Hortense: Horizontal gene transfer detection directly from proteomic MS/MS data

Kathrin Trappe[1], Ben Wulf[1], Joerg Doellinger[2], Sven Halbedel[3], Thilo Muth[1], Bernhard Y. Renard[1]

[1] Bioinformatics Unit (MF1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, Berlin, Germany

[2] Division of Proteomics and Spectroscopy (ZBS 6), Centre for Biological Threats and Special Pathogens, Robert Koch Institute, Berlin, Germany

[3] Division of Enteropathogenic Bacteria and Legionella (FG 11), Robert Koch Institute, Wernigerode, Germany

Corresponding Author:

Thilo Muth[1]

Nordufer 20, Berlin, 13353, Germany

Email address: MuthT@rki.de

Hortense: Horizontal gene transfer detection directly from proteomic MS/MS data

Kathrin Trappe[1], Ben Wulf[1], Joerg Doellinger[2], Sven Halbedel[3], Thilo Muth[1♯*], and Bernhard Y. Renard[1♯]

[1]**Bioinformatics Unit (MF1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, Berlin, Germany**
[2]**Division of Proteomics and Spectroscopy (ZBS 6), Robert Koch Institute, Berlin, Germany**
[3]**Division of Enteropathogenic Bacteria and Legionella (FG 11), Robert Koch Institute, Wernigerode, Germany**
[*]**corresponding author, ♯ joint last author**

ABSTRACT

Horizontal gene transfer (HGT) is a powerful mechanism that allows bacteria to directly transfer long stretches of genomic sequence from one individual to another. The transfer of antimicrobial resistance genes is a prominent example of HGT events in the context of multi-resistant bacteria which pose a high risk to human health. While several approaches for HGT detection exist on the genomic level, to the best of our knowledge, HGT events have not been investigated in a detailed mass spectrometry (MS)-based proteomic study. However, the mere presence of a gene does not necessarily correlate with its expression at the protein level. Consequently, to draw conclusions with respect to the expression of HGT-mediated genes, MS-based proteomics can be employed. We developed a first computational approach - called Hortense - for automated HGT detection directly from shotgun proteomics experiments. We extend the standard database search by a critical cross-validation to unravel potential HGT proteins. A proteogenomic extension gives information about the genomic origin and enables an integration with existing genome-based methods.
We successfully validated our approach on simulated data, and further evaluated it on real data from a transgenic organism and a negative control from an organism not harboring a transferred gene. Our results indicate that our method facilitates MS-based analysis for proteomic evidence of HGT events. Especially as an orthogonal approach to genome-based HGT detection methods, our proposed workflow is a first step toward a systematic and large scale analysis of HGT events in, e.g., antimicrobial resistance context. Hortense is publicly available at https://gitlab.com/rki_bioinformatics/.

Keywords: horizontal gene transfer, proteomics, protein identification, tandem mass spectrometry

1 INTRODUCTION

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 & Szöllősi, 2016). Bacteria have at least three commonly known mechanisms for this transfer (see Figure 1). 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 & Boerlin, 2013). The impact of this powerful mechanism was only recently recognized with the advent of genome sequencing (Daubin & 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

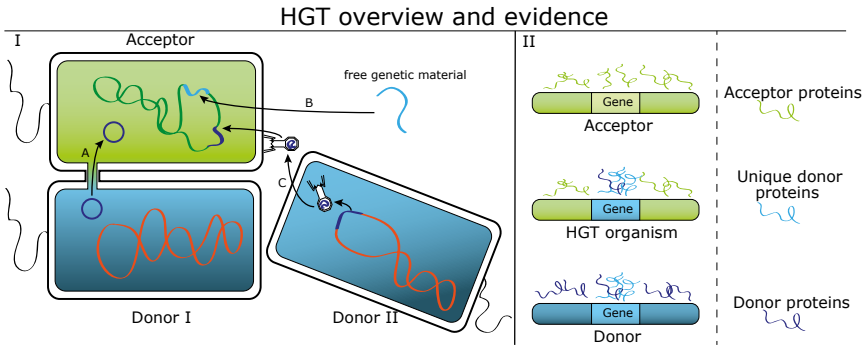


Figure 1. HGT overview and evidence. (I) Via horizontal gene transfer (HGT), genetic material is transferred from the donor cell to the acceptor cell by one of three possible ways. The genetic material can be part of a plasmid (A) that is exchanged directly between the acceptor and donor cell. Bacteria 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).

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 & 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 of microbial species which has been approved for clinical applications (Sauer & 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 & 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 & Aebersold, 2005) and various solutions have been proposed on this topic (Serang & 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 & 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 & 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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3248v1 | CC BY 4.0 Open Access | rec: 13 Sep 2017, publ: 13 Sep 2017

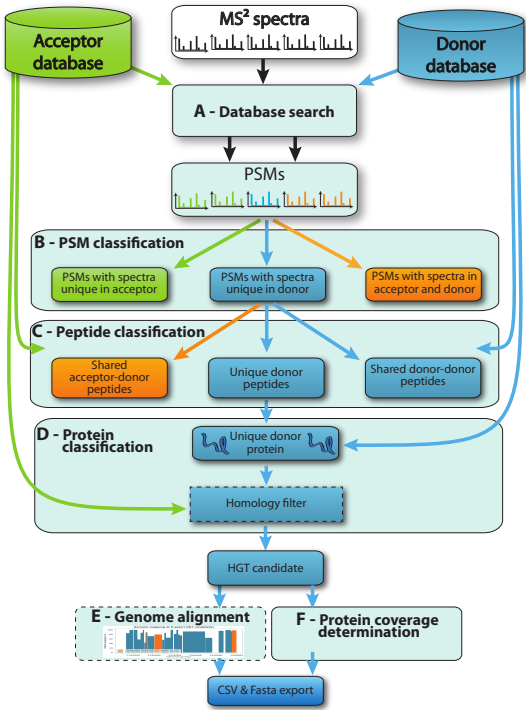


Figure 2. Hortense evidence and workflow. (A) In a first step, the MS/MS spectra are searched separately against the acceptor and donor databases. (B) The resulting PSMs are classified to be either uniquely matching to the acceptor or donor, or shared between both. The unique donor PSMs are filtered 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).

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 & 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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3248v1 | CC BY 4.0 Open Access | rec: 13 Sep 2017, publ: 13 Sep 2017

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 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 & Renard, 2012).

2 METHODS

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). 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.

Database search. The search engine MS-GF+ (Beta (v10089) (7/16/2014)) (Kim & 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 ($-t$), fragmentation method identifier ($-m$), and required memory limits. For now, static MS-GF+ values for decoy database search (true), Orbitrap/FTICR, and enzyme identifier (trypsin) are used ($-tda\,1 - inst\,1 - e\,1$). 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.

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3248v1 | CC BY 4.0 Open Access | rec: 13 Sep 2017, publ: 13 Sep 2017

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.

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).

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.

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.

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 & Rahmann, 2012). Parameter settings are enabled via a configuration file so that the whole pipeline can be automatically executed with one program call.

3 EXPERIMENTAL SETUP

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 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^2PIP (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). 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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3248v1 | CC BY 4.0 Open Access | rec: 13 Sep 2017, publ: 13 Sep 2017

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.

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.

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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3248v1 | CC BY 4.0 Open Access | rec: 13 Sep 2017, publ: 13 Sep 2017

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. 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. 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.

4 RESULTS

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 3 B, *H.pylori*). The naive filtering approach (see Figure 3 A, *H.pylori*) can reduce this number to 284 seemingly unique donor proteins. But since only the 27 transferred proteins should be present, the naive filtering resulted therefore in 257 false positive (FP) reports (including possible one-hit wonders). This means, without further filtering, one would have to investigate 257 protein candidates regarding a possible HGT property. Applying our pipeline, we can drastically 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 3 B and Supplementary Table S1). Figure 4a 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.

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 2317 FP unique donor protein candidates (without filtering for one-hit wonders). 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 (see supplementary result files for a list of all reported proteins). This large list contradicts a single HGT event and would be regarded as evidence for a double infection. The naive filtering again 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.

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 an MS-GF+ search on

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3248v1 | CC BY 4.0 Open Access | rec: 13 Sep 2017, publ: 13 Sep 2017

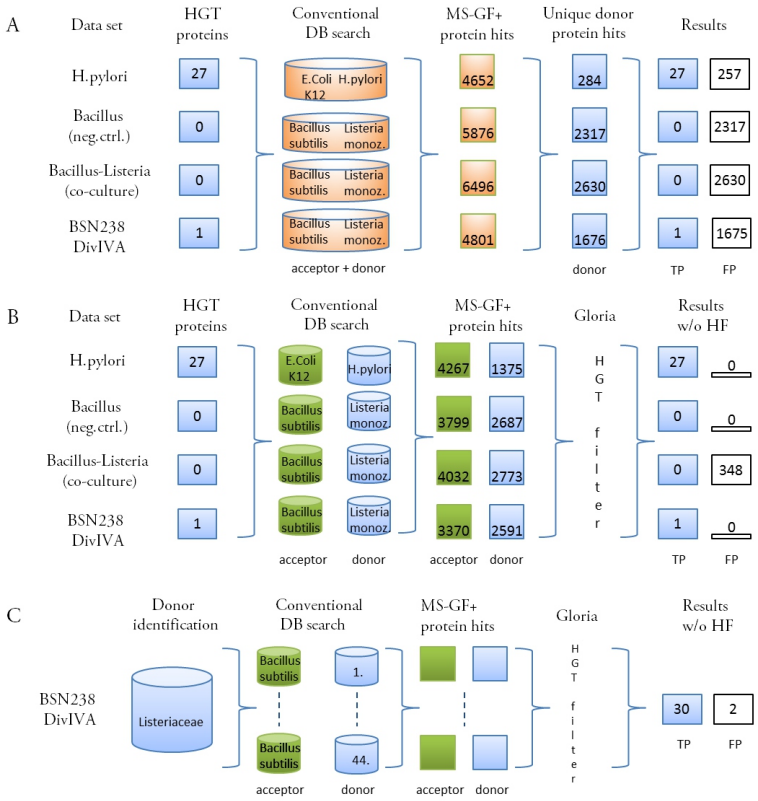
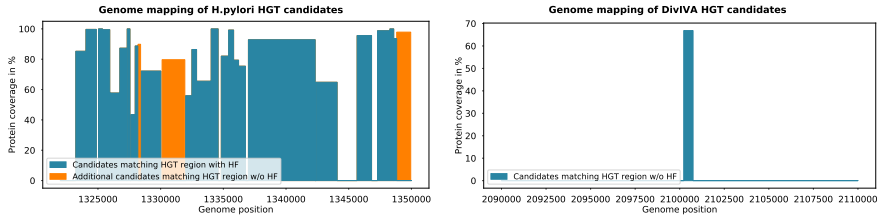


Figure 3. Results for Hortense compared to naive filtering. Column *HGT proteins* states the number of known HGT proteins per data set. (A) For the naive filtering, all spectra were mapped against a combined acceptor+donor database (orange). The resulting *MS-GF+ protein hits* (orange) were filtered for unique donor protein hits (no match on the acceptor for this spectrum) resulting in 284 for the *H.pylori*, 2317 and 2630 for the negative *Bacillus* and co-culture data sets, resp., and 1676 for the DivIVA. 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 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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3248v1 | CC BY 4.0 Open Access | rec: 13 Sep 2017, publ: 13 Sep 2017



(a) Genome alignment *H.pylori*. **(b)** Genome alignment DivIVA.

Figure 4. 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* data set, 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%.

these proteomes and then ran the pipeline on all reported references (see Figure 3 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 3 A and B, and Table S1). The pipeline 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 4b). The naive filtering reports another 1675 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 DISCUSSION

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3248v1 | CC BY 4.0 Open Access | rec: 13 Sep 2017, publ: 13 Sep 2017

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 & 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.

ACKNOWLEDGMENTS

We thank Stephan Fuchs for inspiring discussions, and Samuel Hauf for an excellent contribution to the laboratory work (cultivation and protein extraction). *Funding:* We gratefully acknowledge financial support by Deutsche Forschungsgemeinschaft (DFG), grant number RE3474/2-1 to BYR.

REFERENCES

Clark, C. G., Kruczkiewicz, P., Guan, C., McCorrister, S. J., Chong, P., Wylie, J., van Caeseele, P., Tabor, H. A., Snarr, P., Gilmour, M. W., Taboada, E. N., & Westmacott, G. R. (2013). Evaluation of maldi-tof mass spectroscopy methods for determination of escherichia coli pathotypes. *Journal of microbiological methods*, 94, 180–191.

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

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

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

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

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

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

Ishihama, Y., Rappsilber, J., & Mann, M. (2006). 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.

Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S. J., Selley, J. N., Searle, B. C., Shofstahl, J., Seymour, S. L., Julian, R., Binz, P.-A., Deutsch, E. W., Hermjakob, H., Reisinger, F., Griss, J., Vizcaíno, J. A., Chambers, M., Pizarro, A., & Creasye, D. (2012). The mzIdentML data standard for mass spectrometry-based proteomics results. *Molecular and Cellular Proteomics*, 11(7).

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

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

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

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

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

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

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

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

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

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

Neville, S. A., Lecordier, A., Ziochos, H., Chater, M. J., Gosbell, I. B., Maley, M. W., & van Hal, S. J. (2011). 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table S1. Detailed Hortense results for all datasets (Figure 3 II). Listed are all reported HGT protein candidates (including one-hit wonders) found without homology filter (HF). For the *H.pylori*, all HGT candidates not marked with an * are also reported by Hortense with HF turned on. Column *Protein coverage* states the fraction of the HGT protein covered by all supporting peptides (listed in column *Supporting peptides*. Column *Genomic region* states the genomic position of the HGT candidate determined through Hortenses genome alignment step, if applicable. All one-hit wonders would be dismissed as a valid HGT candidate and are hence marked with a "no" in column *HGT protein*.

Protein HGT candidate	Protein coverage	Supporting peptides	Genomic region	HGT protein
***H. pylori* data set**				
WP_060870030.1 hypothetical protein	99.25	17	1335820, 1335422	yes
WP_060870029.1 hypothetical protein	85.24	45	1324041, 1323232	yes
WP_016059995.1 hypothetical protein	100.0	14	1327555, 1327331	yes
WP_060869215.1 UDP-glucose 4-epimerase GalE	3.20	1	270258, 271289	no
WP_060869824.1 hypothetical protein	93.59	10	1348827, 1348594	yes
WP_060869823.1 hypothetical protein	100.0	24	1348589, 1348284	yes
WP_060869822.1 transcriptional regulator	98.78	59	1348284, 1347298	yes
WP_060869821.1 helicase DnaB	95.58	64	1346821, 1345667	yes
WP_060869820.1 hypothetical protein	64.82	54	1344076, 1342397	yes
WP_060869819.1 hypothetical protein	92.85	276	1342326, 1337002	yes
WP_060869818.1 hypothetical protein	75.41	16	1336763, 1336215	yes
WP_060869817.1 hypothetical protein	79.39	19	1336215, 1335823	yes
WP_060869816.1 hypothetical protein	82.05	20	1335422, 1334838	yes
WP_060869815.1 hypothetical protein	100.0	36	1334586, 1334029	yes
WP_060869814.1 structural protein	65.52	37	1334014, 1332884	yes
WP_060869813.1 hypothetical protein	86.40	13	1332866, 1332492	yes
WP_060869812.1 hypothetical protein	55.94	16	1332426, 1331998	yes
WP_060869810.1 hypothetical protein	72.20	63	1330013, 1328460	yes
WP_060869808.1 hypothetical protein	88.73	5	1328182, 1327970	yes
WP_060869807.1 holin	43.52	4	1327961, 1327641	yes
WP_060869806.1 hypothetical protein	87.29	23	1327286, 1326744	yes
WP_060869805.1 hypothetical protein	57.74	19	1326739, 1325945	yes
WP_060869804.1 hypothetical protein	99.48	29	1325942, 1325367	yes
WP_060869803.1 hypothetical protein	100.0	19	1325364, 1325047	yes
WP_060869802.1 hypothetical protein	99.65	48	1324873, 1324010	yes
WP_060869875.1 methionine ABC transporter ATP-binding protein *	6.12	1	1427838, 1426858	no
WP_001269094.1 phosphoenolpyruvate synthase *	1.11	1	1248049, 1250481	no
WP_060869825.1 site-specific integrase *	97.86	67	1349945, 1348827	yes
WP_060869811.1 hypothetical protein *	79.67	70	1331930, 1330116	yes
WP_060869809.1 hypothetical protein *	89.86	13	1328402, 1328196	yes
BSN238 DivIVA data set				
RS10_LISMO 30S ribosomal protein S10	16.67	1	n.a.	no
Q8Y5N7_LISMO DivIVA protein	66.86	12	2100750, 2100226	yes
SODM_LISMO Superoxide dismutase [Mn]	6.93	1	n.a.	no
***Bacillus* data set (negative control)**				
RL10_LISMO 50S ribosomal protein L10	4.82	1	n.a.	no
Q8Y9N2_LISMO Lmo0493 protein	3.92	1	n.a.	no
Q8Y831_LISMO Lmo1087 protein	7.62	1	n.a.	no
ATPB2_LISMO ATP synthase subunit beta 2	4.02	1	n.a.	no

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3248v1 | CC BY 4.0 Open Access | rec: 13 Sep 2017, publ: 13 Sep 2017

Table S2. Detailed results for the comprehensive HGT search with the DivIVA dataset (Figure 3 III). Listed are all 44 ascertained donor candidates. Donor candidates corresponding to a strain proteome are indented. Column *DivIVA hit* states whether the DivIVA or a homologous protein was reported for the candidate. If it was reported (+), column *Supporting peptides* states the number of observed peptides. The number of other reported false positive HGT proteins - if any - is listed in the column *Hits on other proteins*. *Note: Listeria fleischmannii FSL S10-1203, and Listeria floridensis FSL S10-1187 have two (at least homologous) domains of the DIvIVA protein listed as two separate proteins, both domains were reported and here unified to one protein hit with summed peptide count.

Donor candidate	DivIVA hit	Supporting peptides	Hits on other proteins
Brochothrix campestris FSL F6-1037	+	2	-
Brochothrix thermosphacta	-	-	-
Brochothrix thermosphacta DSM 20171 = FSL F6-1036	+	2	-
Listeria aquatica FSL S10-1188	+	4	-
Listeria booriae	+	3	-
Listeriaceae bacterium FSL A5-0209	+	4	-
Listeria cornellensis FSL F6-0969	+	5	-
Listeria fleischmannii 1991	+	7	1
Listeria fleischmannii FSL S10-1203 *	+	6	-
Listeria fleischmannii subsp. coloradonensis	+	5	1
Listeria floridensis FSL S10-1187 *	+	6	-
Listeria grandensis FSL F6-0971	+	5	-
Listeria grayi	-	-	-
Listeria grayi DSM 20601	+	6	-
Listeria grayi FSL F6-1183	+	6	-
Listeria innocua	-	-	-
Listeria innocua ATCC 33091	+	9	-
Listeria innocua serovar 6a strain ATCC BAA-680 CLIP 11262	+	10	-
Listeria ivanovii	-	-	-
Listeria ivanovii strain ATCC BAA-678 PAM 55	+	12	-
Listeria ivanovii subsp. ivanovii	-	-	-
Listeria ivanovii subsp. londoniensis	+	10	-
Listeria marthii	-	-	-
Listeria marthii FSL S4-120	+	8	-
Listeria monocytogenes	+	12	-
Listeria monocytogenes 36-25-1	-	-	-
Listeria monocytogenes FSL F2-208	+	9	-
Listeria monocytogenes serotype 1/2a strain 10403S	+	12	-
Listeria monocytogenes serotype 4a strain HCC23	-	-	-
Listeria monocytogenes serotype 4a strain M7	+	10	-
Listeria monocytogenes serotype 4b strain CLIP80459	-	-	-
Listeria monocytogenes serotype 4b strain F2365	-	-	-
Listeria monocytogenes serotype 4b str. LL195	+	9	-
Listeria monocytogenes serovar 1/2a strain EGD-e	+	12	-
Listeria newyorkensis	+	5	-
Listeria riparia FSL S10-1204	+	3	-
Listeria rocourtiae	-	-	-
Listeria rocourtiae FSL F6-920	+	5	-
Listeria seeligeri	+	12	-
Listeria seeligeri FSL N1-067	+	10	-
Listeria weihenstephanensis	-	-	-
Listeria weihenstephanensis FSL R9-0317	-	-	-
Listeria welshimeri	-	-	-
Listeria welshimeri serovar 6b strain ATCC 35897 DSM 20650 SLCC5334	+	10	-

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3248v1 | CC BY 4.0 Open Access | rec: 13 Sep 2017, publ: 13 Sep 2017