Project Deliverable
Workpackage

Responsible Partner:

Contributing partners:

GENERAL INFORMATION

European Joint Programme full title	**Promoting One Health in Europe through joint actions on foodborne zoonoses, antimicrobial resistance and emerging microbiological hazards**
European Joint Programme acronym	**One Health EJP – RaDAR: Risk and Disease burden of Antimicrobial Resistance**
Funding	**This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.**
Grant Agreement	**Grant agreement n° 773830**
Start Date	**01/01/2018**
Duration	**24 Months**

DOCUMENT MANAGEMENT

Project deliverable	GWAS-based method for genomic data analysis / Development of regression model for genomic data analysis
Project Acronym	D141-D1.6 & D142-D1.7
Author	Pierre-Emmanuel DOUARRE
Other contributors	
Due month of the report	36
Actual submission month	
Type *R: Document, report* *DEC: Websites, patent filings, videos, etc.;* *OTHER*	R, DEC, other **Save date**: 3-Dec-20
Dissemination level *PU: Public (default)* *CO: confidential, only for members of the consortium (including the Commission Services)*	PU **This is the default setting.** If this project deliverable should be confidential, please add justification here (may be assessed by PMT): ....
Dissemination *Author's suggestion to inform the following possible interested parties.*	OHEJP WP 1 ☐ OHEJP WP 2 ☐ OHEJP WP 3 ☐ OHEJP WP 4 ☐ OHEJP WP 5 ☐ OHEJP WP 6 ☐ OHEJP WP 7 ☐ Project Management Team ☐ Communication Team ☐ Scientific Steering Board ☐ National Stakeholders/Program Owners Committee ☐ EFSA ☐ ECDC ☐ EEA ☐ EMA ☐ FAO ☐ WHO ☐ OIE ☐ Other international stakeholder(s): Social Media: ..... **Other recipient(s):** .....

Background

Genome Wide Association Studies (GWAS) are hypothesis-free methods for identifying genetic variations associated with particular phenotypic traits within a population (Juran and Lazaridis, 2011;Visscher et al., 2017). Microbial genome-wide association studies (mGWAS) are a new and exciting research field that is adapting human GWAS methods to understand how variations in microbial genomes affect host or pathogen phenotypes (Power et al., 2017).

Given the availability of large panels of bacterial genomes combined with phenotypic data in public databases, GWAS have shown promising results for genetic marker discovery and as emerged as a fundamental task in bacterial genomics (Falush, 2016). GWAS will provide microbiologist with an enhanced insight into genotype to phenotype correlations, including complex traits such as virulence, persistence, biofilm formation, epidemicity, host preference and antibiotic resistance (Laabei et al., 2014;Brynildsrud et al., 2016;Lees et al., 2017;Jaillard et al., 2018;Fritsch et al., 2019).

Principle and methods

GWAS aim to identify the genetic basis of phenotypic traits using the variation that exists within natural populations. The genome sequence and phenotypic properties are determined for a collection of strains. Statistical tests are performed to see if particular genetic variants are more common in strains that have the phenotype than in those that do not. Genetic variants that are associated with traits are candidates to explain the variation in the trait that is seen within the population (Visscher et al., 2017).

The most common methods are based on single nucleotide polymorphisms (SNPs), defined by aligning all genomes of the studied panel against a reference genome, and on the gene presence/absence of a collection of genes. However, the use of a reference genome becomes unsuitable when working on bacterial species with a large accessory genome. Focusing on the effects of SNPs alone will miss the acquisition of genes potentially introduce by recombination. On the other hand, methods focusing on genes are unable to cover variants in noncoding regions and some poorly studied species still lack a representative annotation. A third approach has relied on k-mers (all nucleotide substrings of length k found in the genomes) which can account for diverse genetic events such as the acquisition of SNPs, insertions/deletions and recombinations. While k-mers can reflect any genomic variation in a panel, they do not themselves represent biological entities (Power et al., 2017).

Bacterial association mapping is technically challenging due to the unique characteristics of bacterial populations. In comparison to human GWAS, the confounding factors of the microbial GWAS include genome selection, homologous recombination events, population structure, as well as genome wide significance (Vila Nova et al., 2019).

To determine best practices for microbial GWAS, it is essential to compare current GWAS methods in terms of their performance across a range of realistic effect sizes, recombination rates and sample sizes. For this end, Saber et al., recently developed a bacterial GWAS simulator (BacGWASim) to generate bacterial genomes with varying rates of mutation, recombination and other evolutionary parameters, along with a subset of causal mutations underlying a phenotype of interest (Saber and Shapiro, 2020). This simulator could be very useful to assess the performance and optimize different GWAS pipelines.

Tools and Pipelines

Following the first tool computing GWAS with a correction of Eukaryotic population structure based on SNPs (PLINK) in 2007 (Purcell et al., 2007), several computational tools and methods following different approaches have been developed to facilitate the discovery of novel mutations that are associated with the phenotypes of interest. Traditional microbial-based GWAS tools can be broadly categorized into four categories: (a) phylogeny, (b) non phylogeny, (c) hybrid tools that implement a combination of statistical and phylogenetic methods and (d) machine learning . For example, CCTSWEEP and VENN use phylogenetic trees to find correlations between SNPs that are statistically significant (Habib et al., 2007), GWAMAR implements various statistical methods such as mutual information, odds ratio, hypergeometric test and weighted support to associate the phenotypes with point mutations (Wozniak et al., 2014) while the machine learning approach of PhenotypeSeeker identifies phenotype-specific k-mers, generates phenotype prediction model and predicts the phenotype from sequencing data (Aun et al., 2018). With the growing number of different GWAS softwares available, the choice of tool, methods or workflows presents a major challenge to biologists. Here we present in Table 1 a summary of bioinformatics tools and pipelines available for microbial GWAS and highlight their advantages and limtations.

Workflow	Mapping	Analysis	PSC*	Advantages	Limitations	Citation
PLINK	SNPs	Linear and logistic regression of allele count at SNPs	NO			*Purcell, S. et al. 2007*
CCTSWEEP & VENN	SNPs	Phylogenetic trees to find correlations	NO	Consider missing data	Only works well for a large number of SNPs	*Habib et al.,2007*
RoadTrips	SNPs	Association analysis of SNP effect, allowing random variables to account for sample relatedness	NO	Corrects for provided or derived relatedness between samples		*Thornton & McPeek 2010*
PhyC	SNPs	Identify SNPs undergoing recent convergent evolution	YES			*Faraht et al., 2013*
GWAMAR	Genes + SNPs	Computes several statistical scores (mutual information, odds ratio, hypergeometric test, weighted support and TGH	NO		Do not predict epistatic interactions Ignores levels of gene expression	*Wozniak et al., 2014*
Bugwas	SNPs + genes + kmers	Principal components and linear mixed models from GEMMA	YES	Detect polySNP and polygenic effects when multiple low effect variants are responsible for the phenotype		*Earle et al.,2016*
Kover	kmers	Machine learning	NO	Support indels and large-scale genomic rearrangements Several models available on the website	Low sensitivity	*Drouin et al., 2016*
SEER	SNPs + genes + kmers	Linear and logistic regression using kmers, simultaneously testing SNPs and gene presence or absence	YES	Handle large datasets assembled and unassembled output includes effect sizes, direction, and standard error	Complexity	*Lees et al., 2016*
PySEER	SNPs + genes + kmers	Generalized linear models to test for associations between each k-mer	YES	Interactive visualizations (Phandango) Estimation of possible lineage effects / Support InDels		*Lees et al., 2016*
Scoary	Genes	Score the components of the pan-genome	YES	Categorical Phenotype and population structure correction	Not designed to handle large sample	*Brynildsrud et al., 2016*
TreeWAS	SNPs + genes + kmers	Statistical associations between a phenotype and genotype at all loci + confounding effects correction	YES	Recombination inference Supports categorical and continuous phenotype	Complexity	*Collins and Didelot, 2018*
DBGWAS	kmers	Characterize the genomic environment of a k-mer at population level. Relies on GEMMA & Bugwas	YES	Support rearrangements and InDels, Web-based interface Can handle very large datasets	Complexity	*Jaillard et al., 2018*
Phenotype Seeker	kmers	Machine learning (phenotype predictive models pre-trained)	YES	Easy to use and very fast Can handle very large datasets		*Aun et al., 2018*
HAWK	kmers	Component analysis and logistic regression to identify kmers	YES	Fast (Multi-threading) Identify InDels and structural variations such as copy number variations		*Rahman et al., 2018*
microbial-GWAS	Genes + SNPs	Integrate Linear Mixed Model from GEMMA	YES	Support small InDels from the core genome		*Vila Nova et al., 2019*

* PSC = Population Structure correction

Table 1: summary of bioinformatics tools and pipelines available for microbial GWAS

Applications

Over the last 10 years, microbial GWAS had been implemented to explore a diversity of interesting phenotypes. A summary of microbial GWAS applications is presented in Table 2.

Successulf GWAS include the identification of genomic and metabolic signatures in *Salmonella enterica* that were associated to different animal sources (Vila Nova et al., 2019) and the detection of genes encoding vitamin B5 biosynthesis in Campylobacter isolates that were responsible for adaptation to cattle (Sheppard et al., 2013). Another food pathogen was also investigated by Fritsch et al., 2018 who identified a number of genes and SNPs, as well as specific phylogenetic sub-lineages that were associated to cold adaptation of *Listeria monocytogenes*.

In another study, Laabei et al., identified a large number of loci that was significantly associated with toxicity of methicillin resistant *Staphylococcus aureus* (Laabei et al., 2014) while Galardini et al., characterized genetic determinants responsible for extra-intestinal virulence in *Escherichia coli* (Galardini et al., 2020).

Another pan-genome-wide association study identified prophage sequences as being associated with decreased carriage duration of *Streptococcus pneumoniae*, potentially by disruption of the competence mechanism (Lees et al., 2017). Furthermore, mGWAS was used by Davies et al., to determine vaccine candidate coverage from 2083 Group A Streptococcus genomes (Davies et al., 2019).

Microbial GWAS also provide new opportunities to develop insights into the biological mechanisms that underlie antimicrobial resistance. GWAS is a promising tool for identifying common genetic variants associated with antibiotic resistance and rare mutations in candidate gene that can also be associated with resistance phenotypes. Sevaral studies have been used to detect novel mutations associated with drug resistance in *Mycobacterium tuberculosis* (Farhat et al., 2013;Wozniak et al., 2014;Jaillard et al., 2018). In another study, Alam et al., successfully applied GWAS in Vancomycin-intermediate *Staphylococcus aureus* and recovered known mutations of the rpoB gene and also identified rare mutations in a set of candidate genes (walKR, vraSR, graSR, and agrA) associated with intermediary phenotype (Alam et al., 2014). Few other GWAS studies also identified genes acquired by horizontal transfer in *Staphylococcus aureus* (Jaillard et al., 2018) and *Staphylococcus epidermidis* (Brynildsrud et al., 2016).

Phenotype	Species	Sample	Workflow	Reference
Cold persistence	*L. monocytogenes*	51	Scoary + GEMMA	*Fritsch et al., 2018*
Source atribution	*S. enterica*	440	microbial-GWAS	*Vila Nova et al., 2019*
Preferential host	*C. jejuni*	192	Bespoke	*Sheppard et al., 2013*
Virulence	*S. aureus*	90	Plink	*Laabei et al., 2014*
Virulence	*E. coli*	370	PySEER	*Galardini et al. 2019*
Virulence	*K. pneumoniae*	167	Phenotype Seeker	*Aun et al., 2018*
Vaccine candidate	*Group A Streptococcus*	2083	PySEER	*Davies et al., 2019*
Susceptibility	*B. anthrasis*	15	CCTSWEEP & VENN	*Habib et al., 2007*
Carriage	*S. pneumoniae*	2175	fast-lmm & SEER	*Lees et al., 2017*
Antimicrobial resistance	*M. tuberculosis*	123	PLINK	*Chen et al., 2015*
Antimicrobial resistance	*M. tuberculosis*	123	PhyC	*Faraht et al., 2013*
Antimicrobial resistance	*M. tuberculosis - S. aureus*	1398 - 100	GWAMAR	*Wozniak et al., 2014*
Antimicrobial resistance	*S. pneumoniae - S. epidermidis*	3085 -50	Scoary	*Brynildsrud et al., 2016*
Antimicrobial resistance	*M. tuberculosis, S. aureus, P. aeruginosa*	1302, 992, 282	DBGWAS	*Jaillard et al., 2018*

Table 2: Summary of microbial GWAS

GWAS - RADAR – WP1 context

In WP1, we focused our work on the analysis of plasmid as they represent the most important vector for AMR dissemination**.** We first constructed a curated database of unique complete plasmid sequences including a bioinformatic pipeline allowing for systematic classification of plasmids (Douarre et al., 2020). We then used this specific database to conduct a benchmark on plasmid prediction tools to identify *S. enterica* plasmids from short reads and we developed a pipeline for the complete exploration of the plasmidome including identification, reconstruction and annotation of plasmids. Finally, the resistome of the plasmids were identified by aligning all reconstructed plasmid sequences against the Resfinder database (manuscript in preparation).

We successufuly applied this novel workflow on a collection of 2863 *S. enterica* genomes (Salmonella network Anses) and identified 225 plasmids carrying AMR determinant conferring resistance to all the main classes of antibiotics. Working with this particular collection of *S. enterica* genomes was interesting beause of the heterogeneity of the dataset (different sources, serovars). However phenotypic data associated with resistance were not available and the predicted phenotype could not be confirmed. For the same reason GWAS could not be perfomed on this dataset and compared to our genotypes results.

Globally, we believe that searching resistance markers encoded by plasmid in a large dataset is easier and faster through a reference-based method than through a pangenome approach. Moreover, several studies previoulsy reported a strong correlation between the genotype identified using alignment-based program (BLAST-RESFINSDER) and the predicted phenotype. Even though GWAS can detect acquired genes responsible for a resistant phenotype, it stands out from other workflows by its ability to detect of novel or rare mutations explaining an unknwown phenotype.

REFERENCES

Alam, M.T., Petit, R.A., Crispell, E.K., Thornton, T.A., Conneely, K.N., Jiang, Y., Satola, S.W., and Read, T.D. (2014). Dissecting Vancomycin-Intermediate Resistance in Staphylococcus aureus Using Genome-Wide Association. *Genome Biology and Evolution* 6**,** 1174-1185.

Aun, E., Brauer, A., Kisand, V., Tenson, T., and Remm, M. (2018). A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLoS Comput Biol* 14**,** e1006434.

Brynildsrud, O., Bohlin, J., Scheffer, L., and Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* 17**,** 238.

Davies, M.R., Mcintyre, L., Mutreja, A., Lacey, J.A., Lees, J.A., Towers, R.J., Duchene, S., Smeesters, P.R., Frost, H.R., Price, D.J., Holden, M.T.G., David, S., Giffard, P.M., Worthing, K.A., Seale, A.C., Berkley, J.A., Harris, S.R., Rivera-Hernandez, T., Berking, O., Cork, A.J., Torres, R., Lithgow, T., Strugnell, R.A., Bergmann, R., Nitsche-Schmitz, P., Chhatwal, G.S., Bentley, S.D., Fraser, J.D., Moreland, N.J., Carapetis, J.R., Steer, A.C., Parkhill, J., Saul, A., Williamson, D.A., Currie, B.J., Tong, S.Y.C., Dougan, G., and Walker, M.J. (2019). Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics. *Nat Genet* 51**,** 1035-1043.

Douarre, P.E., Mallet, L., Radomski, N., Felten, A., and Mistou, M.Y. (2020). Analysis of COMPASS, a New Comprehensive Plasmid Database Revealed Prevalence of Multireplicon and Extensive Diversity of IncF Plasmids. *Front Microbiol* 11**,** 483.

Falush, D. (2016). Bacterial genomics: Microbial GWAS coming of age. *Nat Microbiol* 1**,** 16059.

Farhat, M.R., Shapiro, B.J., Kieser, K.J., Sultana, R., Jacobson, K.R., Victor, T.C., Warren, R.M., Streicher, E.M., Calver, A., Sloutsky, A., Kaur, D., Posey, J.E., Plikaytis, B., Oggioni, M.R., Gardy, J.L., Johnston, J.C., Rodrigues, M., Tang, P.K., Kato-Maeda, M., Borowsky, M.L., Muddukrishna, B., Kreiswirth, B.N., Kurepina, N., Galagan, J., Gagneux, S., Birren, B., Rubin, E.J., Lander, E.S., Sabeti, P.C., and Murray, M. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nat Genet* 45**,** 1183-1189.

Fritsch, L., Felten, A., Palma, F., Mariet, J.F., Radomski, N., Mistou, M.Y., Augustin, J.C., and Guillier, L. (2019). Insights from genome-wide approaches to identify variants associated to phenotypes at pan-genome scale: Application to L. monocytogenes' ability to grow in cold conditions. *Int J Food Microbiol* 291**,** 181-188.

Galardini, M., Clermont, O., Baron, A., Busby, B., Dion, S., Schubert, S., Beltrao, P., and Denamur, E. (2020). Major role of iron uptake systems in the intrinsic extra-intestinal virulence of the genus Escherichia revealed by a genome-wide association study. *PLoS Genet* 16**,** e1009065.

Habib, F., Johnson, A.D., Bundschuh, R., and Janies, D. (2007). Large scale genotype-phenotype correlation analysis based on phylogenetic trees. *Bioinformatics* 23**,** 785-788.

Jaillard, M., Lima, L., Tournoud, M., Mahe, P., Van Belkum, A., Lacroix, V., and Jacob, L. (2018). A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genet* 14**,** e1007758.

Juran, B.D., and Lazaridis, K.N. (2011). Genomics in the post-GWAS era. *Semin Liver Dis* 31**,** 215-222.

Laabei, M., Recker, M., Rudkin, J.K., Aldeljawi, M., Gulay, Z., Sloan, T.J., Williams, P., Endres, J.L., Bayles, K.W., Fey, P.D., Yajjala, V.K., Widhelm, T., Hawkins, E., Lewis, K., Parfett, S., Scowen, L., Peacock, S.J., Holden, M., Wilson, D., Read, T.D., Van Den Elsen, J., Priest, N.K., Feil, E.J., Hurst, L.D., Josefsson, E., and Massey, R.C. (2014). Predicting the virulence of MRSA from its genome sequence. *Genome Res* 24**,** 839-849.

Lees, J.A., Croucher, N.J., Goldblatt, D., Nosten, F., Parkhill, J., Turner, C., Turner, P., and Bentley, S.D. (2017). Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *Elife* 6.

Power, R.A., Parkhill, J., and De Oliveira, T. (2017). Microbial genome-wide association studies: lessons from human GWAS. *Nat Rev Genet* 18**,** 41-50.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81**,** 559-575.

Saber, M.M., and Shapiro, B.J. (2020). Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb Genom* 6.

Sheppard, S.K., Didelot, X., Meric, G., Torralbo, A., Jolley, K.A., Kelly, D.J., Bentley, S.D., Maiden, M.C., Parkhill, J., and Falush, D. (2013). Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. *Proc Natl Acad Sci U S A* 110**,** 11923-11927.

Vila Nova, M., Durimel, K., La, K., Felten, A., Bessieres, P., Mistou, M.Y., Mariadassou, M., and Radomski, N. (2019). Genetic and metabolic signatures of Salmonella enterica subsp. enterica associated with animal sources at the pangenomic scale. *BMC Genomics* 20**,** 814.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., Mccarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* 101**,** 5-22.

Wozniak, M., Tiuryn, J., and Wong, L. (2014). GWAMAR Genome-wide assessment of mutations associated with drug resistance in bacteria. *BMC Genomics*.