Aquavalens Project

"Protecting the health of Europeans by improving methods for the detection of pathogens in drinking water and water used in food preparation."

Grant agreement number: 311846

Deliverable D5.5

List of selected MST indicators or markers to include in the technological platforms

Authors: Andreas Farnleitner[1], Elisenda Ballesté[2], Anicet R. Blanch[2], Lluis A. Belanche[3],

- [1] TU Wien, Austria
- [2] University of Barcelona, Spain
- [3]Polytechnic University of Catalonia, Spain

Version 1, Jan 2016

1. Table of Contents

2. Validation of the developed MST predictive models

Previous deliverable 5.4 reported the developed MST predictive models to provide answer to eight different scenarios for MST when considering that:

1. Some predictive models, based exclusively on molecular MST markers, should determine a binary outcome (human vs. non-human source of contamination) which are the most important prediction to further quantitative microbial risk health assessment and they are the most significant implication for WPs in the implementation on technological platforms in the next cluster activities of Aquavalens.

2. Other predictive models, based exclusively on molecular MST markers, could be developed to determine a multi-class outcome (four different animal classes: bovine, porcine, poultry, human). These models could be also useful in a secondary level in the implementation on technological platforms in the next cluster activities of the project.

3. The selected predictive models were designed to work under environmental conditions (diluted and aged samples), departing from the provided point source matrix (heavy pollution)

From the statistical point of view, any of the previously obtained and validated models (the simple ones: linear, low complexity, interpretable; against more sophisticated ones: non-linear, high complexity, non-interpretable) during the former WP5 activities has been again considered on the final selection of molecular MST markers.

The selected models in D5.3 were validated and a preliminary subset of nine molecular parameters (D5.4) was considered as candidates for the final list of molecular targets on MST due to the end of the WP5 (Month 32):

The second sampling campaign has allowed to evaluate and provide the final list of selected molecular MST markers to be considered for their application in the technological platforms by Aquavalens partners in cluster 2 and 3. There are 9 molecular markers which constitute the set of parameters used in any of the validated predictive models for the different scenarios evaluated (point source, human vs. non-human, distinguishing 4 faecal sources, dilution and aging of the faecal pollutions).

The validated MST molecular-based models (linear or non-linear) (see D5.4) and the corresponding scenarios providing solutions are:

Linear Discriminating Analysis

Human vs Non-human / point source

HMBif, PLBif, TLBif, Pig2Bac, HF183TaqMan (accuracy 100% by LOOCV)

Four sources / point source

HMBif, BacR, Pig2Bac (accuracy 100% by LOOCV)

Non-linear Analysis (Random Forest)

Human vs Non-human

NoV, PGMit (accuracy 88.5% by 10x10 CV)

Four sources

NoV, PGMit, CWBif, HMBif, TLBif/CWBif (accuracy by 79.1% by 10x10 CV)

Consequently, at least a set of 9 molecular parameters to be analysed during the second and last sampling campaign was identified (D5.4):

- Mitochondrial DNA markers: Pig-associated (PGMit)
- *Bacteroidetes* DNA: Human-associated (HF183Taqman), Ruminant (BacR), Porcine (Pig2Bac)
- *Bifidobacterium* DNA: Human-associated (HMBif), Bovine (CWBif), Poultry (PLBif), Total (TLBif)
- Norovirus RNA (NoV)

Additionally, WP5 partners decided that other molecular markers were also included: Porcine-associated *Neoscardovia* (PGNeo), Total Bacteroidetes (AllBac), molecular Faecal Enterococci (FEqPCR), Cow-associated (CWMit) and poultry-associated (PLMit)

Culture dependent microbial indicators and MST markers were also analysed during the sampling campaign for characterization of faecal loads and complementary information on sources if needed. These parameters were:

- Water microbial indicators: *Escherichia coli* (EC), faecal enterococci (FE), *Clostridium perfringens* spores (CP), somatic coliphages (SomPhg)
- Human-associated *Bacteroides* phages (HMBactPhg)
- Cow-associated *Bacteroides* phages (CWBactPhg)
- Pig-associated *Bacteroides* phages (PGBactPhg)
- Poultry-associated *Bacteroides* phages (PLBactPhg)

3. Analysing diluted and aged faecal polluted water samples to confirm the preliminary list of molecular MST markers.

The confirmation of the list of selected MST indicators or markers to include in the technological platforms was based on the planned (see DoW) second sampling campaign. WP5 partners were using environmental samples or experimentally modified water samples in order to obtain water samples with diluted and aged faecal pollution. This type of samples allow confirming the usefulness of the molecular MST markers preliminary selected by the different predictive models (D5.4).

Second campaign samples have been taken by WP5 partners from February 2015 to September 2015, using the established SOP (D5.1) and following the next criteria and activities:

1. Samples of wastewater (point source) of known faecal origin. If collective (population) faeces were taken as substitutive for wastewater, at least 25 individual faeces were integrated in one sample. In case this number of individual faeces was not constituting the integrated sample, it should be registered and communicate to the partners.

2. Samples representative of populations.

3. Samples were identified by codification as agreed:

   a. 1 character for partner code related to respective country: E for UB, A for TU WIEN, D for DVGW, P for IST, F for UH

   b. 2 character for number of sample: 01, 24, 56, …

   c. plus ENV characters: meaning environmental samples in order to distinguish them from the known fresh point source samples from the first sampling campaign.

   Exemple: P06ENV

4. Wastewater samples could be experimentally modified: diluted up to 1:10.000, aged and/or mixed (maximum two different sources). No indications of the faecal source and modifications was provided to the rest of partners in order to keep samples non-identified by the rest of participants and warranty a blind numerical analysis and prediction of origins by the models.

5. A total of 10 samples by participant were prepared and sent as blind samples. A total of 50 samples were expected which would constitute a testing matrix to be evaluated by the predictive models which were developed using the data base (training matrix) obtained from the first sampling campaign.

5. Each partner measured locally the following parameters in their respective samples: *E. coli* (EC), faecal enterococci (FE), *C. perfringens* spores (CP) and somatic coliphages (SomPhg). Additionally, Human-associated *Bacteroides* phages (HMBactPhg), Cow-associated *Bacteroides* phages (CWBactPhg), Pig-associated *Bacteroides* phages (PGBactPhg) and Poultry-associated *Bacteroides* phages (PLBactPhg) were analysed by UB.

6. Each partner delivered 500 ml (for microbiological or eukaryotic analyses) of each blind sample to the rest of partners analysing the final selected molecular parameters. Samples were sent at 4°C by private courier or any other 24 h delivery service. Recipient laboratories preserved or analysed the samples right after their arrival at their laboratories.

7. The select final molecular MST host-associated markers and the partner in charge for their analysis were:

   a. *Bifidobacterium* DNA: Human-associated (HMBif), Bovine (CWBif), Poultry (PLBif), Total (TLBif) and Porcine-associated *Neoscardovia* (PGNeo) by UB
   b. qPCR specific Bacteroidetes plus total Bacteroidetes: TU WIEN
   c. Mitochondrial DNA markers: Cow-associated (CWMit), Pig-associated (PGMit) and poultry-associated (PLMit) by IST
   d. Norovirus RNA (NoV) by UH
   e. Bacteroidetes DNA: Human-associated (HF183Taqman, US-EPA last protocol), Ruminant (BacR), Porcine (Pig2Bac) and Total Bacteroidetes (AllBac) by TU WIEN.
   f. Molecular Faecal Enterococci (FEqPCR) were analysed by TU WIEN.

8. A data base frame (Excel file) was provided by UB to introduce results along the analyses on Task 5.5. Results were sent to UB, where all results were merged and harmonised. It was essential to avoid missing values – otherwise the whole sample would be omitted for testing using the defined inductive learning predictive methods.

4. Numerical analysis of the testing set of data from the second sampling campaign.

The testing set of data from the second sampling campaign is constituted for a total of 52 samples: 38 samples faecally polluted from a unique source and 14 water samples experimentally prepared by mixtures of several faecal sources.

Institution	Samples
TU WIEN	A01ENV-A12ENV
DVGW	D01ENV-D10ENV
UB	E01ENV-E10ENV
UH	F01ENV-F10ENV
IST	P01ENV-P10ENV
	52/52

There are several departing points and considerations for analysing the data:

1. Only a fraction of parameters (molecular MST markers) should be used.

2. The use of all parameters (culture dependent and molecular targets) could be considered for comparison with the use of only molecular MST markers.

3. The previous defined predictive models (linear versus non-linear) will be used.

4. Analyses are performed attending to two different decision situations: Human versus non-human and distinguishing among four main faecal sources (human, porcine, bovine, poultry).

5. The confirmation of the set of molecular MST markers will be done considering only diluted and aged faecal polluted water samples.

The second campaign samples showed concentrations of microbial indicators at the following ranges:

E. coli, from 0.17 to 5.09 $\log_{10}$ CFU/100 mL
Faecal enterococci, from 1.2 to 4.3 $\log_{10}$ CFU/100 mL
C. perfringens spores, from 0.3 to 3.8 $\log_{10}$ CFU/100 mL
Somatic coliphages, from 0.3 to 4.7 $\log_{10}$ PFU/100 mL

The previously linear and non-linear (random forest) developed models (D5.4) were considered for the numerical analyses. The random forest (Breiman, Leo (2001). Random Forests. Machine Learning 45 (1): 5-32. doi:10.1023/A:1010933404324) is an ensemble approach that consists on a set of randomized decision trees. The main principle behind ensemble methods is that a group of “weak learners” can come together to form a “strong learner”. In a random forest the weak learner is a decision tree. The parameters are the number of trees and the number of variables tried at each split; in our case, we set it to the square root of the total number of variables, as is standard practice. The method is very fast to train, and they is able to deal with unbalanced and missing data.

A detailed description of the learned random forests, as used for this confirmation study, follows:

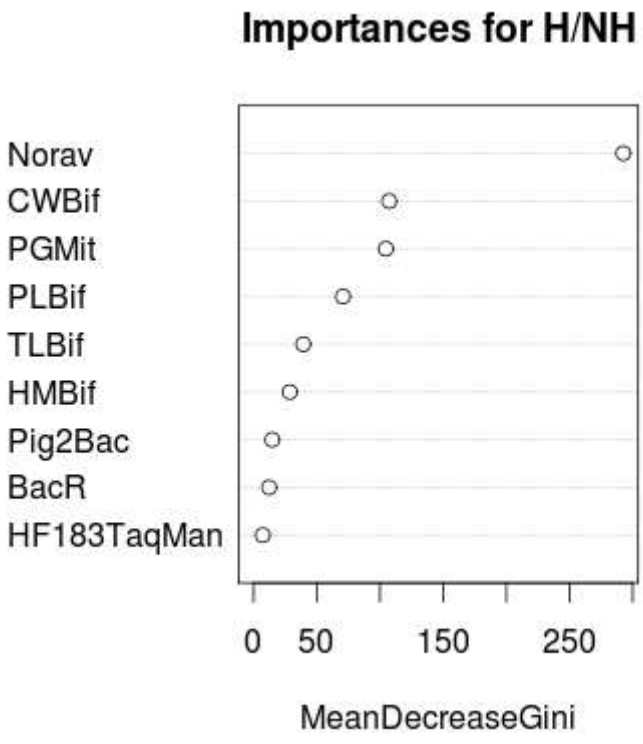
1. ***Determining Human versus non-Human faecal sources (H/NH):***
   - Number of trees: 40
   - No. of variables tried at each split: 2
   - Overall estimate of error rate: 15.46% decomposed as:

	Nonhuman	human	class.error
nonhuman	6177	1108	15.21%
human	438	2277	16.13%

(in all cases, the rows are the true sources, and the columns are the predictions)

using the parameters "PGMit", "BacR", "Norav", "HF183TaqMan", "CWBif", "HMBif", "TLBif", "PLBif", "Pig2Bac".

The importance of each parameter on the prediction is indicated on the following figure:



2. Determining 4 different faecal sources (Human, Bovine, Porcine and Poultry):

- Number of trees: 100
- No. of variables tried at each split: 4
- Overall estimate of error rate: 14.47% decomposed as:

	CW	HM	PG	PL	class.error
CW	1799	37	98	299	19.44%
HM	13	2730	54	351	13.28%
PG	22	227	1862	251	21.17%
PL	6	20	69	2162	4.21%

(in all cases, the rows are the true sources, and the columns are the predictions)

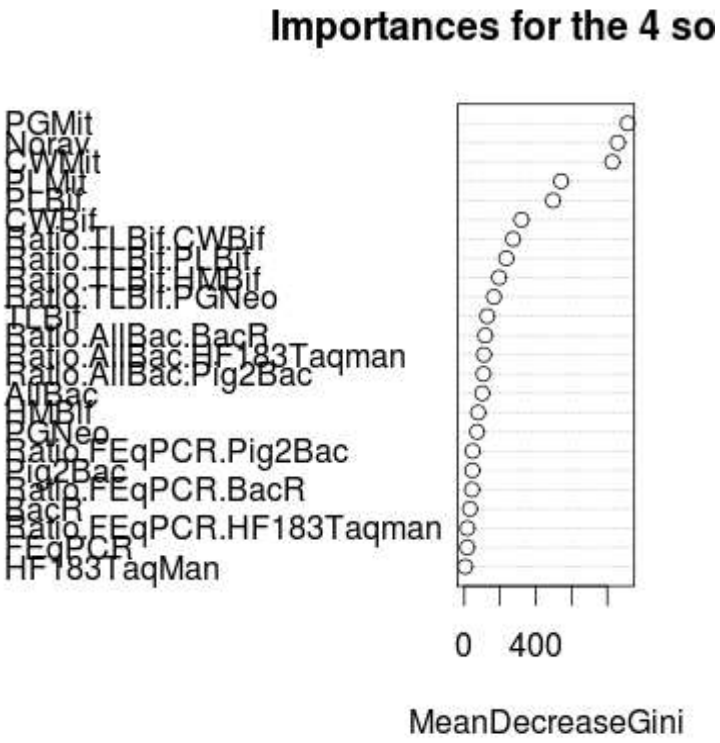
using "HMBif", "CWBif", "PGNeo", "PLBif", "TLBif", "BacR", "Pig2Bac", "AllBac", "HF183TaqMan", "FEqPCR", "CWMit", "PGMit", "PLMit", "Norav"

plus the suggested derived ratios (which seem to be helpful in this case but not on distinguishing Human from Non-human):

"Ratio.TLBif.HMBif", "Ratio.TLBif.CWBif", "Ratio.TLBif.PGNeo", "Ratio.TLBif.PLBif", "Ratio.AllBac.BacR", "Ratio.AllBac.Pig2Bac", "Ratio.AllBac.HF183Taqman", "Ratio.FEqPCR.BacR", "Ratio.FEqPCR.Pig2Bac", "Ratio.FEqPCR.HF183Taqman".

However, these ratios are not increasing the number of parameters to be measured for a prediction.

The importance of each parameter on the prediction is indicated on the following figure:



On the table included in the annex, the prediction for the different studied samples is indicated according to the linear or non-linear method and the goal of MST (Human versus Non-human or 4 faecal sources). A conservative prediction is to take as source the result providing a lower health risk on management.

5. Conclusions

The selected list of molecular MST targets showed predictions up to 86% of accuracy when using random forest predictive models which were developed by the Ichnaea® software (University of Barcelona and Polytechnic University of Catalonia) to distinguish Human versus Non-human (including pig, cow, poultry, cats, horses samples) faecal pollution (see table on Annex). These models are based on the 9 pre-selected molecular markers and allowed to provide predictions for any of the studied scenarios which are accounting for distinguishing Human versus Non-human; at point source or up to 10,000 fold dilutions and aging of the pollution over 300 h.

The accuracy of the predictions decrease when distinguishing the four main faecal sources (human, cos, pig and poultry) in waters. A total of 14 molecular markers plus some ratios are needed allowing to distinguish the 4 main faecal sources up to 68% of accuracy and up to 10,000 fold dilutions and aging of the pollution over 300 h. These developed models for distinguishing 4 sources are just providing 81% of accuracy if applied to differentiate Human versus Non-human sources.

Consequently, the MST markers proposed to be used on the development of technical platforms by Cluster 2 and 3 partners are:

<u>To distinguish Human from Non-human faecal sources only 9 parameters are needed to be included in a technical platform:</u>

- Porcine-associated Mitochondrial (PGMit)
- Human-associated Bacteroidetes (HF183Taqman)
- Ruminant-associated Bacteroidetes (BacR)
- Porcine-associated Bacteroidetes (Pig2Bac)
- Human-associated *Bifidobacterium* (HMBif)
- Cow-associated *Bifidobacterium* (CWBif)
- Poultry-associated *Bifidobacterium* (PLBif)
- Total *Bifidobacterium* (TLBif)
- Norovirus RNA (NoV)

To distinguish 4 faecal sources (human, bovine, porcine, poultry) 14 parameters are needed in a technical platform, but later 10 derived ratios are requested for the best prediction with random forest models as described above:

- Cow-associated Mitochondrial (CWMit)
- Porcine-associated Mitochondrial (PGMit)
- Poultry-associated Mitochondrial (PLMit)
- Human-associated Bacteroidetes (HF183Taqman)
- Ruminant-associated Bacteroidetes (BacR)
- Pig-associated Bacteroidetes (Pig2Bac)
- Human-associated *Bifidobacterium* (HMBif)
- Cow-associated *Bifidobacterium* (CWBif)
- Poultry-associated *Bifidobacterium* (PLBif)
- Pig-associated *Neoscardovia* (PGNeo)
- Total *Bifidobacterium* (TLBif)
- All Bacteroidetes (AllBac)
- Molecular Fecal Enterococci (FEqPCR)
- Norovirus RNA (NoV)

6. Annex. Table of results attending to the linear and non-linear methods applied and to resolve H vs. NH or 4 sources

The table is including the real origins, the percentages of prediction for each possible origin according to the model and the percentage of correct predictions

	SAMPLES	9 previously selected variables HM / NHM				All tested molecular variables HM / NHM						All tested molecular variables 4 Sources						
		Real origin	MODEL 1 (Linear)		MODEL 2 (Random Forest)		%HM	%NHM	Results		Results conservative		Real origin	%CW	%HM	%PG	%PL	4 SOURCES	
UB	E1ENV	NH	H	Incorrect	NH	Correct	25%	75%	NH	Correct	NH	Correct	PG	11%	29%	48%	12%	PG	Correct
	E2ENV	NH	NH	Correct	NH	Correct	25%	75%	NH	Correct	NH	Correct	PG	16%	26%	50%	8%	PG	Correct
	E3ENV	H	H	Correct	H	Correct	70%	30%	H	Correct	H	Correct	HM	20%	54%	15%	11%	HM	Correct
	E4ENV	H	H	Correct	H	Correct	83%	18%	H	Correct	H	Correct	HM	37%	60%	1%	2%	HM	Correct
	E6ENV	H	H	Correct	H	Correct	83%	18%	H	Correct	H	Correct	HM	10%	56%	14%	20%	HM	Correct
	E7ENV	NH	H	Incorrect	NH	Correct	50%	50%	H/NH	Incorrect	H	Incorrect	PL	12%	53%	16%	19%	HM	Incorrect
	E8ENV	H	H	Correct	H	Correct	83%	18%	H	Correct	H	Correct	HM	20%	59%	13%	8%	HM	Correct
	E9ENV	NH	H	Incorrect	H	Incorrect	75%	25%	H	Incorrect	H	Incorrect	PL	12%	55%	12%	21%	HM	Incorrect
TU WIEN	A1ENV	H	H	Correct	H	Correct	83%	18%	H	Correct	H	Correct	HM	21%	48%	20%	11%	HM	Correct
	A2ENV	H	H	Correct	H	Correct	50%	50%	H/NH	Incorrect	H	Correct	HM	26%	46%	20%	8%	HM	Correct
	A3ENV	NH	NH	Correct	NH	Correct	15%	85%	NH	Correct	NH	Correct	CW	66%	13%	14%	7%	CW	Correct
	A4ENV	NH	H	Incorrect	NH	Correct	8%	93%	NH	Correct	NH	Correct	PG	40%	7%	51%	2%	PG	Correct
	A6ENV	H	H	Correct	H	Correct	50%	50%	H/NH	Incorrect	H	Correct	HM	18%	52%	19%	11%	HM	Correct
	A7ENV	H	H	Correct	H	Correct	83%	18%	H	Correct	H	Correct	HM	45%	33%	17%	5%	CW	Incorrect
	A8ENV	NH	H	Incorrect	NH	Correct	15%	85%	NH	Correct	NH	Correct	CW	95%	4%	0%	1%	CW	Correct
	A9ENV	NH	H	Incorrect	NH	Correct	25%	75%	NH	Correct	NH	Correct	PG	20%	27%	43%	10%	PG	Correct
IST	P1ENV	H	H	Correct	NH	Incorrect	50%	50%	H/NH	Incorrect	H	Correct	HM	12%	48%	19%	21%	HM	Correct
	P3ENV	NH	NH	Correct	NH	Correct	15%	85%	NH	Correct	NH	Correct	CW	96%	4%	0%	0%	CW	Correct
	P6ENV	NH	H	Incorrect	NH	Correct	50%	50%	H/NH	Incorrect	H	Incorrect	PL	13%	51%	16%	20%	HM	Incorrect
	P8ENV	NH	NH	Correct	NH	Correct	15%	85%	NH	Correct	NH	Correct	PL	51%	22%	16%	11%	CW	Incorrect
	P10ENV	NH	H	Incorrect	NH	Correct	50%	50%	H/NH	Incorrect	H	Incorrect	PG	13%	51%	16%	20%	HM	Incorrect
DVGW	D1ENV	H	H	Correct	H	Correct	50%	50%	H/NH	Incorrect	H	Correct	HM	9%	45%	12%	34%	HM	Correct
	D2ENV	H	H	Correct	H	Correct	50%	50%	H/NH	Incorrect	H	Correct	HM	18%	46%	14%	22%	HM	Correct
	D3ENV	NH	NH	Correct	NH	Correct	15%	85%	NH	Correct	NH	Correct	CW	66%	13%	14%	7%	CW	Correct
	D4ENV	NH	NH	Correct	NH	Correct	15%	85%	NH	Correct	NH	Correct	CW	66%	13%	13%	8%	CW	Correct
	D6ENV	NH	H	Incorrect	NH	Correct	25%	75%	NH	Correct	NH	Correct	PG	54%	29%	16%	1%	CW	Incorrect
	D7ENV	NH	H	Incorrect	NH	Correct	25%	75%	NH	Correct	NH	Correct	PG	35%	14%	43%	8%	PG	Correct
	D8ENV	H	H	Correct	NH	Incorrect	50%	50%	H/NH	Incorrect	H	Correct	PL	18%	39%	22%	21%	HM	Incorrect
	D9ENV	NH	NH	Correct	H	Incorrect	50%	50%	H/NH	Incorrect	H	Incorrect	PL	45%	45%	0%	10%	HM/CW	Incorrect
UH	F1ENV	H	H	Correct	H	Correct	83%	18%	H	Correct	H	Correct	HM	21%	53%	14%	12%	HM	Correct
	F2ENV	H	H	Correct	H	Correct	83%	18%	H	Correct	H	Correct	HM	18%	62%	13%	7%	HM	Correct
	F3ENV	NH	H	Incorrect	H	Incorrect	50%	50%	H/NH	Incorrect	H	Incorrect	PL	11%	51%	17%	21%	HM	Incorrect
	F4ENV	NH	NH	Correct	NH	Correct	15%	85%	NH	Correct	NH	Correct	CW	74%	10%	12%	4%	CW	Correct
	F6ENV	H	H	Correct	H	Correct	83%	18%	H	Correct	H	Correct	HM	39%	58%	1%	2%	HM	Correct
	F7ENV	H	H	Correct	H	Correct	83%	18%	H	Correct	H	Correct	HM	19%	57%	13%	11%	HM	Correct
	F8ENV	NH	NH	Correct	NH	Correct	50%	50%	H/NH	Incorrect	H	Incorrect	PL	41%	49%	0%	10%	HM	Incorrect
	F9ENV	NH	NH	Correct	NH	Correct	15%	85%	NH	Correct	NH	Correct	CW	72%	13%	10%	5%	CW	Correct
	F10ENV	NH	H	Incorrect	NH	Correct	25%	75%	NH	Correct	NH	Correct	PG	40%	41%	13%	6%	HM/CW	Incorrect
			Correct	26	Correct	33			Correct	25	Correct	31						Correct	26
			Incorrect	12	Incorrect	5			Incorrect	13	Incorrect	7						Incorrect	12
				68,42%		86,84%				65,79%		81,58%							68,42%

Implications of the results of Deliverable Report 5.5

Implications of the results for the Work Package (WP 5)

The research activities of WP5 all along its development during the compromised period have been successfully performed and have identified among the initial list of more than 40 parameters those molecular markers to be considered on the development of technological platforms and devices by other WPs of the project.

Implications of the results for this Cluster 1

The obtained results on WP5 support the achievement of the Objective 2 of the Cluster 1 as specified on the DoW:

Objective 2: to select and define the best molecular tools for microbial source tracking (MST), based on inductive learning and statistical methods, capable of distinguishing between faecal pollution from more than two different sources that can be used routinely in Europe

Implications of the results for the whole project

The 9 selected molecular MST markers are enough to distinguish Human from Non-human faecal pollution in waters. The determination of 4 main faecal pollution sources in water (human, bovine, porcine and poultry) need 14 MST molecular markers plus some ratios. Cluster 2 and 3 partners could consider both combinations for their inclusion on technological platforms.

Indicate key external stakeholders interested in the results of Deliverable Report 5.5

The stakeholders can apply the proposed molecular MST markers to bring to the market solutions on the assessment of the determination of faecal sources in waters by using the identified molecular markers on new technological platforms or devices.

Which internal partners should your deliverable be sent to?

This deliverable should be sent to all the Aquavalens partners and especially to those of cluster 2 and 3 which have to apply the selected molecular MST markers during the performance of activities on cluster 2 and 3.