Tracing the genomic footprint of natural competence in bacteria

Declaration

I hereby confirm that this is my own work and that I have documented all sources used.

Signed: _____

Date: _____

Abstract

Whether natural competence is a common feature across bacteria?

How can we Trace a signal for natural competence?
#Characterize the evolutionary footprint that natural competence leaves in bacterial genomes
#Differentiate it from other means of gene acquisition via horizontal gene transfer.

- The role of DNA uptake for the evolution of A. baumannii
- The role of HGT in the genome innovation of A. baumannii
- How evolutionary history changes along bacterial genome
- Assessing the relevance of natural competence
- To what extent individual bacteria use natural competence for genetic innovation
- Investigating a link between natural competence and the evolution of virulence in Acinetobacter baumannii
- Whether NC is likely to be common feature across A.b or limited to number of them which serve as gateways for new genes to enter
- Effect of competence on the evolution of bacterial genome in general
- Assess the frequency and the distribution of genes with HGT history
- The evolutionary origins of laterally acquired gene investigations
- How does story change along the genome?
- Would natural competency play a main role in A. baumannii genetic innovation and have a significant impact on its' evolution?
- Is there any signature that shows that a gene is directly taken up? (Visualizing distribution pattern along the genome)
-

Zusammenfassung

Table of contenet

INDEX OF FIGURES **III**

INDEX OF TABLES **IV**

ABBREVIATIONS **V**

1 INTRODUCTION **1**

1.1 ACINETOBACTER 1

1.2 ACINETOBACTER BAUMANNII 1

1.3 PHYLOGENETIC TREE 1

1.4 SPECIES TREE & GENE TREE 1

1.5 HOMOLOGS, ORTHOLOGS & PARALOGS 1

1.6 HORIZONTAL GENE TRANSFER 1

1.6.1 Conjugation *1*

1.6.2 Transduction *1*

1.6.3 Transformation *1*

1.7 OBJECTIVE 1

2 MATERIALS AND METHODS **2**

2.1 DATABASE 2

2.2 PHYLOGENETIC TREES 2

2.2.1 Species tree & gene trees *3*

2.2.1.1 Tree test 4

2.2.1.2 Phylogenetic network 4

2.2.1.3 Robinson-Foulds distance 4

2.3 MEGAN 4

2.4 HGTECTOR 5

2.5 IGV AND GVIEW 8

2.6 HAMSTR & FACT 8

2.7 PHYLOPROFILE 9

2.8 ZERO-ONE DISTRIBUTION VECTOR **ERROR! BOOKMARK NOT DEFINED.**

2.9 GENE AGE ESTIMATION 11

2.10 BLAST2GO 12

3 RESULTS **13**

3.1 TRACING A GENOMIC FOOTPRINT OF NATURAL COMPETENCY 13

3.1.1 Identifying horizontally gene transferred candidates *13*

3.1.1.1 Phylogenetics tree reconciliation & tree tests 13

3.1.1.2 Taxonomy assignment approach 13

3.1.1.3 HGT detector tools 13
3.1.2 Detection competence machineries *13*
3.1.3 Distribution of HGT candidates along the genome *13*
3.1.4 0/1 pattern for HGT candidates *13*
3.1.5 Last common ancestor of HGT candidates (Origin) *13*
3.1.6 Recent gene gain and loss events in strains *13*
3.2 PRESENT-ABSENT PATTERNS OF THE CANDIDATES AND THEIR ORTHOLOGS 13
3.3 ROLE OF DNA UPTAKE FOR THE GENOME INNOVATION OF *A. BAUMANNII* 13
3.4 CONTAMINATIONS VERSUS HGTS IN ASSEMBLIES 13
3.5 OUTER MEMBRANE AND EXTRACELLULAR PROTEINS 13
3.6 COMM AND COMC GENES IN TYPEIV PILI MACHINERY 13
4 DISCUSSION 14
5 CONCLUSION & OUTLOOK 15
5.1 CONCLUSION 15
5.2 OUTLOOK 15
5.2.1 Small RNA *15*
5.2.2 Effect of host human products on natural transformation *15*
5.2.3 Analyzing hotspots along the genome (homologous recombinatio *15*
REFERENCES 16
APPENDIX 19
Tables *19*
Figures *20*
ACKNOWLEDGEMENTS 21
CURRICULUM VITAE 21

Index of Figures

Figure 1: *The figure displays the HGTector workflow. Later, the result of HGTector can be visualized by Gview. 7*

Figure 2: The upper part of this example shows three domains and their position in seed and query taxa (ortholog). The lower part presents a schema for how the FAS score is calculated.....9

Figure 3: This example displays how the output looks like in Phyloprofile. The dots indicate presence/absence of ortholog hits and their colors plus background color perform the confidence of the two additional layers.....10

Figure 4: schema of a circle genome in which genes are annotated by 0-1. Green sign shows single event 11

Index of Tables

Abbreviations

DB	Database
A.baumannii	*Acinetobacter baumannii*
OUT	Operational taxonomic unit
OMA	Orthologous matrix
HOG	hierarchical orthologous group
OG	OMA group
ML	Maximum likelihood
SH	Shimodaira-Hasegawa
KH	Kishino-Hasegava
AU	Approximately Unbiased
HGT	Horizontal gene transfer
KDE	Kernel density estimation
OEPs	Outer membrane extracellular proteins
pHMM	Profile hidden Markov model
FAS	Feature architecture similarities
Gff	General feature format
v	Version
LCA	Last common ancestor

1 Introduction

1.1 Acinetobacter

1.2 Acinetobacter baumannii

1.3 Phylogenetic tree

A phylogenetic tree presents the evolutionary relationships among organisms which share a common ancestor. Any operational taxonomic unit (OTU) such as species or genes can form these entities. Nowadays, study of the molecular phylogenies is fundamental step in evolutionary topics. In addition, the phylogenetic analysis performances are improved by the automation of DNA sequencing and plenty of computer programs.

1.4 Species tree & gene tree

move along the bacteria genome and detect whether there are inconsistencies between gene and species evolutionary history or not!

And if there is a difference what can be the reason

1.5 Homologs, Orthologs & Paralogs

1.6 Horizontal gene Transfer

1.6.1 Conjugation

1.6.2 Transduction

1.6.3 Transformation

1.7 Objective

2 Materials and Methods

During this study, we worked with different tools, algorithms and methods. In this section we provide and describe all the necessary information to redo our experiments as well as analyses and reproduce the results.

2.1 Database

The foundation of this study includes a number of databases (DBs) in order to assess an accurate and a comprehensive analysis.

In total, we analyzed 3079 *Acinetobacter* genomes including 2474 *A.baumannii* strains (3052 NCBI RefSeq sequences (O'Leary et al., 2016) (NCBI RefSeq DB, 2018) plus 27 isolated strains). Regarding faster and better performance, we constructed a representative taxon set from the NCBI bacterial DB which is covering the currently known phylogenetic diversity (almost 90%) of the *Acinetobacter* genus. Therefore, individual analyses were run on this subset of 233 representative strains.

Moreover, we extracted hierarchical orthologous groups (HOG) as well as orthologous groups (OG) for this subset from OMA database (2017) (Roth, Gonnet, & Dessimoz, 2008) (Altenhoff, Gil, Gonnet, & Dessimoz, 2013).

Furthermore, the subset of the latest bacterial genomes of NCBI DB (O'Leary et al., 2016) (NCBI RefSeq DB, 2018) were downloaded. This DB represents 116,327 genomes in which 1074 *Acinetobacter* species/strains are placed (738 of them are *A.baumannii* strains).

2.2 Phylogenetic trees

A phylogenetic tree is a graphical representation of the evolutionary relationships among entities that share a common ancestor. It tells us stories about these organisms. There are four key steps to reconstruct the phylogenetic tree: select a sequence of interest, identify orthologs, align sequences and finally calculate phylogeny tree.

2.2.1 Species tree & gene trees

As a primary step, we reconstructed the species tree and the gene trees based on our initial selected 15 *Acinetobacter* orthologous groups (a test set) and then expanded the trees to our whole representative data from this genus (233 species/strains). To this purpose, we first identified the orthologs among the species of interest using HaMStR (2.6) (Ebersberger, Strauss, & von Haeseler, 2009) and aligned the multiple sequences using MAFTT (v7.394) (Katoh, Misawa, Kuma, & Miyata, 2002). Then, we concatenated the alignments using scripts.

Further, Phylip file was formed by submitting the concatenated aligned orthologs to ClustalW (Larkin et al., 2007) for inferring phylogenies. Additionally, ProtTest (v3.4) (Abascal, Zardoya, & Posada, 2005) was applied to estimate the best amino acid substitution model and jModelTest (v2.1.7) (Posada, 2008) was run to select the model of nucleotide substitution. As a result the (L+G+F) model and the (GTR+I+G) model were applied for the protein and nucleotide sequence analysis, respectively. Finally, we ran the RAxML (v8.2.11) (Stamatakis, 2014) to reconstruct the maximum likelihood (ML) tree. In other words, a tree which represents maximum probability of observation of our alignments giving initial tree and the best model.

In order to assess the robustness of the tree as well as estimate the confidence of its branches, we activated the option to make 100 bootstrap replicates (Felsenstein, 1985) from our data in RAxML. As mentioned, we constructed the phylogenetic tree on both the protein sequence level and the nucleotide sequence level. Eventually, we demonstrated our result trees applying FigTree (v1.4.2) (Rambaut, 2009).

Additionally, just for a small group of taxa (15 *Acinetobacter* species/strains), we analyzed only the variable sites of alignments (at protein level) and removed the constant sites in order to speed up the approach.

Following up our analysis, we explored whether the computed gene trees were significantly supported by the species tree or not. To this purpose, we applied

three methods: Tree test assessment, phylogenetic network reconstruction and Robinson-Foulds distance calculation. The three methods are described in greater detail below.

2.2.1.1 Tree test

In this step, we tested tree topology among the gene trees in order to determine whether the gene trees are significantly different from the species tree or they explain the data well, identically. Regarding this, we applied three known likelihood-based tests of topologies in phylogenetics; Kishino-Hasegava (KH) (Goldman, Jon P. Anderson, Allen G., 2000), Shimodaira-Hasegawa (SH) (H. Shimodaira & Hasegawa, 2001) and Approximately Unbiased (AU) (Hidetoshi Shimodaira, 2002). These non-parametric bootstrapping methods compare multiple topologies by calculating p-values to assay the significance of phylogenetic incongruence. Among these tests, AU is the latest one and it is less biased than the other methods.

2.2.1.2 Phylogenetic network

A computed unrooted phylogenetic network was shaped from the gene trees using the consensus network method in SplitsTree4 (v4.13.1) (Huson & Bryant, 2006). For this step only those gene trees were included that showed incongruence in the tree topology test (2.2.1.1) (the calculated p-values for these gene trees were less than 0.05 in the tests). Moreover, the confidence of each branch was computed to discover the strength of these differences.

2.2.1.3 Robinson-Foulds distance

We measured the Robinson-Foulds (Robinson & Foulds, 1981) distance between the set of unrooted *Acinetobacter* gene trees in order to check how significant are the differences among the topology of the trees are.

2.3 MEGAN

MEGAN (v6) (Huson, Auch, Qi, & Schuster, 2007) is a logical, quick and scalable taxonomic assignment approach. We applied this tool to find genes that were

potentially the result of a horizontal gene transfer and to use this information to identify the last common ancestor of these genes.

To proceed with this analysis, we searched for local sequence similarity of our query sequences to a reference database using DIAMOND (v0.8.38.100) (Buchfink, Xie, & Huson, 2015). DIAMOND is a local sequence alignment tool, which is faster than common BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) and has a high sensitivity. Further, we imported the output from DIAMOND into MEGAN, applying a minimum score threshold of 50 and no low complexity filter for the taxonomic assignment, and estimated a taxonomic classification of the sequences.

2.4 HGTector

We scrutinized plenty of tools to identify horizontally acquired genes in bacterial genomes, with respect to cover both parametric and phylogenetic methods. Ultimately, among all these tools, HGTector (v0.2.2) (Zhu, Kosoy, & Dittmar, 2014), offers accurate, fast and genome-wide analysis for a large scale of data, applying an implicit phylogenetic method. And more importantly, it was the one which performed best in our analysis.

The DB applied for this analysis was the whole set of bacterial genomes in NCBI Refseq DB (2018) including in total 116,327 genomes, of which 738 are *A. baumannii* strains, and 336 are other *Acinetobacter* species/strains.

The HGTector workflow begins with an all-against-all BLASTP search of each protein sequence of interest against the DB. The tool records each hit by its bit score and normalizes it later to a value between 0 and 1. Here, HGTector defines three relational hierarchical categories; Self, Close and Distal which are scaled to species, genus and rest of all other Bacteria in our study, respectively.

The application specifies a weight set for each category: Self, Close and Distal. These weight sets are nothing but accumulative normalized bit-scores of genes. In other words, the tool calculates for each gene three bit scores (self, close and distal) in the DB. Afterwards, it selects proper hits considering chosen options,

normalizes them within 0-1 range, sums up all scores belonging to each group and computes corresponding fingerprints. Therefore, each gene has three calculated accumulative normalized scores (self score, close score and distal score) (**Figure 1**) .

At this point, an appropriate cutoff value is calculated for each fingerprint. This value divides the weight distributions into typical and atypical. The cutoff controls the confidence of HGT prediction, it can be strict or relaxed. The program has statistical approaches to calculate the cutoff value or as another option a user is free to apply any other statistics. We chose the kernel density estimation (KDE) cutoff for our investigation. It is necessary to know that each genome analysis generates distinct fingerprints, thus cutoffs may vary. With this in mind, the following rules apply to above results to detect horizontally-derived genes:

1. The gene must have an accumulative bit score below cutoff in the close weight distribution.
2. The gene must have an accumulative bit score equal or above cutoff in the distal weight distribution.

The first rule represents that the orthologs of the query genes are absent in most of the organisms in the species of the taxon of interest (close group). In other words, the hits belonging to this sister group are significantly underrepresented. One possible scenario can be that the gene underwent multiple gene loss events but the lower the bit score (close score) the higher the probability of the second scenario pops up. This scenario declares that the gene was horizontally acquired. The second rule indicates whether the hits from distant organisms are overrepresented for the corresponding gene or not. The higher the bit score in the distal group the more probable the HGT candidate.

In conclusion, when both criteria are fulfilled, the tool infers that the candidate has a putative HGT history. **Error! Reference source not found.** depicts the whole process step by step.

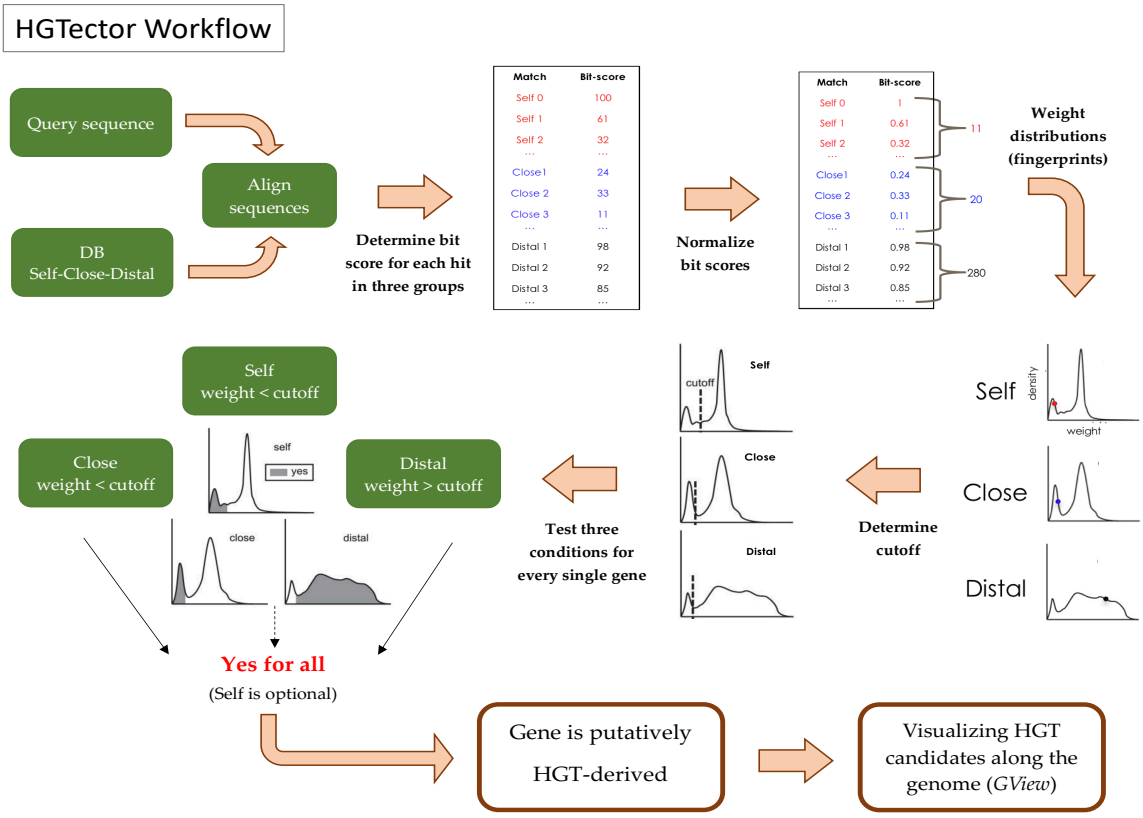

Figure 1: *The figure displays the HGTector workflow. Later, the result of HGTector can be visualized by GView.*

2.5 IGV and GView

One of the important goal during our study was the visualization of a gene or genes of interest along the genome. Initially we deployed IGV (Thorvaldsdóttir, Robinson, & Mesirov, 2013) to view our genomic regions. Since we were working with bacterial genomes which includ only one chromosome and we were interested to have a circular view, hence we continued our analysis with GView (v1.7) (Petkau, Stuart-Edwards, Stothard, & van Domselaar, 2010). In order to create a GView genome map, we just required to upload genome annotation files in the general feature format (GFF) as an input. In this respect, we visualized the position of the HGT candidates along the corresponding taxon. We need to keep in mind that the tool works properly only when the complete genome of the organism is available.

2.6 HaMStR & FACT

In our analysis, HaMStR (v13.2.9) (Ebersberger et al., 2009), an inclusive, targeted orthology prediction tool, played an important role.

We applied this tool to screen for orthologs to the competence apparatus genes plus other interesting proteins such as outer membrane extracellular proteins (OEPs) and HGT-derived genes in taxa of interest.

In simple words, HaMStR uses profile hidden Markov models (pHMM) (Eddy, 1998) to search for orthologs in given taxa. The analysis starts with compiling a core-set of orthologs and generating a pHMM for each sequence cluster in these core orthologs. Subsequently, the pHMMs are applied to search for possible orthologs among query species. Eventually, the candidates must pass a reciprocal best BLAST hit test against the reference species. As a result, reciprocal best hits infer orthologs of corresponding genes and the core-set can be extended using the new information (Figure A 1) (**Figure 2**A).

After tracing the orthologs for genes of interest, we mined deeper to identify their functional equivalents. To this end, we applied FACT (v1.5.1) (Koestler, von Haeseler, & Ebersberger, 2010) and scored feature architecture similarities (FAS)

between protein pairs. These features constitute of Pfam domains, SMART domains, secondary structure elements, transmembrane domains and low complexity regions (**Figure 2**B) . The FAS score, which is a combination of the original multiplicity score and the positional score, is calculated between a seed protein and one of its orthologs (Figure 2C). It lies within the range from 0 (no shared features) to 1 (the ortholog contains all features of the seed protein).

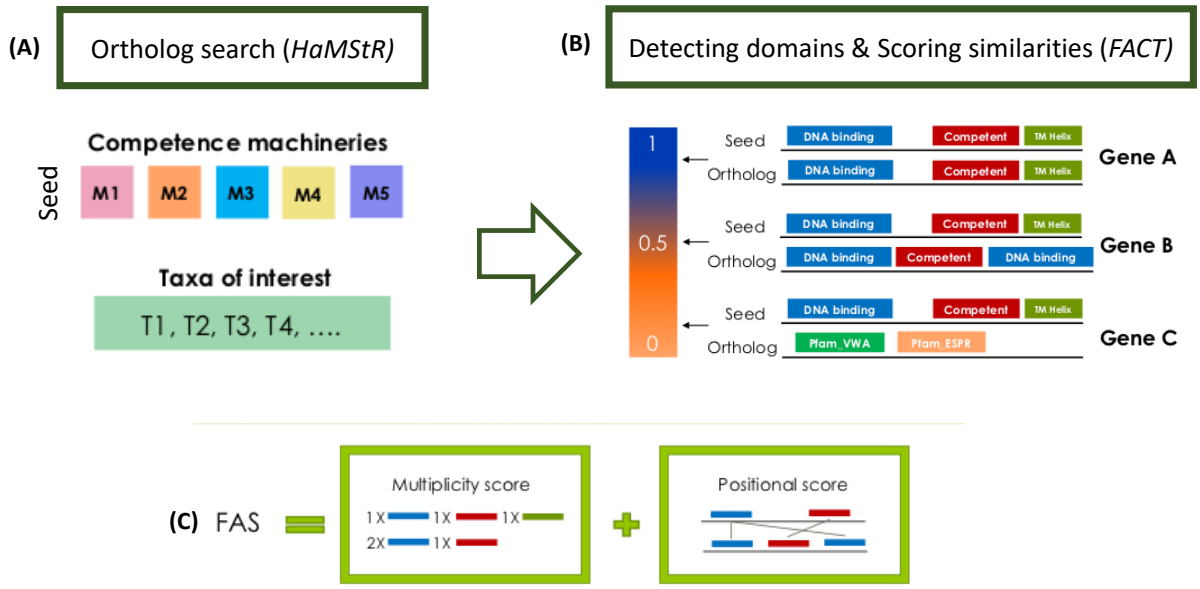

Figure 2: *Here, we present a schema for how HaMStR and FACT work together. In part A we search for orthologs for a corresponding gene and further in part B, we identify the domains for ortholog search result. The lower part (C) presents an example for the FAS score calculation.*

2.7 PhyloProfile

Along our study, we worked with PhyloProfile (Tran, Greshake Tzovaras, & Ebersberger, 2018), an effective, user-friendly and multi-function tool, to visualize and explore FAS phylogenetic profiles as well as HGT score results. The program depicts multiple layers including presence/absence pattern plus two additional layers of information. All details have to be gathered in a proper format which can be either long matrix or wide matrix to feed into PhyloProfile. Moreover, the taxa can be ordered according to a user-defined tree. In profile appearance, the presence or absence of small circles displays the existence or non-existence of orthologs. The color of the present circles shows the confidence of

the first given score category (the darkest blue is the highest range and the lightest orange shows the lowest range). The background tone indicates the certainty of the second given score category (the higher the scores, the darker the yellow tone) (**Figure 3**). One can also provide the tool with domain architecture data and thereby it can visualize domain architecture similarities between orthologs or pairwise equivalence of other protein features.

According to all of our analysis, presence/absence of orthologs (colored circles) provided the first layer of information, while our input for extra layers varied. These additional layers were covered either by bidirectional FAS scores (seed -> ortholog, ortholog -> seed) from the HaMStR/FACT (2.6) analysis or by close and distal scores from the HGTector analysis (2.4).

Also, we uploaded the phylogenetic species tree from 233 representative *Acinetobacter* species/strains to set the output in phylogeny distance order.

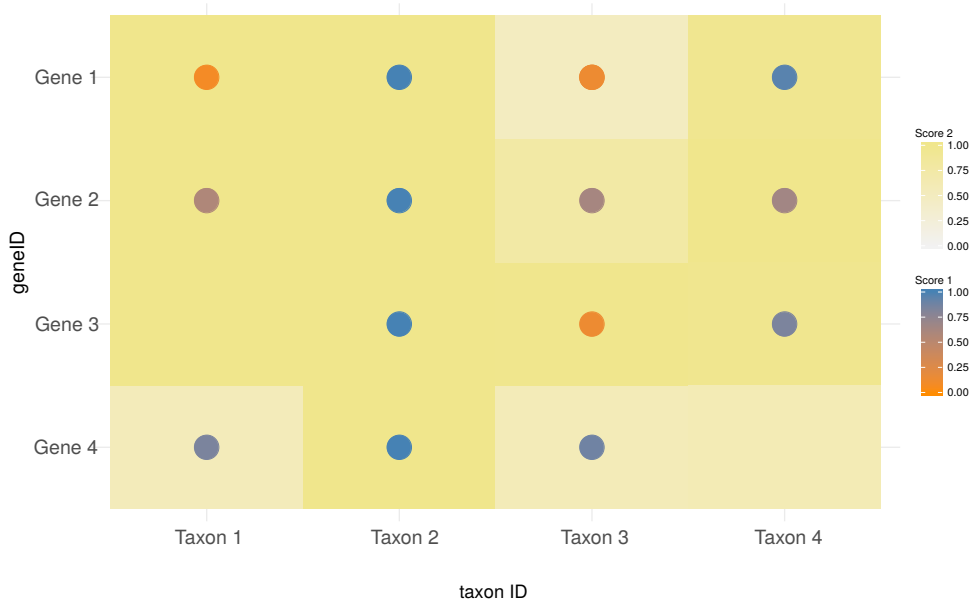

Figure 3: *This example displays how the output looks like in PhyloProfile. The dots indicate presence/absence of ortholog hits and their colors plus background color perform the confidence of the two additional layers.*

2.8 Spatial proximity of HGT candidates

In previous steps, we specified genes with horizontally history (HGTector 2.4) and identified their position along the corresponding genome and the plasmids (GView 2.5). For this part, our main focus was on highlighting the candidates

which integrated into the genome as an isolated HGT event with higher resolution. Therefore, we visualized a pattern in which 1 and 0 represent a gene with horizontal or vertical history, respectively (**Figure 4**).

Accordingly, the isolated HGT event can be one or a few genes which located next to genes inherited from the ancestor but not in an island including many HGT events.

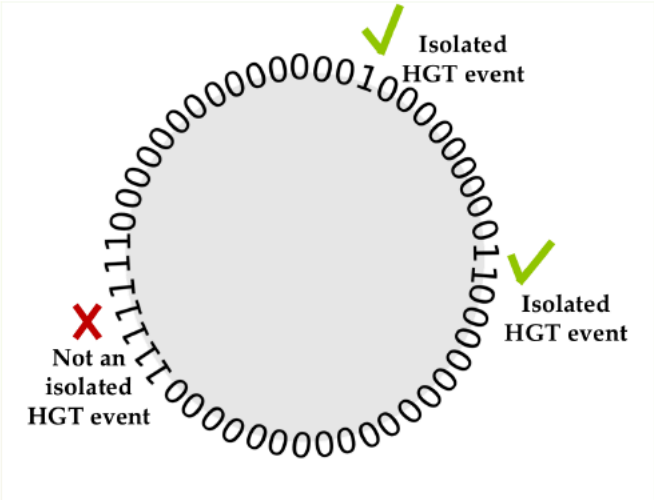

***Figure 4**: schema of a circular genome in which genes are annotated by 0-1, presenting whether the gene was vertically or horizontally transferred, respectively.*

2.9 Age estimation of HGT events

In this part, we were in search for the origin of the HGT-derived genes. Therefore, we dated these candidates. To this end, we computed pairwise orthologs using the OMA algorithm and created orthologous groups (OGs) including core orthologs (Roth et al., 2008; Train, Glover, Gonnet, Altenhoff, & Dessimoz, 2017). Also, we specified hierarchical orthologous groups (HOGs) including orthologs and in-paralogs (Altenhoff et al., 2013). Besides, we applied PhyloProfile (2.7) to visualize phylogenetic presence/absence pattern from the orthologs of the HGT candidates in taxa of interest. At this point, we started to approximate the date of an HGT event by integrating a last common ancestor (LCA) with an interpretation of the phylogenetic profile. This corrects overestimated age estimates due to more than one HGT event in the clade of interest (Figure 5).

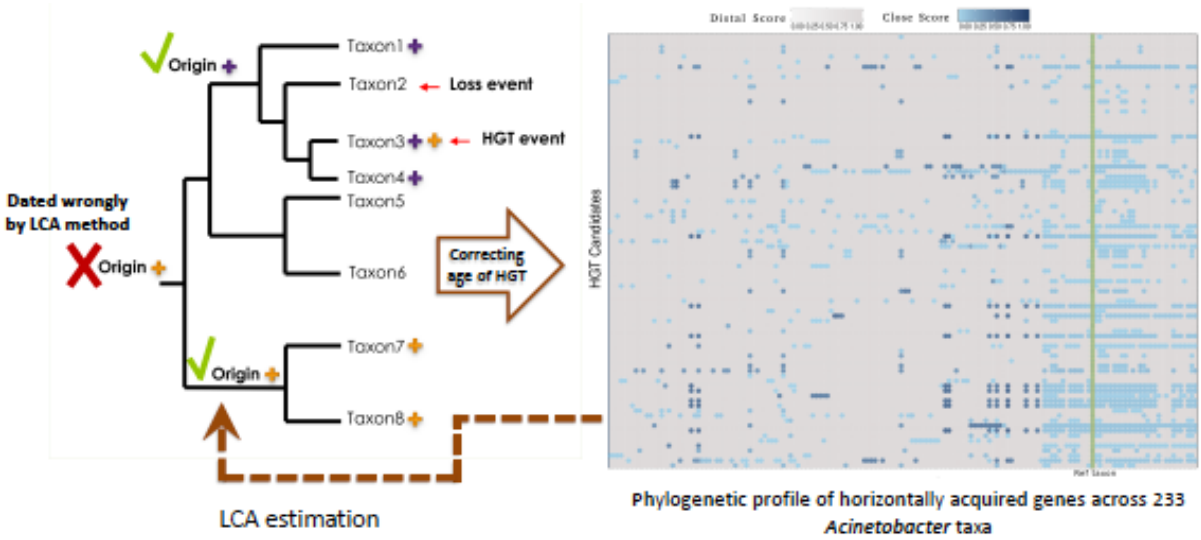

Figure 5: schema for dating the HGT events

2.10 Blast2GO

Eventually, one of the objectives in our analysis was characterizing the main biological role of genes which have been horizontally acquired. Therefore, we primarily considered the NCBI (NCBI RefSeq DB, 2018) functional annotation from Gff files for each gene. Further, we made these annotations more comprehensive by identifying corresponding GO terms. For this purpose, we applied Blast2GO (v5.2) (Conesa & Götz, 2008). To make the process faster, we initially blasted our data against the bacterial DB in NCBI applying DIAMOND (v0.8.38.100) (Buchfink et al., 2015) and later passed the search results on to Blast2GO and continued with gene ontology mapping and annotation.

3 Results

3.1 Tracing a genomic footprint of natural competency

3.1.1 Detection competence machineries

3.1.2 Identifying horizontally gene transferred candidates

3.1.2.1 Phylogenetics tree reconciliation & tree tests

3.1.2.2 Taxonomy assignment approach

3.1.2.3 HGT detector tools

3.1.3 Distribution of HGT candidates along the genome

3.1.4 0/1 pattern for HGT candidates

3.1.5 Last common ancestor of HGT candidates (Origin)

3.1.6 Recent gene gain and loss events in strains

3.2 Present-absent patterns of the candidates and their orthologs

3.3 Role of DNA uptake for the genome innovation of *A. baumannii*

3.4 Contaminations versus HGTs in assemblies

We can talk here about Avian project

3.5 Outer membrane and Extracellular proteins

3.6 ComM and ComC genes in typeIV pili machinery

4 Discussion

5 Conclusion & outlook

5.1 Conclusion

5.2 Outlook

5.2.1 Small RNA

5.2.2 Effect of host human products on natural transformation

5.2.3 Analyzing hotspots along the genome (homologous recombinatio

References

Abascal, F., Zardoya, R., & Posada, D. (2005). ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bti263

Altenhoff, A. M., Gil, M., Gonnet, G. H., & Dessimoz, C. (2013). Inferring Hierarchical Orthologous Groups from Orthologous Gene Pairs. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0053786

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*. https://doi.org/10.1016/S0022-2836(05)80360-2

Buchfink, B., Xie, C., & Huson, D. (2015). Fast and sensitive protein alignment using {DIAMOND}. *Nat Methods*. https://doi.org/10.1038/nmeth.3176

Conesa, A., & Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics*, *2008*, 619832. https://doi.org/10.1155/2008/619832

Ebersberger, I., Strauss, S., & von Haeseler, A. (2009). HaMStR: Profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology*. https://doi.org/10.1186/1471-2148-9-157

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics (Oxford, England)*, *14*(9), 755–763. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9918945

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*. https://doi.org/10.2307/2408678

Goldman, Jon P. Anderson, Allen G., N. (2000). Likelihood-Based Tests of Topologies in Phylogenetics. *Systematic Biology*. https://doi.org/10.1080/106351500750049752

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*. https://doi.org/10.1101/gr.5969107

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in

evolutionary studies. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msj030

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkf436

Koestler, T., von Haeseler, A., & Ebersberger, I. (2010). FACT: Functional annotation transfer between proteins with similar feature architectures. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-11-417

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., McWilliam, H., … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btm404

NCBI RefSeq DB. (2018). Bacterial RefSeq db. Retrieved from ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv1189

Petkau, A., Stuart-Edwards, M., Stothard, P., & van Domselaar, G. (2010). Interactive microbial genome visualization with GView. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq588

Posada, D. (2008). jModelTest: Phylogenetic Model Averaging. *Molecular Biology and Evolution*, *25*(7), 1253–1256. https://doi.org/10.1093/molbev/msn083

Rambaut, A. (2009). Figtree.

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*. https://doi.org/10.1016/0025-5564(81)90043-2

Roth, A. C. J., Gonnet, G. H., & Dessimoz, C. (2008). Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-9-518

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree

selection. *Systematic Biology*. https://doi.org/10.1080/10635150290069913

Shimodaira, H., & Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/17.12.1246

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu033

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbs017

Train, C. M., Glover, N. M., Gonnet, G. H., Altenhoff, A. M., & Dessimoz, C. (2017). Orthologous Matrix (OMA) algorithm 2.0: More robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. In *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btx229

Tran, N.-V., Greshake Tzovaras, B., & Ebersberger, I. (2018). PhyloProfile: dynamic visualization and exploration of multi-layered phylogenetic profiles. *Bioinformatics*. https://doi.org/10.1101/302109

Zhu, Q., Kosoy, M., & Dittmar, K. (2014). HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. *BMC Genomics*. https://doi.org/10.1186/1471-2164-15-717

Appendix

Tables

Figures

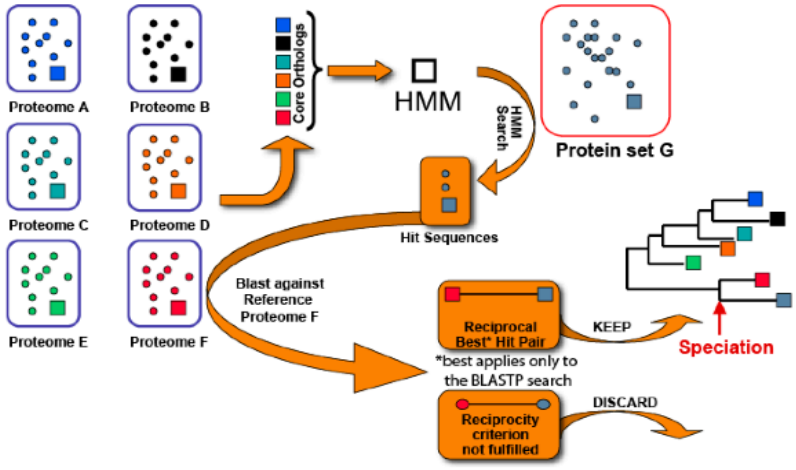

Figure A 1: The HaMStR approach workflow extracted from (Ebersberger et al., 2009)

Acknowledgements

Curriculum Vitae