Microsynth

Genomic Epidemiology Analysis

Determine the sequence type of microbial isolates
Screen for resistance, virulence, and mycotoxin genes
Identify genetic variations that lead to pathogenicity

Introduction

Genomic epidemiology studies are useful to support public health, combat infectious diseases, maintain food safety, or establish sustainable production lines. To support our customers in these diverse applications, Microsynth has developed a genomic epidemiology pipeline that allows addressing numerous scientific topics concerning your epidemiology project:

- Resolve the taxonomy of individual species or sequence type (ST)
- Identify resistance, virulence and mycotoxin genes
- Detect multidrug resistant bacterial pathogens (MDR)
- Detect acquired mutations and assess their potential functional effects
- Establish a population-level profiling and comparison of antimicrobial resistance

Our genomic epidemiology analysis pipeline is based on next generation whole genome sequencing. The generated data and knowledge will support production lines, can be used to track transmission, or used in pathogen outbreak surveillance, to just name a few examples. At Microsynth we will guide you through the whole analysis to assure you will achieve your goals.

Microsynth's Competences and Services

With more than 10 years of experience in the field of next generation sequencing, one of Microsynth's core competences is to provide high quality one-stop services from experimental design to bioinformatics analysis. You may either outsource the entire analysis or only single steps to us as illustrated in ***Figure 1***.

Experimental Design

Microsynth's NGS specialists will help you define suitable experimental setups for your genomic epidemiology analysis projects and discuss possible strategies to address your research questions.

DNA Isolation

You may either perform the DNA extraction yourself or outsource this critical step to Microsynth. We have long-standing experience in processing various sample matrices and DNA/RNA sources.

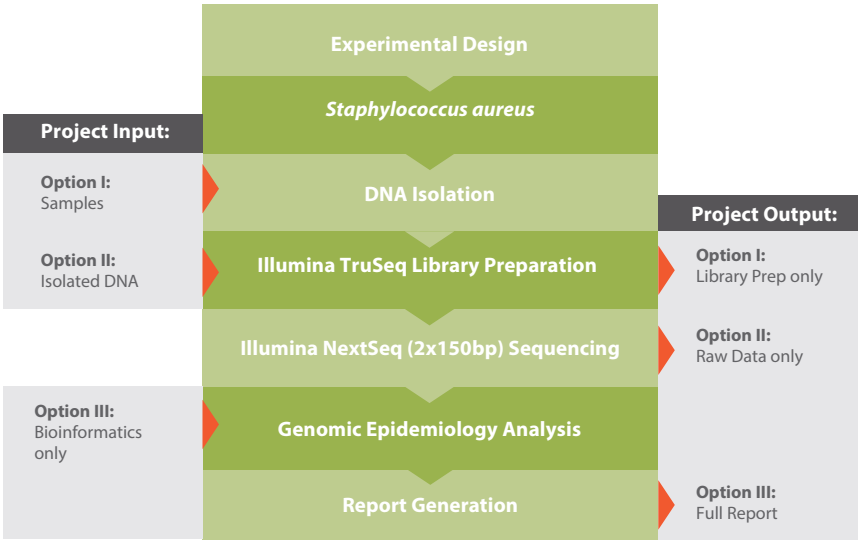

Figure 1. *Microsynth's workflow for genomic epidemiology analysis projects. The workflow can be entered and exited at various steps depending on the requirements of the customer.*

Microsynth AG, Switzerland
Schützenstrasse 15 · P.O. Box · CH-9436 Balgach · Phone +41 71 722 83 33 · Fax +41 71 722 87 58 · info@microsynth.ch · www.microsynth.com

Library Preparation and Sequencing
Following a quality check of your samples, Microsynth will build Illumina libraries including specific adaptors with barcodes. Depending on the experimental design, the libraries are pooled and sequenced either on the Illumina MiSeq or NextSeq 500/550. These flexible platforms allow for optimal sequencing depending on the number of samples and on the required read length.

Bioinformatics Analysis
Thanks to the use of modern methods and algorithms, hundreds of samples can be analyzed in detail for any genomic epidemiology project. Depending on the focus of the project, many different approaches to bioinformatic analysis are possible, but two example scenarios are described in the next section.
In the first example scenario, the genomes contained in the samples come from well-studied model organisms. However, since the composition of the samples is unknown, a shotgun metagenome analysis is performed first to determine whether the samples contain pure strains or whether meta-communities exist, and in both cases the taxonomy is determined (see **Table 1**). Assuming that the samples contain pure strains of a well-studied micro-organism, one of the established Multi Locus Sequence Typing (MLST) schemes can be used to determine its sequence type, which goes beyond the determination of the species alone (see **Table 2**).
Next, in order to reconstruct the strains contained in the samples, the filtered sequencing reads are *de novo* assembled into contigs on which genes are predicted and annotated. The predicted genes are then screened for homologous sequences among the known resistance, virulence, and mycotoxin genes. Single nucleotide variations (SNVs) and small insertions and deletions (InDels) are determined in comparison to the public reference sequences (see **Table 3**).
In the second scenario, the microorganisms to be analyzed are neither well studied nor publicly documented. In this case, determining the similarity of the unknown genomes to any of the genomes stored in the RefSeq [1] database first and second annotate predicted genes with homologues found for instance in the Pfam [2] database (see **Table 4**) are sensible first steps. If no exact phylogenetic relationship is known for the hundreds of samples involved in a genomic epidemiology project, clustering can be used to establish groups of samples that may then be further analyzed on their own [3] (see **Figure 2**).

Example Results

Table 1. *This cutout of a result of a shotgun metagenomics taxonomy assignment, shows the composition of the bacterial community found in the analyzed sample. In this case, the sample contains 96 % of the Staphylococcus genus (Tax Level: G) and 92% are identified as Staphylococcus aureus species (Tax Level: S).*

Percent	Tax Level	Tax ID	Tax Name
100.00	D	2	*Bacteria*
100.00	P	1239	*Firmicutes*
100.00	C	91061	*Bacilli*
96.00	O	1385	*Bacillales*
96.00	F	90964	*Staphylococcaceae*
96.00	G	1279	*Staphylococcus*
92.00	S	1280	*Staphylococcus aureus*

Table 2. *The result of an MLST showing the sequence type of the species found in the sample. In this case, the scheme used for typing included seven genes (arcC, aroE, glpF, gmk, pta, tpi and yqiL) to identify the respective ST and Clonal Complex (CC). The numbers in the gene columns represent different variants of these genes found in the PubMLST database [4].*

Sample	arcC	aroE	glpF	gmk	pta	tpi	yqiL	ST	Clonal Complex
sample1	1	4	1	8	4	4	3	72	CC8

Table 3. *Summary table of the number of observed SNVs and small InDels in the analyzed sample including the type of mutation (silent and non-silent).*

Sample	SNV	InDels	Silent	Non-silent
sample1	236	12	171	53

Table 4. Detail of a table showing the homologous protein domains and their significance found for the predicted genes of the analyzed sample.

Target Name	Accession	Query Name	E-value	Score	Description of Target
TPK_catalytic	PF04263.16	gene_8|401_aa|-|7047|8252	0.00038	20.5	Thiamin pyrophosphokinase, catalytic domain
CRISPR_Cas9_WED	PF18061.1	gene_40|840_aa|-|40414|42936	0.12	12.4	CRISPR-Cas9 WED domain

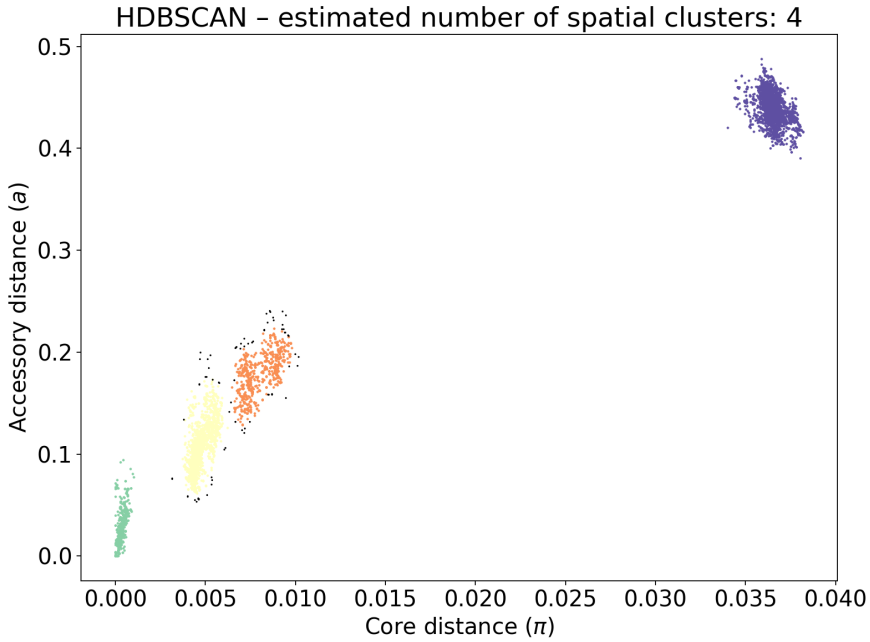

Figure 2. Clustering of a large numbers of samples with unknown phylogenetic relationship resulting in four distinct groups.

Related Services

Microsynth also provides microbial resequencing services that focus on the detection of genetic variations in relation to a reference sequence.

References

[1] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. **Nucleic Acids Res**. 2016 Jan 4;44(D1):D733-45 PubMed

[2] S. El-Gebali, J. Mistry, A. Bateman, S.R. Eddy, A. Luciani, S.C. Potter, M. Qureshi, L.J. Richardson, G.A. Salazar, A. Smart, E.L.L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S.C.E. Tosatto, R.D. Finn. The Pfam protein families database in 2019. **Nucleic Acids Research** (2019) doi: 10.1093/nar/gky995

[3] Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher NJ. Fast and flexible bacterial genomic epidemiology with PopPUNK. **Genome Research** 29:1-13 (2019). doi:10.1101/gr.241455.118

[4] Website (https://pubmlst.org/mlst/) sited at the University of Oxford (Jolley et al. Wellcome Open Res 2018, 3:124 [version 1; referees: 2 approved]). The development of this site has been funded by the Wellcome Trust.