>92.1%, >92%) and Delly (>57.8%,>59.7%), but almost always at the cost of a lower sensitivity. Regarding the single breakpoint evaluation, Gustaf always has both high sensitivity and precision for all tested parameter settings whereas Delly and Pindel have a high variability with varying parameters and between sensitivity and precision.

For deletions, Gustaf has also high sensitivity values (reaching 100% for coverage

	Coverage				Insert Size		Read Length	Large SVs
Tool	5	10	15	30	600,60	1000,100	150 bp	$\geq$ 500, $\leq$ 5000 bp
	S/PPV	S/PPV	S/PPV	S/PPV	S/PPV	S/PPV	S/PPV	S/PPV
Single BP								
Gustaf (se)	0.988/0.955	**0.961**/0.971	**0.988**/0.966	0.999/0.921	**0.988**/0.966	**0.988**/0.966	**0.996**/0.963	0.991/0.948
Gustaf (pe)	**0.993**/0.946	0.948/0.961	0.986/0.962	**1.000**/0.929	0.983/0.962	0.983/0.965	0.993/0.966	0.989/0.966
Delly	0.627/0.578	0.706/0.611	0.715/0.603	0.718/0.578	0.374/0.353	0.191/0.266	0.430/0.414	**0.994**/0.913
Pindel	0.349/**0.992**	0.717/**0.996**	0.837/**0.997**	0.928/**0.981**	0.829/**0.997**	0.830/**0.997**	0.626/**0.987**	0.899/**0.991**
Deletion								
Gustaf (se)	0.938/0.804	0.792/0.864	**0.938**/0.849	**1.000**/0.774	**0.938**/0.849	**0.938**/0.849	**1.000**/**0.923**	**1.000**/**0.920**
Gustaf (pe)	**0.979**/**0.855**	**0.833**/**0.909**	**0.938**/**0.865**	**1.000**/**0.828**	**0.938**/**0.882**	0.917/**0.880**	0.979/0.870	0.957/0.846
Delly	0.583/0.173	0.688/0.205	0.688/0.204	0.646/0.190	0.583/0.204	0.438/0.231	0.458/0.172	**1.000**/0.187
Pindel	0.208/0.164	0.625/0.200	0.833/0.208	0.958/0.198	0.729/0.183	0.812/0.197	0.562/0.205	0.696/0.163
Inversion								
Gustaf (se)	**0.978**/0.968	**0.978**/0.978	**0.978**/0.978	**1.000**/0.920	0.978/0.978	**0.978**/**0.978**	**1.000**/0.968	0.977/0.935
Gustaf (pe)	**0.978**/0.978	0.967/0.967	**0.978**/0.978	**1.000**/0.929	**0.989**/0.978	**0.978**/0.947	**1.000**/0.968	**1.000**/0.978
Delly	0.707/0.596	0.783/0.610	0.772/0.597	0.793/0.598	0.293/0.172	0.011/0.007	0.370/0.239	**1.000**/0.880
Pindel	0.467/**1.000**	0.696/**1.000**	0.783/**0.986**	0.880/**0.976**	0.739/**0.986**	0.707/0.970	0.576/**1.000**	0.955/**1.000**
Dupl. impr.								
Gustaf (se)	0.990/**0.866**	**0.959**/**0.887**	0.990/0.890	**1.000**/0.790	**0.990**/**0.890**	**0.990**/**0.890**	**0.980**/**0.881**	**1.000**/0.796
Gustaf (pe)	**1.000**/0.838	0.939/0.860	**1.000**/**0.891**	**1.000**/**0.803**	0.969/0.856	**0.990**/0.866	**0.980**/0.873	**1.000**/**0.896**
Delly	0.724/0.568	0.806/0.560	0.806/0.534	0.786/0.513	0.245/0.273	0.000/0.000	0.388/0.384	**1.000**/0.589
Pindel	0.265/0.788	0.704/0.726	0.786/0.694	0.898/0.667	0.827/0.692	0.806/0.664	0.551/0.659	0.930/0.645
Duplication								
Gustaf (se)	0.980/**0.941**	**0.939**/**0.920**	0.980/**0.941**	**1.000**/**0.867**	**0.980**/**0.941**	0.980/**0.941**	**0.969**/0.905	**0.977**/0.857
Gustaf (pe)	**0.990**/0.874	0.918/0.882	**0.990**/0.907	**1.000**/0.860	0.949/0.903	**0.990**/0.907	**0.969**/**0.922**	**0.977**/**0.894**
Translocation								
Gustaf (se)	**0.980**/**0.961**	**0.940**/**0.940**	**0.980**/**0.961**	0.960/**0.889**	**0.980**/**0.961**	**0.980**/**0.961**	**0.980**/0.925	**0.929**/**0.963**
Gustaf (pe)	0.920/0.885	0.880/0.898	0.920/0.902	**0.980**/0.875	0.920/0.902	0.920/0.920	0.960/**0.941**	**0.929**/0.897

Table 2.1. Sensitivity (S) and positive-predictive value (PPV) for variants ($\geq$ 30 bp, $\leq$ 500 bp) simulated onto *chr22*, shown for different coverages. Variants $\geq$ 500 bp are shown separately (last column). Read length is 100 bp and mean insert size 300, except in the 'Insert Size' and 'Read Length' column where these parameters are varied separately. The 'Single Breakpoint (BP)' category measures how well the tools can identify individual breakpoints relative to the reference genome with nucleotide precision. Predicted breakpoints are allowed to vary by up to 10 bp from simulated breakpoints. For all other categories, the predicted variant is required to have at least 80% reciprocal overlap with the simulated variant.

30) although at lower PPVs (usually >80%). Delly and Pindel have moderate sensitivities of up to 68.8% (Delly) and up to 95.8% (Pindel) for sufficient coverage. The drastically low PPVs of Delly and Pindel were surprising. A deeper analysis showed that Delly and Pindel call the aforementioned pseudodeletion patterns as deletions. We therefore validated the assumption of called pseudodeletions by including all pseudodeletions into the set of simulated deletions for which Delly and Pindel reached much higher precisions (Delly up to 77.6%, Pindel up to 99.5%, data not shown). This supports our approach of using these deletions to resolve dispersed duplication and translocation patterns. It also explains why the PPVs for the single breakpoints are so much higher (over 91%) and why Rausch et al. (2012) report much higher precision values (over 95%) for Delly and Pindel in their own benchmark where they only simulate deletions, tandem duplications and inversions. As one conclusion from these observations, we can say that especially deletions should be treated with care and be cross examined in case they are part of more complex variants. Maybe this observation even rises an issue with already annotated and not cross validated deletions.

Inversions are generally recovered well by all tools with Gustaf always having the highest sensitivity of 97-100% with very high PPVs (92-97.8%), and Pindel usually having the highest precision of 97-100% although with comparatively low sensitivity (46-88%) depending on the coverage. Delly has robust sensitivities (70-79.3%) and PPVs (around 60%) over the tested coverages but these values decrease with increasing insert size or read length (sensitivities under 40%, PPVs under 25%).

For high coverage ≥ 15, Gustaf always has the highest sensitivity and precision for duplications, even when considering the precise duplications. Pindel can recover almost 90% of the duplications as an imprecise duplication, Delly recovers 80.6%. The precision for both tools, however, is below 78%.

None of the duplications Delly found for an insert size of $(\mu, \sigma) = (1000, 100)$ could be confirmed by the set. Since the recovery rate for duplications is quite high for Delly given sufficient coverage and smaller insert sizes, this must be an artefact for this particular data set given the high insert size.

Gustaf can recover over 94% of the small translocations over the whole tested parameter range with a precision value of almost always over 90%. Gustaf yields the best sensitivity (98%) and precision (94.2%) for the default setting with coverage of 15 and insert size $(\mu, \sigma) = (300, 30)$. On the set of small SVs, Gustaf almost always outperforms Delly and Pindel having equally high sensitivities and precisions. In addition, Gustaf called the dispersed duplications and translocations with high sensitivity and PPVs.

Gustaf can still compete for the large variant set of $500 \leq x \leq 5000\,\text{bp}$ SVs with sensitivity and PPV rates usually over 90%, showing Gustaf's ability to also handle larger SVs quite well. Compared to the small SV set, results for Pindel and Delly improve in terms of sensitivity and precision. Delly is geared towards larger

SVs > 300 bp and has generally much higher sensitivity values and PPVs, having the highest sensitivity (100%) together with Gustaf for single breakpoints, deletions and inversions but with Gustaf having a higher PPV. If we include the pseudodeletions again for validation, both Delly and Pindel here reach a precision of up to 98% (data not shown). Pindel recovers 95.5% of the inversions with full precision (100%).

Delly and Gustaf recover all duplications when evaluated as imprecise where Pindel recovers 93.0% but with Gustaf having a much higher PPV (89.6%) than Delly (58.9%) and Pindel (64.5%). When evaluating duplications including the actual length, Gustaf still recovers 97.7% while keeping a high PPV of 85.7%. Gustaf recovers 92.9% of the large translocations with high precision (96.3%).

In summary, Gustaf compares favorably with Delly and Pindel. In addition, it is to our knowledge the only tool that can call dispersed duplications and translocations including their length. Those two types of SVs are detected with high sensitivity and precision for the tested SV size ranges and parameters, exceeding even the values for Delly and Pindel where both tools call duplications only as an imprecise type. Moreover, Delly and Pindel call the pseudodeletions of these complex variants resulting in a generally low precision for deletions. Considering this benchmark set, Gustaf is well suited for small and large SVs independent of the coverage or read length.

2.4.1 Conclusion

Compared to other state-of-the-art split-read based methods, Gustaf improves detection of small SVs up to 500 bp, including the NGS twilight zone (SVs from 30 to 100 bp). For larger SVs from 500 to 5000 bp, Gustaf's results are comparable. On our simulated data set, Gustaf consistently gives good results on the tested ranges of coverage, fragment size distribution, and read length, with PPV and sensitivity mostly above 90%.

One of Gustaf's unique strengths is its ability to detect SVs that are hard to classify including dispersed duplications and translocations with exact breakpoints. On our high coverage simulated data set, Gustaf recovered up to 100% of the dispersed duplications and 98% of the translocations, both with high specificity.

Our approach is flexible in that it allows multiple splits per read. This feature will gain importance with increasing read lengths or when using Gustaf for mapping contigs. That flexibility even allows an application in mapping RNA-seq reads which may span multiple exons (for preliminary evaluation, see Trappe (2012)).

Performing a local alignment search over a whole reference genome can get computationally expensive depending on the genome size and number of reads. However, local alignment computation can be easily parallelized and can furthermore be run independently of the core of the Gustaf algorithm.

In summary, the benefit of Gustaf is its generic multi-split mapping approach which makes it flexible and versatile in terms of SV types and sizes, and the length, protocol and technology of reads.

3 Mapping-based horizontal gene transfer detection with Daisy

In bacteria, genetic material is commonly exchanged between organisms, a process known as *horizontal gene transfer* (HGT) or lateral gene transfer (Ochman et al., 2000; Boto, 2009; Wiedenbeck and Cohan, 2011). In contrast to vertical gene transfer, i.e. from one generation to the next, HGT enables the exchange of genetic material even between distant species mediated usually by transduction, transformation, or conjugation (Gyles and Boerlin, 2013). Via transduction or conjugation, the foreign DNA is carried in a plasmid or a bacteriophage, respectively, whereas via transformation, the recipient takes up nascent DNA from the environment. By means of HGT, complete genes and functional units, called insertion sequences (IS) or genomic islands (GIs), can be incorporated into the recipients' genome. Each bacterium can also carry several phages at distinct phage insertion sites. Phages of the same type, e.g. λ phages, can also carry diverse genes in their replaceable region with the result that one bacterium can have multiple highly similar phages but with different gene content.

Not surprisingly, HGT greatly contributes to bacteria's ability to adapt to changing environments (Hu et al., 2011; McElroy et al., 2014; Gyles and Boerlin, 2013). It has been demonstrated to play a major role for the acquisition of resistance to antibiotics (Barlow, 2009; Warnes et al., 2012). Moreover, HGT is not limited to bacteria but can also occur in vertebrates, including primates (and humans) (Crisp et al., 2015). However, the focus of the bioinformatics community with respect to HGT has mainly been on methods for detecting past HGT events (Ravenhall et al., 2015) from phylogenetic trees (e.g. Boc et al., 2010; Bansal et al., 2012) or based on genome composition (e.g. Metzler and Kalinina (2014); Jaron et al. (2013)). Composition properties such as GC content or k-mer frequencies usually deviate between different organisms and can therefore be used to detect sequence content of foreign origin. However, over time the foreign sequence signature ameliorates to its new host. Alien_Hunter (Vernikos and Parkhill, 2006), e.g., therefore combines various compositional characteristics or *motifs* in a variable fashion, called Interpolated Variable Order Motifs (IVOM), to improve sensitivity. Their IVOM approach does not require gene annotation or gene position information and can hence be applied to newly sequenced genomes.

Common methods aim to retrace evolutionary history of finished bacterial genomes

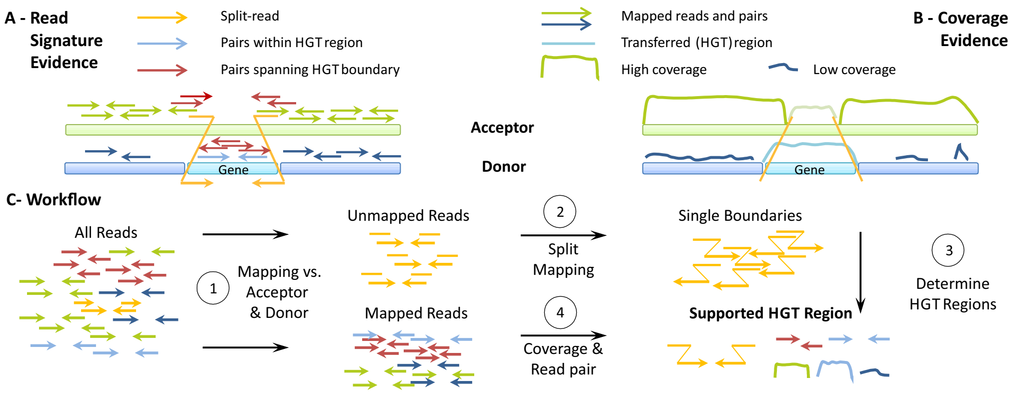


Figure 3.1. Daisy evidence and workflow. (A) Mapping evidence based on read signature: For an acceptor genome (green) and a donor genome (dark blue) with the transferred region (light blue), we evaluate read mapping information from split-reads (yellow arrows) crossing the transfer boundaries, pairs exclusively mapping within the transferred region (light blue arrows), and read pairs spanning the boundary, i.e. they have one read on either side of the boundary (dark red arrows). (B) Mapping evidence based on coverage: We evaluate the coverage based on acceptor reads (green arrows) and donor reads (dark blue arrows). We expect the coverage of the acceptor genome to be high and homogeneous (green lines) except for the HGT insertion site (light green line). The coverage in the transferred region (light blue line) should be comparable to the acceptor genome and higher than the coverage in the remaining donor region (dark blue lines). (C) Workflow overview: After initial read mapping (1), unmapped reads are split mapped to determine single HGT boundaries (2). Single boundaries are paired up to form candidate regions according to size constrains (3), and evaluated in accordance with mapping information regarding coverage and read pair signatures (4).

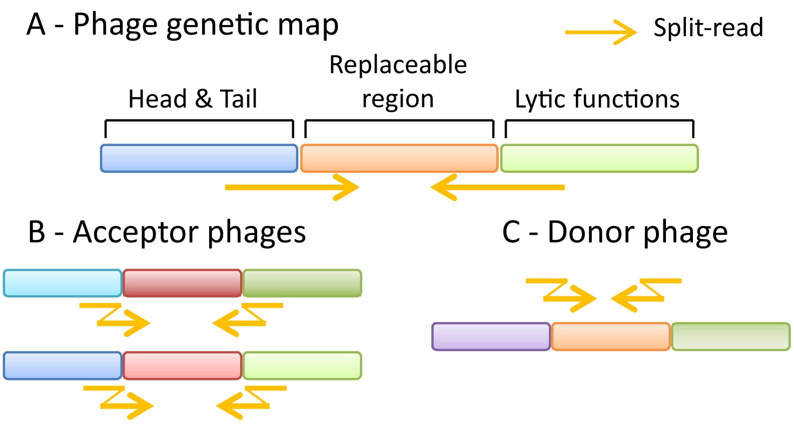


Figure 3.2. Phage composition. (A) Basic genetic map of a λ phage: Each λ phage has mosaic like coding regions for head, tail (blue), and lytic functions (green). In addition, the phage may contain an interchangeable region (orange). HGT related split-reads (yellow) cross the border between these regions. (B) The acceptor is likely to carry similar phages (variants of blue and green) with another replaceable content (red). This can make the split-read mapping ambiguous between the blue and green parts, respectively. (C) The donor very likely carries the transferred HGT region (orange) in a similar phage (blue and green) as the HGT organism.

and have mostly been developed before next-generation sequencing (NGS) became available, and hence, do not directly use NGS data. NGS technologies are well established and widely used by now, and enormous amounts of NGS data are available in public repositories. NGS also offers the chance to detect HGT events early in analyses which can be important in outbreak scenarios. One prominent example is the EHEC outbreak in Germany back in 2011 (Frank et al., 2011). Here, a non-pathogenic strain of *Escherichia coli* bacteria that resides in the gut of every human suddenly acquired two new toxins from another bacterium leading to excessive and often dangerous hemorrhagic gut infections. Especially here, identification and characterization of the pathogen causing the outbreak is highly important. Fast and reliable pathogen identification or detection of antibiotic resistance are generally of particular interest (Byrd et al., 2014). Important applications in diagnostics in the context of HGT are the detection of novel bacterial strains evolved through HGT or the distinction of a single infection with such a strain from a parallel infection by two different strains (Fricke and Rasko, 2013). This distinction is important for treatment and to prevent spreading of the disease, especially with the more frequently occurring cases of antibiotic resistances. With a special focus on these applications, we developed an HGT detection tool that directly uses NGS data.

While methods that directly address the detection of HGT events from NGS data are lacking, various methods for finding structural variations (SVs) in human exist, as for instance reviewed by Medvedev et al. (2009), Alkan et al. (2011), and Pabinger et al. (2014). Furthermore, first systematic attempts are being made to transfer methods for SV discovery to other species, including plants (Leung et al., 2015) and bacteria (Barrick et al., 2014; Hawkey et al., 2015). The latter approaches focus on detecting SVs *within* a genome and do not aim to detect the transfer of genetic material *between* species. To our knowledge, no such method exists to date (Ravenhall et al., 2015).

Conceptually, detecting an HGT event has similarities to identifying an inter-chromosomal translocation in an organism with multiple chromosomes (such as human). Nonetheless, a number of differences render existing methods not directly applicable for the purpose of detecting HGT events. On the one hand, the underlying mechanisms are different, e.g. phage-mediated transfers versus integration of nascent DNA, which potentially leads to other breakpoint signatures. On the other hand, bacteria are subject to much higher mutation rates than humans and can undergo faster evolution (Lee et al., 2012). This usually also implies fast divergence of sequences acquired via HGT (Iranzo et al., 2014). Besides sequence deviation due to evolution, reference databases still contain a number of draft genomes or mis-assemblies (Salzberg and Yorke, 2005; Kuhring et al., 2015), adding sequence deviation due to technical artifacts. Usually, the peformance of methods for calling structural variations from human NGS data deteriorates in the presence of large amounts of sequence divergence. Despite these issues, structural variant detection methods, which we briefly survey below, provide an excellent starting point to approach HGT detection when we combine their individual strengths.

The most commonly used methods to detect structural variants from NGS reads are based on *mapping* reads to reference genomes. To this end, three different paradigms exist: (i) *Coverage* information can be used to detect copy number variants (e.g. Abyzov et al. (2011); Miller et al. (2011)). This allows finding regions that are covered by significantly more or less reads than the background genome in order to predict copy number gains or losses. Such approaches are effective for large events, usually starting from approximately 5 kb, and work best if multiple samples are available for comparison, allowing for properly handling coverage biases (Dohm et al., 2008). (ii) The second class of approaches leverages *read pair* information. Here, the idea is to detect deviation from the expected relative mapping positions of two paired reads generated by mate pair or paired-end sequencing. This technique allows for uncovering also copy neutral events such as inversions, or copy-neutral translocations. The accuracy in terms of breakpoint placement and event length strongly depends on the insert size distribution of the library. In practice, approaches that first classify read pairs as concordant or discordant and then make predictions based on the discordant reads (e.g. Chen et al. (2009); Hormozdiari et al. (2010)) are usually effective for events of approximately 250 bp and larger, while approaches that use all reads (e.g. Lee et al. (2009); Marschall et al. (2012)) can predict variants starting from approximately 30 bp. (iii) Finally, it is possible to align reads across SV breakpoints, which is often referred to as *split alignment* or *split-read mapping* (Trappe et al., 2014; Emde et al., 2012; Ye et al., 2009; Marschall and Schönhuth, 2013; Karakoc et al., 2012). Such approaches can deliver single base pair resolution, but have limitations with respect to repetitive regions: Splitting the reads makes alignment ambiguity even more likely to occur than for full length reads. Especially split-read approaches then have to trade sensitivity for high numbers of false positive calls.

The different paradigms outlined above have different strengths and weaknesses and use different information sources. Therefore, many *hybrid* techniques that use more than one of these ideas have been developed in the past years (e.g. Rausch et al. (2012); Marschall et al. (2013); Jiang et al. (2012)). In contrast to these hybrid approaches, which integrate different techniques into one algorithm, *meta tools* provide a unifying platform to integrate the *results* of complementary methods into a unified variant call set (Lin et al., 2014; Leung et al., 2015).

Besides mapping reads to reference genomes for SV detection, it is also possible to subject them to *de novo* assembly (Luo et al., 2012a; Bankevich et al., 2012; Zerbino and Birney, 2008). There are many advantages to this approach, including that biases due to the choice of reference are avoided, all classes of SVs can be addressed, and 1 bp resolution is attained. However, these advantages only apply *if* the reads can be assembled into sufficiently long contigs, which cannot always be achieved from short read data. Although long read sequencing technologies can drastically improve the ability to assemble difficult regions (Chaisson et al., 2015a), short read technologies are still more prevalent, more cost effective and will thus continue to play a major role in the coming years. This holds in particular for the application to fast evolving genomes such as bacteria since here the low technical error rates of short reads is of clear advantage. Hence, short read technologies are most likely the method of choice in outbreak situations.

We introduce Daisy, a novel mapping-based HGT detection tool using NGS data. Daisy facilitates HGT detection in outbreak scenarios such as the EHEC outbreak 2011 in Germany. Outbreak situations require fast and reliable characterization of novel or unknown pathogens, or the distinction of such a novel pathogen from double infections to prevent disease spreading and to apply proper treatment. We incorporate all three paradigms of mapping-based SV detection: We identify HGT boundaries with split-read mapping and then filter candidate regions using coverage and read pair information (see Figure 3.1). The identification part ensures sensitivity in the presence of sequence divergence whereas the filtering part removes unspecific, non-HGT related events. We show the utility of mapping-based techniques for HGT detection by applying our approach to one simulated data set and two different bacterial case studies, each showing that mapping can help beyond what can be achieved with assembly for HGT detection. With Daisy, we provide an easy to use open source software relying on community standards such as VCF files and readily usable output.

3.1 Determining and sampling HGT regions with Daisy

Daisy is a comprehensive, mapping-based tool for HGT detection using sequencing data of an HGT organism, i.e. an organism with an acquired HGT. The input is a set of reads from the organism with the suspected HGT event, and the references of the acceptor genome (the parent genome of the HGT organism acquiring the HGT sequence) and the donor genome (the parent donating the HGT sequence). Determining acceptor and donor genomes from the read set is a separate pre-step not addressed by Daisy, so for now, we assume that donor and acceptor references are known.

First, we use Yara (successor of Masai, Siragusa et al. (2013)) to simultaneously map the reads against the acceptor genome reference and the donor genome reference (see Figure 3.1). In this read mapping step, Daisy identifies possible split-read candidates and also acquires mapping information around the HGT boundaries that is later incorporated for HGT support (step 1 of the workflow in Figure 3.1C, details below). We use a dedicated split-read mapper to determine the single boundaries (step 2), then pair up the boundaries to HGT regions according to size constraints (step 3), and integrate the mapping information of read pairs spanning and mapping within the HGT region (step 4). We further filter the results by a bootstrap based approach where we resample coverage and the number of reads spanning or within the HGT boundaries from random regions in acceptor and donor. As an additional step, we map the read pairs of candidate donor regions against a bacteriophage database, and flag those candidates having relevant hits. All candidate regions meeting a pre-defined support threshold are reported in VCF format. Daisy has two modes. The automated mode supports single acceptor and donor references with full filtering options. The manual mode gives the possibility to examine multiple donor genomes (see data set KO11FL for an example), although without filtering.

3.1.1 Split-read mapping

In an HGT event, a part of the donor genome has been integrated into the acceptor genome. Given a set of reads of the HGT organism, we expect to see reads mapping across the HGT boundary where one part of a read maps to the acceptor and the other part to the HGT origin in the donor (see yellow reads in Figure 3.1). When these reads are split-read mapped concurrently to both acceptor and donor, we can identify the breakpoints of an HGT event because the signature of an HGT in SV terms then resembles an inter-chromosomal translocation.

These HGT breakpoints are also the main evidence for an actual integration of the possible transferred region in contrast to potential contamination or co-existences of both donor and acceptor. We use the SV detection tool Gustaf (Trappe et al., 2014) for a dedicated split-read mapping of unmapped reads. Gustaf works with single-end data but also incorporates paired-end information from paired-end data. It can handle multiple splits per read and alignment gaps at the read ends or in the middle of the read which is an important property in view of the high bacterial evolutionary rate and common micro-homologies at breakpoint locations. The expected number of split-reads depends on the read coverage and the evolutionary distance between the sequenced organism and its putative acceptor and donor genomes used for analysis. The default value of the user definable parameter for the required number of split-reads is therefore set to 3 (very sensitive but avoiding random split-reads).

3.1.2 Candidate identification

The single breakpoints from the split-read mapping give possible start and end positions, in both acceptor and donor, of an HGT event. The combination of these start and end positions is subject to size constrains regarding the regions delimited in the acceptor and donor genomes in order to sensibly restrict the number of candidate regions. Depending on whether only single genes, operons, or complete bachteriophages are transferred, these regions can vary largely in size and range from a few hundred to several thousand base pairs. The delimited region in the acceptor genome can also be equally large as the designated HGT region if, e.g., another bacteriophage is occupying the destined phage insertion site there. The values for minimal and maximal HGT size are therefore parameterized and user definable. Default values used in the benchmarks are 500 bp and 55,000 bp for minimal and maximal HGT size, respectively. We also reduce duplicate entries. Once we identified a valid candidate, we remove any further identified candidates within a base pair range of a specified tolerance (default 20 bp) around acceptor and donor start and end positions.

3.1.3 Coverage and read pair integration

Each candidate region is then examined for additional mapping support regarding mean coverage, number of pairs spanning and within HGT boundaries (see ”Mapping Evidence” in Figure 3.1). Coverage can vary due to extreme GC content, sequencing efficiency or rearrangement events such as induced by a HGT. Theoretically, the expected coverage of the acceptor genome should be equal to and as homogeneous as the sequencing coverage of the HGT organism (depicted as ”High coverage” in Figure 3.1), except for the HGT insertion site. The coverage of the HGT insertion site should be much lower because the sequence content is unrelated to the HGT organism or donor. Exceptionally, the coverage could be equally high when the insertion site is occupied by another related phage (see below). The coverage of the donor HGT region should, again theoretically, resemble the sequencing coverage of the HGT organism. On the contrary, the coverage of the remaining donor should be low (depicted as ”Low coverage” in Figure 3.1) because the sequence content is unrelated to the HGT organism.

Depending on the evolutionary distance between HGT organism, donor and acceptor, the observed coverage properties of the region can deviate. A direct statistical comparison, e.g. using the framework proposed in Lindner et al. (2013), may lead to insignificantly small values, even for the true HGT regions. We therefore introduced a bootstrap like resampling method where we test the candidate regions compared to equally sized random regions in acceptor and donor. The default sampling size is 100 random regions. As stated earlier, the donor region coverage should be higher than the coverage of the remaining donor. Per default, we require the donor region mean coverage to be higher than the coverage of the random donor regions in at least 95% of the cases, i.e. to have a bootstrap result of ≥ 95. Again, the acceptor region should be unrelated (low coverage) or also have phage origin (high coverage). Hence, we require the mean coverage to be either higher (alternative phage) or lower (unrelated sequence) than the random region coverages in at least 95%, i.e. the bootstrap value has to be (≥ 95 or ≤ 5).

In addition to coverage evidence, we also incorporate mapping evidence from read pairs that are spanning the HGT boundaries (dark red reads in Figure 3.1) and those that map completely within the HGT boundaries of the donor (light blue reads). For the spanning pairs, one mate is mapping on one site of the boundary outside the HGT region in the acceptor whereas the other is mapping on the other site of the boundary inside the HGT region in the donor. For the spanning pairs, we require that both reads have to map within a range of half the defined maximal HGT size from the boundary (i.e. a total range of the maximal HGT size around the boundary). For the pairs within, we compare the number of pairs within the boundaries to the number within equally sized random regions where we expect the HGT region to have more such pairs than the random regions. We apply the same idea of a resampling method from the coverage evidence (using the same random regions) for the evidence from mapped read pairs spanning and within the HGT boundaries. The required resampling value is also 95.

3.1.4 Bacteriophage screening

If the HGT was phage mediated and HGT organism and acceptor contain, and maybe share, several similar phages, the results obtained via split-read mapping can be ambiguous (see Figure 3.2). After filtering the HGT candidates, we therefore screen the EBI phage database (Brooksbank et al. (2014)) from the European Nucleotide Archive (ENA) (Leinonen et al., 2011) for evidence of the candidates' donor HGT regions. We first map all reads against the phage references and during the screening, we evaluate if the reads mapping within or across the donor HGT region also map to any database entry and report this percentage in the TSV output file (see below). This step is not a filter step but intended as an additional flag for each candidate.
The filtered candidates are written to a VCF output file (Danecek et al., 2011), all candidates with bootstrap information are written to a TSV file.

3.2 Experimental setup

Datasets

We tested our method on one simulated and two real data sets, each containing an HGT event with distinct challenges.

H. pylori. The *Helicobacter pylori* data set is a simulated set. Here, we chose *Escherichia coli* K12 as the acceptor and *H. pylori* strain HPML01 (Acc.-Nr.AP014710.1) (Wang et al., 2015) as the donor. *E. coli* K12 substr. DH10B has no λ phages and its *wrbA* insertion site is at 1 120 263 - 1 120 859. *H. pylori* HPML01 has a Helicobacter

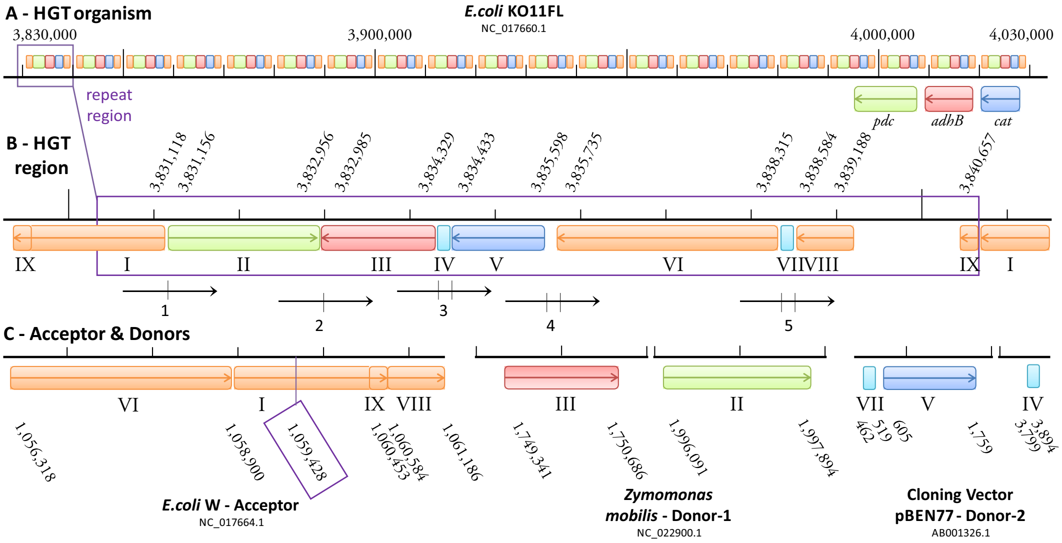


Figure 3.3. KO11FL composition of HGT region. (A) HGT organism *E. coli* KO11FL: The transgenic KO11FL has 20 copies of the transferred HGT region enclosed by the purple rectangle. (B) HGT region composition of KO11FL: Shown are positions of the transgenic genes *pdc* (green), *adhB* (red) and *cat* (blue) and adjacent segments (I-IX) within *E. coli* KO11FL. Reads enumerated 1-5 span HGT related adjacencies of segments I-IX, two dashes on a read imply multiple adjacencies or gaps, resulting in multiple splits of the read. (C) Acceptor & donor HGT region composition: Shown are positions of *pdc*, *adhB*, and *cat* in the donor references *Z. mobilis* and pBEN77, and the HGT insertion site in acceptor reference *E. coli* W. Note the different order and orientation of some of the segments I-IX compared to (B). All positions in (A)-(C) were determined with BLAST.

phage 1961P-like sequence at about 1 322 000 - 1 350 000 (Helicobacter phage 1961P Acc.-No. NC_019512.1, Luo et al. (2012b))).

Before inserting the phage-like sequence at the *E. coli* K12 *wrbA* insertion site, we introduced SNPs (rate 0.01), small indels (1-6 bp, rate 0.001) and large indels (50-1,000 bp, rate 0.00001) using the simulator Mason2 (Holtgrewe, 2010, 2014) into *E. coli* K12 as well as SNPs (rate 0.001) and small indels (1-4 bp, rate 0.0001) into the phage-like sequence. Evolutionary rates and variant sizes were chosen according to Lee et al. (2012) and were also intended to account for multiple generations, i.e. with larger evolutionary distance between HGT organism and acceptor and donor but conserved HGT boundaries. The location of the modified phage within the *modified E. coli* K12 is then 1 117 289 - 1 145 285. We simulated 150 bp paired-end reads with Illumina error profile from 500 bp fragments with 10% sd and 100x coverage with Mason2. The simulated data set is available in the Daisy github repository. The paper results can be reproduced following the usage guidelines and using default parameters as stated in the repository readme.

KO11FL. The first real data set includes *E. coli* W as the acceptor and *Zymomonas mobilis* as the donor genome as well as the cloning vector pBEN77 as a second donor, the resulting genome is the transgenic *E. coli* KO11FL (Turner et al., 2012). The KO11FL is a laboratory version of the original transgenic KO11 (Ohta et al., 1991). *E. coli* W is the parent strain of KO11FL which contains a cloned operon *pcl* including the genes *pdc* and *adhB* from *Z. mobilis*, and a *cat* gene not present in *Z. mobilis*. We therefore chose pBEN77 as a donor genome for the *cat* gene. This transgenic biotechnology scenario resembles a natural HGT event and gives the necessary ground truth on real data.

Figure 3.3 depicts the composition of *E. coli* KO11FL with the transferred genes *pdc* (green), *adhB* (red) and *cat*, the target site of the acceptor genome *E. coli* W (insertion breakpoint framed purple), and excerpts of the donor genomes *Z. mobilis* and cloning vector pBEN77 indicating the positions of the transferred genes. The purple framed HGT region in *E. coli* KO11FL has 20 consecutive copies. The exact order, orientation and positions of the segments enumerated with I-IX has been determined with BLAST (megablast, default parameters). The adjacency of segments I and II defines the first HGT boundary and the adjacency of V and VI defines the second boundary number. The important and challenging part is that II and VI belong to two different donors, i.e. we have a transition within the HGT region and cannot define a single candidate region by pairing up single boundaries as we did for the *H. pylori* data. However, since the transfer was very recent and the boundaries are still clear enough, we will aim to detect all HGT related boundaries via split-read mapping alone.

The read types numbered 1-5 in Figure 3.3 are the expected split reads relevant for or related to HGT detection. Reads 3-5 have multiple splits indicated by two dashes, i.e. the split read mapper must be able to handle multiple split reads. For read 4, the middle part of the read, 137 bp, is covered by neither acceptor nor donor genomes, i.e. the split-read mapper has to also handle such scenarios. Reads 3-5 are multiply split and read 4 spans a gap of over 130 bp, which shows the necessity of a sensitive and versatile split-read mapper. Adjacencies II-III and IV-V reflect intra-chromosomal rearrangements in *Z. mobilis* (II-III) and pBEN77 (IV-V) and are not part of this evaluation. KO11FL has been originally assembled from Roche 454 reads (Turner et al., 2012). Contig gaps have been filled by PCR and Sanger sequencing, and then resequenced Illumina short read paired-end data has been assembled using the Roche 454 assembly as a template. However, only the Roche 454 reads are available via the SRA (SRX022824) and have been used in our benchmark.

EHEC. For the second real data set, we use the *E. coli* O157:H7 Sakai strain as an HGT organism. The *E. coli* O157:H7 serotype is associated with diseases most often and the Sakai strain has been sequenced from an outbreak in Japan (Zhang et al., 2007). *E. coli* O157:H7 arose from the enteropathogenic *E. coli* O55:H7, the acceptor, acquiring Shiga-Toxins (Stx) via HGT of a lamdoid phage in the sequential evolution from its progenitor (Kyle et al., 2012). *E. coli* strains that have both Stx1 and Stx2 have been shown to carry them in two separate and distinct (Herold et al., 2004) lambdoid phages (Allison et al., 2003). The *Shigella dysenteriae*, the assumed donor, is the only Shigella serotype carrying Stx (Yang, 2005). Stx1 is almost identical to the *Shigella dysenteriae* toxin (Shaikh and Tarr, 2003), whereas Stx2 only shares up to 60% with Stx1. Stx in *S. dysenteriae* at positions 1 283 705 - 1 285 203 is carried by the lambdoid stx-phage P27 (all Stx phages are lambdoid bacteriophages (Smith et al., 2012)).

In *E. coli* O157:H7 Sakai, the Stx1 phage Sp15 occupies the insertion site *yehV*, Stx2 phage Sp5 insertion site *wrbA* (Kyle et al., 2012). In *E. coli* O55:H7, *yehV* is occupied by another lambdoid phage (*Cp*10, see Table S1 in Kyle et al. (2012)), whereas *wrbA* is still intact (i.e., there is no phage at this insertion site, and the *wrbA* protein coding gene is intact (Shaikh and Tarr, 2003)).

The reads (SRX172546) are from a Illumina MiSeq paired-end whole genome shotgun sequencing run of *E. coli* O157:H7 Sakai. The read length is 151 bp with fragment length 535, the coverage is approximately 105 x. As acceptor reference, we used the *E. coli* O55:H7 strain RM12579 (Acc.-No. of chromosome is NC_017656.1). The donor reference genome is *S. dysenteriae* Sd197 (Acc.-No. CP000034.1).

Assembly Approach

A comparable approach for HGT detection is de novo assembly with subsequent whole-genome analysis. We chose SOAPdenovo2 (Luo et al., 2012c) as a suitable assembler, in particular for short-read Illumina data (GAGE, Salzberg et al. (2012)). We assembled the reads of all data sets with parameters SOAPdenovo-127mer all -R -F -u -K 31 -m 91.

We applied BWA-MEM (Li and Durbin, 2009) in order to detect possible HGT breakpoints directly on the scaffolds. We further applied Alien_Hunter (Vernikos and Parkhill, 2006) to detect possible GIs on the assembly. Alien_Hunter exploits compositional sequence biases using *k-mer* motifs of variable length. Since Alien_Hunter is designed for fully sequenced genomes, we also applied Alien_Hunter to the HGT organism reference genomes (i.e., the simulated *H. pylori* genome, *E. coli* O157:H7, and *E. coli* KO11FL) for evaluation purposes. In our general use case however, we assume that the HGT organism has not been sequenced yet or is unknown.

The detected regions from Alien_Hunter have a score and are ranked in descending order. The higher the score, the more likely the region matches a foreign genomic island. Daisy and BWA-MEM candidate regions are not scored and therefore not ranked. For Daisy, we consider valid HGT candidates passing the 95% sampling threshold. For BWA-MEM, we consider all regions conforming the same size constraints as for Daisy.

Settings

Daisy and SOAPdenovo2 are run on a 120 Core Linux server with 1024 GB RAM and 1.5 TB SSD with the following parameters. For Daisy, the runtime is around 103 min for the *H. pylori* data, 176 min for the KO11FL data, and 14.3 hours for the large EHEC data set.

Daisy. As part of Daisy, we ran Yara in all-mapper mode (-a) with high error tolerance (-e) of 10%, thread number (-t) of 5, and the appropriate values for library length and library error for paired-end data, i.e. -ll 500 -le 50 for *H. pylori* and -ll 535 -le 50 for EHEC.

We ran Stellar with default parameters except for minimal match length (-l 30). To account for the higher evolutionary rate with frequent small indels, we chose generous gap-related parameters for Gustaf as the default values in Daisy. We allow initial gap lengths of a read of 50 bp (-ith 50) where breakends are only called when larger 70 bp (-bth 70), and gaps up to 100 bp within the read, i.e. between split parts (-gth 100). For the KO11FL, we set -gth 150 to account for the larger artificial gaps. Required read support is set to 3 (-st 3). Library length and error (-ll and -le) as well as thread number for I/O (-nth) are set as in Yara.

The automated HGT evaluation is run with HGT minimal size 500 bp and maximal size 55 000 bp, tolerance for duplicate removal 20 bp, 100 sampling regions and 95% sampling sensitivity (default parameters). For the KO11FL with its two donors, the manual mode of Daisy is run and we manually investigated the single boundaries reported by Gustaf.

3.3 Results

H. pylori

Daisy finds one true positive (TP) HGT candidate with base pair precision without any false positives (FPs) (see Table 3.3 (A) *H. pylori*). Of the spanning read pairs and pairs within, 53% are mapping to the bacteriophage database. This indicates a phage-natured origin of the donor-region which conforms with the ground truth of the inserted Helicobacter phage 1961P-like sequence.

Assembly. Assembly of the simulated reads resulted in 23 936 contigs, 6 484 of them covered by one of the 17 452 scaffolds (N50 of 89 444). BWA-MEM also only reports the correct breakpoints with base pair precision and without FPs. Alien_Hunter detects the region on both the complete genome and the respective assembly scaffold but finds another 63 alternative hits on the assembly and 62 on the complete genome. The regions depicted by Alien_Hunter deviate over 2 200 and up to 5 108 bp regarding the true start and end positions. All three tools detect the HGT region as the best (Alien_Hunter) or only candidate but only Daisy and BWA-MEM with base pair precision.

KO11FL

The transgenic *E. coli* KO11FL is a case study with a very recent artificial transfer and, due to the transgenic origin, two donor references. The composition of the HGT region is shown and explained in Figure 3.3. The read types labeled 1-5 in Figure 3.3 cover the HGT related boundaries which leaves adjacencies I-II, III-IV, V-VI, VI-VII, and VII-VIII as ground truth. To handle two donor genomes, we ran the split-read mapping step with all three genomes, i.e. the acceptor and both

Table 3.1. Daisys KO11FL results. References are coloured according to Figure 3.3. Column *TP* (true positives) states which of the adjacencies between segments I-IX in Figure 3.3 the boundary covers, if any. Column *FP* (false positives) states possible adjacencies when considering alternative repeat region compositions within the 20 copies in *E. coli* KO11FL, empty entries are unrelated FPs. The column *Reads* states the number of split-reads supporting the translocation.

Single Boundaries				TP	FP	Reads
E. coli W	1 058 889	*Z. mobilis*	1 996 084	I-II		3 294
pBEN77	520	*E. coli* W	1 061 186	VII-VIII		872
pBEN77	597	*E. coli* W	1 058 886	V-VI		836
pBEN77	459	*E. coli* W	1 056 308	VI-VII		737
pBEN77	3 795	*Z. mobilis*	1 749 333	III-IV		170
Z. mobilis	1 996 090	*E. coli* W	1 059 365		II-VI	57
Z. mobilis	1 996 090	pBEN77	1 410		II-V	46
E. coli W	1 060 453	*Z. mobilis*	1 750 071		III-IX	35
pBEN77	1 643	*E. coli* W	1 058 666		V-VI	33
pBEN77	828	*Z. mobilis*	1 750 690		III-V	32
E. coli W	2 206 429	*Z. mobilis*	1 996 084			30
Z. mobilis	1 750 302	pBEN77	1 127		III-V	25
E. coli W	1 056 310	*Z. mobilis*	1 996 992		II-VI	24
E. coli W	1 059 006	*Z. mobilis*	1 996 681			24
Z. mobilis	1 996 090	*E. coli* W	1 059 283		II-VI	22
E. coli W	1 060 926	*Z. mobilis*	1 996 277			20

AS	AE	MC	BS:MC	DS	DE	MC	Split	PS-S	PS-W	Phage	BS:MC	BS:PS-S	BS:PS-W
1 736 996	1 739 268	21.21	2	1 320 842	1 322 115	133.9	81	95	229	0.3272	100	100	98
1 741 535	1 744 926	157.74	98	1 283 673	1 288 079	64.72	45	6	820	0.9746	99	100	98
1 958 870	1 982 375	35.74	5	4 034 933	4 035 782	358.52	45	26	677	0.9801	100	100	100
1 992 115	1 992 955	176.61	99	1 320 883	1 322 056	136.67	19	136	227	0.2397	99	100	97
3 607 310	3 630 353	43.35	4	4 189 799	4 198 011	201.3	82	6	4 901	0.9762	100	100	100
3 607 310	3 632 993	39.53	4	4 189 799	4 206 818	99.88	16	4	4 967	0.9624	100	99	100
2 672 402	2 673 467	101.58	87	1 288 585	1 329 814	13.65	7	30	1 143	0.1628	51	94	18
2 672 402	2 675 586	155.7	96	1 288 585	1 329 491	12.45	23	14	1 097	0.117	40	94	12
2 672 402	2 678 767	99.95	95	1 288 585	1 323 371	12.58	25	3	1 044	0.0936	40	97	21
2 672 402	2 679 015	97.95	94	1 288 585	1 323 123	12.38	11	1	1 021	0.0959	39	95	20
2 673 466	2 678 767	99.61	93	1 323 371	1 329 813	19.45	24	0	101	0.69	68	0	20
2 673 466	2 679 015	97.24	92	1 323 123	1 329 813	20.22	10	0	112	0.63	81	0	18
2 675 585	2 678 767	44.2	12	1 323 371	1 329 490	11.66	47	13	55	0.56	50	99	10
2 675 585	2 679 015	44.38	3	1 323 123	1 329 490	12.78	19	13	66	0.48	49	99	9

Table 3.2. Evaluation results of Daisy for EHEC data set: The listed entries show acceptor (AS-AE) and donor (DS-DE) positions, mean coverage (MC), split-read support (Split), pair support - spanning (PS-S) and within (PS-W) and bacteriophage database hits (Phage), bootstrap (BS) results for thresholds ≤ 5 or ≥ 95 for acceptor and ≥ 95 for donor regions. The upper results are filtered candidates of Daisy. Daisy identifies only six candidates where all acceptor positions match alternative phage insertion sites (blue background, Asadulghani et al. (2009); Kyle et al. (2012)). One candidate matches the true donor position (green background) but not the correct acceptor side according to literature. The lower grayed-out entries closest resemble the true insertion site in the acceptor according to literature (green background) but the low sampling values (BS) indicate otherwise. Taken together, these results suggest an alternative Stx-phage insertion in *E. coli* O55:H7 at position 1 741 535-1 744 926 rather than at 2.67 million bp.

donors, and omitted the coverage and read pair filtering.

A summary of the reported breakpoints is listed in Table 3.1 with references colored according to Figure 3.3 for easier reference. The five highest ranked TPs cover all adjacencies stated above, and, likely due to the 20 copies, have distinct high support of over 170 up to 3 294 reads whereas the FPs attain support values only up to 59, allowing perfect separation by a simple cutoff. Furthermore, most FPs can be assigned to an adjacency not expected based on the ground truth shown in Figure 3.3. We cannot assess whether some of these additional adjacencies reflect alternative compositions of the respective components in some of the 20 copies.

Table 3.3 (B) states the total number of breakpoints (16) as the *number of hits* with a total of five true positives. As breakpoint distances on donor and acceptor, we calculated the mean distance of all true positive breakpoints involving acceptor-donor boundaries (adjacencies I-II and V-VI) since these enclose the region and can therefore be compared with Alien_Hunter.

We could successfully detect all of the five possible split-read types including the multiple split ones (reads 3-5 in Figure 3.3) and the one read type covering a gap of over 130 bp (read 4). However, we could not verify the results with the HGT filter due to the two-donor scenario. Still, this case study shows that, given sequencing data of recent HGTs, it is possible to determine the correct boundaries with accurate base pair resolution and high confidence even by split-read mapping alone.

Tool & Data	True Region Detected	Number TP/FP	Breakpoint Distance HGT Organism (start/end)	Breakpoint Distance Acceptor (start/end)	Breakpoint Distance Donor (start/end)
(A) *H. pylori*					
Daisy (reads)	yes	1/0	n/a[1]	0/0	0/0
Alien_Hunter (genome)	yes	1/62	2 289/2 215	n/a[1]	n/a[1]
Alien_Hunter (assembly)	yes	1/63	5 108/2 888	n/a[1]	n/a[1]
BWA-MEM (assembly)	yes	1/0	n/a[1]	0/0	0/0
(B) KO11FL					
Daisy (reads)	yes	5/16	n/a[1]	39/14	22/8
Alien_Hunter (genome)	yes	15/109	1 732/1 174	n/a[1]	n/a[1]
Alien_Hunter (assembly)	assembly failed				
BWA-MEM (assembly)	assembly failed				
(C) EHEC					
Daisy (reads)	yes	1/5	n/a[1]	0/0 *	32/2 876
Alien_Hunter (genome)	yes (Stx1+2)	2/91	0/0 *	n/a[1]	n/a[1]
Alien_Hunter (assembly)	no	0/382	—	—	—
BWA-MEM (assembly)	yes	1/176	n/a[1]	0/0 *	1 864/39 049

Table 3.3. Results for Daisy compared to Alien_Hunter and BWA-MEM: For evaluation purposes, Alien_Hunter has been applied to both the assembly and the full reference genome. Column *True Region Detected* states if the method was able to find the correct HGT region. Column *Number TP/FP* reports the number of true positive (TP) and false positive (FP) hits with regard to the sought HGT event. In the last three columns, we state the precision of the correct candidate in terms of breakpoint distance[1]. For Alien_Hunter, we calculate the base pair distance of the correct candidate region to the ground truth on the HGT organisms reference (column *distance true region*), for Daisy and BWA-MEM, we calculate the base pair distance on acceptor and donor (columns *breakpoint distance acceptor and donor*) (in the form start distance/end distance). Due to the ambiguous positioning of the HGT genes within the phage(s), columns with * state zero distance if breakpoints lie within the designated phage region. (A) For the simulated *H. pylori* data set, all three methods are able to detect the HGT region as the best (Alien_Hunter) or only candidate, but only Daisy and BWA-MEM with base pair precision. (B) For KO11FL, an assembly using SOAPdenovo2 did not produce any scaffolds, and Alien_Hunter was applied to the KO11FL genome only. Both Daisy and Alien_Hunter then both detect the region. (C) For the EHEC data set, Daisy finds the true candidate. Alien_Hunter finds 382 candidate regions when applied to the assembly but none match the scaffold with the HGT region. With BWA-MEM, we find 177 candidate regions, the closest is overlapping the true HGT region but without breakpoint precision. [1]Note that Alien_Hunter reports candidate regions with regard to the reference HGT organism whereas Daisy and BWA-MEM report breakpoints on acceptor and donor.

Assembly. Assembly of the 454 reads from the KO11FL data set with SOAPdenovo2 resulted in 455 419 contigs (singletons) of up to 1 011 bp. No scaffold was constructed, likely due to the repetitive nature of the genome. Turner et al. (2012) also pointed out the long gaps between contigs in their assembly of this data set which they had to fill with PCR and additional Sanger sequencing. Due to the failed assembly, we did not apply Alien_Hunter or BWA-MEM to the contigs. When applied to the finished full KO11FL genome instead of the assembly, Alien_Hunter finds a total of 109 potential GIs where 15 of them overlap with 15 of the 20 copies of the transferred genes *pdc*, *adhB*, and *cat* in KO11FL (see also Figure 3.3 A - HGT organism and B - HGT region). The mean distance for start and end position of this 4 442 bp region, however, is 1 732 (start) and 1 174 (end). This is much higher than the mean distances for Daisy. Since we assume that the full genome is not available, Daisy outperforms the assembly approach even for the longer 454 single-end reads.

EHEC

For the EHEC data set, the true transfer according to literature (Kyle et al., 2012) and BLAST hit examination is 2 643 556-2 694 691 in *E. coli* O55:H7, where the λ phage Cp10 occupies the phage insertion site *yehV*, to positions 1 283 705-1 285 203 in *S. dysenteriae* Sd197 where a defective prophage carries the Stx genes.

Applying Daisy, we created 145 HGT candidates of which six passed the resampling filter. Table 3.2 lists the result values for these six candidates. The true acceptor positions from *E. coli* O55:H7 stated above are not among them, and only one pair of donor positions matches the stated *S. dysenteriae* positions. The true acceptor positions are among the remaining filtered out candidates but have very low bootstrap values (table 3.2 lower, greyed-out reports). So at first glance, it seems as if Daisy created the correct candidates but then too strictly filtered them out while keeping unrelated hits.

However, O55:H7 contains other λ phages (Kyle et al., 2012), two of them (Cp7 and Cp9) have high similarity (up to 99% BLAST identity) to Sp15 (Stx1) at the *yehV* phage insertion site in *EHEC* O157:H7. When we look more closely at the six identified candidates, we observe that all of them have acceptor positions matching an alternative phage insertion site (see Table S1 in Kyle et al. (2012) for details on the phage insertion sites). Among these six candidates, there is the one true candidate regarding the donor positions (1 283 673-1 288 079). The acceptor coordinates (1 741 535-1 744 926) belong to the λ phage Cp7.

The candidate with the true donor positions encloses a region that is 2 876 bp larger than the actual Stx part. A BLAST search of this additional part yields hits on shiga toxin genes and ORFs, and Stx phage and prophage genes, as well as four further hits to *S. dysenteriae* Sd197 (CP000034.1) that match the donor regions of the remaining five candidates. These hits suggest a phage-origin of this additional

2 876 bp (1 285 203-1 288 079) as well as these donor regions. This is supported by the high percentage of donor region read pairs matching an entry in the bacteriophage database (up to 97%). The percentage of phage database hits of the one remaining candidate is also around 97%, suggesting another alternative phage-site in *S. dysenteriae* Sd197 as well. The donor positions of all of the filtered out candidates with matching acceptor positions also all fall within the phage-region ranging from position 1 288 585 to 1 329 490 (data not shown).

So the true challenge in this case study is the fact that the HGT was phage mediated, and that both acceptor and donor have several alternative and occupied phage insertion sites. According to Asadulghani et al. (2009), the same set of bacteriophages can also occupy different phage insertion sites between individuals of the same bacterial strain, making it possible that our candidate (1 741 535-1 744 926 to 1 283 673-1 288 079) actually is the true or most likely candidate in this case. Given the currently available information, we cannot verify that the six phage-related candidates are correct, but there is also sufficient evidence to consider them as such.

Assembly. SOAPdenovo2 assembled 100 601 contigs, 14 897 of them covered by one of the 93 905 scaffolds (N50 of 150). *E. coli* O157:H7 carries Stx1 and Stx2 at distinct locations. However, assembly examination with BLAST suggests that both toxins have been assembled on the same scaffold.

Both toxin HGT regions lie within a phage so it is difficult to ascertain the specific positions of the genes carried by the phage. For Alien_Hunter this is difficult because the tool already (correctly) recognizes the phage sequence itself as a GI. We therefore count *true region found* as yes, if the candidate region is overlapping the phage carrying the toxins. Alien_Hunter finds both Stx regions when applied to the HGT organism, but one is the region with the lowest rank (see Table 3.3, (B) EHEC). The tool finds 382 candidate regions when applied to the assembly but non is matching the scaffold with the HGT region. With BWA-MEM, we find 177 candidate regions. One region is overlapping the true HGT region but without breakpoint precision on the donor. The region is reaching into the repetitive genome part following the shiga-toxin region in *S. dysenteriae* Sd197. The results acquired via BWA-MEM also support our hypothesis of an alternative phage insertion site: The acceptor region of this hit is 1 741 843-1 742 439. For all three tools, the true HGT candidate is integrated into a phage and, hence, we assign breakpoint distance zero (0/0).

3.4 Discussion of results from Daisy

To the best of our knowledge, Daisy is the first approach that allows HGT detection directly from NGS data without requiring a *de novo* assembled genome. It rather relies on detecting HGT boundaries via a split-read mapping approach. It uses the acceptor and donor genomes of the HGT as reference, and integrates coverage and read pair information for HGT candidate evaluation. Daisy facilitates applications related to outbreak scenarios of HGT related pathogens like, e.g., detection of novel bacterial strains evolved through HGT or the distinction of a single infection with such a strain from a parallel infection by two different strains.

We critically evaluated Daisy on three data sets. It has been often noticed that SV detection methods are hard to evaluate since the existence and exact positions of breakpoints are often not known. This is particularly true also for HGT events. In this study, we therefore focused on one simulated and two real data sets for which we also provide partial ground truth for future comparison. The data sets were chosen to both show the power of the approach but also to explore the limitations and provide guidance for other experiments. On the simulated *H. Pylori* data, Daisy produced the correct true positive candidate without false positives. For the real KO11FL data, the five single boundaries with the highest total split-read support already cover all five HGT related boundaries and have a distinctly higher support than the first false positive hit. For the real EHEC data, we called six candidates which all fall into alternative phage insertion sites in both acceptor and donor. The alternative assembly only produced meaningful assemblies for *H. pylori* and EHEC. On the *H. pylori* data set, Alien_Hunter and BWA-MEM both found the HGT region as the best candidate, but Alien_Hunter with low breakpoint precision and many alternative hits. On the EHEC data set, only BWA-MEM found the true candidate on the assembly data but with more FPs than Daisy. Our mapping-based HGT detection approach, which integrates several SV detection methods, is therefore a highly useful strategy. The EHEC example, where we reduced the 145 pure split-based candidates to a few candidates with required HGT signatures, shows how our candidate evaluation successfully filters out false positive hits. Although these use cases were overall successful, the results also show some remaining challenges and need for future development, which we outline below.

One prerequisite of the current approach is that all involved acceptor and donor genomes are known. Selecting these candidate genomes given a set of reads from the HGT carrier genome is a challenging task of its own. It is closely related to the metagenomics problem of finding all occurring species contained in a sample given a set of reads, and is a crucial pre-step for applications in diagnostics. Thus, tools such as MicrobeGPS (Lindner and Renard, 2015) or Kraken (Wood and Salzberg, 2014) can serve to identify candidates for follow-up analysis with our tool.

In this first version of Daisy, we focused on the idea of using mapping-based evidence such as coverage and read pair signatures. Existing parametric HGT detection methods use genome signatures such as differing GC content (Daubin et al., 2003), atypical codon usage (Lawrence and Ochman, 2002) or *k-mer* frequencies (like Alien_Hunter) for identification. In a more comprehensive future version, these genome signatures are possible further filtering options of candidate regions.

Currently, automated filtering and HGT candidate evaluation is only available for a single donor genome. More complex, decomposite HGT regions, consisting of multiple genes from various donor genomes such as in the KO11FL example, require more sophisticated combination and evaluation of candidates and paired support across the donors. An automated extension could benefit the application also in the context of, e.g., genetically modified organisms. While the detection is possible with our approach, as seen in the KO11FL, more manual investigation is required.

As a mapping-based approach, Daisy naturally depends on available reference genomes. Also, in recent HGT events or artificial gene transfers, mapping to reference genomes is easier than for longer evolutionary time spans. As a result, HGT boundaries are more obvious and identifiable with higher confidence without strong influences of evolution. The EHEC data example shows that the ongoing fast evolution of bacteria makes HGT boundaries fuzzy, parallel HGT events obscure boundaries, and HGTs mediated by phages make the area around the target gene ambiguous and evaluation difficult. In general, our approach should be seen as a step towards a more comprehensive analysis pipeline of sequencing data where multiple complementary methods are integrated. The goal of such a pipeline would be a full investigation of a bacterial genome with, e.g., genome annotation and classification, SNP and SV characterization, HGT detection and more.

With our tool Daisy, we present the first mapping-based HGT detection approach known so far. Our approach shows sound results with base pair precision for simulated and real data sets. Alternative assemblies give supportive results but were not successful for all data sets. Daisy was built for and evaluated on bacteria, but should in principle also be applicable for HGT detection in other organisms such as plants.

4 HGT acceptor and donor identification with DaisyGPS

For a long time, evolution in terms of gene transfer was thought to happen only along the tree of life, i.e. from parent to offspring generation. The discovery of horizontal gene transfer (HGT) (Ochman et al., 2005; Boto, 2009; Wiedenbeck and Cohan, 2011; Daubin and Szöllősi, 2016) has revolutionised this dogma, and revealed the mechanism that enables bacteria to quickly adapt to environmental pressure (Hu et al., 2011; McElroy et al., 2014; Gyles and Boerlin, 2013). Via HGT, bacteria can directly transfer one or multiple genes from one individual to another across species boundaries. The known and prominent mechanisms of HGT are transformation (uptake of nascent DNA from the environment), conjugation (direct transfer from cell to cell), and transduction (transfer via bacteriophages) (Gyles and Boerlin, 2013). In all cases, a piece of DNA sequence is - directly or indirectly - transferred from the so called donor organism to the acceptor organism and integrated into the genome (see also Figure 4.1). Especially conjugation and transduction facilitate the transfer of pathogenicity islands and mobile genetic elements involving antimicrobial resistance (AMR) genes (Barlow, 2009; Warnes et al., 2012; Juhas, 2013). Today, we are facing the rise of so called "superbugs" (Juhas, 2013; Perry et al., 2014) as a result of bacterial adaptation and gain of resistance to antibiotic treatment, showing the need for methods to identify, characterise and trace HGT events.

The discrepancy to phylogenetic evolution inspired existing genome-based HGT methods. For a fixed set of species and a potential horizontally transferred gene, these methods detect HGT events by looking at inconsistencies between the gene tree and a phylogenetic tree built for the set of species (Ravenhall et al. (2015)). As a prerequisite, a candidate gene for which to run the calculation and comparison has to be known. Sequence content based methods aim to identify genes of foreign origin in a given genome by exploiting sequence pattern such as *k-mer* frequencies or GC content which vary between different species (Jaron et al. (2013), Metzler and Kalinina (2014)). All methods are based on an assembled HGT organism, meaning they are also prone to the problems of misassemblies. Although AMRs are a prominent example for horizontally transferred genes, methods to directly identify antimicrobial resistance (AMR) genes do not necessarily connect the presence of an AMR gene to an HGT event (e.g., KmerResistance Clausen et al. (2016)).

In previous work, we developed an approach that aims to call HGT events directly

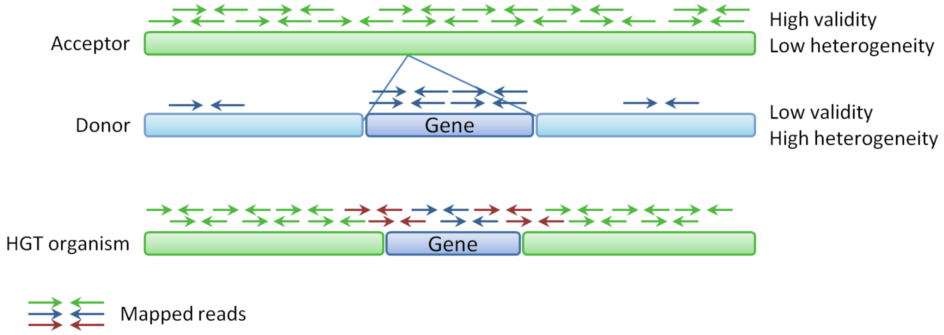


Figure 4.1. HGT overview and evidence. The sequence of an HGT organism consists mainly of the sequence of the acceptor genome (green), and only the transferred part (blue gene) is represented by the donor genome. Hence, reads from the HGT organism should mainly map homogeneously to the acceptor (green arrows), only few reads should map locally to the donor (blue arrows), and some read pairs (red arrows) will span the boundary between the green parts from the acceptor and the blue part from the donor. These mapping patterns can be represented by scores based on the mapping coverage profile. An acceptor with a homogeneous coverage has a high validity score and a low heterogeneity score, a donor has opposite score ranges (low validity and high heterogeneity). Based on these scores, the DaisyGPS *acceptor-score* is $\in [0, 1]$ and *donor-score* is $\in$ [-1, 0).

from next-generation sequencing (NGS) data (Trappe et al., 2016) in a tool called Daisy. Instead of focusing on the sequence content of the HGT organism, Daisy examines the origin of the transfer, namely the prespecified acceptor and the donor organisms, and directly maps the NGS reads to these references. By facilitating structural variant detection methods, we can thereby identify the transferred region from the donor and the insertion site within the acceptor. A prerequisite for Daisy is therefore that both acceptor and donor references are known. This, however, is not always the case, and hence requires methods that are able to infer acceptor and donor candidates from the NGS reads of the HGT organism. Such methods are not yet available.

However, the problem of acceptor and donor identification directly from NGS data of the HGT organism is akin to the problem tackled by metagenomic profiling studies that aim to unravel metagenomic samples. Here, so called metagenomic classification approaches aim at identifying all organisms present in a sample by directly analysing sequencing data with a complex mixture of various organisms (Breitwieser et al., 2017). While in this classical scenario all reads of a single organism in the sample can theoretically be assigned to one reference organism during identification, this is not the case for an organism that carries foreign genes acquired via HGT. Most reads will be assigned to the acceptor genome but only a fraction can map to the donor genome (see mapped reads in Figure 4.1). Hence, we have to account for this

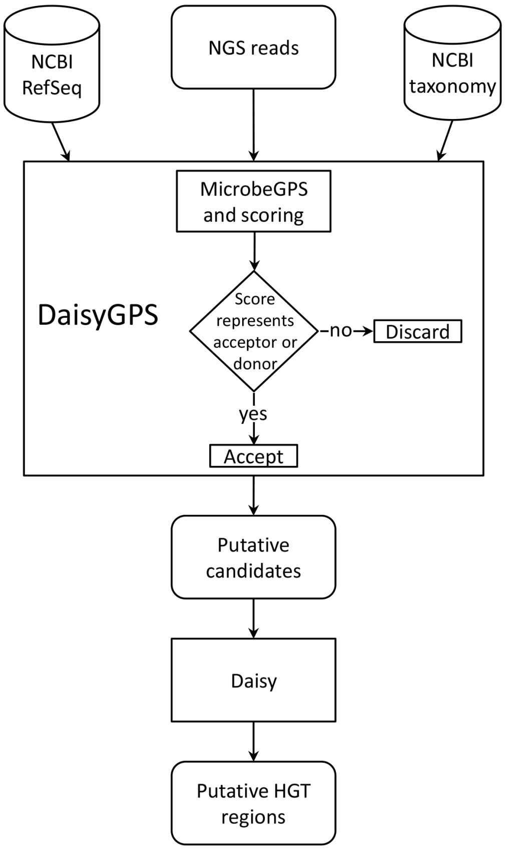


Figure 4.2. Workflow of DaisySuite. The input NGS reads are first processed by DaisyGPS. The reads are mapped to the NCBI RefSeq and then analysed by MicrobeGPS which also incorporates taxonomic information acquired through the NCBI taxonomy database. Based on that, DaisyGPS calculates two scores for acceptor and donor classification (see methods part). Depending on these scores, the highest-ranked candidates are selected as suitable acceptor and donor candidates. Daisy then uses these candidates to identify HGT region candidates.

two mapping properties of the reads during analysis. Another requirement is the resolution of classification on strain level, if possible, since two strains of the same species can already significantly differ in their sequence content.

Metagenomic classification approaches follow either a taxonomy dependent or taxonomy independent approach (Lindgreen et al. (2016), Sedlar et al. (2017)). The general procedure for both approaches is to assign sequencing reads stemming from the same organism in the sample into the same group, a process also referred to as binning. Taxonomic dependent binning approaches assign the reads to specific taxonomic groups, and hereby infer the presence of these taxa in the sample. These methods either also make use of sequence composition patterns, e.g., Kraken (Wood and Salzberg, 2014), or they determine mapping-based sequence similarities for the read assignment, e.g., MEGAN (Huson et al., 2007), Clinical PathoScope (Byrd et al., 2014) or DUDes (Piro et al., 2016). Both approaches will most likely identify the acceptor reference of an HGT organism due to the homogeneous coverage and comparatively high number of reads. The drawback of all read assignment approaches is the limitation in the presence of mobile genetic elements, e.g., integrated via HGT or of hitherto unknown - or unsequenced - organisms in the sample. Reads belonging to these genes or unknown organisms are either assigned to a similar but incorrect taxa or not assigned at all, leading to wrong identifications and biases in abundance estimation. To ensure robustness, many approaches deliberately discard taxonomic candidates with only low and local coverage. Hence these approaches will likely discard any donor candidate references. Composition-based methods such as Kraken would also perform poorly pinpointing the correct donor based on evidence of only few reads given the fairly large number of usually detected species.

In our group, we developed MicrobeGPS (Lindner and Renard, 2015), a metagenomics approach that accounts for sequences not yet present in the database. Instead of reporting fixed taxa with assigned reads, MicrobeGPS in turn uses the candidate taxa to describe the organisms in the sample in terms of a genomic distance measure. That is, it uses available references to model the composition of the organisms present in the sample in terms of coverage profiles and continuity, instead of directly assigning reference organisms to characterize the sample. If the organism in the sample is present in the database and covered homogeneously then the distance approximates to zero. If not, MicrobeGPS identifies the closest relatives by positioning the organism among references with the lowest genomic distance. Hence, the tool considers scores and metrics that reflect a donor-like, in-homogeneous coverage but filters out false positive candidates with inhomogeneous coverage for the purpose of species assignment. From the perspective of HGT detection, these may be highly relevant and should not be excluded.

Here we present DaisyGPS, a pipeline building on concepts of MicrobeGPS and tailored to the identification of acceptor and donor candidates from sequencing reads of an HGT organism. DaisyGPS uses genome distance metrics to define a score that allows the classification into acceptor and donor among the reported organisms. Owing to the properties of these scores, we still find the closest relatives of acceptor and donor in case these references are not present in the database. DaisyGPS further offers optional blacklists and a species filter to refine the search space for acceptor and donor candidates. DaisyGPS and Daisy are integrated into one pipeline called DaisySuite to offer a comprehensive HGT detection, and publically available at `https://gitlab.com/rki_bioinformatics/DaisySuite`. We validate DaisySuite on a large scale simulation where we show sensitivity and specificity of our approach and the robustness when applied to non-HGT samples. On a real data set from an MRSA outbreak, we demonstrate the ability of the DaisySuite to distinguish between the outbreak associated and unassociated samples in terms of sequenced content potentially acquired through HGT events.

4.1 Identifying acceptor and donor candidate identification with DaisyGPS

The problem of mapping-based HGT detection from NGS data is twofold: First, the acceptor (organism that receives genetic information) and donor (organism that the information is transferred from) references have to be identified. Based on that, the precise HGT region and its insertion site within the acceptor can be characterised. We presented a method to solve the second task in Trappe et al. (2016). Here, we propose the tool DaisyGPS (see also Figure 4.2) with the objective to identify possible acceptor and donor candidates given reads of a potential HGT organism. We provide Daisy and DaisyGPS in an integrated pipeline that we call DaisySuite.

The genome of the HGT organism consists mainly of the acceptor genome (see Figure 4.1). When the reads of the HGT organism are mapped against the acceptor reference, most reads should map properly. Therefore a high and continuous mapping coverage pattern of the acceptor genome can be expected. In contrast to that, only a small part of the donor genome is present within the genome of the HGT organism, hence only a small fraction of the reads should map against the donor reference and then only within a zoned part (i.e. the part that has been transferred). This results in a discontinuous mapping coverage pattern where only a small part of the reference shows a high mapping coverage (see Figure 4.1).

In a first step, we need to define metrics that represent the expectations we have, i.e. how much of the genome is covered by reads (mapping coverage) and how uniformly these reads are distributed across the genome (discontinuous vs. continuous patterns). Given only the reads of the HGT organism, the acceptor and donor candidate identification problem is similar to aspects of metagenomic profiling. A standard problem in metagenomics is the identification of organisms in a sample using a read dataset of this sample. At first glance, it may appear that the methods designed to solve this problem can also be applied to our identification objective, i.e. we have the read dataset of the HGT organism and we are looking for two organisms (acceptor and donor) that are in the sample. However, because the HGT organism consists mainly of the acceptor genome, such an approach works only well for the identification of the acceptor. For the donor, additional information is needed to guarantee a reliable identification because references with only local or discontinuous coverage are usually dismissed by the profiler. We use the metagenomic profiling tool MicrobeGPS to obtain a coverage profile of our given HGT organism from mapping coverage metrics. MicrobeGPS fits our requirements as it can be configured to not filter any organisms and reports additional metrics that we use to represent acceptor and donor attributes. Next, we evaluate the gathered metrics and establish a score that reflects our defined acceptor or donor coverage properties. Then, the candidates are ranked by this score and a list of acceptor and donor candidates is generated. These acceptor and donor candidates can then be further analysed with tools such as Daisy.

DaisyGPS scores. For the purpose of HGT detection, we aim to define a scoring that reflects the mapping coverage properties of the acceptor and donor references: The acceptor has a continuous, homogeneous coverage over the complete length of the genome. The donor has a local, but still homogeneous coverage in the area where the transferred genes are originated but should have nearly no coverage at all otherwise. The score should further allow a clear distinction between acceptor and donor candidates and provide a meaningful ranking according to the likelihood of being the most suitable candidate.

As a basis for our scoring, we use the *Genome Dataset Validity* defined in Lindner et al. (2013) and *homogeneity* metric defined in Lindner and Renard (2015). The Genome Dataset Validity, or short validity, describes the fraction of the reference genome for which there is read evidence. In contrast, the homogeneity reflects how evenly the reads are distributed. Both have a range $\in [0, 1]$. The validity is defined such that a genome that is covered - either low or high - over the full length has a high validity (≈ 1). We define a *heterogeneity* metric based on the Kolmogorov-Smirnov test statistic defined in Lindner and Renard (2015) such that an evenly covered genome has a low heterogeneity (≈ 0) and a genome with local, high coverage a high heterogeneity (≈ 1).

An acceptor is a genome with a continuous, high coverage that then has a high validity (≈ 1) and a low heterogeneity (≈ 0) score whereas a distantly related donor genome with only local, discontinuous coverage has a low validity (≈ 0) and a high heterogeneity (≈ 1) score.

As can be seen above, both validity and heterogeneity are complementary for acceptors and donors, and hence the relation of both metrics infers the property of

a candidate between being an acceptor or a donor candidate.
We define:
$score = validity - heterogeneity$ with $score \in$ [-1, 1]
Therefore, the value for a completely covered acceptor with uniform read distribution would approach +1. Likewise, the value for a donor that is only covered in a small region would approach -1. In addition to the coverage profile, there is a high evidence by sheer read numbers for acceptors:

$acceptor\text{-}score = w * score$, $w = \frac{\#mapped\,reads}{\#total\,reads}$

where w is the fraction of all mapped reads that mapped to the specific acceptor candidate. For the donor, however, the size of the transferred region is not known in advance. Hence, we do not expect a specific read number evidence and therefore omit the weighting and define
$donor\text{-}score = score$
Both *acceptor-score* and *donor-score* are determined for every candidate and they have a codomain of [-1, 1]. Acceptor candidates have a homogeneous coverage and hence high validity and low heterogeneity, i.e. $validity > heterogeneity$. Hence, we classify the candidates with $acceptor\text{-}score \geq 0$ as acceptor and rank them from highest to lowest score. Donor candidates have a high heterogeneity and low validity, i.e. $validity < heterogeneity$. Therefore, we classify candidates with $donor\text{-}score <$ 0 as donor candidates and rank them from lowest to highest score.

There is a special case if acceptor and donor are very similar. Here, the donor might not express the attributes we are looking for. In particular, the donor might have a significant read number evidence arising from acceptor reads also mapping to the donor. These shared reads lead to more regions of the donor genome being covered (higher validity) and to a less local, more homogeneous coverage pattern across the donor genome (lower heterogeneity), hence $validity \approx heterogeneity$ and $donor\text{-}score \approx 0$. We classify candidates with a $donor\text{-}score > 0$ as acceptor-like donors and rank them from lowest to highest.

Candidate selection with blacklist filter (optional). There are scenarios where it is necessary to exclude certain results from being reported. For example, in a reanalysis case, the assembled sequence from the sample reads might already been added to the reference set of your choice. For HGT detection from such reads, however, there is no information gain if DaisyGPS reports this entry as a suitable acceptor. Other examples include cases, where one can exclude certain species or taxa due to preanalysis information that nevertheless could be reported by DaisyGPS due to their high sequence similarity to the sampled organism or the presumed acceptor or donor candidates. To make the search for acceptor and donor candidates adaptable for such cases, DaisyGPS features the blacklisting of certain taxa. It is possible to exclude single taxa, a complete species taxon or a complete subtree below a specified taxon. For a default run, the filter is turned off.

Candidate selection with species filter (optional). DaisyGPS generally considers candidates on different taxonomic levels, e.g. species and strain level, and reports the candidate level with the best scores. Often the strain references contain additional sequences compared to the species level reference representative, and hence, the species reference will mostly have a homogeneous coverage that will then lead to a high acceptor score. Usually identification on species level is sufficient. There are however species such as, e.g., *E.coli*, where a high number of strains have been sequenced already and differ in their properties such as pathogenicity among the strains (e.g. *E.coli* K12 versus EHEC strain O157:H7). In these cases, a mere detection of the acceptor or donor on a species level might not be precise enough. For these situations, we implemented a species filter. If this filter is activated, only candidates below species level are reported. In case no candidate would be reported with an active species filter, the filter is disabled and the user informed that for further analysis also candidates on species level are used. For a default run, this filter is also turned off.

Daisy inference and integration with Snakemake. Snakemake is a common workflow management system (Köster and Rahmann, 2012) which we used to implement the different steps of DaisyGPS. We generated the alignment file required for MicrobeGPS by mapping the reads of the HGT organism against the NCBI RefSeq (complete RefSeq, no plasmids, downloaded March 15th 2017) (O'Leary et al., 2016) using Yara (Siragusa et al., 2013; Dadi et al., 2018). To ensure compatibility, we reimplemented the Daisy workflow in Snakemake as well, and integrated both into a combined suite (called DaisySuite, see also Figure 4.2). DaisyGPS yields a configurable number of acceptors, donors and acceptor-like donors (default: 2, 3, 2). For each possible pair of acceptor and donor, a Daisy call is inferred. Both pipelines can still be run independently. To unburden installation, we provide a setup script and provide DaisySuite components as Conda (Con) packages. The simulations are also integrated into the DaisySuite pipeline (see DaisySuite documentation for details).

4.2 Experimental setup

Data sets

We tested the complete DaisySuite on three types of data sets to validate both DaisyGPS and the integration with Daisy. The first type comprises the *H.pylori* data set, the KO11FL data set and the EHEC data set. All three were used in the Daisy publication (see Trappe et al. (2016) for detailed data set description) and are chosen

		True condition (ground truth)	
		Simulation contains HGT (positive setting)	Simulation does not contain HGT (negative setting)
Predicted condition	Run reports HGT	TP	FP
(DaisyGPS)	Run does not report any HGT	FN	TN

Table 4.1. Confusion matrix for DaisyGPS classifications. If the simulation contains an HGT and DaisyGPS reports at least one candidate pair that corresponds to the correct acceptor/donor pair, the run is considered a TP. If DaisyGPS fails to report the correct acceptor or donor, the run is deemed a FN (note that all reported, wrong pairs will still undergo follow up analysis by Daisy). In a negative test setting, a FP occurs if DaisyGPS reports any pair where the acceptor does not equal the donor and a TN means that either no pair was reported or acceptor and donor of the pair are the same organism.

		True condition (ground truth)	
		Pair represents HGT (DaisyGPS TP)	Pair does not represent HGT (DaisyGPS FP)
Predicted condition	Pair reports HGT	TP	FP
(Daisy)	Pair does not report any HGT	FN	TN

Table 4.2. Confusion matrix for Daisy classifications. If a pair represents an HGT and Daisy reports an HGT, the pair is classified as TP, otherwise as FN. If a pair is a DaisyGPS FP, a FP occurs if Daisy reports any HGT and a TN whenever nothing is reported.

as suitable ground truth and for the purpose of showing reproducibility. The second type comprises a large-scale simulation analogous to the *H.pylori* simulation. Both positive (simulated HGT) and negative (no HGT) simulations are used to estimate sensitivity and specificity of the DaisySuite. In a third part, we use real data from an outbreak data set with 14 MRSA samples to elucidate further applicability of both DaisySuite. The details of the data sets and *in silico* experiments are explained below.

H. pylori. The data set *Helicobacter pylori* presents a simulated data set for a proof of principle already used for validation in the Daisy paper (see Trappe et al. (2016) for details of genomic simulation). The acceptor is *Escherichia coli* K12 substr. DH10B (NC_010473.1), the donor is *H. pylori* strain M1 (NZ_AP014710.1). The *in silico* transferred phage region of the *H. pylori* comprises genomic positions 1 322 000 - 1 350 000.

EHEC. The HGT organism in the EHEC data set is *E.coli* O157:H7 Sakai (Zhang et al., 2007) that derived from *E.coli* O55:H7 and is assumed to have acquired the Shiga-Toxins (Stx) via transduction from *Shigella dysenteriae*. According to literature, the bacteriophage carrying Stx is supposedly positioned at 2 643 556 - 2 694 691 in *E.coli* O55:H7. In Trappe et al. (2016) we proposed an alternative phage insertion site at 1 741 535 - 1 744 926.

KO11FL. The KO11FL data set comprises the transgenic *E.coli* KO11FL (Turner et al., 2012). The acceptor is *E.coli* W, and the two donors are *Zymomonas mobilis* and the cloning vector pBEN77.

Large-scale simulation. We designed a large-scale simulation analogous to the *H.pylori* data set with positive and negative simulations. For each positive simulation, first an acceptor and a donor organism are randomly chosen among the available RefSeq sequences (date of retrieval: March 21, 2017, plasmids are ignored for sake of size consistency). A random 28 Kbp region is selected from the donor and inserted at a random position in the acceptor. SNPs and indels are introduced into acceptor and donor region (SNP rate: 0.01 , indel rate: 0.001). For each negative simulation, only an acceptor is randomly chosen, and SNPs and indels are introduced with the same rates as above. 150 bp reads are simulated from 500 bp fragments with 50 bp standard deviation with the Mason simulator (Holtgrewe, 2014). The positive and negative simulations are repeated automatically 100 times.

MRSA outbreak. The MRSA data set consists of 14 samples of methicillin resistant *Staphylococcus aureus* strains obtained during a MRSA outbreak at a neonatal intensive care unit (ENA accession number ERP001256, Köser et al. (2012)). Seven samples are associated with the outbreak, labeled O1-O7 in this manuscript, the other seven samples N1-N7 are not associated with the outbreak. Sample description and run accession numbers are stated in Table 4.6. Phylogenetic analysis by Köser et al. (2012) separated the 14 samples into distinct groups according to their outbreak association. The reference isolate used in that study is the EMRSA-15 representative HO 5096 0412, and we use this as ground truth for acceptor candidates reported by DaisyGPS. The seven outbreak related MRSA samples have a distinct antimicrobial resistance pattern, and it is believed that the related resistance genes have been introduced via HGT. With DaisySuite we want to investigate if the outbreak strains share the same HGT regions and if they can be distinguished from the non-outbreak strains.

Structure of validation

The setup of the validation is according to the types of data sets. In a first phase, we want to show a proof of concept given data with sufficient ground truth. The aim is to predict the correct acceptor and donor candidates with DaisyGPS and at the same time to reproduce the results obtained from Daisy. We therefore use the data sets already shown in the Daisy paper for sake of consistency. We set DaisyGPS to report a total of two acceptor candidates, four donor candidates, and two acceptor-like donor candidates for every data set and we evaluate if the correct acceptor and donor candidates are among them. For incorrect candidates of acceptor and donor, Daisy should not report HGT candidates unless the transferred region is present in multiple strains or there are multiple possible acceptors present with high sequence similarities as, e.g., among *E.coli* strains. For the EHEC data set, we activate the species filter since we are interested in strain candidates, and further blacklist taxa from the HGT organism to be analysed (*E.coli* O157:H7, taxon 83334) and the complete O157 lineage (parent taxon 1045010). For the KOFL11 data set, the HGT organism is blacklisted as well (*E.coli* KOFL11, taxon 595495). In a second part, we want to estimate the rate of sensitivity and specificity of the DaisySuite. We designed a large-scale simulation analogous to the *H.pylori* data set with positive and negative simulations (100 simulations each). From the positive simulations, we calculate the sensitivity for both DaisyGPS and Daisy (see below for definitions on metrics). DaisyGPS is designed with high sensitivity in mind and always reports the closest fitting candidates given sequencing data, even for non-HGT organisms. Hence, also for the negative simulations, DaisyGPS will report candidates and we expect a low specificity here. Daisy, however, should then report only few - if any - HGT candidates from the acceptor-donor pairs. In the last evaluation part, we test the DaisySuite on real data with unknown or uncertain ground truth. The MRSA outbreak data set consists of 14 samples, seven outbreak related and seven unrelated. Here we want to test if DaisySuite is able to distinguish between the outbreak and non-outbreak samples according to their reported acceptor, donor and HGT region candidates.

Definition of evaluation metrics

The interpretation of various statistics depends on the hypothesis to be tested. In our analysis in the large-scale simulations, we differentiate between two scenarios: in the first one we expect to detect an HGT event (positive test), while in the other one we assume the absence of an HGT event (negative test). For each simulation or run, a DaisyGPS call will lead to multiple pairs to be evaluated by Daisy. We therefore distinguish between statistics on runs and statistics on pairs that we will explain in the following.

For DaisyGPS, we consider during a positive test a single run as a true positive (TP) if the correct acceptor/donor pair is reported. Accordingly, a false negative (FN) occurs when the correct pair is not reported. Since the number of reported pairs is set by our settings, we will almost always have a fixed number of downstream verifications (except if there are not enough candidates to report) and thus we report the number of runs instead of pairs. Consequently, we can define the sensitivity as TP / #Runs. In a negative test setting, we deem those runs as true negatives (TNs) where either no pairs are reported or acceptor and donor of the pair are the very same organism. All other pairs are regarded as FP that will each trigger an unnecessary verification in the downstream tools. Since we are interested in how many runs did not cause verifications, we can characterize the specificity by TN / #Runs. While it is obvious in both settings to rely on an exact match of the reported results and the ground truth, a reported organism still may be very close to the ground truth organism in terms of sequence similarity (negative and positive settings) and even include the very regions involved in the HGT event (positive setting). To account for this, we also use BLAST in the case that no TP was reported and compare the FP to the ground truth. If the Blast identity of the FP to the ground truth is above 80% we change the classification from FP to BLAST-supported TP (Blast TP) since Daisy might still be able to infer the correct HGT region from these Blast TPs given the sufficient sequence similarity.

In Daisy, we evaluate acceptor/donor pairs and therefore the statics are defined based on the condition of a pair reported by DaisyGPS. In a positive simulation, Daisy TP pairs are those that represent the correct pair and are detected by Daisy. It directly follows that each correct pair that is not supported by Daisy can be seen as a false negative (FN). Given that the pair is incorrect, i.e. a FP from DaisyGPS where the acceptor or donor is wrong, we count a rightly not supported pair as true negative (TN) and an erroneously detected pair as FP. To measure how many pairs are correctly identified, we define the sensitivity as (TP + TN) / #Pairs. Considering a negative test setting, we are mainly interested in the pairs that are wrongly reported as being involved in an HGT event. We declare those pairs as FP and describe the specificity as (#Pairs - FP) / #Pairs. It also follows that all the pairs that are not detected are TN.

Lastly, in the context of the complete DaisySuite pipeline, we evaluate the combined results of DaisyGPS and Daisy. Each pair reported by DaisyGPS for a single simulation induces an evaluation by Daisy. Since the overall result of the pipeline should indicate whether a simulation contains an HGT event or not, the classification of a DaisySuite run depends exclusively on the consolidated results of each Daisy evaluation for a single simulation. In a positive test setting, we want to find exactly the one pair that represents the HGT event. From that follows that a complete DaisySuite run can be classified as TP if Daisy supports solely the correct pair, i.e. Daisy reports the TP and no FP. This also implies that DaisyGPS needs to detect the TP. Similarly, in a negative test setting, a TN occurs if Daisy reports no HGT candidates at all.

Settings and pre-/post-processing

DaisySuite is run with default parameters as of version 0.0.1 unless stated otherwise. The parameter to combine potentially overlapping HGT candidates within Daisy is set to 20 bp, hence, overlapping regions with start and end positions differing by more than 20 bp are reported as separate candidates. For the comparison of the number and content of HGT sequences, we clustered overlapping HGT candidates

DaisyGPS				DaisySuite						
TP	Blast TP	FP	sensitivity	TP	Blast TP	TN	FP	Blast FP	FN	sensitivity
79	22	21	0.79	55	13	14	27	27	4	0.69

Table 4.3. Positive HGT simulation. DaisyGPS calls correct acceptor and donor candidates with a sensitivity of 79%. The total sensitivity for DaisySuite from 100 HGT simulations regarding correct acceptor and donor candidates with a follow up correct HGT site call is 69%.

DaisyGPS pairs	TP	Blast TP	TN	FP	Blast FP	FN	Blast FN	sensitivity
818	74	22	656	32	32	56	51	0.89

Table 4.4. Positive HGT simulation. Daisy evaluates 818 pairs reported by DaisyGPS and calls the correct HGT region or correctly no HGT region with a sensitivity of 89%.

with the tool usearch9 (v9.1.13_i86linux32) with identity 1.0 (Edgar, 2010).

For validation, we determine the true presence of a HGT region in the samples by mapping the sample reads to all suggested, clustered regions with Bowtie2 (version 2.2.4). For comparison, we take the mean coverage of every region and apply a sigmoidal function to map all mean coverages to the [0.5,1] space for displaying a meaningful heatmap. The application of a sigmoidal function and the heatmap is computed in R (Rscript version 3.3.3). The heatmap function in R uses a hierarchical clustering with complete linkage as default, and we turned of the dendrogram for the columns. In addition, we perform a whole-genome alignment using the Mauve plugin (version 2.3.1) as part of the Geneious software (version 10.0.5) to to establish shared HGT regions among the samples. To do this, we concatenate all HGT regions of a sample and separate the regions with segments of 1000*'N' to avoid fragmented regions or overlapping LCBs.

DaisyGPS		DaisySuite		Daisy		
TN	specificity	FP	specificity	DaisyGPS pairs	FP	specificity
6	0.06	3	0.97	743	6	0.99

Table 4.5. Negative HGT simulation. For the 100 negative simulations, DaisyGPS correctly reports no acceptor and donor candidates for six simulations. From the 94 simulations causing a downstream evaluation with Daisy, only three lead to a FP call considering all outcomes from DaisySuite (summarised over the 100 simulations). Daisy evaluates 743 pairs and only has six FP HGT region calls in total over all those pairs.

4.3 Results

4.3.1 Acceptor and donor identification with DaisyGPS.

In the first part of the validation, we test DaisyGPS on three data sets from simulated and real data with sufficient ground truth and already previously evaluated with Daisy. Since DaisySuite combines both tools, DaisyGPS and Daisy, the aim is to support our previous results even when now the donor and acceptor are not prespecified.

The *H.pylori* data set was simulated from *E.coli* K12 substr. DH10B as acceptor and *H. pylori* strain M1 as donor. DaisyGPS successfully reports both as such (see Supplement Tables A.1 and A.2), and the subsequent Daisy run also reports the true HGT site. In addition to the only true HGT candidate previously already reported in the Daisy paper, DaisySuite reports another, FP HGT site for a region from *Haemophilus ducreyi.* The HGT region reported for *H. ducreyi* strain GHA9 has no continuous similarity with the HGT region from *H.pylori* (no blast hits longer than 15 bp, data not shown). However, the region on *H. ducreyi* shares the first 1200 bp and the last 1300 bp with the acceptor *E.coli* K12 substr. DH10B on multiple sites, and since beginning and end of the region are covered, almost six times as many split-reads are found as for the true acceptor site. The total coverage of the region is relatively low with 30x compared to 95x of the *H.pylori* but obviously high enough to pass the coverage filter.

The EHEC *E.coli* O157:H7 Sakai is supposedly derived by an HGT event where a defective prophage has been transferred from *Shigella dysenteriae* to *E.coli* O55:H7. Both are reported by DaisyGPS as candidates (see Supplement Table A.3). In line with its strong sequence similarity to the *E.coli* species, *S.dysenteriae* is labeled as an acceptor-like donor candidate. The proposed alternative HGT insertion site from our previous Daisy paper is still reported (see Supplement Table A.4).

The KO11FL data set comprises a transgenic *E.coli* W variant with transferred genes from *Zymomonas mobilis* and a plasmid that was not analysed here. DaisyGPS successfully reports *E.coli* W and *Zymomonas mobilis* as acceptor and donor candidates (see Supplement Table A.5). Daisy does not report any FP HGT candidates.

4.3.2 Estimating sensitivity, specificity and robustness of DaisySuite through large-scale simulations.

After validating DaisyGPS on data previously evaluated with Daisy as a proof of principle, we analyse DaisySuite in terms of robustness and sensitivity by performing a large-scale simulation. We perform the simulation for the *H.pylori* data set in a randomised and automated fashion generating 100 simulations with a transferred HGT region. To evaluate robustness, we also perform 100 negative simulations where an acceptor genome is simulated but no HGT region is inserted. With the positive simulations, we can estimate the sensitivity of the complete DaisySuite. For DaisyGPS, we evaluate how many from the 100 simulations have the correct acceptor and donor genome identified. Since DaisyGPS reports more than one potential acceptor-donor pair, we count a TP hit if the true pair is among them, and only count a FN if the true pair was not reported at all. In addition, we consider pairs with Blast sequence identity $> 80\%$ also as a potential HGT candidate pair, and also count them as a TP. To evaluate Daisy, we consider all pairs proposed by DaisyGPS.

For a true pair reported by DaisyGPS, Daisy can either report a TP HGT region or a FN if the region could not be identified. For an acceptor-donor pair wrongly proposed by DaisyGPS, Daisy can either report no HGT candidate region (TN) or a FP hit. When we summarise the DaisySuite results over all pairs of one simulation, we only count a TP for that simulation if Daisy did not report any FPs (despite any TPs or TNs).

Table 4.3 states the resulting counts for DaisyGPS and for the complete DaisySuite summarised over the 100 simulations. DaisyGPS yields a sensitivity of 79%. From the 79 TPs, 22 are based on either a wrong acceptor, or donor, or both but have still sufficient Blast similarity to the original acceptor or donor to be counted as TP according to our scoring. 69% of the TPs and FPs resulted in a TP or TN call from Daisy. It is noticeable that all DaisySuite FPs are Blast FPs.

Table 4.4 states the number of reported pairs proposed by DaisyGPS and a detailed count based on each pair for Daisy. From the resulting 818 pairs, Daisy then reports the correct HGT region, or correctly no HGT region from a DaisyGPS FPs, with a sensitivity of 89%.

In addition to the positive simulations, we performed another 100 negative simulations where we randomly selected and variated an acceptor genome but did not insert any foreign region from a donor. DaisyGPS can now either produce a TN hit, i.e. report no candidates at all, or FP candidates. Since DaisyGPS is very sensitive by design, we expect it to report candidates most of the time and, hence, we want to estimate if these negative HGTs trigger reports by a Daisy follow-up call. As expected, the specificity for DaisyGPS is very low with 6% (see Table 4.5). However, Daisy reports only six FPs on all pairs in total, i.e. three simulations produced a FP HGT report.

From these results we can infer that DaisySuite is able to distinguish HGT from non-HGT organisms and is very robust if no HGT is present.

4.3.3 Exploration of HGT detection with DaisySuite from MRSA outbreak data

MRSA strains are generally assumed to undergo HGT events frequently (Lindsay, 2010, 2014). The MRSA data set considered here consists of 14 samples with seven of them related to an MRSA outbreak (O1-O7) and seven MRSA samples not asso-

Label	Isolate	Accession	EMRSA-15	HGT regions	EMRSA-15 as acceptor
			as acceptor	HGT regions	for HGT regions
O1	1B	ERR103401	x	4	4
O2	6C	ERR103403	x	4	3
O3	7C	ERR103404	x	5	3
O4	8C	ERR103405	x	3	3
O5	10C	ERR101899	x	4	4
O6	11C	ERR101900	x	1	1
O7	12C	ERR103394	x	5	3
N1	14C	ERR103395	-	5	-
N2	15C	ERR103396	x	2	2
N3	16B	ERR103397	-	4	-
N4	17B	ERR103398	-	4	-
N5	18B	ERR159680	-	5	-
N6	19B	ERR103400	x	7	5
N7	20B	ERR103402	x	2	2

Table 4.6. Acceptor and number of HGT region candidates. For 10 of the 14 samples, EMRSA-15 (HO 5096 0412) was reported as acceptor candidate. This includes all outbreak samples. Column *HGT regions* states the number of reported HGT regions, and column *EMRSA-15 as acceptor for HGT regions* the respective number that were reported with HO 5096 0412 as acceptor.

	Reported donors
Outbreak and non-outbreak	*S.pseudointermedius* ED99 and HKU10-03
	S.warneri SG1
	S.epidermidis RP62A
	S.haemolyticus JCSC1435
	S.aureus COL
	S.lugdunensis HKU09-01
Non-outbreak only	*S.epidermidis* ATCC 12228 (N1,N6 only) and PM221 (N4 only)
	E.faecium Aus0004 (N1 only)

Table 4.7. Reported donors summarised for all samples. Both outbreak associated and unassociated samples mostly report the same donor candidates with only few variations (see supplementary tables A.6-A.33 for details). The only unique donors are reported for the unassociated samples N1, N4 and N6.

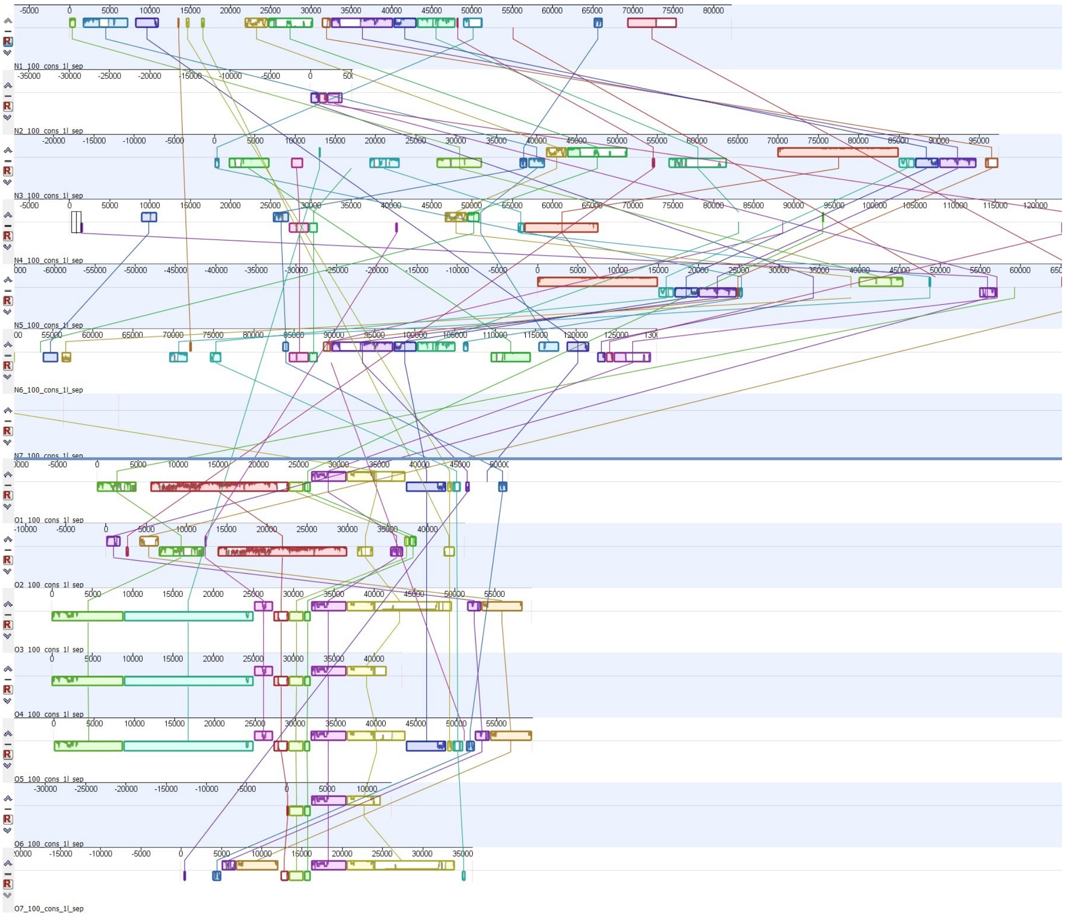


Figure 4.3. Mauve alignment of concatenated HGT regions. The HGT regions of all samples are aligned with Mauve to establish shared regions between them. The outbreak associated samples (O1-O7) in the lower part share most of their regions whereas the unassociated samples (N1-N7) in the upper part do not.

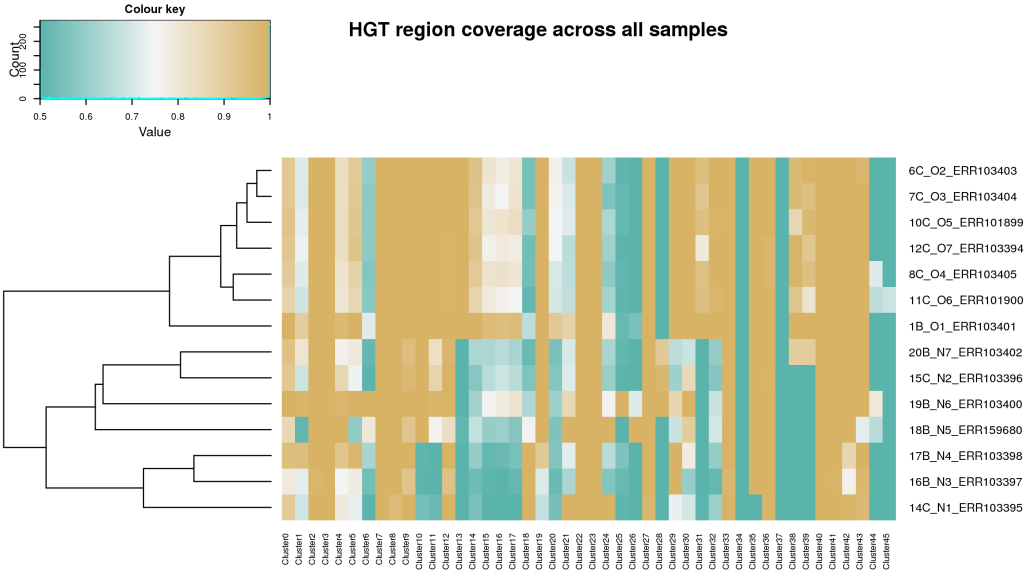


Figure 4.4. Heatmap of HGT region coverages. The mean coverages of HGT regions from all samples are calculated across every sample, and compared after application of a sigmoidal function. Solid green spots indicate no coverage, solid ochre high coverage. Regions 34 and 37 are not covered in any sample and hence FP calls. Sample O6 shows presence of multiple HGT regions called by DaisySuite for other samples but missed here. There is a distinct presence of HGT regions between the outbreak samples in the upper part and the unassociated samples in the lower part.

ciated with the outbreak (N1-N7) but that occurred in the same time frame (Köser et al., 2012). Köser et al. (2012) analysed all 14 samples and compared them to the EMRSA-15 representative HO 5096 0412 as the supposedly closest relative of the outbreak strains. We first evaluate acceptor and donor candidates reported by DaisyGPS in relation to the proposed HO 5096 0412 reference and then investigate HGT region candidates reported by Daisy regarding a possible distinction of outbreak vs. non-outbreak samples. We activate the species filter as we are again interested in strain level candidates.

For all outbreak samples O1-O7, *S.aureus* HO 5096 0412 was reported as acceptor candidate by DaisyGPS (see Table 4.6 and supplementary tables A.6 - A.33). The same acceptor was also reported for non-outbreak samples N2, N6 and N7. Acceptor candidates for sample N1 are *S.aureus* ECT-R-2 and N315, for N3 and N4 *S.aureus* MSSA476 and MW2, and for N5 *S.aureus* MRSA252. Although not associated with the outbreak, samples N3 and N4 are from patients that shared the same room in the hospital where the outbreak occurred and hence are possibly related (Köser et al., 2012).

The reported donors are largely the same for both outbreak and non-outbreak samples (see Table 4.7). No donor was reported exclusively for the outbreak samples but three donors only for non-outbreak strains N1, N4 and N6. These are *S.epidermidis* strains ATCC 12228 and PM221 as well as *Enterococcus faecium* Aus0004. Although *S.aureus* HO 5096 0412 was reported for all outbreak samples, there is no clear distinction in acceptor and donor candidates reported by DaisyGPS apart from the non-outbreak only donors.

Table 4.6 states the total number of clustered HGT regions and the number of the clustered regions where HO 5096 0412 is the acceptor that are found by DaisySuite. Most HGT regions hence have the EMRSA-15 representative as acceptor.

Figure 4.3 shows a Mauve alignment of the concatenated HGT regions of all 14 samples. There is a clear connection between the HGT regions from the lower seven samples O1-O7 that are the outbreak related samples. Samples N1-N7 also share some regions but do not have a clear connection as among the outbreak related strains. The overlap between outbreak and non-outbreak HGT regions is also low.

Figure 4.4 shows the presence of the 41 HGT regions determined by mapping coverage called by Daisy among all samples. The purpose of the coverage analysis is to evaluate again if the HGT regions differ between the outbreak and non-outbreak strains but also to estimate if there are regions shared by all outbreak strains that are FN candidates of Daisy, or regions not covered at all that are likely FP candidates.

The clustering of samples according to the dendrogram shown in figure 4.4 was done automatically (see settings part), and hence reflects the relation of the samples according to the mapping coverage of the proposed HGT regions.

All outbreak strains are clustered together and share most of their HGT regions. All non-outbreak strains for which DaisyGPS did not report EMRSA-15 as an acceptor candidate are clustered away furthest from the outbreak strains (N1, N3 - N5). The likely related samples N3 and N4 are clustered together. Regarding a distinction of outbreak and non-outbreak strains, DaisySuite is able to determine the outbreak-related HGT regions which differ from the HGT candidates for the non-outbreak strains. Hence, a distinction is possible. Although DaisySuite only called one HGT region for O6, we can deduce from the coverage profile that more HGT regions called for the other outbreak samples are present as well but were missed by DaisySuite. As can be seen in the heatmap, clusters 34 and 37 are not covered by any sample and hence likely FPs. We detected the AMR gene *mecA* on Cluster 0, however, resistance is shared among all 14 samples according to Köser et al. (2012). No further AMR genes tested by Köser et al. (2012) are detected on the other clusters. However, most of these AMR genes are on plasmids that were not analysed here.

4.4 Discussion of results from DaisyGPS

We presented DaisyGPS, a pipeline that facilitates metagenomic profiling strategies to identify acceptor and donor candidates from NGS reads of a potential HGT organism. DaisyGPS, together with Daisy, is part of the comprehensive HGT detection suite DaisySuite. We successfully validated DaisyGPS on simulated and real data previously analysed in Trappe et al. (2016). We further demonstrated robustness of the DaisySuite on a large-scale simulation with 100 negative HGT tests, showing that DaisySuite correctly reports no HGT events with a specificity of 97%. On a large-scale simulation with 100 positive HGT simulations, DaisySuite reports the correct HGT event with a total sensitivity of 69%. From the 818 pairs reported by DaisyGPS among the 100 simulations, Daisy called the TP and TN regions with a sensitivity of 89%. Lastly, we evaluated DaisySuite on an MRSA outbreak data set with seven outbreak associated samples and seven not associated with the outbreak but that occurred during the same time frame. Here we could show that DaisySuite successfully distinguishes between associated and not associated samples regarding their suggested HGT regions, i.e. the outbreak samples show a distinct number and content of reported HGT regions.

One has to acknowledge that all outbreak strains have a high sequence similarity to the EMRSA-15 strain, which is not necessarily the case for the non-outbreak strains. This is also reflected in the results from DaisyGPS where *S.aureus* HO 5096 0412 is the best acceptor candidate for all outbreak strains but not reported at all for some non-outbreak strains. It directly follows that a sequence comparison based analysis as done with DaisySuite will likely find different patterns for the outbreak and non-outbreak strains, and a difference in HGT region candidates might seem obvious. However, starting from having established such a difference, there is value in then analysing the shared HGT region candidates among the outbreak-related strains. For this proof of concept, we performed a relatively simple evaluation by performing a coverage analysis of all HGT regions across all samples and investigating the presence of AMR genes within the HGT regions. But a future thorough follow-up analysis of the origin and functionality provided by the potential HGT sites could benefit our understanding of the risk and pathogenicity of these outbreak strains.

The observed FP and FN candidates, however, also reveal weaknesses of the sequence comparison approach. DaisyGPS is designed with a focus on sensitivity and hence inevitably leads to FP acceptor and donor candidate pairs to be examined by Daisy. Since these FPs are still due to a sufficient degree of mapping coverage, spurious split-reads and spanning reads can cause downstream FP calls as observed for the simulated data set from *E.coli K12* DH10 and *H.pylori*. The reported HGT site from *H.ducreyi* has only similarities in the start and end part of the proposed region compared to the transferred *H.pylori* region though. Insertion sites can also lie within repeat regions which enhances the negative impact of ambiguous mappings. This emphasises that a critical evaluation of HGT predictions is always crucial.

From the missing HGT region calls for sample O6 that could be inferred from the coverage analysis, we can deduce that DaisySuite does not detect all HGT regions due to insufficient evidence. A potential cause could be that DaisyGPS did not report the correct donor reference. Even if DaisyGPS could find an appropriate donor genome, it is still likely that the genome content differs between the region present in the donor and the region actually present in the HGT organism. An alternative, complementary approach to cope with this problem of a lack of a suitable donor candidate could be to facilitate local, insertion sequence assembly. By offering identified insertion sequences, we can still provide the content of a potential HGT sequence and thereby enable downstream analysis. This approach would also support the detection of novel HGT sequences not present in current reference databases, and therefore also the detection of, e.g., novel antimicrobial resistance genes. Popins (Kehr et al., 2015) is a tool for population-based insertion calling developed for human sequencing data (see, e.g., Kehr et al. (2017)). Popins only locally assembles unmapped reads (same input as for Daisy) with Velvet guided by a reference, thereby minimising the risk of potential misassemblies. On top of the assembly, Popins first uses spanning pairs (see red read pairs in Figure 4.1) to place an insertion in the (acceptor) reference, and then performs a local split-read alignment around the potential breakpoint. If multiple samples are provided, Popins merges contigs across samples into supercontigs, assuming that the same insertion is present in multiple samples. Although different bacterial samples do not represent a population as given for human populations, outbreak related samples still resemble a population such that one could use Popins for this purpose and gain valuable information. However, local insertion assembly only gives evidence for an insertion compared to the chosen acceptor reference, that does not necessarily mean that the insertion resulted from an HGT event. Hence, means to sophistically include insertion assembly results into the HGT context need to be defined first. Despite the evidence for an HGT event that DaisySuite can provide, the results should always be tested for alternative causations such as gene loss. With DaisyGPS, we present a tool for acceptor and donor identification from NGS reads of an HGT organism. To do that, DaisyGPS refines metrics already defined and used for metagenomic profiling purposes to account for the acceptor and donor specific coverage profiles. We integrated DaisyGPS with Daisy into a comprehensive HGT detection suite, called DaisySuite, that provides an automatic workflow to first determine acceptor and donor candidates and then identify and characterise HGT regions from the suggested acceptor-donor pairs. We successfully evaluated DaisyGPS on data previously analysed with Daisy, and demonstrated sensitivity and robustness of the DaisySuite in a large-scale simulation with 100 simulated positive and negative HGT events. We could further show the benefits of an HGT analysis with DaisySuite on an MRSA outbreak data set where DaisySuite reported HGT candidates that help to distinguish between outbreak associated and unassociated samples and therefore also provide information for outbreak strain characterisation.

5 Horizontal gene transfer detection from MS/MS data with Hortense

The recognition of horizontal gene transfer (HGT), also called lateral gene transfer, has changed the way we regard evolution. Compared to the established notion of parent to offspring inheritance of genes and functions, HGT enables the direct transfer between individuals of the same generation, and, more importantly, across species boundaries (Ochman et al., 2005; Daubin and Szöllősi, 2016). Bacteria have at least three commonly known mechanisms for this transfer (see Figure 1.4). They can take up naked DNA from the environment (transformation), transfer DNA directly from cell to cell via a pilus (conjugation), or receive DNA through an infection by a bacteriophage (transduction) (Gyles and Boerlin, 2013). The impact of this powerful mechanism was only recently recognized with the advent of genome sequencing (Daubin and Szöllősi, 2016). While HGT has been previously assumed to be a sporadic event with low relevance to the recipient organism, nowadays, it is common knowledge that HGT occurs frequently, and that pathogenic components such as toxins and antimicrobial resistance genes are prominent examples for HGT (Liu et al., 2012; Juhas, 2013; Perry et al., 2014). In the era of ”superbugs” and fast spreading resistances (Juhas, 2013), methods are urgently required that can identify, characterize and also trace the origin of HGT events.

Still, there is only a limited number of HGT detection methods and they focus on the genomic level (Ravenhall et al., 2015) so far, since for the screening and classification of bacteria, whole genome sequencing technologies have been established. Only recently, we developed a first HGT detection method based directly on next-generation sequencing (NGS) data (Trappe et al., 2016).

However, the genomic level does not reveal any information about gene expression and involved metabolic pathways. This, in turn, motivates the use of orthogonal post-genomic analysis methods, such as transcriptomics and proteomics (Radhouani et al., 2012). In particular, the field of proteomics has recently experienced various significant developments with respect to accuracy and speed of mass spectrometry (MS) instrumentation (Van Oudenhove and Devreese, 2013). MS-based proteomics therefore becomes an increasingly suitable tool which enables to detect and identify expressed proteins in bacteria. As a prominent example, matrix-assisted laser desorption ionization–time of flight (MALDI-TOF), although in use for several decades, has fairly recently emerged as a rapid and cost-saving method for the identification

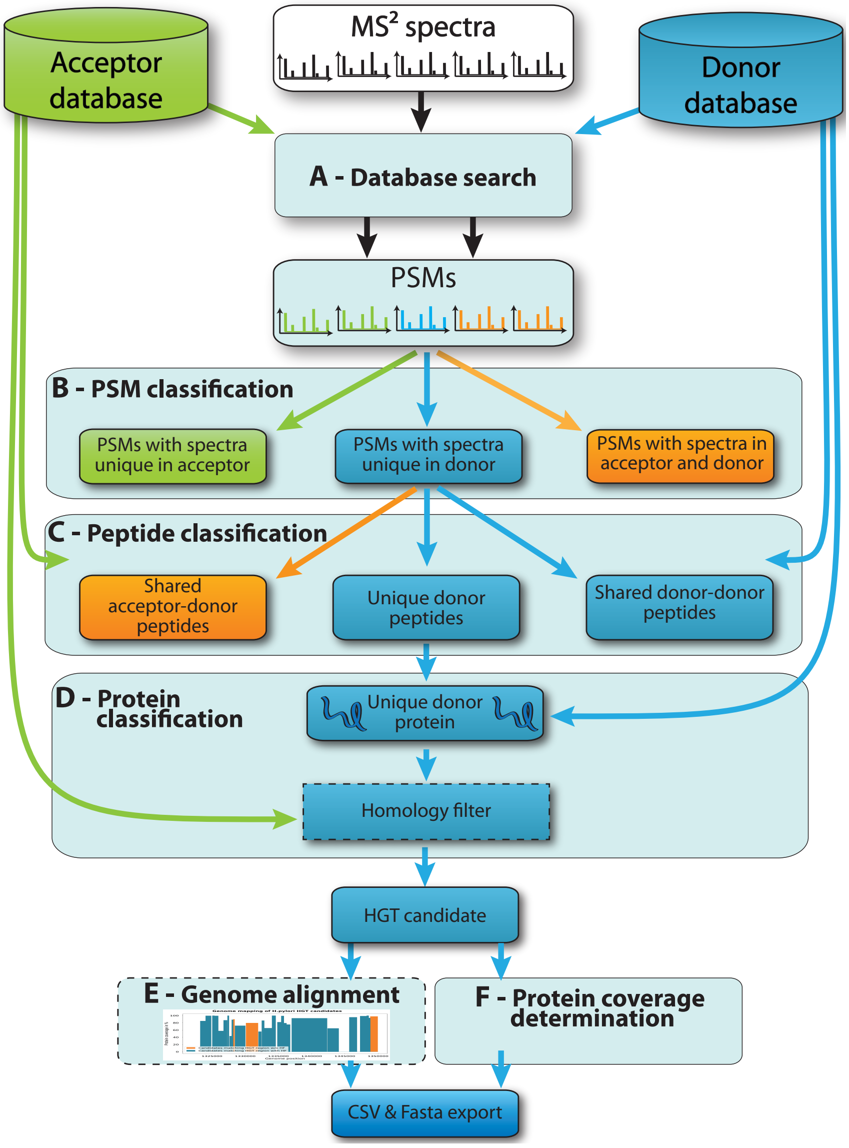

Acceptor database
MS² spectra
Donor database
A - Database search
PSMs
B - PSM classification
PSMs with spectra unique in acceptor
PSMs with spectra unique in donor
PSMs with spectra in acceptor and donor
C - Peptide classification
Shared acceptor-donor peptides
Unique donor peptides
Shared donor-donor peptides
D - Protein classification
Unique donor protein
Homology filter
HGT candidate
E - Genome alignment
F - Protein coverage determination
CSV & Fasta export

Figure 5.1. Hortense evidence and workflow. (A) In a first step, the MS/MS spectra are searched separately against the acceptor and donor databases. (B) The resulting PSMs are classified to be either uniquely matching to the acceptor or donor, or shared between both. The unique donor PSMs are filtered further in the following peptide classification (C). Here, the identified peptides are checked if they are unique within the donor - i.e. can identify only one protein. The peptides are also cross-validated against the acceptor database to filter shared peptides that were missed in the previous step. Only unique donor peptides are used for protein identification. (D) Identified proteins can be optionally filtered further by the homology filter in case the acceptor has a homologous HGT protein. All HGT candidate proteins can be optionally aligned to the genome to determine their genomic position (F), and are reported with information on protein coverage (F) (number of peptides and fraction of protein being covered by the peptides).

of microbial species which has been approved for clinical applications (Sauer and Kliem (2010), Neville et al. (2011), Clark et al. (2013)). While the latter approach processes information at the MS1 level, tandem mass spectrometry (MS/MS)-based proteome analysis techniques decipher amino acid sequences from their fragmentation pattern by matching tandem mass spectra against a provided sequence reference database. Besides mere protein identification, MS/MS-based proteomics enables to detect the taxonomic origin of bacterial species and to infer functional information of the expressed proteins, e.g. their molecular function or role in enzymatic pathways (Muth et al., 2016). For example, bacterial proteins can be identified which are linked to antibiotic resistance or which constitute virulence factors (Pérez-Llarena and Bou, 2016). In addition to the important feature of functional annotation, for accurate HGT detection, the higher resolution at the MS/MS level is required to unambiguously identify one or multiple proteins of which their genomic templates have been transferred between different bacterial species. Another important problem in proteomic workflows presents the occurrence of shared peptide sequences: such peptides are found in multiple proteins within a proteome database, e.g. linking to sequences which belong to closely related organisms or well-conserved protein families. Therefore, the identification of shared peptides leads to ambiguities making it difficult to determine the actual presence of a specific protein within a sample. This so-called protein inference issue has been previously described (Nesvizhskii and Aebersold, 2005) and various solutions have been proposed on this topic (Serang and Noble, 2012). Finally, a recommended practice in proteomics is to disregard so-called one-hit wonders which refer to protein identifications that are confirmed by a single peptide hit only. It should be considered, however, that a significant proportion of identified proteins in an MS experiment might be affected by such a rigorous filtering and previous studies have shown that a high amount of one-hit wonders are actually expressed (Gupta and Pevzner, 2009).

The rising importance to investigate antimicrobial resistance led to an increased number of proteomics studies in which virulence properties and involved molecular mechanisms of bacterial pathogens have been investigated (Radhouani et al., 2012; Pérez-Llarena and Bou, 2016). Tomazella et al. (2012), and dos Santos et al. (2010), e.g., studied relevant mechanisms of multi-resistant Escherichia coli, the most common bacterial pathogen. Multi-resistant Staphylococcus aureus strains are a severe issue in hospital related infections with resistant pathogens and hence also subject of numerous studies investigating resistance targets and patterns (e.g., Sirichoat et al. (2016)). An important approach to investigate protein functions from - but not limited to - resistance or resistance related genes presents the creation and use of transgenic bacteria. In this method, bacteria are engineered through an artificial HGT event, i.e. genes are deliberately transferred into another organism to characterise protein functionality under known conditions. For instance, Kaval and Halbedel (2012), investigate the role of the DivIVA protein homologues in different species. They replaced the DivIVA protein in Bacillus subtilis by a homolog DivIVA variant from the facultative human pathogen Listeria monocytogenes and discovered a species-specific, diverse role within the cell. Such transgenic bacteria could also serve as a realistic model organism to study mechanisms and characteristics of HGT. More recently, transgenic bacteria also gain importance as therapeutic agents, e.g. for human microbiome related diseases (Mimee et al., 2016).

While the above mentioned studies investigate potential HGT organisms, i.e. an organism harboring a transferred gene, in terms of gain of pathogenicity or antimicrobial resistance, to our knowledge, HGT detection and characterisation has not been investigated on the proteomic level yet. Such a characterisation involves (i) to determine the acceptor (the organism acquiring the novel sequence) and the donor (the organism donating the sequence, see Figure 1), and (ii) to establish evidence through protein identification that the gain of function did indeed arise from an HGT event. Such evidence in turn can help understanding the mechanisms and constraints behind HGT.

In this manuscript, we present a novel approach for MS-based HGT detection. The main objective is to find unique proteomic evidence of the transferred protein in the HGT organism. For a proteome analysis, any conventional database search of MS data from a HGT sample against a comprehensive bacterial reference proteome can identify the expressed proteins. This strategy, however, lacks information about whether these proteins have been involved in a HGT event. To investigate this property, we examine the origin of the HGT organism, namely the acceptor and the potential donor proteomes (see Figure 1 I). Given that the acceptor proteome and at least a potential donor proteome candidate is known, the goal is to determine proteins that can be solely attributed to the donor proteome while all remaining protein identifications have to be linked to the acceptor (see Figure 1.4 II). The presence of other donor proteins could be an indicator for a mixed probe of acceptor and donor (like, e.g., in a double infection or co-culture) rather than for a single HGT organism. In a naive filtering approach, one would try to filter all unique donor protein hits from a search against a combined acceptor-donor database. This, however, can lead to a high amount of false positive reports if, e.g., acceptor and donor share at least part of their proteome. Beyond classic database searching, the post-processing features of our pipeline and an optional homology-based filtering method remove false positive detections and thereby ensure the robustness of the approach. In an optional step, we map identified proteins to their genomic counterpart and thereby connect our approach to existing genome-based approaches which enables a joint analysis in proteogenomic fashion, e.g. using iPiG (Kuhring and Renard, 2012).

5.1 Identifying unique donor proteins as HGT proteins

The objective for developing our pipeline was to identify unique proteins that support a previously occurring HGT event. To achieve this goal, we define the HGT detection problem as follows: In terms of sequence and hence proteome content, an HGT organism consists primarily of the acceptor organism, i.e., the organism that has acquired the novel gene(s) (see Figure 1.4). These novel gene(s) stem from the donor organism, and should not have been present in the acceptor organism before the transfer. Using MS data acquired from samples of the potential HGT organism, the goal is to identify proteins that can be solely attributed to the donor proteome whereas the remaining protein identifications should be assigned to the acceptor proteome. For the sake of specificity, we only regard unique donor proteins, and hence, disregard ambiguous protein groups.

Our method is based on database searches against the acceptor proteome and the donor proteome. The aim is to first identify peptides not belonging to the acceptor proteome that can be linked to the donor proteome. Protein identifications from these peptide spectrum matches (PSMs) should only lead to unique donor proteins. At the same time, no identifications assigned to the remaining donor proteome should be detected. This uniqueness property corresponds to the characteristic of a HGT protein, hence any shared peptides are unlikely to identify a HGT protein or to add further information to characterise such a protein. To ensure the uniqueness property, filter criteria are applied to the identified PSMs, peptides and proteins, and the results may be refined with an optional homology filter. Finally, identified proteins are mapped to their genomic counterpart to pinpoint the genomic region of the HGT. We explain the steps of our method in more detail in the following paragraphs.

5.1.1 Database search

The search engine MS-GF+ (Beta (v10089) (7/16/2014)) (Kim and Pevzner, 2014) is used to search MS/MS spectra acquired from HGT organism samples against two protein databases, derived from both acceptor and donor proteome. The databases are searched separately to ensure shared peptides are reported in unbiased fashion for both acceptor and donor. Our pipeline currently accepts Mascot Generic Format (MGF) input format, and supports MS-GF+-specific settings, such as parent mass tolerance ($-t$, default 10ppm), fragmentation method identifier ($-m$, default CID), and required memory limits. For now, static MS-GF+ values for decoy database search (true), Orbitrap/FTICR, and enzyme identifier (trypsin) are used ($-tda\,1 - inst\,1\ - e\,1$). We use the default number of tolerable (tryptic) termini, $-ntt\,2$, i.e. fully-tryptic peptides only are considered. However, if MS-GF+ is executed outside our pipeline, any parameter settings are possible and the pipeline can be run on provided *mzIdentML* files (Jones et al., 2012). All database search hits, i.e. all PSMs, are examined in various filtering steps which are described in the following paragraphs.

5.1.2 Unique donor peptides and proteins

The goal of the uniqueness filter is to identify proteins from the spectra that can be uniquely assigned to the donor proteome. For this purpose, the following filter criteria are applied to the resulting PSMs and peptides. All identification steps during the filtering are done by the Hortense pipeline without using another external search engine such as MS-GF+. After filtering by a stringent false discovery rate (FDR) threshold ($< 1\%$), the PSMs are first classified into either acceptor or donor or shared by both. Only peptides from unique donor PSMs are used for protein identification. Ideally, the unique donor PSMs should lead to only unique donor peptides being identified. Due to some FDR artifacts, e.g. in case the donor and acceptor database differ in size, this is not always the case. This might result in a protein being identified by the supposedly unique donor peptides that can then be assigned to a protein from the acceptor (whose PSM was filtered out by the FDR applied to the acceptor database, see e.g. Renard et al. (2010)). Hence, all supposedly unique peptides are filtered further in a cross validation step: All peptides are mapped against the set of all possible tryptic peptides derived from the acceptor proteome. The *in silico* digestion is also done by the Hortense pipeline and tryptic peptides have a length between 6 and 40 amino acids. In addition, each isoleucin is replaced by leucin, and missing start codons are always ignored.

It is also required that an identified peptide is unique within the donor proteome. Thus, if one peptide can infer multiple proteins, it cannot be assured which of them is the supposed HGT candidate, and, hence, such a non-unique peptide is excluded from the following protein identification step. Since only single proteins can hence be identified, we do not regard protein groups among the reported HGT candidates. These ambiguous proteins are again identified in another round of cross validation against the set of all tryptic peptides of the donor proteome. All remaining proteins are reported as HGT candidates.

5.1.3 Homology filter

The homology filter presents an optional filtering step for the case that the acceptor and donor organism share homologous proteins that have not been detected by the previous filtering. Per se, it is unlikely that a suggested HGT candidate actually is a HGT protein if a homologous counterpart exists in both references. As a default in our use case, a protein is defined as homologue to another if both share at least three peptides. This number of shared peptides can also be defined by the user. In some cases, however, it makes sense to turn this filter off (see DivIVA data set for an example).

5.1.4 Genome alignment

To determine the genomic origin of identified proteins, the protein sequences are mapped to the six frame translation of the donor genome sequence. In case of multiple transferred proteins, e.g., we can thereby examine if these proteins are colocated on the genome, and hence may be involved in the same HGT event.

5.1.5 Output of HGT candidates

All proteins that pass the previously described filtering criteria (FDR filter, uniqueness, homology) are reported as HGT candidates in a custom CSV format featuring their protein header information, genomic location (if available), protein coverage (in percentage of sequence content covered by observed peptides), and number of supporting peptides along with their sequence. All protein sequences are also provided in FASTA format for convenience. Please note that, in the interest of sensitivity and completeness, we report all candidates including those candidates that are supported by only one peptide. We leave it to the user to critically evaluate those candidates.

5.1.6 Snakemake wrapper

All pipeline steps are implemented in Python3. To ensure better usability, we wrapped the single program calls into one pipeline file using the workflow management system Snakemake (Köster and Rahmann, 2012). Parameter settings are enabled via a configuration file so that the whole pipeline can be automatically executed with one program call.

5.2 Experimental setup

5.2.1 Data sets

To validate Hortense, the pipeline is tested on four data sets. *H.pylori* presents a simulated data set for a proof of principle. The DivIVA is a real data set from a transgenic organism. The non-HGT *Bacillus* data set and a set of mixed spectra from *B.subtilis* and *Listeria* that emulates a co-culture serve as negative controls. The details of the experiments are explained below.

H. pylori. The *Helicobacter pylori* data set is a simulated set from a genomic HGT simulation (see Trappe et al. (2016) for details of genomic simulation). The acceptor is *Escherichia coli* K12 substr. DH10B (NC_010473.1), *H. pylori* strain M1

(NZ_AP014710.1) the donor. The *in silico* transferred phage region (genomic positions 1'322'000-1'350'000) contains a total of 27 proteins. These proteins together with all *E.coli* K12 substr. DH10B proteins (retrieved from NCBI 08/06/2016) built up the HGT proteome, and are digested *in silico* to tryptic peptides. We defined the digested peptides to have minimal length six, maximal length 30 and to have at most three missed cleavages. Using the tool MS^{2}PIP (Degroeve et al., 2015), all peptides were converted to simulated spectra, yielding a total of 295.539 MS^2 scans (i.e., one spectrum for every created peptide). The pipeline was tested with and without homology filter.

DivIVA. The HGT organism in this data set is *Bacillus subtilis* BSN238, a transgenic organism that is a chimera of *B. subtilis* 168 where the DivIVA protein has been replaced with the DivIVA from *Listeria monocytogenes* strain EGD-e (van Baarle et al., 2012). Proteomes of *B. subtilis* 168 and *L. monocytogenes* strain EGD-e were retrieved vom UniProt on 15/11/2016. The Listeria DivIVA protein is located on the complement strand at positions 2'100'224-2'100'751 (NC_003210.1). Bacterial cultivation, protein extraction and proteomic sample measurements were performed in house. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaíno et al., 2016) partner repository with the dataset identifier PXD007242 and 10.6019/PXD007242. For details on the isolation of cellular proteins and the nLC-MS/MS, please see Supplement B.

Bacillus negative control. As a negative control, *B. subtilis* 168 - the acceptor in the DivIVA data set - is utilised. This Bacillus still has its original DivIVA protein and no HGT event should be detected in the same setting as for the above DivIVA data set. Existing MS data from the PRIDE archive is used: project number PXD003764, raw data files 20130707_VR_Bsu_pWTPtkAPtpZreplicate4_F01-6. Acceptor proteome is again *B.subtilis* 168, donor proteome *L.monocytogenes* EGD-e.

Bacillus-Listeria mixed spectra. As a second approach of a negative control, an *in silico* experiment was conducted with input spectra that stem from a simulated co-culture of acceptor and donor instead of a pure culture of the HGT organism. This data set was created from the *B.subtilis* 168 spectra used in the first negative control above and *L.monocytogenes* EGD-e spectra (PRIDE project PXD001108). The expected outcome is that all *L.monocytogenes* EGD-e proteins not shared with *B.subtilis* 168 should be reported, as they are represented by the spectra but not present in the acceptor proteome.

5.2.2 Experimental Design and Statistical Rationale

The experimental rationale of datasets PXD003764 and PXD001108 has been described elsewhere (Shi et al., 2016; Misra et al., 2014). For the DivIVA dataset, as a proof of principle analysis without quantitative or differential analyses a single replicate was considered sufficient. The successful transfer of DivIVA was already shown in van Baarle et al. (2012) and the expected behavior of the pipeline for negative controls and mixed cultures was shown using the aforementioned existing MS/MS data.

Setup of in silico Experiments

Our (*in silico*) experiments are based on the aforementioned four data sets and are separated in two parts. First, we conduct a proof of principle with the simulated *H.pylori* data set, and also validate our approach on the two negative control *Bacillus* data sets. Here, acceptor and donor references are regarded as known and fixed in these settings. In the first negative control with a single non-HGT, no HGT proteins should be reported since there should be no spectra in the data set covering a foreign protein. In contrast to that, in the second negative control with a simulated co-culture, many spectra cover the presumed donor. Since our pipeline always assumes that the data represents a HGT organism, we expect our pipeline to report all proteins from the presumed donor that are represented by spectra and not present in the acceptor proteome. The goal of the second *in silico* experiment is to demonstrate that it is possible to distinguish a pure culture from a (accidental) co-culture.

To show the advantage of Hortense, we compare our results to a naive filtering approach. In this case, one would search the spectra from the HGT organism against a combined database of acceptor and donor, and then filter for the unique donor protein hits. Here, it can be assumed that all HGT proteins are identified, but the number of false positive identifications cannot be assessed.

Using a more comprehensive analysis approach in the second part, we want to emulate a real use case scenario by applying our workflow to the DivIVA data set under the assumption that only little is known about the transfer in advance. In a first attempt, one might opt for searching against a comprehensive bacterial reference database to identify potential references. Once potential acceptor and donor candidates are known, the search space can be reduced to their respective proteomes. To account for all possible proteomes, we would aim to search against a combined database of UniProtKB/Swiss-Prot and UniProtKB/TrEMBL, i.e. the complete set of available protein sequences. Due to current limitations regarding database size by MS-GF+, this database had to be reduced to the Listeriaceae taxonomy level (128'445 proteins, retrieval date 10/03/2017). Thus, we assume that the acceptor - *B. subtilis* - is known and that the donor is contained within the

Listeria lineage. The pipeline is then executed on all pairs of potential Listeria donor proteomes paired with *B. subtilis.*

This search is analogous to the database search described in the Methods section above. We show these results compared to the filtered results of our complete pipeline. In our experiments, we regard only those reported HGT candidates as (true) positive that are supported by more than one peptide.

Settings

We run all data sets with default parameters described in the Methods section. MS-GF+ parameters are as stated in the database search paragraph, and default otherwise. As per default in MS-GF+, Carboamidomethylation of C (C+57) was used as fixed modification and no variable modifications were considered. For the DivIVA and Bacillus data sets, we deactivate the homology filter since the DivIVA protein in *L. monocytogenes* is a homolog of the natural *B. subtilis* 168 DivIVA protein. For the naive filtering approach, we use the same MS-GF+ settings as for the evaluation of our pipeline. All reported numbers of protein hits are without one-hit wonders unless stated otherwise.

5.3 Results

5.3.1 Precision of Hortense for HGT protein detection

The simulated *H.pylori* data set is based on a genomically simulated HGT organism for which a phage with 27 proteins was transferred *in silico* from an *H.pylori* to an *E.coli* K12. The theoretical proteins of this artificial HGT organism were digested *in silico*, and the simulated spectra were used for a proof of concept for our pipeline. The conventional database search yields 4267 protein hits on the acceptor proteome (*E.coli* K12), and 1375 on the donor proteome (see Figure 5.2 B, *H.pylori*). The naive filtering approach (see Figure 5.2 A, *H.pylori*) can reduce this number to 78 seemingly unique donor proteins. But since only the 27 transferred proteins should be present, the naive filtering resulted therefore in 51 false positive (FP) reports. This means, without further filtering, one would have to investigate 78 protein candidates regarding a possible HGT property. Applying our pipeline, we can reduce this number to only true positive HGT proteins. From the 27 possible HGT proteins, Hortense detected 24 with the homology filter turned on, and all 27 with the homology filter turned off (see Figure 5.2 B and Supplementary Table S1). Figure 5.3a shows the successful mapping of the HGT proteins to their genomic positions. No additional protein candidates except one-hit wonders were reported. This proof of concept shows that our pipeline is able to successfully detect HGT proteins as such without reporting unwanted non-HGT proteins.

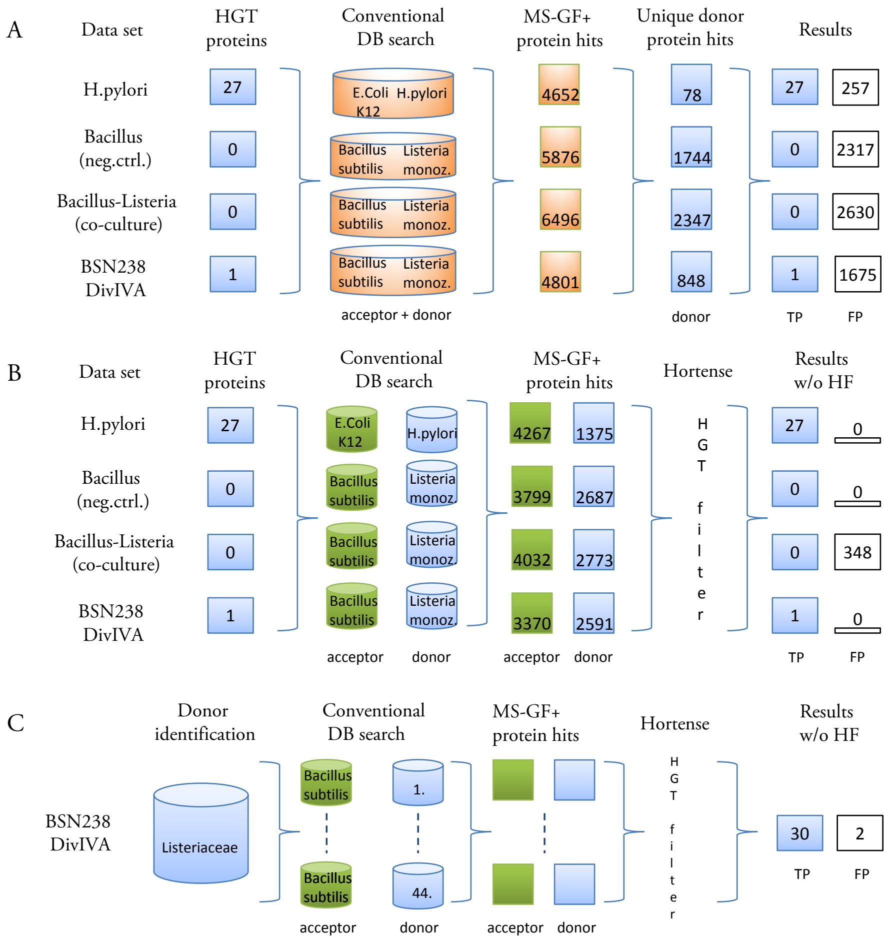
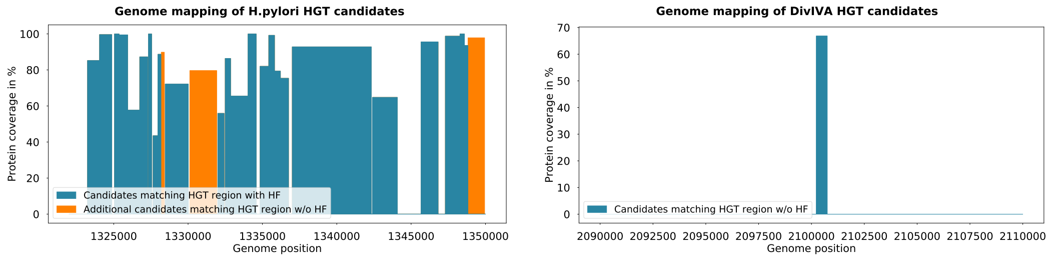


Figure 5.2. Results for Hortense compared to naive filtering. Column *HGT proteins* states the number of known HGT proteins per data set. (A) For the naive filtering, all spectra were mapped against a combined acceptor+donor database (orange). The resulting *MS-GF+ protein hits* (orange) were filtered for unique donor protein hits (no match on the acceptor for this spectrum) resulting in 78 for the *H.pylori*, 1744 and 2347 for the negative *Bacillus* and co-culture data sets, resp., and 848 for the DivIVA (no one-hit wonder). This means, a lot of FP remained in addition to the HGT proteins. (B) For Hortense, spectra were matched against separate databases of acceptor and donor (green and blue), and the *MS-GF+ protein hits* were filtered by the pipeline. For the HGT organisms in the *H.pylori* and DivIVA data sets, only the true HGT proteins were reported without FP hits without homology filtering. For the negative *Bacillus*, no candidates were reported. For the simulated co-culture, a high number of candidates was reported, marking the result as a non-HGT mix of - likely - acceptor and donor. (C) For the DivIVA data set, a more comprehensive HGT search was emulated. Spectra were first matched against the Listeriaceae proteome to determine donor proteome candidates. Hortense was applied to all 44 candidates, 30 of them carrying the DivIVA protein. Hortense reported all 30 with two FP. Nothing was reported for the 14 non-DivIVA candidates.



(a) Genome alignment *H.pylori*.

(b) Genome alignment DivIVA.

Figure 5.3. Genome alignment of Hortense HGT candidates. Shown is a fraction containing only the genomic HGT region. Protein coverage of all candidates is plotted. (a) For the *H.pylori* dataset, all HGT proteins could be successfully aligned to the correct genomic region. The 24 candidates with homology filter (HF) are marked in blue, the three additional candidates found without HF are marked in orange. All HGT proteins were reported with a peptide support covering at least 40% of the protein. (b) The single DivIVA HGT protein has the correct genomic mapping position and a protein coverage of 67%.

5.3.2 Robustness of Hortense for non-HT organisms

In addition to the proof of concept, we want to show the robustness of our approach via negative controls, i.e., with data from non-HGT organisms. In the first negative control, *Bacillus*, with MS data from *B.subtilis* 168, database searches against acceptor (*B.subtilis* 168) and donor (*L.monocytogenes* EGD-e) yield 3799 and 2687 protein hits. When removing one-hit wonders (no hit on DivIVA), no HGT candidate proteins are reported by our pipeline. The naive filtering approach reports 1744 FP unique donor protein candidates. In the second negative control, we simulated a double infection by mixing MS data sets from two experiments from *B.subtilis* 168 and *L.monocytogenes* EGD-e. We assumed a HGT organism concurrent with the DivIVA HGT organism, and ran the pipeline with *B.subtilis* 168 as acceptor and *L.monocytogenes* EGD-e as donor. Compared to a single non-HGT run, our pipeline should report all covered donor proteins that are not also present in the acceptor proteome. As expected, Hortense reports a plethora of *L.monocytogenes* EGD-e proteins (348 without homology filter, 194 with homology filter) as HGT candidates. This large list contradicts a single HGT event and would be regarded as evidence for a double infection. The naive filtering reports over 2000 unique donor proteins. Most are likely present but regarding the question if a HGT event has occurred and the HGT organism is present, this outcome cannot be distinguished from the pure negative control, and hence, it cannot be directly identified as a non-HGT co-culture or double infection.

5.3.3 Application of Hortense in a real HGT detection process

With the DivIVA data set, we wanted to emulate the process of HGT detection given MS data where only little is known about the transfer and involved acceptor and donor candidates. That is, in a first step, the goal is to identify potential acceptor and donor proteomes in a metaproteomic fashion, i.e. by searching against a large collection of proteome references. Due to MS-GF+ memory limitations, we had to reduce the reference database to the Listeriaceae taxonomy level, i.e. we had to assume the acceptor *B. subtilis* is known. The aim is to identify potential donor candidates among the Listeriaceae lineage. We ran a MS-GF+ search on these proteomes and then ran the pipeline on all reported references (see Figure 5.2 C). Supplementary Table S2 lists all donor candidates together with the pipeline results, i.e., if the DivIVA protein could be reported, the number of supporting peptides, and how many hits on other proteins were reported (false positive HGT reports). A total of 44 donor candidates were processed, 12 proteomes at the species level and 32 proteomes at the strain level. These 12 correspond to the species level proteome of at least one reported strain. For eight of these 12 candidates, our pipeline did not report any HGT protein meaning that the DivIVA protein is not part of the species core proteome. Another six proteomes at the strain level also turned out to not have the DivIVA protein or any corresponding homologous protein. The real donor, Listeria monocytogenes serovar 1/2a strain EGD-e, is among the DivIVA positive hits. Here, the DivIVA protein was reported with a support of 12 peptides. For all remaining donor candidates, our pipeline reports the DivIVA protein also with 12 or fewer supporting peptides. For only two donor candidates, our pipeline reported the same additional false HGT protein. This protein is the GMP synthase (glutamine-hydrolyzing) for both *L. fleischmannii* 1991 and subsp.coloradonensis (Uniprot IDs A0A0J8JA30 and H7F4C6). So even among multiple donor candidates, we could successfully identify the HGT protein with almost no false positive hits. The number of donor candidates with a positive DivIVA hit, however, already illustrates the difficulty that arises if the HGT protein is present in multiple organisms. Although the real donor was among the candidates with the highest peptide support, this property alone is not sufficient to distinguish the true donor candidates. For the true donor candidate Listeria monocytogenes serovar 1/2a strain EGD-e, we examined the pipeline results in more detail in consistency with our remaining data sets (see Figure 5.2 A and B, and Table S1). The pipeline drastically reduced the number of acceptor (3370) and donor (2591) protein hits to one HGT protein candidate without any additional false positives. The 12 supporting peptides cover 67% of the protein, and the determined genomic positions 2'100'750 - 2'100'226 from the genome alignment correspond to the DivIVA protein location (see Figure 5.3b). The naive filtering reports another 847 FP unique donor proteins in addition to the DivIVA protein. As a conclusion, also for the real DivIVA data set we could successfully apply our pipeline to reduce the number of conventional database hits to single out the correct HGT protein without reporting false positive hits.

5.4 Discussion of results from Hortense

In the era of multi-resistant bacteria, which frequently acquire specific traits via horizontal gene transfer, it is important to be able to detect and characterise such HGT events on a proteomic level. We defined two objectives for such a detection and characterisation process. First, the acceptor and the donor of the HGT organism have to be determined. Secondly, we want to establish evidence through protein identification for the presence of horizontally transferred proteins. Given that acceptor and donor are known, one would assume that a conventional database search on a combined acceptor+donor proteome with a following naive filter that reports only unique donor proteins should be sufficient. We showed that such a naive filter indeed identifies the HGT proteins but at the cost of many false positive reports. Even for a non-HGT organism for which no unique donor proteins should be found, the naive filter reports several 100 false positives. Using an adapted database search approach as presented in Hortense can be advantageous to pinpoint HGT proteins represented in the sample. Hortense is able to precisely detect HGT proteins with few - if any - false positives, and, at the same time, is robust for non-HGT samples. It should be noted that our results for the simulated data may be somewhat overly optimistic regarding the number of peptides. This can become problematic if the detected HGT protein is only found at a low abundance. As with all database approaches, the limitation is the availability of suitable reference proteomes which should, however, become less prominent as more and more proteomes are made available. If the donor proteome is not available at all, one could still opt for a *de novo* peptide sequencing approach to assemble the presumed HGT protein from the spectra that could not be mapped to the acceptor proteome (Muth and Renard, 2017). However, although *de novo* sequencing has been successfully applied for assembling full-length antibody sequences (Tran et al., 2016), the technique is still not as reliable as database searching and requires MS/MS spectra of high resolution and -even better- of different fragmentation modes to achieve a sufficient performance (Guthals et al., 2013).

In the application of Hortense to a real HGT detection scenario, we addressed the first objective of acceptor and donor proteome selection. These proteomes can be identified in a metaproteomic fashion from a database search against a comprehensive database. Due to current limitations by MS-GF+, we had to reduce the search space for the identification from the complete UniProtKB/Swiss-Prot and UniProtKB/TrEMBL to the Listeriaecae lineage for the donor. Here, we were still able to successfully identify the DivIVA protein among multiple lineages. Still, this application shows difficulties that arise when we allow homologous proteins. We gain many hits on different strains and the true donor could not be clearly distinguished from other, biological relevant, hits. Performing an additional functional analysis by inferring phenotypic knowledge for such ambiguous protein candidates may help to further refine the reported results. The metaproteomic problem of identifying different organisms within a sample is not HGT specific and has already been addressed in various studies (see review article by Muth et al. (2016)). While computational approaches evolve, we can expect an increase in the resolution of the bacterial composition, and also be better able to handle larger databases. A possible alternative to the metaproteomic approach could be to determine acceptor and donor candidates on the genomic level first. If also NGS sequencing data of the HGT organism is available, one could, e.g., leverage metagenomic profiling tools to identify acceptor and donor candidates. Here, we show results from simulated and transgenic organisms where ground truth is clear and without doubt. Few proteomic studies (e.g. Tomazella et al. (2012), dos Santos et al. (2010), or Sirichoat et al. (2016)) explicitly investigate potential HGT organisms. Since they often have very specific objectives and since the data is either not suitable for our generic HGT question or is simply not available, verification is hard to obtain. Better data sets could thus help to further improve HGT algorithm engineering. By detecting and characterising horizontal transfers, Hortense can help to increase our general understanding of HGT events and its implications for public health.

6 Summary and further discussion

6.1 Summary of contributions

Structural variations (SVs) in general and horizontal gene transfer (HGT) events in particular have a huge impact on both human health and disease. Especially in the latter case, SVs play an important role in diseases such as cancer or for the spread of severe functionality like in the form of antimicrobial resistance genes. It is important to study such events both on the genomic and proteomic level: the genome elucidates the enormous potential from the gene content but the presence of a gene alone does not inevitably correlate with its expression on the protein level.

Existing methods for SV detection from NGS data were limited in their resolution of SV type, size or complexity. HGT events, a special kind of SV in some terms, have been investigated from fully assembled genomes by methods that find compositional patterns in genome sequences corresponding to transferred genes or via phylogenetic discrepancies. However, so far they have not been detected directly from NGS data, which can provide a faster and complementary approach to available methods. On the proteomic level, there are studies investigating potentially transferred genes like antibiotic resistance genes but HGT events have not been investigated and characterised as such.

This thesis describes four computational methods for SV detection in general and HGT events in particular both from genomic NGS data and proteomic mass spectrometry (MS) shotgun data. In Chapter 2, we introduced a generic SV detection method called Gustaf. Gustaf is a split-read aligner that uses paired-end information as a second stage of support. Gustaf can identify even complex variants such as translocations and duplications with base pair resolution and further improves former available means of SV detection by filling the size gap of SV resolution termed the NGS twilight zone at that time. A special contribution lies in the identification of duplication and translocation events as the combination of simpler copy or deletion events with an insertion at a distinct location. This means that the distantly related simpler events are usually not spanned by the same read, and hence, separately called events have to be identified as belonging to the same complex variation.

We then applied and adapted the concepts of SV detection for the special HGT event. In an abstract point of view, a transferred gene has the same pattern as an inter-chromosomal translocation, if acceptor and donor genome are treated as chromosomes that can undergo rearrangements. The problem of HGT identification is actually two-staged. First, the acceptor and donor references, if not known in advance, have to be identified. Based on that, these two references can be used in a mapping-based SV detection approach tailored to HGTs for HGT site characterisation and gene identification. In Chapter 3, we presented Daisy, a method that aims to answer the second question of HGT site characterisation. Here, we use Gustaf to find HGT candidates as translocation events between acceptor and donor, and then do a thorough analysis of coverage and paired-end information around the HGT site candidate to establish further evidence for the transfer. The split-read evidence provided by Daisy is the first HGT site evidence with base pair resolution while existing methods all remain to some degree probabilistic. We followed up on that with the development of a method to answer the question of acceptor and donor identification. We presented this method, called DaisyGPS, in Chapter 4. The problem of acceptor and donor identification from NGS reads is akin to the problem of species identification in metagenomics. Hence, we adapted and applied the tool MicrobeGPS to the specialised purpose to identify acceptor and donor references from NGS reads. Both Daisy and DaisyGPS provide novel evidence for the HGT detection problem since these are the first tools available that enable mapping-based HGT detection directly from NGS data.

As mentioned above, the incorporation of a novel gene does not inevitably lead to its expression. Hence, we further investigated HGT events on the protein level. To provide complementary evidence on an orthogonal level to our NGS approach, we developed a proteomics HGT detection method from shotgun MS data called Hortense that we introduced in Chapter 5. Hortense extends the standard database search routine of MS spectra with a thorough cross-validation to ensure the properties of the identified proteins that we defined to be characteristic for a transferred protein and its acceptor and donor proteomes. The proteogenomic feature of Hortense provides the possibility to integrate Daisy, DaisyGPS and Hortense and their results. Together, all three HGT detection and characterisation methods provide means of analysis that was not possible before.

6.2 Recent developments, challenges and future research

6.2.1 Gustaf in the context of current SV detection research and findings

In 2012, Evan Eichler said that "the strength will be when we can go in and integrate across all the variation, irrespective of class, type, variant and frequency. [...] I hope in ten years, people will just be studying the full spectrum of genetic variation" (Baker, 2012). In my opinion, there is still a long way to go to reach that goal.

The problem of complex variants is still not fully grasped, and the spectrum and extent of variants are far from resolved. Hence, the discovery of the SV landscape is still in its beginnings (Baker, 2012). Almost every tool among the plethora of SV detection methods available today still tends to only call a fraction of - known - SVs, specific in size and type. Albeit specialised methods are meaningful when focussing on single events, hybrid methods such as the new Delly version, meta callers such as LUMPY (Layer et al., 2014), CLOVE (Schröder et al., 2017) or SV^2 (Antaki et al., 2017), or integrative callers such GRIDSS (Cameron et al., 2017) promise more sensitive and specific SV calling for large scale SV detection. In addition, tools are developed to compare different call sets of SVs for the same data set (e.g., Sedlazeck et al. (2017)), or efforts to create searchable databases of SVs that can also be directly screened with NGS data (e.g. Lam et al. (2009)).

To really pinpoint the current state of the art and to unveil the next important directions in SV calling, benchmarks are needed. However, these are not easily done in an extensive way as ground truth for combined, complex variants is rare and simulated data tends to not reflect reality very well when it comes to SVs. As an additional hurdle for benchmarking efforts, different variant call formats have been used throughout the years, and even the now commonly used VCF format offers different types of data entries for the same variant.

Another challenge in SV detection regards the choice of the reference. In a pure mapping based approach and with only a single reference being used, of course only SVs apparent from this single reference can be detected. With the steadily increasing number of sequenced genomes, the single reference approach will not suffice and there is need for novel approaches to handle the large amount of data. The concept of pan-genomes, that was originally defined as the sum of all core genes of all the strains within a clade, has been revised and expanded to be "any collection of genomic sequences to be analysed jointly or to be used as a reference" (The Computational Pan-Genomics Consortium., 2016). Until the advent of pan-genome methods, multiple references were usually only handled via whole-genome alignment methods. Structures such as journaled string trees (Rahn et al., 2014) or PanCake (Ernst and Rahmann, 2013) first introduced the concept of analysing multiple references at once (apart from whole-genome alignment methods) and methods like seq-seq-pan (Jandrasits et al., 2018) promise to offer efficient, flexible handling and mapping of a high number of reference genomes. The first SV detection methods taking advantage of pan-genome structures and methods are already on their way, e.g. Valenzuela et al. (2015). Other efforts go towards population based variant calling, e.g. Popins (Kehr et al., 2017), where Popins focuses on insertions so far.

The resolution of translocation and duplication events by Gustaf is also just one example of possible combinations of simple SVs to form more complex types. Another example that has been observed and also incorporated in Gustaf since is the combination of inversion and deletion events. That means, part of an inverted sequence has been deleted around one of the breakpoints during the inversion event, obscuring the observed SV type around that breakpoint. This is bound to be just another example and it is highly like that also more than two types of events could be combined in a complex SV.

Nevertheless, SV detection is limited for NGS data because of constraints from short read lengths that prevent full characterisation of SV breakpoints, especially within the context of repetitive DNA. Chaisson et al. (2015b) even made the point that short read NGS data is not suited for both full genome assembly and SV detection because of the lack of long range connectivity. This fact also prevents haplotype resolution for SV calls. So in the end, everything comes down to incorporating long reads.

Gustaf was developed when 454 was still rising on the market and it seemed clear that long read sequencing would lead the market at some point in the future. Although 454 is not continued any more, Illumina can now produce longer reads as well. Illumina however, focusses on deep sequencing and produces billions of reads. Gustaf was not designed for that and does not have the necessary efficiency to handle this huge amount of data, so at the moment Gustaf is used in combination with a specific read mapper on umapped reads. Still, the technology landscape is bound to change again with improved products from both MinION and PacBio already on the market. While Illumina short-read technologies with their deep coverage approach are still prominent, third generation long-read technologies - although still more expensive with less throughput - gain popularity again.

However, the current computational methods are optimised for Illumina and have to be adapted for the various challenges like, e.g., different sequencing error profiles, of the long-read technologies. The first tools focussing on long reads are already out, e.g. PBHoney (English et al., 2015). Norris et al. (2016) detected SVs in cancer using nanopore sequencing, Huddleston et al. (2016) explored long read data and discovered the importance for haplotype resolution. In the most recent study by Chaisson et al. (2017), the authors combined long and short read technologies for haplotype resolved SV detection in a trio data set, and also provided a comprehensive real data set available as ground truth for future benchmarking efforts. Only with improved methods for and further studies from long read incorporations, the community will reach the next milestones concerning meaningful SV detection.

6.2.2 HGT in microbiome communities and further applications

Jiang et al. (2017) applied a method similar to Daisy to gut microbiome samples from the Human Microbiome Project, and identified 5600 putative mobile genetic elements (MGEs) that they also provide via the database ImmeDB. This study shows the applicability of our HGT detection approach in a large scale study. Yet, in this large scale approach the focus can only lie in collecting MGEs present - or now absent - from the bacteria present in the microbiome. According to the authors, there is also more value for functional analysis and characterisation of MGEs in longitudinal studies that analyse microbiomes over a range of time instead of focusing on static snapshots from onetime samplings. In this context, approaches like Daisy would benefit from an integration with further evidence, e.g., with a large scale insert assembly as we tested with Popins for a cohort of related outbreak strains. A follow-up cross-validation between these insertions and HGT candidates from Daisy could help to characterise differences between MGEs that relocate within the same genome and foreign integrated genes.

Such an approach becomes feasible since more and more genomes are sequenced, hence more references are available. On the downside, massive sequencing can also become a source for bias. Poorly assembled reference genomes can mislead in the metagenomic profiling and hence identification of interaction partners when looking for suitable acceptors and donors. Especially when acceptor and donor are closely related and have a high sequence similarity, mapping-based identification is challenging. Unfinished assemblies that are missing sequence content can also lead to false positive HGT identifications.

For future research, there are bound to be further possibilities and applications of our HGT detection methods. The next logical step would be to really integrate our genomics methods with our proteomics approach. Also, Daisy has been used in a study to report horizontally transferred genes in the Daphnia iridescent virus 1 (Toenshoff et al., 2018), showing that the method is further transferable to other organisms apart from bacteria. Lab technology wise, breakthroughs such as CRISPR-cas9 will give new fields of applications. CRISPR-cas9 allows the specific and precise alteration of genomic sequences. Originally, it is a bacterial defense mechanism to avoid viral reinfection. From an abstract point of view, introducing DNA with CRIPSR-cas9 creates somewhat of an artificial HGT event that should be detectable with the same means. The same holds true for genetically modified organisms. The former has implications also in the context of human infection biology that will revolutionise modern medicine at some point. But this also brings up the need for methods for quality and risk assessment of such artificial transfers, most likely on all three levels - DNA, RNA and protein.

A Appendix - DaisySuite

Table A.1. Acceptor and donor candidates for sim1HP run with yara, no species filter and no samflag filter. Sampling sensitivity = 90. No taxon blacklist. No parent blacklist. No species blacklist. -0.000* represents absolute values $< 0.000\bar{4}$. [1]Salmonella enterica subsp. enterica serovar Anatum str. USDA-ARS-USMARC-1676

Candidate			**MicrobeGPS metrics**			**DaisyGPS metrics**	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Escherichia coli str. K-12 substr. DH10B	NC_010473.1	197800	0.254	0.082	0.173	0.003
Acceptor	Escherichia coli K-12	NZ_CP010445.1	187050	0.237	0.075	0.162	0.003
Donor	[Haemophilus] ducreyi	NZ_CP015434.1	322	0.001	0.926	-0.924	-0.000*
Donor	Salmonella enterica [...] USDA-ARS-USMARC-1676[1]	NZ_CP014620.1	126	0.001	0.919	-0.918	-0.000*
Donor	Klebsiella oxytoca KONIH1	NZ_CP008788.1	1791	0.001	0.795	-0.794	-0.000*
Donor	Helicobacter pylori	NZ_AP014710.1	9154	0.018	0.79	-0.782	-0.001
Acceptor-like Donor	Escherichia coli	NZ_CP016182.1	74580	0.094	0.088	0.006	0.000*

Table A.2. Results for sim1HP run with yara, gustaf, no species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		**Acceptor**			**Donor**			**Read Evidence**			**Evidence Filter**			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NZ_CP010445.1	NZ_AP014710.1	1880235	1880237	44.0	1322002	1350000	94.62	152	182	8712	7	100	100	100
NZ_CP010445.1	NZ_CP015434.1	3904873	3904886	40.54	114928	126957	30.41	871	156	884	3	100	100	100
NC_010473.1	NZ_AP014710.1	1120261	1120263	43.0	1322002	1350000	94.62	154	182	8712	3	100	100	100

Table A.3. Acceptor and donor candidates for real1B run with yara, species filter and no samflag filter. Taxon blacklist: [83334, 1045010]. Parent blacklist: [83334]. No species blacklist. (-)0.000* represents absolute values $< 0.000\bar{4}$.

Candidate			**MicrobeGPS metrics**			**DaisyGPS metrics**	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Escherichia coli Xuzhou21	NC_017906.1	1040394	0.846	0.054	0.792	0.018
Acceptor	Escherichia coli O55:H7 str. RM12579	NC_017656.1	816492	0.723	0.040	0.683	0.012
Donor	Cronobacter sakazakii CMCC 45402	NC_023032.1	201	0.006	0.861	-0.855	-0.000*
Donor	Enterobacter hormaechei subsp. hormaechei	NZ_CP010377.1	206	0.002	0.78	-0.778	-0.000*
Donor	Citrobacter freundii CFNIH1	NZ_CP007557.1	1443	0.001	0.743	-0.742	-0.000*
Donor	Citrobacter koseri ATCC BAA-895	NC_009792.1	93	0.004	0.560	-0.557	-0.000*
Acceptor-like Donor	Corynebacterium humireducens NBRC 106098 = DSM 45392	NZ_CP005286.1	117	0.444	0.078	0.366	0.000*
Acceptor-like Donor	Shigella dysenteriae Sd197	NC_007606.1	148868	0.193	0.041	0.152	0.001

Table A.4. Results for real1B run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 95. Split read threshold = 3. Taxon blacklist: [83334, 1045010]. Parent blacklist: [83334]. No species blacklist. Results (139 HGT candidates) for NC_017656.1 (acceptor) and NZ_CP007557.1 (donor) are omitted here for sake of simplicity. For all other pairs no HGT candidates were reported.

Organism		**Acceptor**			**Donor**			**Read Evidence**			**Evidence Filter**			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NC_017656.1	NC_007606.1	314439	334641	27.39	2213697	2214454	63.18	39	3	102	0	100	100	100
NC_017656.1	NC_007606.1	1570633	1580081	138.85	1282007	1320884	7.51	9	1	714	100	97	96	98
NC_017656.1	NC_007606.1	1570633	1584983	141.99	1282007	1329491	11.14	11	12	973	99	97	98	97
NC_017656.1	NC_007606.1	1580080	1584983	148.04	1320883	1329491	27.6	8	12	261	99	99	99	99
NC_017656.1	NC_007606.1	1589216	1618452	247.73	4032919	4035786	110.69	107	10	576	100	100	100	100
NC_017656.1	NC_007606.1	1738741	1739271	30.87	1321240	1322115	88.45	42	73	60	4	100	100	98
NC_017656.1	NC_007606.1	1738741	1739785	157.15	1321240	1322656	58.2	17	5	72	95	100	100	99
NC_017656.1	NC_007606.1	1738741	1740010	134.9	1321240	1322870	51.13	50	3	72	96	100	100	98
NC_017656.1	NC_007606.1	1738741	1740078	129.54	1321240	1322973	49.81	17	6	81	100	98	100	98
NC_017656.1	NC_007606.1	1738741	1745278	119.31	1321240	1331304	23.91	9	52	202	96	98	100	98
NC_017656.1	NC_007606.1	1739270	1739785	287.13	1322114	1322656	9.33	56	5	13	99	96	100	99
NC_017656.1	NC_007606.1	1739270	1740477	130.81	1322114	1323341	21.27	28	3	42	96	98	99	98
NC_017656.1	NC_007606.1	1739270	1745278	127.11	1322114	1331304	17.77	24	52	143	97	99	99	96
NC_017656.1	NC_007606.1	1739784	1741539	10.67	1283675	1322655	11.22	19	294	897	4	97	100	97
NC_017656.1	NC_007606.1	1739784	1745278	112.11	1322655	1331304	18.29	16	51	130	95	100	100	100
NC_017656.1	NC_007606.1	1740009	1740477	6.25	1322869	1323341	42.62	20	3	28	5	97	100	96
NC_017656.1	NC_007606.1	1740009	1745278	115.53	1322869	1331304	18.65	17	52	129	98	99	100	96
NC_017656.1	NC_007606.1	1740077	1740477	2.25	1322972	1323341	46.64	16	3	25	4	100	100	100
NC_017656.1	NC_007606.1	**1741538**	**1744925**	164.13	**1283674**	**1288080**	59.4	18	9	692	99	100	100	100
NC_017656.1	NC_007606.1	1741538	1745278	159.71	1283674	1331304	12.51	9	166	1031	100	97	99	95
NC_017656.1	NC_007606.1	1957909	1958879	132.94	4032919	4035786	110.69	41	7	576	99	99	98	99
NC_017656.1	NC_007606.1	1957909	1982375	118.01	4032919	4035782	110.56	17	12	576	97	100	100	100
NC_017656.1	NC_007606.1	1958870	1982375	117.37	4034933	4035782	356.29	22	35	576	98	100	100	100
NC_017656.1	NC_007606.1	1986050	1986053	726.33	1288361	1331322	7.47	10	335	319	100	97	100	95
NC_017656.1	NC_007606.1	1986050	1992463	155.63	1321775	1331322	25.25	126	72	197	99	98	100	97
NC_017656.1	NC_007606.1	1986234	1992463	146.03	1321775	1329808	28.06	261	80	190	100	98	100	96
NC_017656.1	NC_007606.1	1986234	1992955	155.68	1320887	1329808	32.55	35	126	308	99	100	100	99
NC_017656.1	NC_007606.1	1992462	1992955	277.57	1320887	1321774	73.17	131	91	106	100	99	100	99
NC_017656.1	NC_007606.1	2431977	2443616	15.53	1282008	1322832	10.76	17	60	897	0	96	100	96
NC_017656.1	NC_007606.1	2435781	2443492	8.8	1282069	1320883	7.51	193	62	714	3	98	98	98
NC_017656.1	NC_007606.1	2469232	2481815	49.5	4032919	4035785	110.66	81	24	576	2	99	100	99
NC_017656.1	NC_007606.1	2486033	2488461	149.98	4298967	4301718	16.95	23	5	67	95	97	100	96
NC_017656.1	NC_007606.1	2486033	2488662	150.6	4298967	4301905	16.19	65	10	68	99	98	100	98
NC_017656.1	NC_007606.1	2486203	2488662	153.24	4299043	4301905	16.62	47	10	68	99	97	100	95
NC_017656.1	NC_007606.1	2486203	2488723	152.86	4299043	4301977	17.39	29	10	69	98	96	100	95
NC_017656.1	NC_007606.1	2487505	2489413	150.49	953376	956244	23.61	10	3	119	98	98	99	99
NC_017656.1	NC_007606.1	2488461	2489413	130.37	953376	954653	52.39	12	4	119	95	99	100	97
NC_017656.1	NC_007606.1	2488601	2488723	136.13	4301842	4301977	32.39	8	11	2	98	97	100	96
NC_017656.1	NC_007606.1	2678766	2679015	44.61	1323123	1323370	29.6	18	4	5	5	97	100	96
NC_017656.1	NC_007606.1	3607310	3629241	31.35	4189699	4189800	491.86	42	24	12	0	100	100	98
NC_017656.1	NC_007606.1	3615738	3630353	153.33	4195901	4198011	700.99	149	6	4245	97	100	98	100
NC_017656.1	NC_007606.1	3615738	3632904	131.04	4195901	4206697	139.78	21	4	4245	96	100	98	100
NC_017656.1	NC_007606.1	3615738	3632993	130.65	4195901	4206818	138.38	19	4	4250	98	100	98	100
NC_017656.1	NC_007606.1	3629240	3630353	1409.38	4189698	4198011	184.22	222	38	4278	100	100	100	100
NC_017656.1	NC_007606.1	3629240	3632904	430.46	4189698	4206697	91.85	30	36	4278	100	100	100	100
NC_017656.1	NC_007606.1	3629240	3632993	421.57	4189698	4206818	91.3	27	36	4283	100	100	100	100

Table A.5. Acceptor and donor candidates for real4 run with yara, no species filter and no samflag filter. Taxon blacklist: [595495]. No parent blacklist. No species blacklist. (-)0.000* represents absolute values $< 0.000\bar{4}$.

Candidate			**MicrobeGPS metrics**			**DaisyGPS metrics**	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Escherichia coli W	NC_017635.1	221389	0.852	0.024	0.829	0.026
Acceptor	Escherichia coli W	NC_017664.1	221570	0.853	0.025	0.828	0.026
Donor	Salmonella enterica subsp. enterica serovar Infantis	NZ_CP016410.1	83	0.005	0.943	-0.938	-0.000*
Donor	[Haemophilus] ducreyi	NZ_CP015434.1	119	0.001	0.920	-0.919	-0.000*
Donor	Zymomonas mobilis subsp. mobilis NRRL B-12526	NZ_CP003709.1	3067	0.002	0.876	-0.874	-0.000*
Acceptor-like Donor	Shigella boydii CDC 3083-94	NC_010658.1	23506	0.150	0.047	0.104	0.000*
Acceptor-like Donor	Shigella sonnei 53G	NC_016822.1	29127	0.168	0.073	0.095	0.000*

Table A.6. Acceptor and donor candidates for ERR103401 run with yara, species filter and no samflag filter. No taxon blacklist. No parent blacklist. No species blacklist. (-)0.000* represents absolute values < 0.000$\bar{4}$.

Candidate			MicrobeGPS metrics			DaisyGPS metrics	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Staphylococcus aureus subsp. aureus HO 5096 0412	NC_017763.1	440076	0.832	0.04	0.792	0.041
Acceptor	Staphylococcus aureus subsp. aureus	NZ_CP007659.1	439586	0.824	0.041	0.783	0.040
Donor	Staphylococcus pseudintermedius ED99	NC_017568.1	1089	0.002	0.691	-0.689	-0.000*
Donor	Staphylococcus warneri SG1	NC_020164.1	523	0.003	0.631	-0.628	-0.000*
Donor	Staphylococcus epidermidis RP62A	NC_002976.3	5512	0.006	0.540	-0.534	-0.000*
Donor	Staphylococcus haemolyticus JCSC1435	NC_007168.1	3614	0.005	0.291	-0.285	-0.000*
Donor	Staphylococcus aureus subsp. aureus COL	NC_002951.2	49889	0.106	0.233	-0.127	-0.001
Acceptor-like Donor	Staphylococcus aureus subsp. aureus	NZ_CP012011.1	54992	0.11	0.109	0.001	0.000*

Table A.7. Results for ERR103401 run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		Acceptor			Donor			Read Evidence			Evidence Filter			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NZ_CP007659.1	NC_020164.1	37045	37048	413.0	121379	123703	36.08	5	36	123	100	100	100	100
NZ_CP007659.1	NC_020164.1	37045	37176	220.86	111790	121379	2.26	20	7	47	100	99	100	100
NZ_CP007659.1	NC_020164.1	37047	37125	272.67	121461	123702	24.58	7	36	103	100	100	100	100
NZ_CP007659.1	NC_020164.1	37047	37176	217.84	111790	123702	8.84	19	38	160	100	100	100	100
NZ_CP007659.1	NC_020164.1	37124	37176	134.96	111790	121460	5.17	22	7	72	100	100	98	100
NC_017763.1	NZ_CP012011.1	1525462	1554768	130.8	1228987	1251487	14.6	4	26	826	100	97	100	97
NC_017763.1	NZ_CP012011.1	1525488	1554768	130.8	1228987	1251477	14.59	10	26	826	100	100	100	100
NC_017763.1	NC_020164.1	37044	37047	412.0	121379	123703	36.18	5	36	124	100	100	100	100
NC_017763.1	NC_020164.1	37044	37175	220.35	111790	121379	2.26	20	7	47	100	100	100	100
NC_017763.1	NC_020164.1	37046	37124	271.83	121461	123702	24.69	7	36	104	100	100	100	100
NC_017763.1	NC_020164.1	37046	37175	217.34	111790	123702	8.86	19	38	161	100	100	100	100
NC_017763.1	NC_020164.1	37123	37175	134.96	111790	121460	5.17	22	7	72	100	100	100	100
NZ_CP007659.1	NC_002951.2	1568261	1575973	129.75	359692	369382	5.66	9	3	42	98	98	99	95
NZ_CP007659.1	NC_002951.2	1568261	1576904	131.71	358442	369382	6.99	41	5	87	97	97	100	98
NZ_CP007659.1	NC_002951.2	1568261	1579141	126.48	356047	369382	11.01	5	3	264	99	99	100	99
NZ_CP007659.1	NC_002951.2	1568286	1575973	129.72	359692	369358	5.64	7	3	42	100	99	100	97
NZ_CP007659.1	NC_002951.2	1568286	1576904	131.69	358442	369358	6.97	39	5	87	100	98	100	98
NZ_CP007659.1	NC_002951.2	1568286	1579141	126.45	356047	369358	11.0	3	3	264	97	98	100	98
NZ_CP007659.1	NC_002951.2	1568948	1575973	132.04	359692	369170	5.6	6	1	40	100	99	99	96
NZ_CP007659.1	NC_002951.2	1568948	1576904	133.9	358442	369170	6.95	22	3	85	100	95	100	97
NZ_CP007659.1	NC_002951.2	1568948	1579141	127.84	356047	369170	11.04	4	1	262	100	99	98	97
NZ_CP007659.1	NC_002951.2	1575972	1576904	147.92	358442	359691	17.26	58	2	45	100	98	100	99
NZ_CP007659.1	NC_002951.2	1576903	1579141	106.26	356047	358441	29.37	13	29	177	94	99	100	100
NC_017763.1	NC_002976.3	37130	37175	108.33	2256184	2258869	26.44	25	206	76	92	99	100	99
NZ_CP007659.1	NZ_CP012011.1	1539648	1568954	130.85	1228987	1251487	14.6	4	26	826	100	94	99	93
NZ_CP007659.1	NZ_CP012011.1	1539674	1568954	130.85	1228987	1251477	14.59	10	26	826	100	97	100	97
NC_017763.1	NC_002951.2	1554075	1561787	129.75	359692	369382	5.66	9	3	42	99	98	100	94
NC_017763.1	NC_002951.2	1554075	1562718	131.71	358442	369382	6.99	41	5	87	100	96	100	97
NC_017763.1	NC_002951.2	1554075	1564955	126.48	356047	369382	11.01	5	3	264	97	97	100	97
NC_017763.1	NC_002951.2	1554100	1561787	129.72	359692	369358	5.64	7	3	42	98	99	99	94
NC_017763.1	NC_002951.2	1554100	1562718	131.69	358442	369358	6.97	39	5	87	99	96	100	98
NC_017763.1	NC_002951.2	1554100	1564955	126.45	356047	369358	11.0	3	3	264	99	99	99	99
NC_017763.1	NC_002951.2	1554762	1561787	132.04	359692	369170	5.6	6	1	40	99	100	100	96
NC_017763.1	NC_002951.2	1554762	1562718	133.9	358442	369170	6.95	22	3	85	99	98	99	97
NC_017763.1	NC_002951.2	1554762	1564955	127.84	356047	369170	11.04	4	1	262	98	97	99	97
NC_017763.1	NC_002951.2	1561786	1562718	147.92	358442	359691	17.26	58	2	45	100	97	100	98
NC_017763.1	NC_002951.2	1562717	1564955	106.26	356047	358441	29.37	13	29	177	95	98	100	99
NZ_CP007659.1	NC_002976.3	37131	37176	108.33	2256184	2258869	26.44	25	206	76	93	100	100	100

Table A.8. Acceptor and donor candidates for ERR103403 run with yara, species filter and no samflag filter. Sampling sensitivity = 85. No taxon blacklist. No parent blacklist. No species blacklist. (-)0.000* represents absolute values < 0.000$\bar{4}$.

Candidate			MicrobeGPS metrics			DaisyGPS metrics	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Staphylococcus aureus subsp. aureus HO 5096 0412	NC_017763.1	206493	0.813	0.063	0.750	0.039
Acceptor	Staphylococcus aureus subsp. aureus	NZ_CP007659.1	206231	0.806	0.066	0.74	0.038
Donor	Staphylococcus warneri SG1	NC_020164.1	196	0.003	0.639	-0.636	-0.000*
Donor	Staphylococcus pseudintermedius HKU10-03	NC_014925.1	705	0.001	0.582	-0.581	-0.000*
Donor	Staphylococcus epidermidis RP62A	NC_002976.3	2171	0.005	0.537	-0.532	-0.000*
Donor	Staphylococcus haemolyticus JCSC1435	NC_007168.1	1398	0.005	0.287	-0.283	-0.000*
Donor	Staphylococcus aureus subsp. aureus TW20	NC_017331.1	27837	0.096	0.364	-0.268	-0.002
Acceptor-like Donor	Staphylococcus aureus CA-347	NC_021554.1	31231	0.148	0.146	0.003	0.000*

Table A.9. Results for ERR103403 run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		Acceptor			Donor			Read Evidence			Evidence Filter			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NZ_CP007659.1	NC_021554.1	1568897	1578954	57.77	1567257	1577035	1.38	3	4	35	98	97	100	97
NC_017763.1	NC_017331.1	1525080	1525467	58.54	413136	417103	24.3	11	6	302	96	98	100	99
NC_017763.1	NC_017331.1	1525080	1525489	59.9	413114	417103	24.24	17	8	302	99	100	100	100
NC_017763.1	NC_017331.1	1525080	1559823	62.98	382945	417103	15.53	6	7	1608	99	99	98	100
NC_017763.1	NC_017331.1	1525466	1559823	63.03	382945	413135	14.37	5	18	1306	100	100	99	100
NC_017763.1	NC_017331.1	1525466	1561786	62.78	381925	413135	13.92	25	11	1306	98	97	100	100
NC_017763.1	NC_017331.1	1525488	1559823	63.02	382945	413113	14.37	13	18	1306	99	100	100	100
NC_017763.1	NC_017331.1	1525488	1561786	62.77	381925	413113	13.92	23	11	1306	97	99	99	100
NC_017763.1	NC_014925.1	36951	37132	244.55	906205	906387	96.3	26	10	11	100	100	100	100
NC_017763.1	NC_014925.1	36951	37151	233.42	906205	906409	93.52	20	10	11	96	100	100	100
NC_017763.1	NC_014925.1	37044	37151	163.06	906300	906409	116.82	4	9	11	100	100	100	100
NZ_CP007659.1	NC_014925.1	36952	37133	244.55	906205	906387	96.3	26	10	11	100	100	100	100
NZ_CP007659.1	NC_014925.1	36952	37152	233.42	906205	906409	93.52	20	10	11	100	100	100	100
NZ_CP007659.1	NC_014925.1	37045	37152	163.06	906300	906409	116.82	4	9	11	100	100	100	100
NZ_CP007659.1	NC_017331.1	1539266	1539653	58.54	413136	417103	24.3	11	6	302	98	98	100	99
NZ_CP007659.1	NC_017331.1	1539266	1539675	59.9	413114	417103	24.24	17	8	302	100	100	100	100
NZ_CP007659.1	NC_017331.1	1539266	1574009	62.99	382945	417103	15.53	6	7	1608	100	100	97	100
NZ_CP007659.1	NC_017331.1	1539652	1574009	63.04	382945	413135	14.37	5	18	1306	99	100	100	100
NZ_CP007659.1	NC_017331.1	1539652	1575972	62.79	381925	413135	13.92	25	11	1306	95	98	99	100
NZ_CP007659.1	NC_017331.1	1539674	1574009	63.02	382945	413113	14.37	13	18	1306	98	99	100	100
NZ_CP007659.1	NC_017331.1	1539674	1575972	62.77	381925	413113	13.92	23	11	1306	99	99	99	100
NC_017763.1	NC_021554.1	1554711	1564768	57.77	1567257	1577035	1.38	3	4	35	100	98	99	98

Table A.10. Acceptor and donor candidates for ERR103404 run with yara, species filter and no samflag filter. Sampling sensitivity = 85. No taxon blacklist. No parent blacklist. No species blacklist. (-)0.000* represents absolute values $< 0.000\bar{4}$.

Candidate			MicrobeGPS metrics			DaisyGPS metrics	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Staphylococcus aureus subsp. aureus HO 5096 0412	NC_017763.1	193345	0.812	0.043	0.769	0.041
Acceptor	Staphylococcus aureus subsp. aureus	NZ_CP007659.1	193065	0.805	0.044	0.761	0.041
Donor	Staphylococcus pseudintermedius ED99	NC_017568.1	459	0.001	0.702	-0.700	-0.000*
Donor	Staphylococcus warneri SG1	NC_020164.1	244	0.003	0.631	-0.627	-0.000*
Donor	Staphylococcus epidermidis RP62A	NC_002976.3	2256	0.005	0.536	-0.531	-0.000*
Donor	Staphylococcus haemolyticus JCSC1435	NC_007168.1	1441	0.005	0.299	-0.295	-0.000*
Donor	Staphylococcus aureus subsp. aureus COL	NC_002951.2	20891	0.101	0.233	-0.133	-0.001
Acceptor-like Donor	Staphylococcus aureus subsp. aureus DSM 20231	NZ_CP011526.1	16400	0.102	0.084	0.018	0.000*

Table A.11. Results for ERR103404 run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		Acceptor			Donor			Read Evidence			Evidence Filter			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NC_017763.1	NZ_CP011526.1	1554767	1561786	52.96	846400	854250	11.47	24	1	254	100	100	100	100
NZ_CP007659.1	NZ_CP011526.1	1568953	1575972	52.96	846400	854250	11.47	24	1	254	100	100	100	100
NZ_CP007659.1	NC_002951.2	1568275	1575973	50.88	359692	369368	2.48	7	1	24	98	98	100	95
NZ_CP007659.1	NC_002951.2	1568275	1576904	52.23	358442	369368	2.96	23	2	43	98	98	100	98
NZ_CP007659.1	NC_002951.2	1575972	1576904	63.34	358442	359691	6.72	34	1	19	100	100	100	100
NC_017763.1	NC_002951.2	1554089	1561787	50.88	359692	369368	2.48	7	1	24	98	98	100	94
NC_017763.1	NC_002951.2	1554089	1562718	52.23	358442	369368	2.96	23	2	43	100	98	100	99
NC_017763.1	NC_002951.2	1561786	1562718	63.34	358442	359691	6.72	34	1	19	100	98	100	100
NC_017763.1	NC_002951.2	2045963	2074149	30.15	369125	397269	12.8	12	10	330	6	100	97	100

Table A.12. Acceptor and donor candidates for ERR103405 run with yara, species filter and no samflag filter. Sampling sensitivity = 85. No taxon blacklist. No parent blacklist. No species blacklist. (-)0.000* represents absolute values $< 0.000\bar{4}$.

Candidate			MicrobeGPS metrics			DaisyGPS metrics	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Staphylococcus aureus subsp. aureus HO 5096 0412	NC_017763.1	192851	0.811	0.03	0.781	0.041
Acceptor	Staphylococcus aureus subsp. aureus	NZ_CP007659.1	192626	0.804	0.031	0.773	0.040
Donor	Staphylococcus pseudintermedius ED99	NC_017568.1	459	0.001	0.698	-0.696	-0.000*
Donor	Staphylococcus warneri SG1	NC_020164.1	236	0.003	0.658	-0.655	-0.000*
Donor	Staphylococcus epidermidis RP62A	NC_002976.3	2006	0.005	0.543	-0.538	-0.000*
Donor	Staphylococcus haemolyticus JCSC1435	NC_007168.1	1278	0.005	0.293	-0.289	-0.000*
Donor	Staphylococcus aureus subsp. aureus COL	NC_002951.2	21599	0.097	0.227	-0.13	-0.001
Acceptor-like Donor	Staphylococcus aureus subsp. aureus	NZ_CP018205.1	20618	0.100	0.091	0.009	0.000*

Table A.13. Results for ERR103405 run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		Acceptor			Donor			Read Evidence			Evidence Filter			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NC_017763.1	NZ_CP018205.1	1559883	1562718	58.73	1959491	1961823	7.95	12	1	38	100	100	100	100
NC_017763.1	NZ_CP018205.1	1561784	1562718	66.49	1960572	1961823	9.45	66	1	31	100	100	100	100
NZ_CP007659.1	NZ_CP018205.1	1574069	1576904	58.73	1959491	1961823	7.95	12	1	38	99	99	100	98
NZ_CP007659.1	NZ_CP018205.1	1575970	1576904	66.49	1960572	1961823	9.45	66	1	31	100	100	100	100
NZ_CP007659.1	NC_002951.2	1568261	1576904	56.11	358442	369382	3.13	10	1	50	100	98	100	100
NZ_CP007659.1	NC_002951.2	1575976	1576904	66.66	358442	359692	9.25	19	1	29	100	99	100	99
NZ_CP007659.1	NC_002951.2	2059982	2087935	30.68	369359	397269	12.87	5	12	341	10	100	100	100
NZ_CP007659.1	NC_002951.2	2059982	2088169	30.68	369125	397269	12.79	21	12	341	6	100	98	100
NC_017763.1	NC_002951.2	1554075	1562718	56.11	358442	369382	3.13	10	1	50	100	99	100	100
NC_017763.1	NC_002951.2	1561790	1562718	66.66	358442	359692	9.25	19	1	29	100	100	100	100
NC_017763.1	NC_002951.2	2045963	2073915	30.7	369359	397269	13.02	5	12	355	6	100	99	100
NC_017763.1	NC_002951.2	2045963	2074149	30.7	369125	397269	12.94	21	12	355	8	100	97	100

Table A.14. Acceptor and donor candidates for ERR101899 run with yara, species filter and no samflag filter. Sampling sensitivity = 85. No taxon blacklist. No parent blacklist. No species blacklist. (-)0.000* represents absolute values $< 0.000\bar{4}$.

Candidate			MicrobeGPS metrics			DaisyGPS metrics	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Staphylococcus aureus subsp. aureus HO 5096 0412	NC_017763.1	206272	0.814	0.047	0.767	0.040
Acceptor	Staphylococcus aureus subsp. aureus	NZ_CP007659.1	206076	0.807	0.049	0.759	0.04
Donor	Staphylococcus pseudintermedius ED99	NC_017568.1	536	0.001	0.707	-0.705	-0.000*
Donor	Staphylococcus warneri SG1	NC_020164.1	263	0.003	0.658	-0.655	-0.000*
Donor	Staphylococcus epidermidis RP62A	NC_002976.3	2226	0.005	0.537	-0.532	-0.000*
Donor	Staphylococcus haemolyticus JCSC1435	NC_007168.1	1378	0.004	0.296	-0.291	-0.000*
Donor	Staphylococcus aureus subsp. aureus COL	NC_002951.2	22973	0.098	0.236	-0.139	-0.001
Acceptor-like Donor	Staphylococcus aureus subsp. aureus DSM 20231	NZ_CP011526.1	18223	0.099	0.085	0.014	0.000*

Table A.15. Results for ERR101899 run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		Acceptor			Donor			Read Evidence			Evidence Filter			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NZ_CP007659.1	NC_002951.2	1568261	1575972	53.51	359694	369382	2.62	3	2	15	99	100	100	97
NZ_CP007659.1	NC_002951.2	1568261	1576904	55.07	358442	369382	3.05	8	2	34	99	99	99	99
NZ_CP007659.1	NC_002951.2	1568287	1575972	53.49	359694	369357	2.61	3	2	15	98	100	99	99
NZ_CP007659.1	NC_002951.2	1568287	1576904	55.06	358442	369357	3.04	8	2	34	100	100	99	99
NZ_CP007659.1	NC_002951.2	2059982	2087936	31.07	369358	397269	13.23	9	15	395	10	100	100	100
NZ_CP007659.1	NC_002951.2	2059982	2088169	31.08	369125	397269	13.15	31	15	396	7	100	94	100
NZ_CP007659.1	NC_020164.1	37045	37177	87.31	111789	121379	0.85	4	2	20	100	97	100	100
NZ_CP007659.1	NZ_CP011526.1	1568903	1575972	56.42	846397	854374	12.62	9	1	267	100	100	100	100
NC_017763.1	NZ_CP011526.1	1554717	1561786	56.42	846397	854374	12.62	9	1	267	98	100	100	100
NC_017763.1	NC_002951.2	1554075	1561786	53.51	359694	369382	2.62	3	2	15	97	98	100	94
NC_017763.1	NC_002951.2	1554075	1562718	55.07	358442	369382	3.05	8	2	34	99	99	100	100
NC_017763.1	NC_002951.2	1554101	1561786	53.49	359694	369357	2.61	3	2	15	98	99	99	93
NC_017763.1	NC_002951.2	1554101	1562718	55.06	358442	369357	3.04	8	2	34	99	98	100	97
NC_017763.1	NC_002951.2	2045963	2073916	31.1	369358	397269	13.45	9	15	415	8	100	91	100
NC_017763.1	NC_020164.1	37044	37176	87.31	111789	121379	0.85	4	2	20	100	100	99	100

Table A.16. Acceptor and donor candidates for ERR101900 run with yara, species filter and no samflag filter. Sampling sensitivity = 85. No taxon blacklist. No parent blacklist. No species blacklist. (-)0.000* represents absolute values $< 0.000\bar{4}$.

Candidate			MicrobeGPS metrics			DaisyGPS metrics	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Staphylococcus aureus subsp. aureus HO 5096 0412	NC_017763.1	162488	0.801	0.049	0.752	0.04
Acceptor	Staphylococcus aureus subsp. aureus	NZ_CP007659.1	162328	0.794	0.050	0.744	0.039
Donor	Staphylococcus pseudintermedius ED99	NC_017568.1	1521	0.002	0.706	-0.704	-0.000*
Donor	Staphylococcus warneri SG1	NC_020164.1	215	0.004	0.654	-0.650	-0.000*
Donor	Staphylococcus epidermidis RP62A	NC_002976.3	3028	0.005	0.560	-0.555	-0.001
Donor	Staphylococcus lugdunensis HKU09-01	NC_013893.1	53	0.002	0.358	-0.356	-0.000*
Donor	Staphylococcus haemolyticus JCSC1435	NC_007168.1	1116	0.005	0.254	-0.25	-0.000*
Donor	Staphylococcus aureus subsp. aureus COL	NC_002951.2	17868	0.103	0.242	-0.139	-0.001
Acceptor-like Donor	Staphylococcus aureus subsp. aureus NCTC 8325	NC_007795.1	16873	0.107	0.089	0.018	0.000*

Table A.17. Results for ERR101900 run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		Acceptor			Donor			Read Evidence			Evidence Filter			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NC_017763.1	NC_002951.2	1554089	1562718	45.31	358442	369368	2.81	15	1	31	100	100	99	98
NC_017763.1	NC_002951.2	1554762	1562718	46.77	358442	369170	2.78	8	1	29	100	100	98	98
NC_017763.1	NC_002951.2	1561790	1562718	53.62	358442	359696	5.56	8	1	15	98	99	100	100
NZ_CP007659.1	NC_007795.1	1575971	1576904	53.59	1961777	1963027	6.06	57	1	16	99	99	100	99
NC_017763.1	NC_007795.1	1561785	1562718	53.59	1961777	1963027	6.06	57	1	16	99	97	99	97
NZ_CP007659.1	NC_002951.2	1568275	1576904	45.31	358442	369368	2.81	15	1	31	99	99	100	99
NZ_CP007659.1	NC_002951.2	1568948	1576904	46.77	358442	369170	2.78	8	1	29	99	99	100	99
NZ_CP007659.1	NC_002951.2	1575976	1576904	53.62	358442	359696	5.56	8	1	15	100	99	100	98

Table A.18. Acceptor and donor candidates for ERR103394 run with yara, species filter and no samflag filter. Sampling sensitivity = 85. No taxon blacklist. No parent blacklist. No species blacklist. (-)0.000* represents absolute values $< 0.000\bar{4}$.

Candidate			MicrobeGPS metrics			DaisyGPS metrics	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Staphylococcus aureus subsp. aureus HO 5096 0412	NC_017763.1	183503	0.807	0.048	0.759	0.040
Acceptor	Staphylococcus aureus subsp. aureus	NZ_CP007659.1	183292	0.801	0.05	0.751	0.04
Donor	Staphylococcus warneri SG1	NC_020164.1	250	0.004	0.656	-0.653	-0.000*
Donor	Staphylococcus pseudintermedius HKU10-03	NC_014925.1	747	0.001	0.584	-0.582	-0.000*
Donor	Staphylococcus epidermidis RP62A	NC_002976.3	2358	0.005	0.546	-0.541	-0.000*
Donor	Staphylococcus haemolyticus JCSC1435	NC_007168.1	1541	0.005	0.301	-0.296	-0.000*
Donor	Staphylococcus aureus subsp. aureus COL	NC_002951.2	20650	0.100	0.246	-0.146	-0.001
Acceptor-like Donor	Staphylococcus aureus subsp. aureus DSM 20231	NZ_CP011526.1	16141	0.102	0.091	0.011	0.000*

Table A.19. Results for ERR103394 run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		Acceptor			Donor			Read Evidence			Evidence Filter			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NZ_CP007659.1	NC_014925.1	36953	37046	319.63	906200	906301	89.36	6	17	18	100	100	100	100
NZ_CP007659.1	NC_014925.1	36953	37133	263.84	906200	906387	137.22	13	20	20	100	100	100	100
NZ_CP007659.1	NC_014925.1	36953	37152	254.4	906200	906409	135.31	9	19	20	100	100	100	100
NZ_CP007659.1	NC_014925.1	36999	37133	225.07	906256	906387	169.08	11	18	20	100	100	100	100
NZ_CP007659.1	NC_014925.1	36999	37152	217.61	906256	906409	161.88	7	18	20	100	100	100	100
NZ_CP007659.1	NC_014925.1	37045	37152	197.65	906300	906409	178.28	5	12	19	100	100	100	100
NC_017763.1	NC_002951.2	1554089	1562718	54.38	358442	369368	2.46	16	3	25	99	98	100	95
NC_017763.1	NC_002951.2	1554762	1562718	55.43	358442	369170	2.46	29	3	25	99	99	100	97
NC_017763.1	NC_014925.1	36952	37045	299.67	906200	906301	76.57	7	15	19	100	100	100	100
NC_017763.1	NC_014925.1	36952	37132	228.17	906200	906387	109.47	21	17	21	100	100	100	100
NC_017763.1	NC_014925.1	36952	37151	217.36	906200	906409	105.63	15	16	21	100	100	100	100
NC_017763.1	NC_014925.1	36998	37132	183.81	906256	906387	134.34	11	16	21	99	100	100	100
NC_017763.1	NC_014925.1	36998	37151	175.26	906256	906409	125.52	8	16	21	100	100	100	100
NC_017763.1	NC_014925.1	37044	37151	145.74	906300	906409	132.93	6	10	20	100	100	100	100
NZ_CP007659.1	NZ_CP011526.1	1568903	1575973	55.17	846399	854374	10.79	13	3	234	97	99	100	99
NZ_CP007659.1	NZ_CP011526.1	1568903	1579178	54.74	842251	854374	19.62	5	3	747	98	100	99	100
NZ_CP007659.1	NZ_CP011526.1	1568953	1575973	55.21	846399	854250	10.95	26	3	234	99	98	100	98
NZ_CP007659.1	NZ_CP011526.1	1568953	1579178	54.76	842251	854250	19.82	10	3	747	100	100	100	100
NC_017763.1	NZ_CP011526.1	1554717	1561787	55.17	846399	854374	10.79	13	3	234	99	100	100	100
NC_017763.1	NZ_CP011526.1	1554717	1564992	54.74	842251	854374	19.62	5	3	747	99	100	100	100
NC_017763.1	NZ_CP011526.1	1554767	1561787	55.21	846399	854250	10.95	26	3	234	99	100	100	100
NC_017763.1	NZ_CP011526.1	1554767	1564992	54.76	842251	854250	19.82	10	3	747	97	100	100	100
NZ_CP007659.1	NC_002951.2	1568275	1576904	54.38	358442	369368	2.46	16	3	25	99	98	100	96
NZ_CP007659.1	NC_002951.2	1568948	1576904	55.43	358442	369170	2.46	29	3	25	99	99	100	97

Table A.20. Acceptor and donor candidates for ERR103395 run with yara, species filter and no samflag filter. Sampling sensitivity = 85. No taxon blacklist. No parent blacklist. No species blacklist. (-)0.000* represents absolute values $< 0.000\bar{4}$.

Candidate			MicrobeGPS metrics			DaisyGPS metrics	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Staphylococcus aureus subsp. aureus ECT-R 2	NC_017343.1	120322	0.591	0.070	0.521	0.013
Acceptor	Staphylococcus aureus subsp. aureus N315	NC_002745.2	121110	0.576	0.069	0.507	0.013
Donor	Enterococcus faecium Aus0004	NC_017022.1	471	0.001	0.974	-0.973	-0.000*
Donor	Staphylococcus epidermidis ATCC 12228	NC_004461.1	391	0.001	0.971	-0.97	-0.000*
Donor	Staphylococcus pseudintermedius HKU10-03	NC_014925.1	470	0.001	0.806	-0.805	-0.000*
Donor	Staphylococcus lugdunensis HKU09-01	NC_013893.1	59	0.003	0.765	-0.762	-0.000*
Donor	Staphylococcus warneri SG1	NC_020164.1	294	0.011	0.693	-0.683	-0.000*
Donor	Staphylococcus haemolyticus JCSC1435	NC_007168.1	362	0.002	0.556	-0.554	-0.000*
Acceptor-like Donor	Staphylococcus aureus subsp. aureus	NZ_CP009554.1	14824	0.093	0.091	0.002	0.000*

Table A.21. Results for ERR103395 run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		Acceptor			Donor			Read Evidence			Evidence Filter			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NC_002745.2	NC_013893.1	2060607	2069048	16.44	2073055	2083555	11.08	28	8	339	1	100	100	100
NC_002745.2	NC_013893.1	2060607	2069067	16.47	2073055	2083576	11.07	18	6	339	2	100	99	100
NC_002745.2	NC_013893.1	2060762	2069048	16.74	2073192	2083555	11.11	7	8	339	1	100	100	100
NC_002745.2	NZ_CP009554.1	1142176	1142913	0.26	685582	686374	22.86	8	11	52	3	100	100	100
NC_002745.2	NZ_CP009554.1	1142176	1142913	0.26	685582	717267	0.66	12	14	53	5	94	97	94
NC_002745.2	NZ_CP009554.1	1142912	1142913	1.0	685581	716475	0.66	11	12	53	4	93	98	94
NC_002745.2	NZ_CP009554.1	2056699	2058174	0.02	2150234	2162636	2.65	10	6	43	2	98	99	94
NC_002745.2	NZ_CP009554.1	2056699	2060475	5.31	2150276	2162636	2.66	13	6	43	1	100	99	99
NC_002745.2	NZ_CP009554.1	2056699	2069076	10.31	2158985	2162636	3.14	24	34	8	2	99	100	90
NC_002745.2	NZ_CP009554.1	2058173	2069076	11.7	2150233	2158985	2.44	51	5	34	0	97	100	97
NC_002745.2	NZ_CP009554.1	2058173	2069105	11.7	2150233	2159011	2.48	7	5	34	3	100	100	98
NC_002745.2	NZ_CP009554.1	2058173	2069324	11.5	2150233	2159202	3.5	27	26	43	0	100	100	99
NC_002745.2	NZ_CP009554.1	2058173	2069355	11.57	2150233	2159253	3.62	7	26	43	0	100	100	96
NC_002745.2	NZ_CP009554.1	2060474	2069076	12.5	2150275	2158985	2.45	52	5	34	0	99	100	99
NC_002745.2	NZ_CP009554.1	2060474	2069105	12.5	2150275	2159011	2.49	8	5	34	2	100	100	98
NC_002745.2	NZ_CP009554.1	2060474	2069324	12.23	2150275	2159202	3.51	28	26	43	5	98	100	97
NC_002745.2	NZ_CP009554.1	2060474	2069355	12.31	2150275	2159253	3.63	8	26	43	1	100	100	99
NC_002745.2	NZ_CP009554.1	2060607	2065052	19.41	364874	369569	10.13	14	5	133	2	100	100	100
NC_002745.2	NZ_CP009554.1	2060607	2068738	12.04	361186	369569	7.19	18	1	154	0	100	100	100
NC_002745.2	NZ_CP009554.1	2065051	2068738	3.16	361186	364873	3.44	9	3	21	3	98	99	97
NC_002745.2	NZ_CP009554.1	2069075	2069324	2.76	2158984	2159202	45.94	89	32	8	4	100	100	100
NC_002745.2	NZ_CP009554.1	2069075	2069355	6.49	2158984	2159253	41.82	9	38	8	2	100	100	99
NC_002745.2	NZ_CP009554.1	2069104	2069324	1.65	2159010	2159202	50.14	12	32	8	3	100	100	99
NC_017343.1	NZ_CP009554.1	1100697	1101434	0.42	685582	686374	22.86	8	11	52	1	100	100	100
NC_017343.1	NZ_CP009554.1	1100697	1101434	0.42	685582	717267	0.66	15	14	53	1	99	99	98
NC_017343.1	NZ_CP009554.1	1101433	1101434	1.0	685581	716475	0.66	11	12	53	0	98	96	96
NC_002745.2	NC_004461.1	61651	61779	4.23	37793	55322	7.62	30	73	392	4	100	99	100
NC_002745.2	NC_004461.1	61651	61799	3.8	37814	55322	7.62	14	73	392	4	100	100	100
NC_002745.2	NC_004461.1	61651	61851	2.83	37866	55322	7.61	18	73	392	2	100	99	100
NC_002745.2	NC_004461.1	61755	61779	3.04	37793	55383	7.59	12	73	392	2	100	100	100
NC_002745.2	NC_004461.1	61755	61799	2.14	37814	55383	7.59	8	73	392	5	100	99	100
NC_002745.2	NC_004461.1	61755	61851	1.01	37866	55383	7.58	9	73	392	3	100	100	100
NC_002745.2	NC_004461.1	61778	62058	2.57	37792	57274	6.86	7	73	392	1	100	100	100
NC_002745.2	NC_004461.1	61778	62354	7.14	37792	57575	6.75	7	68	392	2	100	100	100
NC_002745.2	NC_004461.1	61778	62414	7.02	37792	57608	6.74	8	64	392	1	100	99	100
NC_002745.2	NC_004461.1	61798	62058	2.68	37813	57274	6.85	3	73	392	5	100	100	100
NC_002745.2	NC_004461.1	61798	62354	7.35	37813	57575	6.75	3	68	392	1	100	99	100
NC_002745.2	NC_004461.1	61798	62414	7.21	37813	57608	6.74	4	64	392	1	100	100	100
NC_002745.2	NC_004461.1	61850	62058	3.34	37865	57274	6.84	4	73	392	5	100	99	100
NC_002745.2	NC_004461.1	61850	62354	8.11	37865	57575	6.74	4	68	392	4	100	100	100
NC_002745.2	NC_004461.1	61850	62414	7.87	37865	57608	6.73	7	64	392	2	100	100	100

Table A.22. Acceptor and donor candidates for ERR103396 run with yara, species filter and no samflag filter. Sampling sensitivity = 85. No taxon blacklist. No parent blacklist. No species blacklist. (-)0.000* represents absolute values $< 0.000\bar{4}$.

Candidate			MicrobeGPS metrics			DaisyGPS metrics	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Staphylococcus aureus subsp. aureus HO 5096 0412	NC_017763.1	222016	0.817	0.042	0.775	0.043
Acceptor	Staphylococcus aureus subsp. aureus	NZ_CP007659.1	223952	0.815	0.049	0.767	0.043
Donor	Staphylococcus pseudintermedius ED99	NC_017568.1	536	0.002	0.708	-0.707	-0.000*
Donor	Staphylococcus warneri SG1	NC_020164.1	267	0.003	0.696	-0.693	-0.000*
Donor	Staphylococcus epidermidis RP62A	NC_002976.3	1067	0.003	0.582	-0.579	-0.000*
Donor	Staphylococcus haemolyticus JCSC1435	NC_007168.1	370	0.003	0.492	-0.489	-0.000*
Donor	Staphylococcus aureus subsp. aureus COL	NC_002951.2	21752	0.098	0.156	-0.058	-0.000*
Acceptor-like Donor	Staphylococcus aureus subsp. aureus	NZ_CP012012.1	21332	0.097	0.094	0.003	0.000*

Table A.23. Results for ERR103396 run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		Acceptor			Donor			Read Evidence			Evidence Filter			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NC_017763.1	NZ_CP012012.1	98589	98635	95.67	125862	126004	35.02	3	20	5	100	100	100	100
NC_017763.1	NC_017568.1	409730	409775	16.98	2481624	2485653	3.95	14	5	35	1	100	100	100

Table A.24. Acceptor and donor candidates for ERR103397 run with yara, species filter and no samflag filter. Sampling sensitivity = 85. No taxon blacklist. No parent blacklist. No species blacklist. (-)0.000* represents absolute values $< 0.000\bar{4}$.

Candidate			MicrobeGPS metrics			DaisyGPS metrics	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Staphylococcus aureus subsp. aureus MSSA476	NC_002953.3	84971	0.634	0.094	0.540	0.017
Acceptor	Staphylococcus aureus subsp. aureus MW2	NC_003923.1	83556	0.621	0.089	0.531	0.017
Donor	Staphylococcus pseudintermedius HKU10-03	NC_014925.1	3645	0.002	0.744	-0.742	-0.001
Donor	Staphylococcus warneri SG1	NC_020164.1	168	0.003	0.69	-0.697	-0.000*
Donor	Staphylococcus haemolyticus JCSC1435	NC_007168.1	2650	0.004	0.604	-0.600	-0.001
Donor	Staphylococcus epidermidis RP62A	NC_002976.3	1082	0.002	0.583	-0.581	-0.000*
Donor	Staphylococcus lugdunensis HKU09-01	NC_013893.1	3709	0.004	0.356	-0.352	-0.001
Donor	Staphylococcus aureus subsp. aureus	NZ_CP009554.1	19819	0.092	0.314	-0.222	-0.002
Acceptor-like Donor	Staphylococcus aureus subsp. aureus	NZ_CP009361.1	9253	0.097	0.092	0.005	0.000*

Table A.25. Results for ERR103397 run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		Acceptor			Donor			Read Evidence			Evidence Filter			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NC_003923.1	NC_007168.1	44986	45306	40.24	66689	67028	5.47	6	1	2	98	100	100	100
NC_003923.1	NC_002976.3	44776	44988	12.98	2520640	2520803	10.01	4	4	1	6	100	100	100
NC_003923.1	NC_002976.3	44987	45380	36.37	2520639	2561294	0.64	14	58	46	96	95	98	95
NC_003923.1	NC_002976.3	44987	45606	31.5	2520639	2561094	0.63	4	58	44	95	96	100	97
NC_003923.1	NC_002976.3	45026	45380	37.73	2561294	2561636	10.28	14	3	3	100	99	100	99
NC_003923.1	NC_002976.3	45026	45606	32.0	2561094	2561636	7.1	4	3	4	92	98	100	98
NC_003923.1	NC_002976.3	45026	45870	27.86	2560793	2561636	6.17	6	4	4	91	99	100	100
NC_003923.1	NC_002976.3	45070	45307	38.1	2561337	2561580	12.23	4	5	3	94	100	100	100
NC_003923.1	NC_002976.3	45070	45380	38.28	2561294	2561580	12.12	40	5	3	98	100	100	100
NC_002953.3	NC_007168.1	41508	57483	26.01	67036	120082	2.7	20	4	471	94	98	95	98

Table A.26. Acceptor and donor candidates for ERR103398 run with yara, species filter and no samflag filter. Sampling sensitivity = 85. No taxon blacklist. No parent blacklist. No species blacklist. (-)0.000* represents absolute values $< 0.000\bar{4}$.

Candidate			MicrobeGPS metrics			DaisyGPS metrics	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Staphylococcus aureus subsp. aureus MSSA476	NC_002953.3	192949	0.671	0.11	0.562	0.017
Acceptor	Staphylococcus aureus subsp. aureus MW2	NC_003923.1	189418	0.658	0.103	0.555	0.016
Donor	Staphylococcus pseudintermedius HKU10-03	NC_014925.1	16866	0.002	0.745	-0.742	-0.002
Donor	Staphylococcus warneri SG1	NC_020164.1	461	0.003	0.69	-0.697	-0.000*
Donor	Staphylococcus epidermidis PM221	NZ_HG813242.1	4779	0.001	0.656	-0.655	-0.001
Donor	Staphylococcus haemolyticus JCSC1435	NC_007168.1	12023	0.004	0.636	-0.632	-0.001
Donor	Staphylococcus lugdunensis HKU09-01	NC_013893.1	16800	0.004	0.356	-0.351	-0.001
Donor	Staphylococcus aureus subsp. aureus	NZ_CP009554.1	70966	0.095	0.398	-0.304	-0.003
Acceptor-like Donor	Staphylococcus aureus CA-347	NC_021554.1	18666	0.098	0.090	0.007	0.000*

Table A.27. Results for ERR103398 run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		Acceptor			Donor			Read Evidence			Evidence Filter			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NC_002953.3	NC_007168.1	27843	41291	63.48	67243	97433	4.35	18	2	436	95	97	99	99
NC_002953.3	NC_007168.1	27843	41907	63.94	66699	97433	4.54	3	4	446	91	99	99	99
NC_002953.3	NC_007168.1	27843	57484	69.06	94688	97433	47.35	7	2	431	99	100	100	100
NC_002953.3	NC_007168.1	41290	41907	73.94	66699	67242	15.3	9	6	14	96	100	100	100
NC_002953.3	NC_007168.1	41290	57483	73.69	67242	120082	6.24	77	6	1105	99	94	99	97
NC_002953.3	NC_007168.1	41290	57484	73.69	30072	67242	0.28	25	7	16	100	93	100	97
NC_002953.3	NC_007168.1	41508	41605	95.51	66925	67036	6.53	17	6	1	97	100	100	100
NC_002953.3	NC_007168.1	41508	41907	90.78	66699	67036	7.76	5	1	2	98	100	100	100
NC_002953.3	NC_007168.1	41508	57483	74.11	67036	120082	6.32	39	10	1111	98	91	94	95
NC_002953.3	NC_007168.1	41508	57484	74.11	67036	94688	0.25	13	8	11	100	93	97	95
NC_002953.3	NC_007168.1	41604	41907	89.3	66699	66924	8.38	9	2	1	97	99	100	99
NC_002953.3	NC_007168.1	41604	57483	73.98	66924	120082	6.32	77	15	1112	100	92	99	94
NC_002953.3	NC_007168.1	41906	57483	73.68	66698	120082	6.33	21	15	1115	100	91	100	94
NC_002953.3	NC_007168.1	41906	57484	73.68	66698	94688	0.34	8	13	15	100	92	100	96
NC_002953.3	NZ_HG813242.1	34149	41829	68.51	35795	85587	4.98	78	19	287	99	99	97	99
NC_002953.3	NZ_HG813242.1	34149	41907	68.57	35721	85587	4.97	50	19	287	95	97	98	97
NC_002953.3	NZ_HG813242.1	34149	57484	71.98	54292	85587	6.1	34	3	280	97	97	97	97
NC_002953.3	NZ_HG813242.1	34149	57484	71.98	57395	85587	5.91	42	2	207	100	97	100	97
NC_002953.3	NZ_HG813242.1	34180	41829	68.6	35795	85647	4.97	240	19	287	93	96	97	96
NC_002953.3	NZ_HG813242.1	34180	41907	68.66	35721	85647	4.96	142	19	287	97	95	96	95
NC_002953.3	NZ_HG813242.1	34180	57484	72.02	54292	85647	6.09	86	3	280	100	96	100	96
NC_002953.3	NZ_HG813242.1	34180	57484	72.02	57395	85647	5.9	114	2	207	100	97	100	97
NC_002953.3	NZ_HG813242.1	41828	56986	72.4	35794	85689	4.98	81	19	287	99	94	97	94
NC_002953.3	NZ_HG813242.1	41828	57484	73.68	35794	57395	3.75	43	2	70	99	94	100	93
NC_002953.3	NZ_HG813242.1	41828	57484	73.68	35794	62757	9.18	15	3	286	100	97	97	97
NC_002953.3	NZ_HG813242.1	41906	56986	72.39	35720	85689	4.97	143	19	287	100	93	98	93
NC_002953.3	NZ_HG813242.1	41906	57484	73.68	35720	57395	3.74	67	2	70	100	98	98	98
NC_002953.3	NZ_HG813242.1	41906	57484	73.68	35720	62757	9.16	11	3	286	100	99	98	99
NC_002953.3	NZ_HG813242.1	56985	57484	112.78	54292	85688	6.1	64	3	280	100	97	98	97
NC_002953.3	NZ_HG813242.1	56985	57484	112.78	57395	85688	5.91	80	2	207	100	98	99	98
NC_003923.1	NC_007168.1	44606	45306	69.87	66689	67411	10.45	29	2	16	95	100	100	100
NC_003923.1	NC_007168.1	45026	45143	97.32	66870	66987	2.7	9	4	2	100	99	100	99
NC_003923.1	NC_007168.1	45026	45306	93.51	66689	66987	17.21	25	6	13	98	100	100	100
NC_003923.1	NC_007168.1	45062	45306	93.32	66689	66930	20.85	19	3	11	99	99	100	100
NC_003923.1	NC_007168.1	45082	45306	92.39	66689	66929	20.93	10	4	11	98	100	100	100
NC_003923.1	NC_007168.1	45142	45306	90.77	66689	66869	26.71	10	5	11	97	99	100	99
NC_002953.3	NC_021554.1	41508	41593	91.24	60998	61108	10.47	19	2	5	96	99	100	99
NC_002953.3	NC_021554.1	41508	41884	81.12	60998	61399	20.48	7	1	11	95	99	100	99
NC_002953.3	NC_021554.1	41548	41884	80.93	61045	61399	21.66	4	2	9	96	100	100	100
NC_002953.3	NC_021554.1	41592	41884	78.23	61107	61399	24.23	22	3	8	92	99	100	99
NC_003923.1	NC_021554.1	44986	45384	76.26	61006	61391	27.57	55	2	20	90	100	100	100
NC_003923.1	NC_021554.1	45026	45306	80.71	61045	61347	29.43	7	3	18	95	100	100	100
NC_003923.1	NC_021554.1	45026	45384	76.52	61045	61391	29.23	113	3	18	94	100	100	100
NC_003923.1	NC_021554.1	45062	45306	78.63	61102	61347	35.75	3	1	14	92	100	100	99
NC_003923.1	NC_021554.1	45062	45384	74.48	61102	61391	34.55	109	1	14	92	100	100	100
NC_003923.1	NC_021554.1	45082	45384	72.6	61105	61391	34.9	55	1	14	91	100	100	100

Table A.28. Acceptor and donor candidates for ERR159680 run with yara, species filter and no samflag filter. Sampling sensitivity = 85. No taxon blacklist. No parent blacklist. No species blacklist. (-)0.000* represents absolute values $< 0.000\bar{4}$.

Candidate			MicrobeGPS metrics			DaisyGPS metrics	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Staphylococcus aureus subsp. aureus MRSA252	NC_002952.2	236631	0.892	0.047	0.845	0.043
Acceptor	Staphylococcus aureus subsp. aureus	NZ_CP009554.1	227305	0.871	0.046	0.825	0.041
Donor	Staphylococcus pseudintermedius ED99	NC_017568.1	780	0.003	0.946	-0.944	-0.000*
Donor	Streptococcus pasteurianus ATCC 43144	NC_015600.1	397	0.001	0.828	-0.827	-0.000*
Donor	Staphylococcus epidermidis RP62A	NC_002976.3	6553	0.019	0.804	-0.785	-0.001
Donor	Streptococcus gallolyticus UCN34	NC_013798.1	453	0.001	0.752	-0.751	-0.000*
Donor	Staphylococcus haemolyticus JCSC1435	NC_007168.1	1295	0.005	0.516	-0.511	-0.000*
Donor	Staphylococcus lugdunensis HKU09-01	NC_013893.1	494	0.003	0.356	-0.353	-0.000*
Acceptor-like Donor	Staphylococcus aureus subsp. aureus	NZ_AP014652.1	17647	0.096	0.085	0.011	0.000*

Table A.29. Results for ERR159680 run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		**Acceptor**			**Donor**			**Read Evidence**			**Evidence Filter**			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NZ_CP009554.1	NC_002976.3	34120	34123	12.67	2536574	2584194	38.98	30	19	5752	6	100	98	100
NZ_CP009554.1	NC_002976.3	859613	866305	63.76	1398260	1404973	0.3	27	4	5	100	98	100	98
NZ_CP009554.1	NC_013893.1	2130925	2133716	7.9	2343346	2345047	5.05	17	1	10	3	100	100	100
NZ_CP009554.1	NC_013893.1	2131388	2133716	4.04	2343670	2345047	6.23	16	1	10	0	100	100	100
NC_002952.2	NC_002976.3	906791	906792	12.0	1398259	1404972	0.3	18	4	5	1	94	100	97
NZ_CP009554.1	NZ_AP014652.1	414814	417301	46.9	438237	438358	11.31	4	2	5	96	99	100	99
NZ_CP009554.1	NZ_AP014652.1	2110903	2123964	16.67	2007772	2020977	17.83	6	4	762	0	98	99	98
NZ_CP009554.1	NZ_AP014652.1	2110903	2131317	11.83	2007772	2029781	21.15	3	4	1493	0	100	97	100
NZ_CP009554.1	NZ_AP014652.1	2110903	2134197	10.62	2007772	2030266	20.77	25	3	1493	0	99	99	99
NZ_CP009554.1	NZ_AP014652.1	2123963	2131021	1.86	2020976	2029466	27.01	5	1	729	0	98	100	98
NZ_CP009554.1	NZ_AP014652.1	2123963	2131317	3.23	2020976	2029781	26.14	7	2	731	1	98	100	98
NZ_CP009554.1	NZ_AP014652.1	2123963	2134197	2.91	2020976	2030266	24.94	51	1	731	0	100	99	100
NZ_CP009554.1	NZ_AP014652.1	2125253	2131317	1.84	2022297	2029781	30.59	3	3	731	0	98	100	98
NZ_CP009554.1	NZ_AP014652.1	2125253	2134197	1.92	2022297	2030266	28.92	25	2	731	0	99	100	99
NZ_CP009554.1	NZ_AP014652.1	2131020	2134197	5.24	2029465	2030266	2.97	25	1	2	2	97	100	97
NZ_CP009554.1	NZ_AP014652.1	2131316	2134197	2.09	2029780	2030266	3.19	49	6	2	1	97	100	96
NC_002952.2	NC_013893.1	413772	417366	53.37	2079996	2083590	0.78	5	4	2	100	99	100	100

Table A.30. Acceptor and donor candidates for ERR103400 run with yara, species filter and no samflag filter. Sampling sensitivity = 85. No taxon blacklist. No parent blacklist. No species blacklist. (-)0.000* represents absolute values $< 0.000\bar{4}$.

Candidate			**MicrobeGPS metrics**			**DaisyGPS metrics**	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Staphylococcus aureus subsp. aureus HO 5096 0412	NC_017763.1	484936	0.835	0.037	0.798	0.041
Acceptor	Staphylococcus aureus subsp. aureus	NZ_CP007659.1	489699	0.832	0.048	0.784	0.041
Donor	Staphylococcus haemolyticus JCSC1435	NC_007168.1	3222	0.006	0.799	-0.792	-0.000*
Donor	Staphylococcus pseudintermedius ED99	NC_017568.1	1398	0.002	0.701	-0.699	-0.000*
Donor	Staphylococcus warneri SG1	NC_020164.1	583	0.003	0.695	-0.692	-0.000*
Donor	Staphylococcus epidermidis ATCC 12228	NC_004461.1	3245	0.005	0.483	-0.479	-0.000*
Donor	Staphylococcus lugdunensis HKU09-01	NC_013893.1	69	0.005	0.342	-0.337	-0.000*
Donor	Staphylococcus aureus subsp. aureus	NZ_CP009554.1	132861	0.21	0.254	-0.044	-0.001
Acceptor-like Donor	Staphylococcus aureus subsp. aureus T0131	NC_017347.1	50347	0.104	0.103	0.001	0.000*

Table A.31. Results for ERR103400 run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		Acceptor			Donor			Read Evidence			Evidence Filter			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NZ_CP007659.1	NC_017347.1	36952	63749	105.32	2780055	2782476	52.24	24	43	358	92	99	100	99
NC_017763.1	NC_007168.1	44772	58518	112.86	67396	74115	52.21	29	1	1039	96	98	100	98
NC_017763.1	NC_007168.1	44772	58518	112.86	67396	74141	52.36	9	1	1045	94	99	100	99
NC_017763.1	NC_007168.1	44772	58661	113.09	67396	73961	51.35	79	1	1007	94	99	100	100
NC_017763.1	NC_007168.1	44772	58729	113.16	67396	73859	51.65	49	1	991	94	99	100	100
NC_017763.1	NC_007168.1	44772	58751	113.17	67396	73849	51.68	9	1	991	97	100	100	100
NC_017763.1	NC_007168.1	44772	63969	109.85	67396	68656	13.93	9	1	31	96	99	100	99
NC_017763.1	NC_007168.1	45010	58518	113.61	67122	74115	50.25	41	1	1040	96	99	100	99
NC_017763.1	NC_007168.1	45010	58518	113.61	67122	74141	50.39	13	1	1046	97	100	100	100
NC_017763.1	NC_007168.1	45010	58661	113.84	67122	73961	49.37	111	1	1008	96	99	100	100
NC_017763.1	NC_007168.1	45010	58729	113.91	67122	73859	49.63	69	1	992	95	99	100	100
NC_017763.1	NC_007168.1	45010	58751	113.91	67122	73849	49.66	13	1	992	97	100	100	100
NC_017763.1	NC_007168.1	45010	63969	110.35	67122	68656	11.81	13	1	32	92	99	100	99
NC_017763.1	NC_007168.1	45149	45440	132.36	66689	67061	16.42	94	4	5	95	100	100	100
NC_017763.1	NC_007168.1	45149	58518	114.17	67061	74115	49.88	22	1	1040	97	99	100	99
NC_017763.1	NC_007168.1	45149	58518	114.17	67061	74141	50.02	10	1	1046	98	100	100	100
NC_017763.1	NC_007168.1	45149	58661	114.4	67061	73961	49.0	52	1	1008	97	99	100	99
NC_017763.1	NC_007168.1	45149	58729	114.46	67061	73859	49.25	34	1	992	93	98	99	99
NC_017763.1	NC_007168.1	45149	58751	114.47	67061	73849	49.29	10	1	992	98	100	100	100
NC_017763.1	NC_007168.1	45149	63969	110.72	67061	68656	11.64	10	1	32	91	99	100	99
NC_017763.1	NC_007168.1	45439	58518	113.77	66688	74115	48.2	27	8	1045	94	99	99	100
NC_017763.1	NC_007168.1	45439	58518	113.77	66688	74141	48.35	7	8	1051	95	100	100	100
NC_017763.1	NC_007168.1	45439	58661	114.0	66688	73961	47.34	77	8	1013	97	100	99	100
NC_017763.1	NC_007168.1	45439	58729	114.07	66688	73859	47.55	47	8	997	97	99	100	99
NC_017763.1	NC_007168.1	45439	58751	114.07	66688	73849	47.58	7	8	997	97	96	100	96
NC_017763.1	NC_007168.1	45439	63969	110.38	66688	68656	12.57	7	8	37	94	100	100	100
NZ_CP007659.1	NC_004461.1	34160	34165	33.4	95612	110079	29.54	150	87	1321	0	100	97	100
NZ_CP007659.1	NC_004461.1	34160	36402	120.83	70358	110079	10.93	27	76	1321	95	100	99	100
NZ_CP007659.1	NC_004461.1	34164	36402	120.99	70358	95611	0.28	24	42	3	94	98	100	97
NZ_CP007659.1	NC_004461.1	44952	44985	50.7	37902	55503	0.48	4	2	6	5	94	99	94
NC_017763.1	NZ_CP009554.1	80759	82440	679.04	690422	696668	404.85	5	179	7590	100	100	100	100
NC_017763.1	NZ_CP009554.1	82439	82964	358.88	690421	696666	405.02	3	22	7591	99	100	99	100
NC_017763.1	NC_004461.1	34159	34164	33.4	95612	110079	29.54	150	87	1321	3	100	100	100
NC_017763.1	NC_004461.1	34159	36401	120.83	70358	110079	10.93	27	76	1321	95	100	98	100
NC_017763.1	NC_004461.1	34163	36401	120.99	70358	95611	0.28	24	42	3	95	99	99	95
NC_017763.1	NC_004461.1	44951	44984	50.7	37902	55503	0.48	4	2	6	5	96	100	92
NZ_CP007659.1	NC_007168.1	44773	58519	112.86	67396	74115	52.21	29	1	1039	96	99	98	100
NZ_CP007659.1	NC_007168.1	44773	58519	112.86	67396	74141	52.36	9	1	1045	97	99	100	99
NZ_CP007659.1	NC_007168.1	44773	58662	113.09	67396	73961	51.35	79	1	1007	97	100	100	100
NZ_CP007659.1	NC_007168.1	44773	58730	113.16	67396	73859	51.65	49	1	991	97	100	100	100
NZ_CP007659.1	NC_007168.1	44773	58752	113.17	67396	73849	51.68	9	1	991	98	100	100	100
NZ_CP007659.1	NC_007168.1	44773	63970	109.85	67396	68656	13.93	9	1	31	92	100	100	100
NZ_CP007659.1	NC_007168.1	45011	58519	113.61	67122	74115	50.25	41	1	1040	95	99	99	100
NZ_CP007659.1	NC_007168.1	45011	58519	113.61	67122	74141	50.39	13	1	1046	96	99	99	100
NZ_CP007659.1	NC_007168.1	45011	58662	113.84	67122	73961	49.37	111	1	1008	96	100	100	100
NZ_CP007659.1	NC_007168.1	45011	58730	113.91	67122	73859	49.63	69	1	992	90	99	100	99
NZ_CP007659.1	NC_007168.1	45011	58752	113.91	67122	73849	49.66	13	1	992	98	100	100	100
NZ_CP007659.1	NC_007168.1	45011	63970	110.35	67122	68656	11.81	13	1	32	98	97	100	99
NZ_CP007659.1	NC_007168.1	45150	45441	132.36	66689	67061	16.42	94	4	5	93	100	100	100
NZ_CP007659.1	NC_007168.1	45150	58519	114.17	67061	74115	49.88	22	1	1040	96	100	100	100
NZ_CP007659.1	NC_007168.1	45150	58519	114.17	67061	74141	50.02	10	1	1046	96	100	100	100
NZ_CP007659.1	NC_007168.1	45150	58662	114.4	67061	73961	49.0	52	1	1008	95	100	99	100
NZ_CP007659.1	NC_007168.1	45150	58730	114.46	67061	73859	49.25	34	1	992	99	99	100	99
NZ_CP007659.1	NC_007168.1	45150	58752	114.47	67061	73849	49.29	10	1	992	96	98	100	99
NZ_CP007659.1	NC_007168.1	45150	63970	110.72	67061	68656	11.64	10	1	32	91	100	100	100
NZ_CP007659.1	NC_007168.1	45440	58519	113.77	66688	74115	48.2	27	8	1045	98	100	100	100
NZ_CP007659.1	NC_007168.1	45440	58519	113.77	66688	74141	48.35	7	8	1051	96	100	100	99
NZ_CP007659.1	NC_007168.1	45440	58662	114.0	66688	73961	47.34	77	8	1013	99	100	100	100
NZ_CP007659.1	NC_007168.1	45440	58730	114.07	66688	73859	47.55	47	8	997	96	100	100	100
NZ_CP007659.1	NC_007168.1	45440	58752	114.07	66688	73849	47.58	7	8	997	94	100	100	100
NZ_CP007659.1	NC_007168.1	45440	63970	110.38	66688	68656	12.57	7	8	37	98	100	100	100
NC_017763.1	NC_017568.1	409726	409769	39.98	2481629	2485653	7.69	41	4	51	1	100	100	100
NC_017763.1	NC_017568.1	409747	409769	42.36	2481607	2485653	7.78	30	4	51	2	100	100	99
NZ_CP007659.1	NZ_CP009554.1	80760	82441	678.48	690422	696668	404.87	5	179	7590	100	100	100	100

Table A.32. Acceptor and donor candidates for ERR103402 run with yara, species filter and no samflag filter. Sampling sensitivity = 85. No taxon blacklist. No parent blacklist. No species blacklist. (-)0.000* represents absolute values $< 0.000\bar{4}$.

Candidate			MicrobeGPS metrics			DaisyGPS metrics	
Type	Name	Accession.Version	Number Reads	Validity	Heterogeneity	Property	Property Score
Acceptor	Staphylococcus aureus subsp. aureus	NZ_CP007659.1	169032	0.804	0.05	0.754	0.04
Acceptor	Staphylococcus aureus subsp. aureus HO 5096 0412	NC_017763.1	167480	0.806	0.052	0.754	0.039
Donor	Staphylococcus warneri SG1	NC_020164.1	231	0.003	0.69	-0.697	-0.000*
Donor	Staphylococcus pseudintermedius ED99	NC_017568.1	1176	0.002	0.657	-0.655	-0.000*
Donor	Staphylococcus epidermidis RP62A	NC_002976.3	786	0.003	0.578	-0.575	-0.000*
Donor	Staphylococcus lugdunensis HKU09-01	NC_013893.1	676	0.001	0.357	-0.355	-0.000*
Donor	Staphylococcus haemolyticus JCSC1435	NC_007168.1	1123	0.003	0.351	-0.348	-0.000*
Donor	Staphylococcus aureus subsp. aureus str. JKD6008	NC_017341.1	18272	0.097	0.19	-0.103	-0.001
Acceptor-like Donor	Staphylococcus aureus subsp. aureus	NZ_CP009423.1	17888	0.096	0.085	0.011	0.000*

Table A.33. Results for ERR103402 run with yara, gustaf, species filter and no samflag filter. Sampling sensitivity = 90. Split read threshold = 3. No taxon blacklist. No parent blacklist. No species blacklist.

Organism		**Acceptor**			**Donor**			**Read Evidence**			**Evidence Filter**			
Acceptor	Donor	Start	End	Coverage	Start	End	Coverage	Split	Spanning	Within	A-Cov	D-Cov	Spanning	Within
NZ_CP007659.1	NC_020164.1	2038921	2038922	83.0	121511	123832	2.22	48	24	12	100	100	100	100
NC_017763.1	NC_020164.1	2024903	2024904	83.0	121511	123832	2.22	52	24	12	99	100	99	100
NZ_CP007659.1	NC_017341.1	2036785	2038062	62.16	2760130	2761402	144.91	6	48	540	99	100	100	100
NZ_CP007659.1	NC_013893.1	2036709	2038062	66.05	1722127	1723472	67.35	10	10	2	100	100	100	100
NZ_CP007659.1	NC_013893.1	2036785	2038063	62.02	949399	950670	74.87	9	1	3	99	100	100	100
NZ_CP007659.1	NC_013893.1	2036785	2038062	62.03	1722127	1723395	70.11	18	51	2	100	100	100	100
NC_017763.1	NC_013893.1	2022691	2024044	66.05	1722127	1723472	67.46	10	10	2	100	100	100	100
NC_017763.1	NC_013893.1	2022767	2024045	62.02	949399	950670	74.75	9	1	3	99	100	100	100
NC_017763.1	NC_013893.1	2022767	2024044	62.03	1722127	1723395	70.23	18	51	2	100	100	100	100
NC_017763.1	NC_017341.1	2022767	2024044	62.16	2760130	2761402	144.91	6	48	540	99	100	100	100
NC_017763.1	NC_007168.1	2022767	2024044	62.16	1828214	1829486	144.91	10	48	540	99	100	100	100
NZ_CP007659.1	NC_007168.1	2036785	2038062	62.16	1828214	1829486	144.91	10	48	540	100	100	100	100

B Appendix - Hortense

Experimental details DivIVA dataset

The HGT organism in this dataset is *Bacillus subtilis* BSN238, a transgenic organism which is a chimera of *B. subtilis* 168 where the DivIVA protein has been replaced with the DivIVA from *Listeria monocytogenes* strain EGD-e (van Baarle et al., 2012). The Listeria DivIVA protein is located on the complement strand at positions 2'100'224-2'100'751 (NC_003210.1). Bacterial cultivation, protein extraction and proteomic sample measurements were performed in house.

Isolation of cellular proteins

B. subtilis strain BSN238 (ΔdivIVA::tet amyE::Pxyl-divIVALmo spc) was cultivated in LB broth containing 0.5% xylose at 37 °C and harvested by centrifugation at an optical density (λ=600 nm) of 1.0. Cells were washed with ZAP buffer (10 mM Tris/HCl pH 7.5 and 200 mM NaCl), resuspended in 1 ml ZAP buffer also containing 1 mM phenylmethylsulfonyl fluoride and disrupted by sonication. Cell debris was removed by centrifugation (1 min, 13000 rpm in a table top centrifuge). The resulting supernatant was used as total cellular protein extract.

nLC-MS/MS

Proteins were precipitated at −20 °C for 24 h using four volumes of acetone. Pellets were resuspended in 1 M Urea, 50 mM Tris-HCl (pH 8.5) and digested for 18 h at 37 °C using Trypsin Gold, Mass Spectrometry Grade (Promega, Fitchburg, WI, USA) at a protein/enzyme ratio of 50:1. The peptides were desalted using 200 μL StageTips packed with four Empore^{TM} SPE Disks C18 (3 M Purification, Inc., Lexington, USA) (Ishihama et al., 2006) and were further quantified by measuring the absorbance at 280 nm using a Nanodrop 1000 (Thermo Fisher Scientific, Rockford, IL, USA). Proteome analysis was performed on an Easy-nanoLC (Proxeon, Odense, Denmark) coupled online to an LTQ Orbitrap Discovery^{TM} mass spectrometer (Thermo Fisher Scientific, Rockford, IL, USA). 1 μg peptides were loaded directly on a Reprosil-Pur 120 C18-AQ, 2.4 μm, 300 mm x 75 μm fused silica capillary column (Dr. Maisch, Ammerbuch-Entringen, Germany), which was kept at 60 °C using a butterfly heater (Phoenix S&T, Chester, PA, USA). Peptides were separated using a linear 240 min gradient of acetonitrile in 0.1% formic acid and 3% DMSO from 0 to 29% at 200 nL/min flow rate. The mass spectrometer was operated in a data-dependent manner in the m/z range of 400–1400 with a resolution of 30000 in the orbitrap. Up to the seven most intense 2+ and 3+ charged ions were selected for low-energy CID type fragmentation in the ion trap with a normalized collision energy of 35% using an activation time of 10 ms. The m/z isolation width for MS/MS fragmentation was set to 2 Th. Once fragmented, up to 500 isolated peaks were dynamically excluded from precursor selection for 90 s within a 20 ppm window. The ion selection threshold for MS/MS spectra was 1000 counts, and the maximum allowed ion accumulation times were 500 ms for full scans and 100 ms for MS/MS spectra. Automatic gain control was set to a target value of 1e6 for full scans and 5e3 for MS/MS.

Bibliography

Conda website. URL `https://conda.io/docs/index.html`.

H. J. Abel and E. J. Duncavage. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer Genetics*, 206(12):432–440, 2013.

A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 21(6):974–984, 2011.

T. Akiba, K. Koyama, Y. Ishiki, S. Kimura, and T. Fukushima. On the mechanism of the development of multiple-drug-resistant clones of Shigella. *Japanese Journal of Microbiology*, 4(2):219–227, 1960.

C. Alkan, B. P. Coe, and E. E. Eichler. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12:363–376, 2011.

H. E. Allison, M. J. Sergeant, C. E. James, J. R. Saunders, D. L. Smith, R. J. Sharp, T. S. Marks, and A. J. McCarthy. Immunity Profiles of Wild-Type and Recombinant Shiga-Like Toxin-Encoding Bacteriophages and Characterization of Novel Double Lysogens. *Infection and Immunity*, 71(6):3409–3418, 2003.

A. Altmann, P. Weber, D. Bader, M. Preuß, E. B. Binder, and B. Müller-Myhsok. A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Human Genetics*, 131(10):1541–1554, 2012.

D. Antaki, W. M. Brandler, and J. Sebat. SV^2 : Accurate structural variation genotyping and de novo mutation detection. *bioRxiv*, 2017.

S. Ardui, A. Ameur, J. R. Vermeesch, and M. S. Hestand. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research*, 46(5):2159–2168, 2018.

M. Asadulghani, Y. Ogura, T. Ooka, T. Itoh, A. Sawaguchi, A. Iguchi, K. Nakayama, and T. Hayashi. The defective prophage pool of Escherichia coli O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. *PLoS Pathogens*, 5(5):e1000408, 2009.

F. Backhed. Host-Bacterial Mutualism in the Human Intestine. *Science*, 307(5717): 1915–1920, 2005.

M. Baker. Structural variation: the genome's hidden architecture. *Nature Methods*, 9(2):133–q137, 2012.

A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477, 2012.

M. S. Bansal, E. J. Alm, and M. Kellis. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, 28 (12):i283–i291, 2012.

S. Barik, S. Das, and H. Vikalo. QSdpR: Viral quasispecies reconstruction via correlation clustering. *Genomics*, 2017.

M. Barlow. What antimicrobial resistance has taught us about horizontal gene transfer. *Methods in molecular biology (Clifton, N.J.)*, 532:397–411, 2009.

J. E. Barrick, G. Colburn, D. E. Deatherage, C. C. Traverse, M. D. Strand, J. J. Borges, D. B. Knoester, A. Reba, and A. G. Meyer. Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. *BMC Genomics*, 15(1):1039, 2014.

M. F. Berger, M. S. Lawrence, F. Demichelis, Y. Drier, K. Cibulskis, A. Y. Sivachenko, A. Sboner, R. Esgueva, D. Pflueger, C. Sougnez, et al. The genomic complexity of primary human prostate cancer. *Nature*, 470(7333):214–220, 2011.

A. Boc, H. Philippe, and V. Makarenkov. Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Systematic Biology*, 59(2):195–211, 2010.

M.-L. Bondeson, N. Dahl, H. Malmgren, W. J. Kleijer, T. Tönnesen, B.-M. Carlberg, and U. Pettersson. Inversion of the IDS gene resulting from recombination with IDS-related sequences in a common cause of the hunter syndrome. *Human Molecular Genetics*, 4(4):615–621, 1995.

D. I. Boomsma, C. Wijmenga, E. P. Slagboom, M. A. Swertz, L. C. Karssen, A. Abdellaoui, K. Ye, V. Guryev, M. Vermaat, F. van Dijk, et al. The genome of the netherlands: design, and project goals. *European Journal of Human Genetics*, 22 (2):221–227, 2013.

L. Boto. Horizontal gene transfer in evolution: facts and challenges. *Proceedings of the Royal Society B: Biological Sciences*, 277(1683):819–827, 2009.

F. Boulund, R. Karlsson, L. Gonzales-Siles, A. Johnning, N. Karami, O. AL-Bayati, C. Åhrén, E. R. B. Moore, and E. Kristiansson. Typing and Characterization of Bacteria Using Bottom-up Tandem Mass Spectrometry Proteomics. *Molecular & Cellular Proteomics*, 16(6):1052–1063, 2017.

N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34:525–527, 2016.

F. P. Breitwieser, J. Lu, and S. L. Salzberg. A review of methods and databases for metagenomic classification and assembly. *Briefings in bioinformatics*, 2017.

C. Brooksbank, M. T. Bergman, R. Apweiler, E. Birney, and J. Thornton. The European Bioinformatics Institute's data resources 2014. *Nucleic Acids Research*, 42(Database issue):D18–D25, 2014.

M. Brudno, S. Malde, A. Poliakov, C. B. Do, O. Couronne, I. Dubchak, and S. Batzoglou. Glocal alignment: finding rearrangements during alignment. *Bioinformatics*, 19 Suppl 1:i54–i62, 2003.

M. Burrows and D. J. Wheeler. A block-sorting lossless data compression algorithm. Technical report, Palo Alto, 1994.

J. Butler, I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, and D. B. Jaffe. ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Research*, 18(5):810–820, 2008.

A. L. Byrd, J. F. Perez-Rogers, S. Manimaran, E. Castro-Nallar, I. Toma, T. McCaffrey, M. Siegel, G. Benson, K. A. Crandall, and W. E. Johnson. Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. *BMC Bioinformatics*, 15 (1):262, 2014.

D. L. Cameron, J. Schröder, J. S. Penington, H. Do, R. Molania, A. Dobrovic, T. P. Speed, and A. T. Papenfuss. GRIDSS: sensitive and specific genomic rearrangement detection using positional de bruijn graph assembly. *Genome Research*, 27 (12):2050–2060, 2017.

B. Canard and R. S. Sarfati. DNA polymerase fluorescent substrates with reversible 3′-tags. *Gene*, 148(1):1–6, 1994.

M. J. Chaisson and P. A. Pevzner. Short read fragment assembly of bacterial genomes. *Genome Research*, 18(2):324–330, 2008.

M. J. Chaisson, A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. Rodriguez, L. Guo, R. L. Collins, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv*, 2017.

M. J. P. Chaisson, J. Huddleston, M. Y. Dennis, P. H. Sudmant, M. Malig, F. Hormozdiari, F. Antonacci, U. Surti, R. Sandstrom, M. Boitano, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517 (7536):608–611, 2015a.

M. J. P. Chaisson, R. K. Wilson, and E. E. Eichler. Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics*, 16(11):627–640, 2015b.

B. T. Chait. Mass Spectrometry: Bottom-Up or Top-Down? *Science*, 314(5796): 65–66, 2006.

K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath, M. C. Wendl, Q. Zhang, D. P. Locke, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, 6(9):677–681, 2009.

K. Chen, L. Chen, X. Fan, J. Wallis, L. Ding, and G. Weinstock. TIGRA: A targeted iterative graph routing assembler for breakpoint assembly. *Genome Research*, 24 (2):310–317, 2013a.

K. Chen, L. Chen, X. Fan, J. Wallis, L. Ding, and G. Weinstock. TIGRA: A Targeted Iterative Graph Routing Assembler for breakpoint assembly. *Genome Research*, 2013b.

X. Chen, O. Schulz-Trieglaff, R. Shaw, B. Barnes, F. Schlesinger, M. Källberg, A. J. Cox, S. Kruglyak, and C. T. Saunders. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8):1220–1222, 2015.

D. Y. Chiang, G. Getz, D. B. Jaffe, M. J. T. O'Kelly, X. Zhao, S. L. Carter, C. Russ, C. Nusbaum, M. Meyerson, and E. S. Lander. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods*, 6(1): 99–103, 2008.

C. G. Clark, P. Kruczkiewicz, C. Guan, S. J. McCorrister, P. Chong, J. Wylie, P. van Caeseele, H. A. Tabor, P. Snarr, M. W. Gilmour, et al. Evaluation of MALDI-TOF mass spectroscopy methods for determination of escherichia coli pathotypes. *Journal of microbiological methods*, 94:180–191, 2013.

J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, 4(4):265–270, feb 2009. doi: 10.1038/nnano.2009.12.

P. T. L. C. Clausen, E. Zankari, F. M. Aarestrup, and O. Lund. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *The Journal of Antimicrobial Chemotherapy*, 71:2484–2488, 2016.

D. Cortez, L. Delaye, A. Lazcano, and A. Becerra. Composition-based methods to identify horizontal gene transfer. In *Horizontal Gene Transfer*, pages 215–225. Humana Press, 2009.

R. Craig and R. C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20:1466–1467, 2004.

A. Crisp, C. Boschetti, M. Perry, A. Tunnacliffe, and G. Micklem. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biology*, 16(1), 2015.

T. H. Dadi, E. Siragusa, V. Piro, A. Andrusch, E. Seiler, B. Renard, and K. Reinert. Dream-yara: An exact read mapper for very large databases with short update time. *Bioinformatics*, 2018.

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011.

V. Daubin and G. J. Szöllősi. Horizontal gene transfer and the history of life. *Cold Spring Harbor Perspectives in Biology*, 8(4):a018036, jan 2016.

V. Daubin, E. Lerat, and G. Perrière. The source of laterally transferred genes in bacterial genomes. *Genome Biology*, 4(9):R57, 2003.

A. P. J. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genetics*, 7 (12):e1002384, 2011.

S. Degroeve, D. Maddelein, and L. Martens. MS2pip prediction server: compute and visualize MS2peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Research*, 43(W1):W326–W330, 2015.

C. Dessimoz, D. Margadant, and G. H. Gonnet. DLIGHT - lateral gene transfer detection using pairwise evolutionary distances in a statistical framework. In *Research in Computational Molecular Biology. RECOMB 2008. Lecture Notes in Computer Science.*, pages 315–330. Springer Berlin Heidelberg, 2008. doi: 10.1007/978-3-540-78839-3_27.

A. Doerr. Mass spectrometry-based targeted proteomics. *Nature Methods*, 7(23), 2013.

J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16):e105–e105, 2008.

A. Döring, D. Weese, T. Rausch, and K. Reinert. SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, 9:11, 2008.

K. V. dos Santos, C. G. Diniz, L. de Castro Veloso, H. M. de Andrade, M. da Silva Giusta, S. da Fonseca Pires, A. V. Santos, A. C. M. Apolônio, M. A. R. de Carvalho, and L. de Macêdo Farias. Proteomic analysis of escherichia coli with experimentally induced resistance to piperacillin/tazobactam. *Research in Microbiology*, 161(4):268–275, 2010.

J. P. Dworzanski and A. P. Snyder. Classification and identification of bacteria using mass spectrometry-based proteomics. *Expert Review of Proteomics*, 2(6):863–878, 2005.

J. P. Dworzanski, A. P. Snyder, R. Chen, H. Zhang, D. Wishart, and L. Li. Identification of Bacteria Using Tandem Mass Spectrometry Combined with a Proteome Database and Statistical Scoring. *Analytical Chemistry*, 76(8):2355–2366, 2004.

D. Earl, K. Bradnam, J. S. John, A. Darling, D. Lin, J. Fass, H. O. K. Yu, V. Buffalo, D. R. Zerbino, M. Diekhans, et al. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, 21(12):2224–2241, 2011.

H. A. Ebhardt, A. Root, C. Sander, and R. Aebersold. Applications of targeted proteomics in systems biology and translational medicine. *Proteomics*, 15(18): 3193–3208, 2015.

P. B. Eckburg. Diversity of the human intestinal microbial flora. *Science*, 308(5728): 1635–1638, 2005.

R. C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, 2010.

J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, jan 2009. doi: 10.1126/science.1162986.

J. A. Eisen. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Current Opinion in Genetics & Development*, 10(6): 606–611, 2000.

A.-K. Emde, M. H. Schulz, D. Weese, R. Sun, M. Vingron, V. M. Kalscheuer, S. A. Haas, and K. Reinert. Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS. *Bioinformatics*, 28(5): 619–627, 2012.

A. C. English, W. J. Salerno, O. A. Hampton, C. Gonzaga-Jauregui, S. Ambreth, D. I. Ritter, C. R. Beck, C. F. Davis, M. Dahdouli, S. Ma, et al. Assessing structural variation in a personal genome—towards a human reference diploid genome. *BMC Genomics*, 16(1), 2015.

C. Ernst and S. Rahmann. PanCake: A Data Structure for Pangenomes. In *German Conference on Bioinformatics 2013*, pages 35–45. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2013.

J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246:64–71, 1989.

P. Ferragina and G. Manzini. Opportunistic Data Structures with Applications. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, FOCS '00, page 390. IEEE Computer Society, 2000.

M. Fischer, B. Strauch, and B. Y. Renard. Abundance estimation and differential testing on strain level in metagenomics data. *Bioinformatics*, 33:i124–i132, 2017.

R. Fleischmann, M. Adams, O. White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J. Tomb, B. Dougherty, J. Merrick, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, 1995.

A. Frank and P. Pevzner. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry*, 77:964–973, 2005.

C. Frank, D. Werber, J. P. Cramer, M. Askar, M. Faber, M. an der Heiden, H. Bernard, A. Fruth, R. Prager, A. Spode, et al. Epidemic Profile of Shiga-Toxin-Producing Escherichia coli O104:H4 Outbreak in Germany. *New England Journal of Medicine*, 365(19):1771–1780, 2011.

R. E. Franklin and R. G. Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171(4356):740–741, 1953.

W. F. Fricke and D. A. Rasko. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nature Reviews Genetics*, 15(1):49–55, 2013.

T. Frickey. PhyloGenie: automated phylome generation and analysis. *Nucleic Acids Research*, 32(17):5231–5238, 2004.

S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, 108(4):1513–1518, 2010.

A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, et al. Life with 6000 genes. *Science*, 274(5287):546–567, 1996.

F. Guarner and J.-R. Malagelada. Gut flora in health and disease. *The Lancet*, 361 (9356):512–519, 2003. doi: 10.1016/s0140-6736(03)12489-0.

N. Gupta and P. A. Pevzner. False discovery rates of protein identifications: a strike against the two-peptide rule. *Journal of proteome research*, 8:4173–4181, 2009.

A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013.

A. Guthals, K. R. Clauser, A. M. Frank, and N. Bandeira. Sequencing-grade de novo analysis of ms/ms triplets (CID/HCD/ETD) from overlapping peptides. *Journal of proteome research*, 12:2846–2857, 2013.

C. Gyles and P. Boerlin. Horizontally Transferred Genetic Elements and Their Role in Pathogenesis of Bacterial Disease. *Veterinary Pathology*, 51(2):328–340, 2013.

I. Hajirasouliha, F. Hormozdiari, C. Alkan, J. M. Kidd, I. Birol, E. E. Eichler, and S. C. Sahinalp. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics*, 26(10):1277–1283, 2010.

R. E. Handsaker, J. M. Korn, J. Nemesh, and S. A. McCarroll. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet*, 43(3):269–276, Mar 2011.

J. Hawkey, M. Hamidian, R. R. Wick, D. J. Edwards, H. Billman-Jacobe, R. M. Hall, and K. E. Holt. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. *BMC Genomics*, 16(1), 2015.

J. M. Heather and B. Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, 2016. doi: 10.1016/j.ygeno.2015.11.003.

S. Herold, H. Karch, and H. Schmidt. Shiga toxin-encoding bacteriophages–genomes in motion. *International Journal of Medical Microbiology*, 294(2-3):115–121, 2004.

E. Hilario and J. P. Gogarten. Horizontal transfer of atpase genes–the tree of life becomes a net of life. *Biosystems*, 31:111–119, 1993.

R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick, and A. Zamir. Structure of a ribonucleic acid. *Science*, 147: 1462–1465, 1965.

M. Holtgrewe. Mason - a read simulator for second generation sequencing data. Technical Report TR-B-10-06, Institut für Mathematik und Informatik, Freie Universität Berlin, 2010.

M. Holtgrewe. Mason: a tool suite for simulating nucleotide sequences. unpublished, 2014.

M. Holtgrewe, L. Kuchenbecker, and K. Reinert. Methods for the detection and assembly of novel sequence in high-throughput sequencing data. *Bioinformatics*, 31(12):1904–1912, 2015.

F. Hormozdiari, C. Alkan, E. E. Eichler, and S. C. Sahinalp. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research*, 19(7):1270–1278, 2009.

F. Hormozdiari, I. Hajirasouliha, P. Dao, F. Hach, D. Yorukoglu, C. Alkan, E. E. Eichler, and S. C. Sahinalp. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, 26(12):i350–i357, 2010.

B. Hu, G. Xie, C.-C. Lo, S. R. Starkenburg, and P. S. G. Chain. Pathogen comparative genomics in the next-generation sequencing era: genome alignments, pangenomics and metagenomics. *Briefings in Functional Genomics*, 10(6):322–333, 2011.

J. Huddleston, M. J. Chaisson, K. M. Steinberg, W. Warren, K. Hoekzema, D. Gordon, T. A. Graves-Lindsay, K. M. Munson, Z. N. Kronenberg, L. Vives, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*, 27(5):677–685, 2016.

D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster. MEGAN analysis of metagenomic data. *Genome Research*, 17:377–386, 2007.

A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee. Detection of large-scale variation in the human genome. *Nature Genetics*, 36(9):949–951, 2004.

N. H. G. R. Institute. https://www.genome.gov/27541954/dna-sequencing-costs-data/, 07 2018.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

J. Iranzo, M. J. Gómez, F. J. López de Saro, and S. Manrubia. Large-scale genomic analysis suggests a neutral punctuated dynamics of transposable elements in bacterial genomes. *PLoS Computational Biology*, 10(6):e1003680, 2014.

Y. Ishihama, J. Rappsilber, and M. Mann. Modular stop and go extraction tips with stacked disks for parallel and multidimensional peptide fractionation in proteomics. *Journal of Proteome Research*, 5(4):988–94, 2006.

R. E. Jabbour, S. V. Deshpande, M. M. Wade, M. F. Stanford, C. H. Wick, A. W. Zulich, E. W. Skowronski, and A. P. Snyder. Double-Blind Characterization of Non-Genome-Sequenced Bacteria by Mass Spectrometry-Based Proteomics. *Applied and Environmental Microbiology*, 76(11):3637–3644, 2010.

R. E. Jabbour, S. V. Deshpande, M. F. Stanford, C. H. Wick, A. W. Zulich, and A. P. Snyder. A Protein Processing Filter Method for Bacterial Identification by Mass Spectrometry-Based Proteomics. *Journal of Proteome Research*, 10(2): 907–912, 2011.

C. Jandrasits, P. W. Dabrowski, S. Fuchs, and B. Y. Renard. seq-seq-pan: building a computational pan-genome data structure on whole genome alignment. *BMC Genomics*, 19(1), 2018.

K. S. Jaron, J. C. Moravec, and N. Martínková. SigHunt: horizontal gene transfer finder optimized for eukaryotic genomes. *Bioinformatics*, 30(8):1081–1086, 2013.

D. Jayasundara, I. Saeed, S. Maheswararajah, B. Chang, S.-L. Tang, and S. K. Halgamuge. ViQuaS: an improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing. *Bioinformatics*, 31(6):886–896, 2014.

X. Jiang, A. B. Hall, R. J. Xavier, and E. J. Alm. Comprehensive analysis of mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. *bioRxiv*, nov 2017. doi: 10.1101/214213.

Y. Jiang, Y. Wang, and M. Brudno. PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics*, 28(20):2576–2583, 2012.

A. R. Jones, M. Eisenacher, G. Mayer, O. Kohlbacher, J. Siepen, S. J. Hubbard, J. N. Selley, B. C. Searle, J. Shofstahl, S. L. Seymour, et al. The mzIdentML data standard for mass spectrometry-based proteomics results. *Molecular and Cellular Proteomics*, 11(7), 2012.

M. Juhas. Horizontal gene transfer in human pathogens. *Critical Reviews in Microbiology*, 41(1):101–108, 2013.

E. Karakoc, C. Alkan, B. J. O'Roak, M. Y. Dennis, L. Vives, K. Mark, M. J. Rieder, D. A. Nickerson, and E. E. Eichler. Detection of structural variants and indels within exome data. *Nature Methods*, 9(2):176–178, 2012.

M. Karas and F. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, 60:2299–2301, 1988.

S. Karlin. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends in Microbiology*, 9:335–343, 2001.

R. Karlsson, L. Gonzales-Siles, F. Boulund, L. Svensson-Stadler, S. Skovbjerg, A. Karlsson, M. Davidson, S. Hulth, E. Kristiansson, and E. R. Moore. Proteotyping: Proteomic characterization, classification and identification of microorganisms - A prospectus. *Systematic and Applied Microbiology*, 38(4):246–257, 2015.

K. G. Kaval and S. Halbedel. Architecturally the same, but playing a different game. *Virulence*, 3(4):406–407, 2012.

J. Kececioglu. The maximum weight trace problem in multiple sequence alignment. In *Proceedings of the 4th Symposium on Combinatorial Pattern Matching (CPM)*, volume 684 of *Lecture Notes in Computer Science*, pages 106–119. Springer-Verlag, 1993.

B. Kehr, D. Weese, and K. Reinert. STELLAR: fast and exact local alignments. In *Ninth Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics*, volume 12(Suppl 9), 2011.

B. Kehr, K. Reinert, and A. E. Darling. Hidden breakpoints in genome alignments. In B. Raphael and J. Tang, editors, *Algorithms in Bioinformatics*, volume 7534 of *Lecture Notes in Computer Science*, pages 391–403. Springer Berlin Heidelberg, 2012.

B. Kehr, K. Trappe, M. Holtgrewe, and K. Reinert. Genome alignment with graph data structures: a comparison. *BMC Bioinformatics*, 15(1):99, 2014.

B. Kehr, P. Melsted, and B. V. Halldórsson. PopIns: population-scale detection of novel sequence insertions. *Bioinformatics*, 32(7):961–967, 2015.

B. Kehr, A. Helgadottir, P. Melsted, H. Jonsson, H. Helgason, A. Jonasdottir, A. Jonasdottir, A. Sigurdsson, A. Gylfason, G. H. Halldorsson, et al. Diversity in non-repetitive human sequences not found in the reference genome. *Nature Genetics*, 49(4):588–593, 2017.

S. Kim and P. A. Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, 5:5277, 2014.

W. P. Kloosterman, L. C. Francioli, F. Hormozdiari, T. Marschall, J. Y. Hehir-Kwa, A. Abdellaoui, E.-W. Lameijer, M. H. Moed, V. Koval, I. Renkens, et al. Characteristics of de novo structural changes in the human genome. *Genome Research*, 25(6):792–801, 2015.

E. V. Koonin. Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Research*, 5:1805, 2016.

E. V. Koonin, K. S. Makarova, and L. Aravind. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annual Review of Microbiology*, 55(1): 709–742, 2001.

J. O. Korbel, A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero, L. Du, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–426, 2007.

J. O. Korbel, A. Abyzov, X. Mu, N. Carriero, P. Cayting, Z. Zhang, M. Snyder, and M. B. Gerstein. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*, 10(2):R23, 2009.

S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy. Canu: scalable and accurate long-read assembly via adaptivek-mer weighting and repeat separation. *Genome Research*, 27(5):722–736, 2017.

C. U. Köser, M. T. Holden, M. J. Ellington, E. J. Cartwright, N. M. Brown, A. L. Ogilvy-Stuart, L. Y. Hsu, C. Chewapreecha, N. J. Croucher, S. R. Harris, et al. Rapid whole-genome sequencing for investigation of a neonatal mrsa outbreak. *New England Journal of Medicine*, 366(24):2267–2275, 2012.

J. Köster and S. Rahmann. Snakemake - scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

J. Kriegsmann, M. Kriegsmann, and R. Casadonte. MALDI TOF imaging mass spectrometry in clinical pathology: A valuable tool for cancer diagnostics (Review). *International Journal of Oncology*, 46(3):893–906, 2014.

M. Kuhring and B. Y. Renard. iPiG: Integrating peptide spectrum matches into genome browser visualizations. *PLOSE ONE*, 7(12):e50246, 2012.

M. Kuhring, P. W. Dabrowski, V. C. Piro, A. Nitsche, and B. Y. Renard. SuRankCo: supervised ranking of contigs in de novo assemblies. *BMC Bioinformatics*, 16(1), 2015.

P. Kumar, M. Al-Shafai, W. A. Muftah, N. Chalhoub, M. F. Elsaid, A. Aleem, and K. Suhre. Evaluation of SNP calling using single and multiple-sample calling algorithms by validation against array base genotyping and mendelian inheritance. *BMC Research Notes*, 7(1):747, 2014.

J. L. Kyle, C. A. Cummings, C. T. Parker, B. Quiñones, P. Vatta, E. Newton, S. Huynh, M. Swimley, L. Degoricija, M. Barker, et al. Escherichia coli serotype O55:H7 diversity supports parallel acquisition of bacteriophage at Shiga toxin phage insertion sites during evolution of the O157:H7 lineage. *J Bacteriol*, 194 (8):1885–1896, 2012.

D. Lakich, H. H. Kazazian, S. E. Antonarakis, and J. Gitschier. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia a. *Nature Genetics*, 5(3):236–241, 1993.

H. Y. K. Lam, X. J. Mu, A. M. Stütz, A. Tanzer, P. D. Cayting, M. Snyder, P. M. Kim, J. O. Korbel, and M. B. Gerstein. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature Biotechnology*, 28(1): 47–55, 2009.

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

J. G. Lawrence. Gene transfer in bacteria: speciation without species? *Theoretical Population Biology*, 61(4):449–460, 2002.

J. G. Lawrence and H. Ochman. Amelioration of bacterial genomes: rates of change and exchange. *Journal of molecular evolution*, 44:383–397, 1997.

J. G. Lawrence and H. Ochman. Molecular archaeology of the escherichia coli genome. *Proceedings of the National Academy of Sciences of the United States of America*, 95:9413–9417, 1998.

J. G. Lawrence and H. Ochman. Reconciling the many faces of lateral gene transfer. *Trends in Microbiology*, 10(1):1–4, 2002.

R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, 15(6):R84, 2014. doi: 10.1186/gb-2014-15-6-r8.

H. Lee, E. Popodi, H. Tang, and P. L. Foster. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proceedings of the National Academy of Sciences*, 109(41): E2774–E2783, 2012.

S. Lee, F. Hormozdiari, C. Alkan, and M. Brudno. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nature Methods*, 6(7): 473–474, 2009.

W.-P. Lee, M. P. Stromberg, A. Ward, C. Stewart, E. P. Garrison, and G. T. Marth. MOSAIK: A hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS ONE*, 9(3):e90581, 2014.

R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, et al. The European Nucleotide Archive. *Nucleic Acids Research*, 39(Database issue):D28–D31, 2011.

W. Y. Leung, T. Marschall, Y. Paudel, L. Falquet, H. Mei, A. Schönhuth, and T. Y. Maoz. SV-AUTOPILOT: optimized, automated construction of structural variation discovery and benchmarking pipelines. *BMC Genomics*, 16(1):238, 2015.

D. Levy, M. Ronemus, B. Yamrom, Y. ha Lee, A. Leotta, J. Kendall, S. Marks, B. Lakshmi, D. Pai, K. Ye, et al. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron*, 70(5):886–897, 2011.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

R. Li, Y. Li, K. Kristiansen, and J. Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, 2008.

R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, and J. Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, 2009a.

R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2):265–272, 2009b. doi: 10.1101/gr.097261.109.

S. Li, R. Li, H. Li, J. Lu, Y. Li, L. Bolund, M. H. Schierup, and J. Wang. SOAPindel: Efficient identification of indels from short paired reads. *Genome Research*, 23(1): 195–200, 2012a.

Y. Li, J. Chien, D. I. Smith, and J. Ma. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics*, 27(12):1708–1710, 2011.

Y. Li, W. Chen, E. Y. Liu, and Y.-H. Zhou. Single nucleotide polymorphism (SNP) detection and genotype calling from massively parallel sequencing (MPS) data. *Statistics in Biosciences*, 5(1):3–25, 2012b.

T. B. Lima, M. F. S. Pinto, S. M. Ribeiro, L. A. de Lima, J. C. Viana, N. G. Júnior, E. de Souza Cândido, S. C. Dias, and O. L. Franco. Bacterial resistance mechanism: what proteomics can elucidate. *The FASEB Journal*, 27(4):1291–1303, 2013.

K. Lin, S. Smit, G. Bonnema, G. Sanchez-Perez, and D. d. Ridder. Making the difference: integrating structural variation detection tools. *Briefings in Bioinformatics*, page bbu047, 2014.

S. Lindgreen, K. L. Adair, and P. P. Gardner. An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*, 6:19233, 2016.

M. S. Lindner and B. Y. Renard. Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Research*, 41:e10, 2013.

M. S. Lindner and B. Y. Renard. Metagenomic Profiling of Known and Unknown Microbes with MicrobeGPS. *PLoS ONE*, 10(2):e0117711, 2015.

M. S. Lindner, M. Kollock, F. Zickmann, and B. Y. Renard. Analyzing genome coverage profiles with applications to quality control in metagenomics. *Bioinformatics*, 29(10):1260–1267, 2013.

J. A. Lindsay. Genomic variation and evolution of staphylococcus aureus. *International Journal of Medical Microbiology*, 300(2):98 – 103, 2010.

J. A. Lindsay. Staphylococcus aureus genomics and the impact of horizontal gene transfer. *International Journal of Medical Microbiology*, 304(2):103 – 109, 2014.

B. Liu, J. M. Conroy, C. D. Morrison, A. O. Odunsi, M. Qin, L. Wei, D. L. Trump, C. S. Johnson, S. Liu, and J. Wang. Structural variation discovery in the cancer genome using next generation sequencing: Computational solutions and perspectives. *Oncotarget*, 6(8), 2015.

L. Liu, X. Chen, G. Skogerbø, P. Zhang, R. Chen, S. He, and D.-W. Huang. The human microbiome: A hot spot of microbial horizontal gene transfer. *Genomics*, 100(5):265–270, 2012.

J. Lu, F. P. Breitwieser, P. Thielen, and S. L. Salzberg. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science*, 3:e104, 2017.

G. Lunter and M. Goodson. Stampy: A statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome Research*, 21(6):936–939, 2010.

C. Luo, D. Tsementzi, N. C. Kyrpides, and K. T. Konstantinidis. Individual genome assembly from complex community short-read metagenomic datasets. *The ISME Journal*, 6(4):898–901, 2012a.

C.-H. Luo, P.-Y. Chiou, C.-Y. Yang, and N.-T. Lin. Genome, Integration, and Transduction of a Novel Temperate Phage of Helicobacter pylori. *Journal of Virology*, 86(16):8781–8792, 2012b.

R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Giga Science*, 1(1):18, 2012c.

J. R. Lupski, R. M. de Oca-Luna, S. Slaugenhaupt, L. Pentao, V. Guzzetta, B. J. Trask, O. Saucedo-Cardenas, D. F. Barker, J. M. Killian, C. A. Garcia, et al. Dna duplication associated with charcot-marie-tooth disease type 1a. *Cell*, 66:219–232, 1991.

L. E. MacConaill and L. A. Garraway. Clinical implications of the cancer genome. *Journal of Clinical Oncology*, 28(35):5219–5228, 2010.

C. A. Maher, C. Kumar-Sinha, X. Cao, S. Kalyana-sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy, and A. M. Chinnaiyan. Transcriptome sequencing to detect gene fusions in cancer. *Nature Letters*, 458, 2009.

M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

T. Marschall and A. Schönhuth. Sensitive Long-Indel-Aware Alignment of Sequencing Reads. *arXiv*, 1303.3520, Mar. 2013. URL `http://arxiv.org/abs/1303.3520`.

T. Marschall, I. G. Costa, S. Canzar, M. Bauer, G. W. Klau, A. Schliep, and A. Schönhuth. CLEVER: clique-enumerating variant finder. *Bioinformatics*, 28 (22):2875–2882, 2012.

T. Marschall, I. Hajirasouliha, and A. Schönhuth. MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics*, 29(24):3143–3150, 2013.

P. Mazodier and J. Davies. Gene transfer between distantly related bacteria. *Annual Review of Genetics*, 25(1):147–171, 1991.

K. McElroy, T. Thomas, and F. Luciani. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microbial Informatics and Experimentation*, 4(1):1, 2014.

L. McHugh and J. W. Arthur. Computational Methods for Protein Identification from Mass Spectrometry Data. *PLoS Computational Biology*, 4(2):e12, 2008.

A. McPherson, F. Hormozdiari, A. Zayed, R. Giuliany, G. Ha, M. G. F. Sun, M. Griffith, A. H. Moussavi, J. Senz, N. Melnyk, et al. deFuse: An algorithm for gene fusion discovery in tumor RNA-seq data. *PLoS Computational Biology*, 7(5): e1001138, 2011.

P. Medvedev, M. Stanciu, and M. Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, 6(11s): S13–S20, 2009.

P. Medvedev, M. Fiume, M. Dzamba, T. Smith, and M. Brudno. Detecting copy number variation with mated short reads. *Genome Research*, 20(11):1613–1622, 2010.

M. L. Metzker. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

S. Metzler and O. V. Kalinina. Detection of atypical genes in virus families using a one-class SVM. *BMC Genomics*, 15:913, 2014.

J. J. Michaelson and J. Sebat. forestSV: structural variant discovery through statistical learning. *Nature Methods*, 9(8):819–821, 2012.

M. Mielczarek and J. Szyda. Review of alignment and SNP calling algorithms for next-generation sequencing data. *Journal of Applied Genetics*, 57(1):71–79, 2015.

A. Mikheenko, A. Prjibelski, V. Saveliev, D. Antipov, and A. Gurevich. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, 34(13):i142–i150, 2018.

C. A. Miller, O. Hampton, C. Coarfa, and A. Milosavljevic. ReadDepth: A parallel r package for detecting copy number alterations from short sequencing reads. *PLoS ONE*, 6(1):e16327, 2011.

R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, et al. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, 2011.

M. Mimee, R. J. Citorik, and T. K. Lu. Microbiome therapeutics - advances and challenges. *Advanced Drug Delivery Reviews*, 105:44–54, 2016.

S. K. Misra, F. M. D. Aké, Z. Wu, E. Milohanic, T. N. Cao, P. Cossart, J. Deutscher, V. Monnet, C. Archambaud, and C. Henry. Quantitative proteome analyses identify prfa-responsive proteins and phosphoproteins in listeria monocytogenes. *Journal of Proteome Research*, 13(12):6046–6057, 2014.

V. Moncunill, S. Gonzalez, S. Beà, L. O. Andrieux, I. Salaverria, C. Royo, L. Martinez, M. Puiggròs, M. Segura-Wang, A. M. Stütz, et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nature Biotechnology*, 32(11):1106–1112, 2014.

M. Morey, A. Fernández-Marmiesse, D. Castiñeiras, J. M. Fraga, M. L. Couce, and J. A. Cocho. A glimpse into past, present, and future DNA sequencing. *Molecular Genetics and Metabolism*, 110(1-2):3–24, 2013.

J. M. Mullaney, R. E. Mills, W. S. Pittard, and S. E. Devine. Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, 19(R2): R131–R136, 2010.

T. Muth and B. Y. Renard. Evaluating *de novo* sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Briefings in Bioinformatics*, 2017.

T. Muth, A. Behne, R. Heyer, F. Kohrs, D. Benndorf, M. Hoffmann, M. Lehtevä, U. Reichl, L. Martens, and E. Rapp. The MetaProteomeAnalyzer: A Powerful Open-Source Software Suite for Metaproteomics Data Analysis and Interpretation. *Journal of Proteome Research*, 14(3):1557–1565, 2015.

T. Muth, B. Y. Renard, and L. Martens. Metaproteomic data analysis at a glance: advances in computational microbial community proteomics. *Expert Review of Proteomics*, 13(8), 2016.

T. Muth, F. Kohrs, R. Heyer, D. Benndorf, E. Rapp, U. Reichl, L. Martens, and B. Y. Renard. MPA Portable: A Stand-Alone Software Package for Analyzing

Metaproteome Samples on the Go. *Analytical Chemistry*, 90(1):685–689, 2017. doi: 10.1021/acs.analchem.7b03544.

E. W. Myers. A whole-genome assembly of drosophila. *Science*, 287(5461):2196–2204, 2000.

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48:443–453, 1970.

A. I. Nesvizhskii and R. Aebersold. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & cellular proteomics : MCP*, 4:1419–1440, 2005.

S. A. Neville, A. Lecordier, H. Ziochos, M. J. Chater, I. B. Gosbell, M. W. Maley, and S. J. van Hal. Utility of matrix-assisted laser desorption ionization-time of flight mass spectrometry following introduction for routine laboratory bacterial identification. *Journal of clinical microbiology*, 49:2980–2984, 2011.

S. B. Ng, D. A. Nickerson, M. J. Bamshad, and J. Shendure. Massively parallel sequencing and rare disease. *Human Molecular Genetics*, 19(R2):R119–R124, 2010.

A. L. Norris, R. E. Workman, Y. Fan, J. R. Eshleman, and W. Timp. Nanopore sequencing detects structural variants in cancer. *Cancer Biology & Therapy*, 17 (3):246–253, 2016.

H. Ochman, J. G. Lawrence, and E. A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 2000.

H. Ochman, E. Lerat, and V. Daubin. Examining bacterial species under the specter of gene transfer and exchange. *Proceedings of the National Academy of Sciences*, 102(Supplement 1):6595–6599, 2005.

K. Ohta, D. S. Beall, J. P. Mejia, K. T. Shanmugam, and L. O. Ingram. Genetic improvement of Escherichia coli for ethanol production: chromosomal integration of Zymomonas mobilis genes encoding pyruvate decarboxylase and alcohol dehydrogenase ii. *Applied Environmental Microbiology*, 57(4):893–900, 1991.

N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, 44(D1):D733–745, Jan 2016.

L. Olendzenski and J. P. Gogarten. Gene transfer: Who benefits? In *Horizontal Gene Transfer*, pages 3–9. Humana Press, 2009. doi: 10.1007/978-1-60327-853-9_1.

M. Onishi-Seebacher and J. O. Korbel. Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond. *Bioessays*, 33(11):840–850, 2011.

S. Pabinger, A. Dander, M. Fischer, R. Snajder, M. Sperk, M. Efremova, B. Krabichler, M. R. Speicher, J. Zschocke, and Z. Trajanoski. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, 15(2):256–278, 2014.

J. Parkhill, B. W. Wren, N. R. Thomson, R. W. Titball, M. T. G. Holden, M. B. Prentice, M. Sebaihia, K. D. James, C. Churcher, K. L. Mungall, et al. Genome sequence of yersinia pestis, the causative agent of plague. *Nature*, 413(6855): 523–527, 2001. doi: 10.1038/35097083.

B. Paten, J. Herrero, K. Beal, S. Fitzgerald, and E. Birney. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research*, 18(11):1814–1828, 2008.

B. Paten, M. Diekhans, D. Earl, J. S. John, J. Ma, B. Suh, and D. Haussler. Cactus graphs for genome comparisons. *Journal of Computational Biology*, 18(3):469–481, 2011a.

B. Paten, D. Earl, N. Nguyen, M. Diekhans, D. Zerbino, and D. Haussler. Cactus: algorithms for genome multiple sequence alignment. *Genome Research*, 21(9): 1512–1528, 2011b.

A. Penzlin, M. S. Lindner, J. Doellinger, P. W. Dabrowski, A. Nitsche, and B. Y. Renard. Pipasic: similarity and expression correction for strain-level identification and quantification in metaproteomics. *Bioinformatics*, 30(12):i149–i156, 2014.

F. J. Pérez-Llarena and G. Bou. Proteomics as a tool for studying bacterial virulence and antimicrobial resistance. *Frontiers in Microbiology*, 7, 2016.

J. A. Perry, E. L. Westman, and G. D. Wright. The antibiotic resistome: what's new? *Current Opinion in Microbiology*, 21:45–50, 2014.

J. S. Peters, B. Calder, G. Gonnelli, S. Degroeve, E. Rajaonarifara, N. Mulder, N. C. Soares, L. Martens, and J. M. Blackburn. Identification of quantitative proteomic differences between mycobacterium tuberculosis lineages with altered virulence. *Frontiers in Microbiology*, 7:813, 2016.

P. A. Pevzner, H. Tang, and M. S. Waterman. An eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17): 9748–9753, 2001.

P. A. Pevzner, H. Tang, and G. Tesler. De novo repeat classification and fragment assembly. *Genome Res*, 14(9):1786–1796, 2004.

D. Pinkel and D. G. Albertson. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics*, 37(6s):S11–S17, 2005.

D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W.-L. Kuo, C. Chen, Y. Zhai, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20(2):207–211, 1998.

D. Pinto, E. Delaby, D. Merico, M. Barbosa, A. Merikangas, L. Klei, B. Thiruvahindrapuram, X. Xu, R. Ziman, Z. Wang, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *The American Journal of Human Genetics*, 94(5):677–694, 2014.

V. C. Piro, M. S. Lindner, and B. Y. Renard. DUDes: a top-down taxonomic profiler for metagenomics. *Bioinformatics*, 32:2272–2280, 2016.

S. Podell and T. Gaasterland. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biology*, 8(2):R16, 2007.

P. Portin. The birth and development of the DNA theory of inheritance: sixty years since the discovery of the structure of DNA. *Journal of Genetics*, 93:293–302, 2014.

C. Putonti, Y. Luo, C. Katili, S. Chumakov, G. E. Fox, D. Graur, and Y. Fofanov. A computational tool for the genomic identification of regions of unusual compositional properties and its utilization in the detection of horizontally transferred sequences. *Molecular Biology and Evolution*, 23(10):1863–1868, 2006.

J. Qi and F. Zhao. inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic Acids Research*, 39(suppl_2): W567–W575, 2011.

H. Radhouani, L. Pinto, P. Poeta, and G. Igrejas. After genomics, what proteomics tools could help us understand the antimicrobial resistance of escherichia coli? *Journal of Proteomics*, 75(10):2773–2789, 2012.

R. Rahn, D. Weese, and K. Reinert. Journaled string tree-a scalable data structure for analyzing thousands of similar genomes on your laptop. *Bioinformatics*, 30 (24):3499–3505, 2014.

K. R. Rasmussen, J. Stoye, and E. W. Myers. Efficient q-gram filters for finding all epsilon-matches over a given length. *Journal of Computational Biology*, 13(2): 296–308, 2006.

T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 2012.

M. Ravenhall, N. Škunca, F. Lassalle, and C. Dessimoz. Inferring Horizontal Gene Transfer. *PLoS Computational Biology*, 11(5):e1004095, 2015.

R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, et al. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, 2006.

B. Y. Renard, W. Timm, M. Kirchner, J. A. J. Steen, F. A. Hamprech, and H. Steen. Estimating the confidence of peptide identifications without decoy databases. *Analytical Chemistry*, 88(11):4314–4318, 2010.

J. A. Reuter, D. V. Spacek, and M. P. Snyder. High-throughput sequencing technologies. *Molecular Cell*, 58(4):586–597, 2015. doi: 10.1016/j.molcel.2015.05.004.

G. Rizk, A. Gouin, R. Chikhi, and C. Lemaitre. MindTheGap: integrated detection and assembly of short and long insertions. *Bioinformatics*, 30(24):3451–3457, 2014.

J. M. Rothberg, W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, 2011.

L. Salmela and E. Rivals. LoRDEC: accurate and efficient long read error correction. *Bioinformatics*, 30(24):3506–3514, 2014.

S. L. Salzberg and J. A. Yorke. Beware of mis-assembled genomes. *Bioinformatics*, 21(24):4320–4321, 2005.

S. L. Salzberg, A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 22 (3):557–567, 2012.

S. J. Sanders, A. G. Ercan-Sencicek, V. Hus, R. Luo, M. T. Murtha, D. Moreno-De-Luca, S. H. Chu, M. P. Moreau, A. R. Gupta, S. A. Thomson, et al. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 williams syndrome region, are strongly associated with autism. *Neuron*, 70(5):863–885, 2011.

S. Sauer and M. Kliem. Mass spectrometry tools for the classification and identification of bacteria. *Nature Reviews Microbiology*, 8(1):74–82, 2010.

L. Schaeffer, H. Pimentel, N. Bray, P. Melsted, and L. Pachter. Pseudoalignment for metagenomic read assignment. *Bioinformatics*, 2017.

J. A. Schloss. How to get genomes at one ten-thousandth the cost. *Nature Biotechnology*, 26(10):1113–1115, 2008.

J. Schröder, A. Hsu, S. E. Boyle, G. Macintyre, M. Cmero, R. W. Tothill, R. W. Johnstone, M. Shackleton, and A. T. Papenfuss. Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics*, 30(8):1064–1072, 2014.

J. Schröder, A. Wirawan, B. Schmidt, and A. T. Papenfuss. CLOVE: classification of genomic fusions into structural variation events. *BMC Bioinformatics*, 18(1), 2017.

J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, et al. Strong Association of De Novo Copy Number Mutations with Autism. *Science*, 316(5823):445–449, 2007.

K. Sedlar, K. Kupkova, and I. Provaznik. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Computational and Structural Biotechnology Journal*, 15:48–55, 2017.

F. J. Sedlazeck, A. Dhroso, D. L. Bodian, J. Paschall, F. Hermes, and J. M. Zook. Tools for annotation and comparison of structural variation. *F1000Research*, 6: 1795, 2017.

O. Serang and W. Noble. A review of statistical methods for protein identification using tandem mass spectrometry. *Statistics and its interface*, 5:3–20, 2012.

H. M. B. Seth-Smith. Spi-7: Salmonella's vi-encoding pathogenicity island. *Journal of infection in developing countries*, 2:267–271, 2008.

N. Shaikh and P. I. Tarr. Escherichia coli O157:H7 Shiga Toxin-Encoding Bacteriophages: Integrations, Excisions, Truncations, and Evolutionary Implications. *Journal of Bacteriology*, 185(12):3596–3605, 2003.

L. Shi, V. Ravikumar, A. Derouiche, B. Macek, and I. Mijakovic. Tyrosine 601 of bacillus subtilis dnak undergoes phosphorylation and is crucial for chaperone activity and heat shock survival. *Frontiers in Microbiology*, 7:533, 2016.

T. Sicheritz-Ponten. A phylogenomic approach to microbial evolution. *Nucleic Acids Research*, 29(2):545–552, 2001.

J. L. Siefert. Defining the mobilome. In *Horizontal Gene Transfer*, pages 13–27. Humana Press, 2009.

J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol. ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19 (6):1117–1123, 2009.

S. Sindi, E. Helman, A. Bashir, and B. J. Raphael. A geometric approach for classification and comparison of structural variants. *Bioinformatics*, 25(12):i222–i230, 2009.

E. Siragusa, D. Weese, and K. Reinert. Fast and accurate read mapping with approximate seeds and multiple backtracking. *Nucleic Acids Research*, 41(7):e78, 2013.

A. Sirichoat, A. Lulitanond, R. Kanlaya, R. Tavichakorntrakool, A. Chanawong, S. Wongthong, and V. Thongboonkerd. Phenotypic characteristics and comparative proteomics of staphylococcus aureus strains with different vancomycin-resistance levels. *Diagnostic Microbiology and Infectious Disease*, 86(4):340–344, 2016.

D. L. Smith, D. J. Rooks, P. C. Fogg, A. C. Darby, N. R. Thomson, A. J. McCarthy, and H. E. Allison. Comparative genomics of Shiga toxin encoding bacteriophages. *BMC Genomics*, 13(1):311, 2012.

S. D. Smith, J. K. Kawash, and A. Grigoriev. GROM-RD: resolving genomic biases to improve read depth detection of copy number variants. *PeerJ*, 3:e836, 2015.

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, 1981.

M. Snyder, J. Du, and M. Gerstein. Personal genome sequencing: current approaches and challenges. *Genes & Development*, 24(5):423–431, 2010.

G. F. Sprague. Genetic exchange between kingdoms. *Current Opinion in Genetics & Development*, 1(4):530–533, 1991.

P. Stankiewicz and J. R. Lupski. Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61(1):437–455, 2010.

H. Steen and M. Mann. The abc's (and xyz's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5(9):699–711, 2004.

P. J. Stephens, C. D. Greenman, B. Fu, F. Yang, G. R. Bignell, L. J. Mudie, E. D. Pleasance, K. W. Lau, D. Beare, L. A. Stebbings, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):27–40, 2011.

P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015. doi: 10.1038/nature15394.

S. Suzuki, T. Yasuda, Y. Shiraishi, S. Miyano, and M. Nagasaki. ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics*, 12(Suppl 14):S7, 2011.

M. Syvanen. Cross-species gene transfer; implications for a new theory of evolution. *Journal of Theoretical Biology*, 112:333–343, 1985.

The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, oct 2010. doi: 10.1038/nature09534.

The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

The Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, page bbw089, oct 2016. doi: 10.1093/bib/bbw089.

B. Timmermann, M. Kerick, C. Roehr, A. Fischer, M. Isau, S. T. Boerno, A. Wunderlich, C. Barmeyer, P. Seemann, J. Koenig, et al. Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. *PLoS ONE*, 5(12):e15661, 2010.

E. R. Toenshoff, P. D. Fields, Y. X. Bourgeois, and D. Ebert. The end of a 60-year riddle: Identification and genomic characterization of an iridovirus, the causative agent of white fat cell disease in zooplankton. *G3: Genes,Genomes,Genetics*, page g3.300429.2017, 2018.

G. G. Tomazella, K. Risberg, H. Mylvaganam, P. C. Lindemann, B. Thiede, G. A. de Souza, and H. G. Wiker. Proteomic analysis of a multi-resistant clinical escherichia coli isolate of unknown genomic background. *Journal of Proteomics*, 75 (6):1830–1837, 2012.

A. Töpfer, T. Marschall, R. A. Bull, F. Luciani, A. Schönhuth, and N. Beerenwinkel. Viral quasispecies assembly via maximal clique enumeration. *PLoS Computational Biology*, 10(3):e1003515, 2014.

N. H. Tran, M. Z. Rahman, L. He, L. Xin, B. Shan, and M. Li. Complete de novo assembly of monoclonal antibody sequences. *Scientific reports*, 6:31730, 2016.

K. Trappe. Multi-Split-mapping of NGS reads for variant detection. Master's thesis, Department of Computer Science, Freie Universität Berlin, Berlin, Germany, March 2012. URL `http://www.mi.fu-berlin.de/en/inf/groups/abi/theses/master_dipl/trappe/index.html`.

K. Trappe, A.-K. Emde, H.-C. Ehrlich, and K. Reinert. Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. *Bioinformatics*, 30(24):3484–3490, 2014.

K. Trappe, T. Marschall, and B. Y. Renard. Detecting horizontal gene transfer by mapping sequencing reads across species boundaries. *Bioinformatics*, 32(17): i595–i604, 2016.

K. Trappe, B. Wulf, J. Doellinger, S. Halbedel, T. Muth, and B. Y. Renard. Hortense: Horizontal gene transfer detection directly from proteomic ms/ms data. *PeerJ Preprints*, 5:e3248v1, 2017.

D. T. Truong, E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, and N. Segata. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12:902–903, 2015.

P. C. Turner, L. P. Yomano, L. R. Jarboe, S. W. York, C. L. Baggett, B. E. Moritz, E. B. Zentz, K. T. Shanmugam, and L. O. Ingram. Optical mapping and sequencing of the Escherichia coli KO11 genome reveal extensive chromosomal rearrangements, and multiple tandem copies of the Zymomonas mobilis pdc and adhB genes. *Journal of Industrial Microbiology and Biotechnology*, 39(4):629–639, 2012.

E. Tuzun, A. J. Sharp, J. A. Bailey, R. Kaul, V. A. Morrison, L. M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, et al. Fine-Scale structural variation of the human genome. *Nature Genetics*, 37(7):727–732, 2005.

M. Tyers and M. Mann. From genomics to proteomics. *Nature*, 422(6928):193–197, 2003.

D. Valenzuela, N. Välimäki, E. Pitkänen, and V. Mäkinen. On enhancing variation detection through pan-genome indexing. *bioRxiv*, 2015. doi: 10.1101/021444.

A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J. A. Malek, G. Costa, K. McKernan, et al. A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning. *Genome Research*, 18(7):1051–1063, 2008.

S. van Baarle, I. N. Celik, K. G. Kaval, M. Bramkamp, L. W. Hamoen, and S. Halbedel. Protein-protein interaction domains of bacillus subtilis DivIVA. *Journal of Bacteriology*, 195(5):1012–1021, 2012.

L. Van Oudenhove and B. Devreese. A review on recent developments in mass spectrometry instrumentation and quantitative tools advancing bacterial proteomics. *Applied Microbiology and Biotechnology*, 97:4749–4762, 2013.

M. van Passel, A. Bart, H. Thygesen, A. Luyf, A. van Kampen, and A. van der Ende. An acquisition account of genomic islands based on genome signature comparisons. *BMC Genomics*, 6(1):163, 2005.

M. Vaudel, A. Sickmann, and L. Martens. Current methods for global proteome identification. *Expert Review of Proteomics*, 9(5):519–532, 2012.

M. Vaudel, J. M. Burkhart, R. P. Zahedi, E. Oveland, F. S. Berven, A. Sickmann, L. Martens, and H. Barsnes. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature Biotechnology*, 33(1):22–24, 2015.

J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

G. S. Vernikos and J. Parkhill. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the salmonella pathogenicity islands. *Bioinformatics*, 22(18):2196–2203, 2006.

J. Vizcaíno, A. Csordas, N. del Toro, J. Dianes, J. Griss, I. Lavidas, G. Mayer, Y. Perez-Riverol, F. Reisinger, T. Ternent, et al. 2016 update of the pride database and related tools. *Nucleic Acids Research*, 44(D1):D447–D456, 2016.

T. Walsh, J. M. McClellan, S. E. McCarthy, A. M. Addington, S. B. Pierce, G. M. Cooper, A. S. Nord, M. Kusenda, D. Malhotra, A. Bhandari, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, 320(5875):539–543, 2008.

H.-C. Wang, F.-C. Cheng, M.-S. Wu, H.-Y. Shu, H. S. Sun, Y.-C. Wang, I.-J. Su, and C.-J. Wu. Genome Sequences of Three Helicobacter pylori Strains from Patients with Gastric Mucosa-Associated Lymphoid Tissue Lymphoma. *Genome Announcements*, 3(2):e00229–15, 2015.

J. Wang, C. G. Mullighan, J. Easton, S. Roberts, S. L. Heatley, J. Ma, M. C. Rusch, K. Chen, C. C. Harris, L. Ding, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nature Methods*, 8(8):652–654, 2011.

W. Wang, H. Xi, M. Huang, J. Wang, M. Fan, Y. Chen, H. Shao, and X. Li. Performance of mass spectrometric identification of bacteria and yeasts routinely isolated in a clinical microbiology laboratory using MALDI-TOF MS. *Journal of Thoracic Disease*, 6:524–533, 2014.

S. L. Warnes, C. J. Highmore, and C. W. Keevil. Horizontal Transfer of Antibiotic Resistance Genes on Abiotic Touch Surfaces: Implications for Public health. *MBio*, 3(6):e00489–12, 2012.

J. D. Watson and F. H. Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171:964–967, 1953.

N. I. Weisenfeld, S. Yin, T. Sharpe, B. Lau, R. Hegarty, L. Holmes, B. Sogoloff, D. Tabbaa, L. Williams, C. Russ, et al. Comprehensive variation discovery in single human genomes. *Nature Genetics*, 46(12):1350–1355, 2014.

J. Wiedenbeck and F. M. Cohan. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiology Reviews*, 35(5):957–976, 2011.

M. H. F. Wilkins, A. R. Stokes, and H. R. Wilson. Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. *Nature*, 171(4356): 738–740, 1953.

P. Wilmes and P. L. Bond. Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends in Microbiology*, 14(2):92–97, 2006.

K. Wong, T. M. Keane, J. Stalker, and D. J. Adams. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biology*, 11(12):R128, 2010.

D. E. Wood and S. L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014.

P. Worning, L. J. Jensen, K. E. Nelson, S. Brunak, and D. W. Ussery. Structural analysis of dna sequence: evidence for lateral gene transfer in thermotoga maritima. *Nucleic acids research*, 28:706–709, 2000.

J. D. Wuitschick and K. M. Karrer. Analysis of genomic g + c content, codon usage, initiator codon context and translation termination sites in tetrahymena thermophila. *The Journal of eukaryotic microbiology*, 46:239–247, 1999.

R. Xi, A. G. Hadjipanayis, L. J. Luquette, T.-M. Kim, E. Lee, J. Zhang, M. D. Johnson, D. M. M. ande David A. Wheeler, R. A. Gibbs, R. Kucherlapati, et al. Copy number variation detection in whole-genome sequencing data using the bayesian information criterion. In *Proceeding of the National Academy of Science of the United States of America*, volume 108, pages E1128–36, 2011.

C. Xie and M. T. Tammi. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, 10(1):80, 2009.

F. Yang. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. *Nucleic Acids Research*, 33(19):6445–6458, 2005.

L. Yang, L. J. Luquette, N. Gehlenborg, R. Xi, P. S. Haseley, C.-H. Hsieh, C. Zhang, X. Ren, A. Protopopov, L. Chin, et al. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, 153(4):919–929, 2013.

K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, 2009.

S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*, 19(9): 1586–1592, 2009a.

S. Yoon, Z. Xuan, K. Ye, and J. Sebat. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*, 19:1586–92, 2009b.

Y. Yu, H. S. Kim, H. Chua, C. Lin, S. Sim, D. Lin, A. Derr, R. Engels, D. DeShazer, B. Birren, et al. Genomic patterns of pathogen evolution revealed by comparison of burkholderia pseudomallei, the causative agent of melioidosis, to avirulent burkholderia thailandensis. *BMC Microbiology*, 6(1):46, 2006. doi: 10.1186/1471-2180-6-46.

O. Zagordi, A. Bhattacharya, N. Eriksson, and N. Beerenwinkel. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics*, 12(1):119, 2011.

B. Zeitouni, V. Boeva, I. Janoueix-Lerosey, S. Loeillet, P. Legoix-né, A. Nicolas, O. Delattre, and E. Barillot. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, 26 (15):1895–1896, 2010.

D. R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, 2008.

J. Zhang and Y. Wu. SVseq: an approach for detecting exact breakpoints of deletions with low-coverage sequence data. *Bioinformatics*, 2011.

Y. Zhang, C. Laing, M. Steele, K. Ziebell, R. Johnson, A. K. Benson, E. Taboada, and V. P. Gannon. Genome evolution in major Escherichia coli O157:H7 lineages. *BMC Genomics*, 8(1):121, 2007.

Q. Zhu, M. Kosoy, and K. Dittmar. HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. *BMC Genomics*, 15 (1):717, 2014.

Zusammenfassung

Strukturvariationen (SVs) haben eine immense Bedeutung im Genom sämtlicher Spezies. Sie sind das Ergebnis fundamentaler Rekonstruktionsmechanismen und verleihen gleichzeitig Bakterien die Fähigkeit, sich an ihre Umgebung anzupassen. In Bakterien gibt es zudem das Phänomen des horizontalen Gentransfers (HGT), bei dem Gene über Speziesgrenzen hinweg von einem Donor-Individuum zu einem anderen Akzeptor übertragen werden. Die Integration eines neuen Gens kann auf genomischer Ebene untersucht werden. Die Aktivität und Expression hingegen lässt sich nur auf Proteinebene bestimmen.

In dieser Doktorarbeit werden bioinformatische Methoden zur Detektion von komplexen SVs unterschiedlichen Typs und Größe anhand von *Next- generation Sequencing* Daten und proteomischen Massenspektrometriedaten mit einem Fokus auf HGT-Events vorgestellt. Bei einem HGT-Event muss zunächst bestimmt werden, zwischen welchen Organismen der Transfer stattgefunden hat und welche Gene aus dem Donor an welcher Stelle im Akzeptor eingefügt wurden. Anschließend kann man untersuchen, ob das transferierte potentielle Protein auch funktionell ist.

Als erstes wird das SV-Detektionstool Gustaf vorgestellt, welches eine bessere Auflösung bezogen auf Größe und Typ von SVs im Vergleich zu vorherigen Methoden ermöglicht. Einen besonderen Vorteil bietet Gustaf in der Charakterisierung von komplexen Translokationen und Duplikationen als Kombination von simpleren, im Genom voneinander entfernten Varianten. Mit dieser generischen Methode als Basis wurden zwei mapping-basierte Methoden, Daisy und DaisyGPS, zur HGT-Detektion entwickelt. Daisy verwendet Gustaf und weitere SV-Detektionsstrategien um die transferierte Region im Donorgenom und ihre Insertionsstelle im Akzeptorgenom präzise zu bestimmen. DaisyGPS verwendet etablierte Strategien für die metagenomische Bestimmung von Mikroorganismen in einer Probe, um eine passende Akzeptor- und Donorreferenz zu identifizieren. Daisy und DaisyGPS basieren auf Sequenzvergleichen und heben sich damit von den bisher existierenden Methoden ab, welche HGTs anhand von Sequenzkompositionsmustern und phylogenetischen Inkonsistenzen bestimmen. Im letzten Projekt wird die proteomische Methode Hortense vorgestellt. Hortense erweitert die Standarddatenbanksuche von Spektren um eine umfassende Kreuzvalidierung, um definierte Eigenschaften eines HGT-Proteins sicher zu stellen. Alle drei Methoden zur HGT-Detektion ermöglichen eine ganzheitliche Analyse von HGT-Events, welche vorher oder nur mit einer einzelnen der drei Methoden nicht möglich wäre.

Eigenständigkeitserklärung

Ich versichere, dass ich die hier vorgelegte Dissertation selbstständig angefertigt habe und die benutzten Quellen und Hilfsmittel vollständig angegeben sind.

Ein Promotionsverfahren wurde zu keinem früheren Zeitpunkt an einer anderen in- oder ausländischen Hochschule oder bei einem anderen Fachbereich beantragt. Die Bestimmungen der Promotionsordnung sind mir bekannt.

Kathrin Trappe, Berlin, 27.08.2018