



Bioinformatik (AP4)

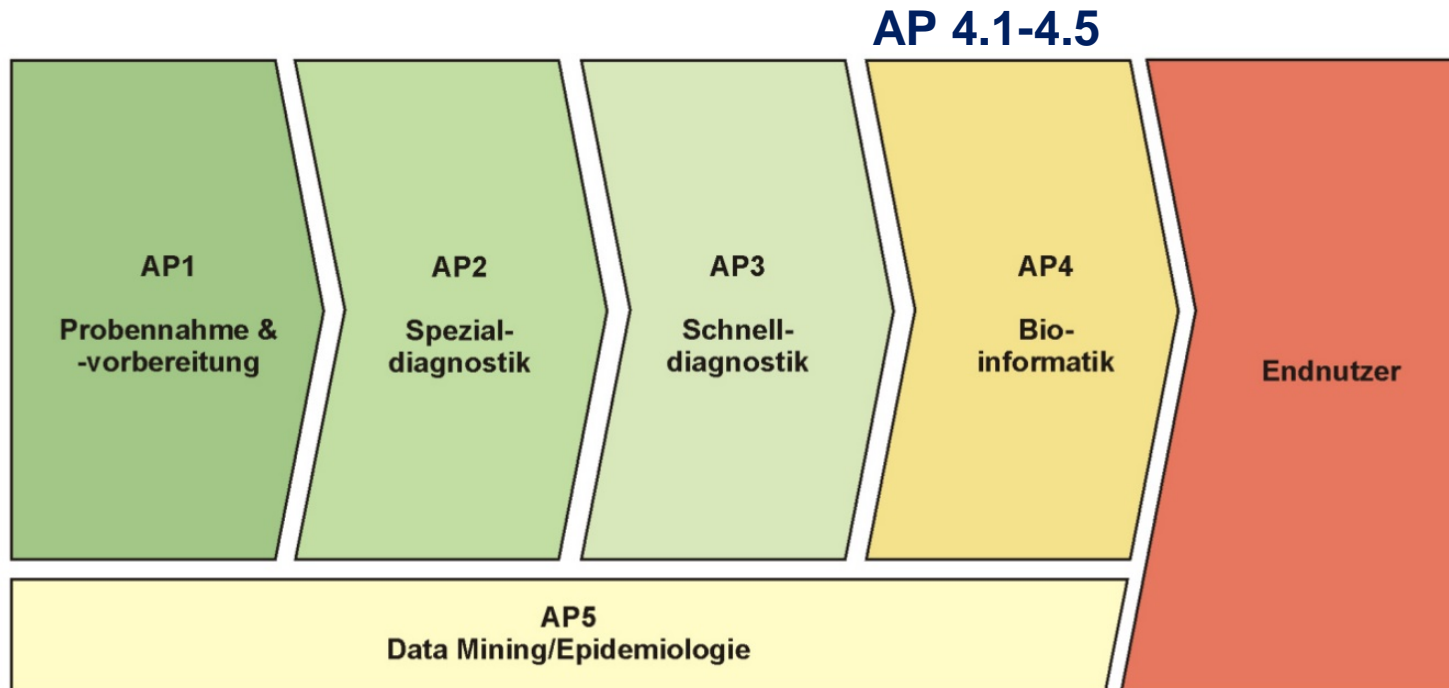
Knut Reinert

Sandro Andreotti

Kerstin Neubert

20.11.2017

Ess-BAR Arbeitspakete



Ess-BAR Bioninformatik-Arbeitspakete

AP 4.1: Aufbau einer Referenzdatenbank am Beispiel von drei Pathogenen (*Brucella spp.*, *F. tularensis* and *B. anthracis*) und nahe verwandten Spezies

AP 4.2: Referenz-basierte Assemblierung von Bakterienisolaten

AP 4.3: Identifikation/ Klassifizierung von Pathogenen mittels NGS-Daten

AP 4.4: Auswertung von proteomischen Daten (LC/MS and MRM/PRM)

AP 4.5: Integration von Methoden aus 4.1-4.4 in KNIME

RefDB-workflow I (externe DBs)

- Genome/ Proteome ausgewählter Erreger
- Weitere Annotationen: Virulenzfaktoren, Resistenzen, Phagen, Sequenztypen
- Genom-Sequenzen und -Index aller bakteriellen Erreger aus NCBI oder PATRIC

Phylogenetische Analyse

Taxonomische Profilierung

- Identifikation der Spezies in einem Isolat
- Quantifizierung der Spezies in einer metagenomischen Probe

Speziess Selektion

Analyse von Sub-Spezies und Stämmen

- Charakterisierung des Isolates z.B. anhand von Virulenzfaktoren, Resistenzen
- Intraspezifische Phylogeny (cgSNP)
- MLST-Analyse

Assembly-workflow

- Berechnung der Genomsequenzen und Annotationen der Referenzstämme

RefDB-workflow II (Ess-BAR-DB)

- Annotierte Genomsequenzen
- Proteinsequenzen mit Orthologen

Proteomische Analyse

LC/MS Analyse

- Identifikation von spezies-spezifischen Referenzpeptid-Profilen
- Charakterisierung der Proteome

Referenzmarker

Fingerprinting mit MRM/PRM-Methoden

- Identifizierung und Charakterisierung von Erreger-spezifischen Peptidmarkern

AP 4.1

AP 4.2

AP 4.3

AP 4.4

Ess-BAR workflows (Meilenstein)

RefDB-workflow I (externe DBs)

- Genome/ Proteome ausgewählter Erreger
- Weitere Annotationen: Virulenzfaktoren, Resistenzen, Phagen, Sequenztypen
- Genom-Sequenzen und -Index aller bakteriellen Erreger aus NCBI oder PATRIC



Assembly-workflow

- Berechnung der Genomsequenzen und Annotationen der Referenzstämme



RefDB-workflow II (Ess-BAR-DB)

- Annotierte Genomsequenzen
- Proteinsequenzen mit Orthologen



Phylogenetische Analyse



Taxonomische Profilierung

- Identifikation der Spezies in einem Isolat
- Quantifizierung der Spezies in einer metagenomischen Probe



Speziess Selektion

Analyse von Sub-Spezies und Stämmen

- Charakterisierung des Isolates z.B. anhand von Virulenzfaktoren, Resistenzen
- Intraspezifische Phylogeny (cgSNP)
- MLST-Analyse

Proteomische Analyse

LC/MS Analyse

- Identifikation von spezies-spezifischen Referenzpeptid-Profilen
- Charakterisierung der Proteome



Referenzmarker

Fingerprinting mit MRM/PRM-Methoden

- Identifizierung und Charakterisierung von Erreger-spezifischen Peptidmarkern

AP 4.1

AP 4.2

AP 4.3

AP 4.4

RefDB-workflow I (externe DBs)

- Genome/ Proteome ausgewählter Erreger
- Weitere Annotationen: Virulenzfaktoren, Resistenzen, Phagen, Sequenztypen
- Genom-Sequenzen und -Index aller bakteriellen Erreger aus NCBI oder PATRIC



Assembly-workflow

- Berechnung der Genomsequenzen und Annotationen der Referenzstämme



RefDB-workflow II (Ess-BAR-DB)

- Annotierte Genomsequenzen
- Proteinsequenzen mit Orthologen



Phylogenetische Analyse



Taxonomische Profilierung

- Identifikation der Spezies in einem Isolat
- Quantifizierung der Spezies in einer metagenomischen Probe



Speziess Selektion

Analyse von Sub-Spezies und Stämmen

- Charakterisierung des Isolates z.B. anhand von Virulenzfaktoren, Resistenzen
- Intraspezifische Phylogeny (cgSNP)
- MLST-Analyse

Proteomische Analyse

LC/MS Analyse

- Identifikation von spezies-spezifischen Referenzpeptid-Profilen
- Charakterisierung der Proteome



Referenzmarker

Fingerprinting mit MRM/PRM-Methoden

- Identifizierung und Charakterisierung von Erreger-spezifischen Peptidmarkern

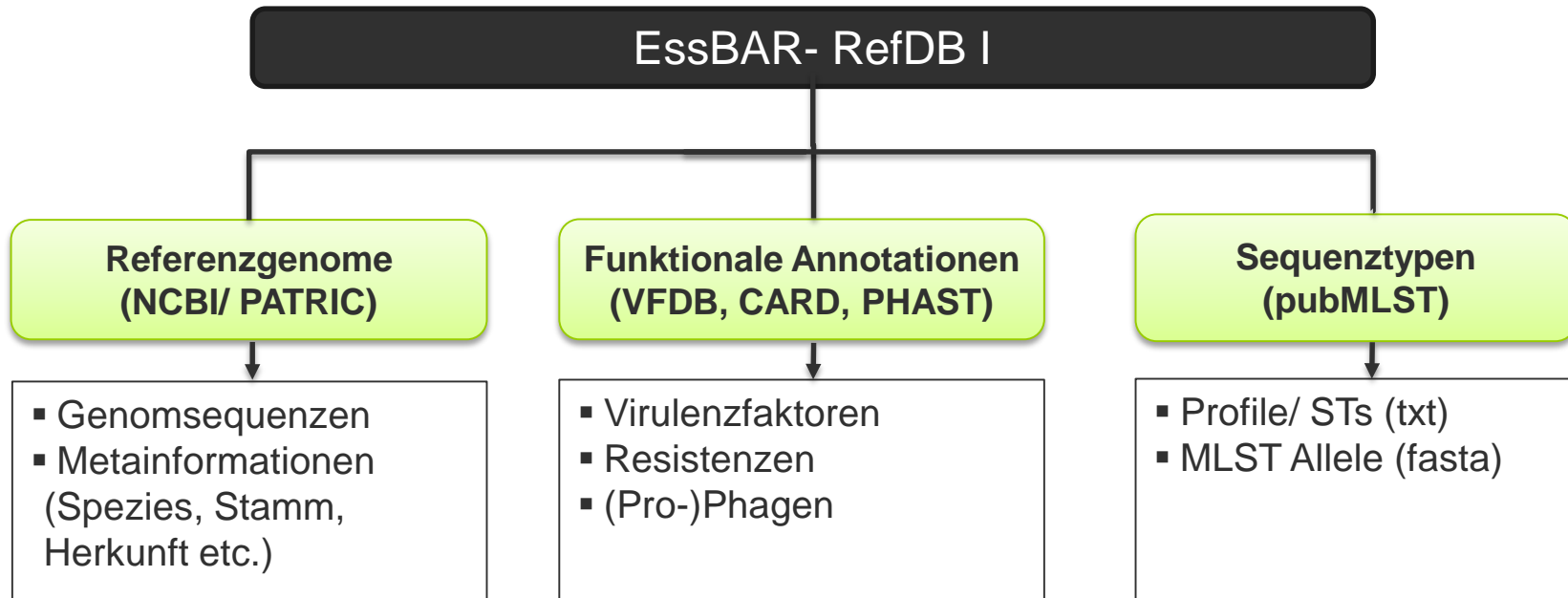
AP 4.1

AP 4.2

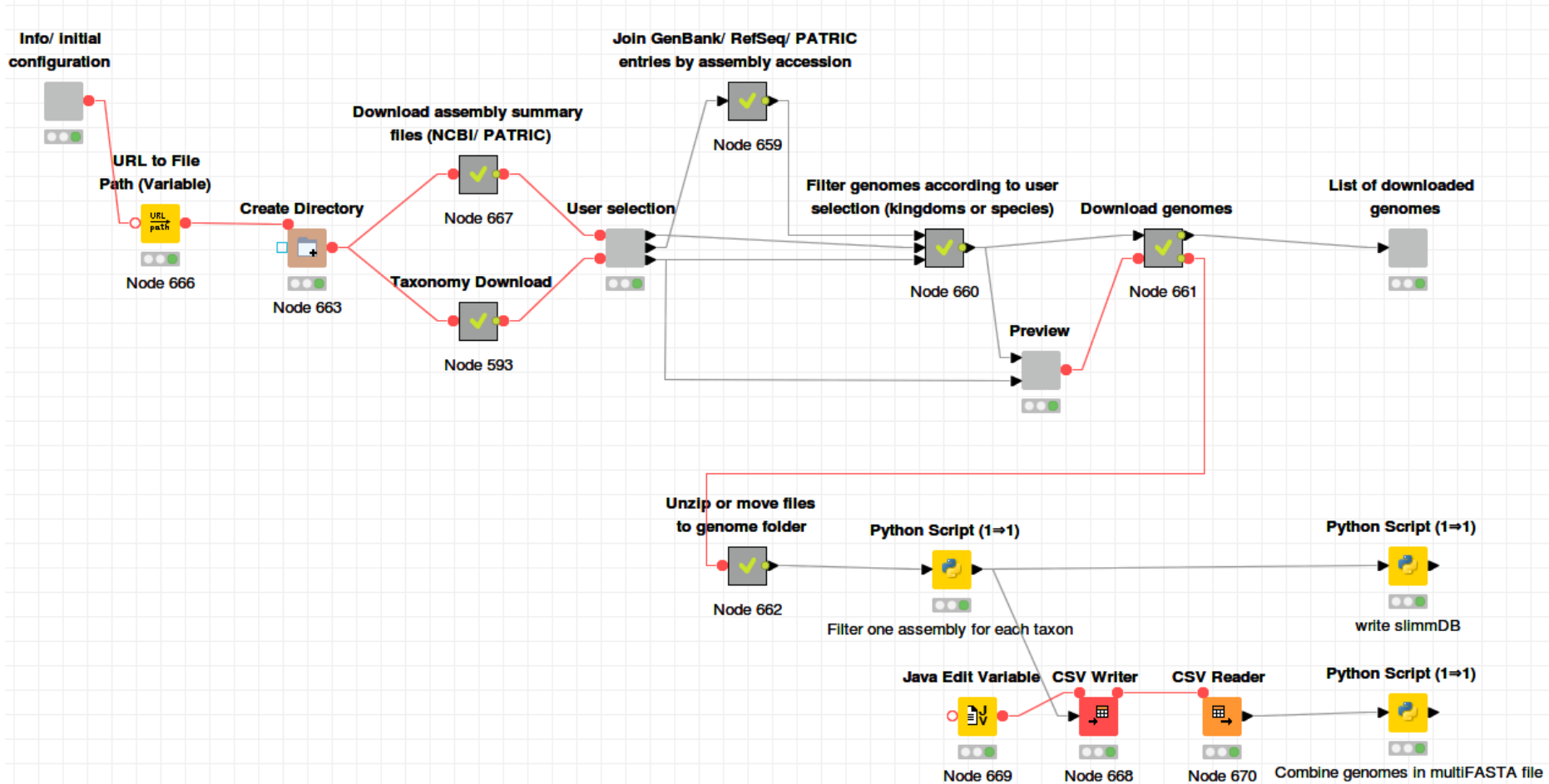
AP 4.3

AP 4.4

RefDB I (externe DB)



RefDB-workflow I (externe DB)



RefDB-workflow I (externe DB)



RefDB-workflowI_v1.1a2 2017-11-01 19.59.49

Species

Brucella
Francisella tularensis
Bacillus anthracis

Select database(s)

- RefSeq (NCBI's Reference Sequence collection)
- GenBank (annotated collection of all publicly available DNA sequences)
- PATRIC (Bacterial Bioinformatics Resource Center)

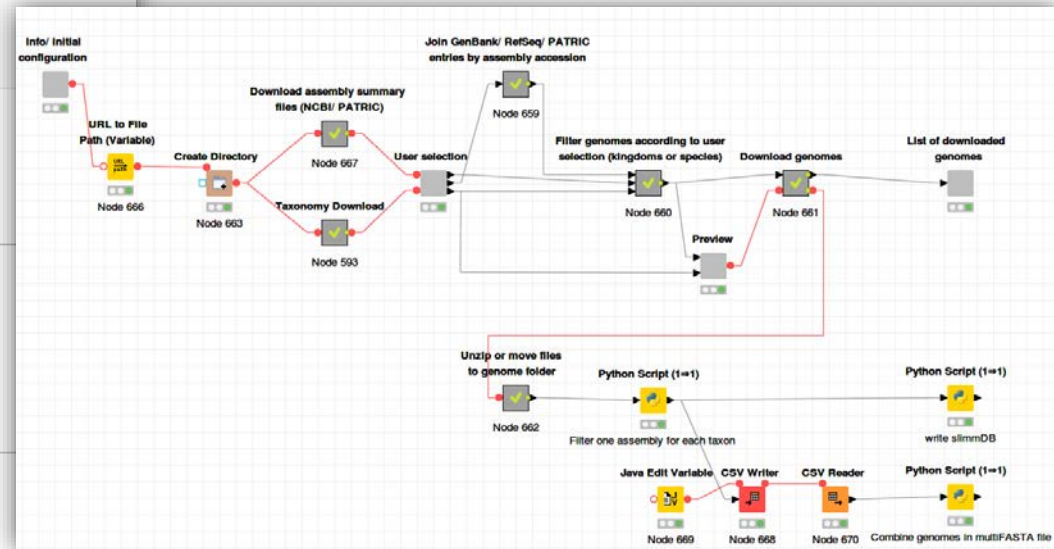
Select kingdoms(s)

- Archae
- Bacteria
- Viruses
- Eukaryota

Include draft assemblies

Species-level multiFASTA

©2017 KNIME.com AG - Version 4.5.1



RefDB-workflow I (externe DB)

Open for Innovation
KNIME WebPortal

Summary of assemblies

	database ↑↓	species_level_taxid ↑↓	number ↑↓	strains
■	RefSeq	1392	54	Bacillus anthracis(39), Bacillus anthracis 52-G(1), Bacillus anthracis 8903-G(1), Bacillus anthracis 9080-G(1), Bacillus anthracis str. 'Ames Ancestor'(1), Bacillus anthracis str. A0248(1), Bacillus anthracis str. anthracis str. H9401(1), Bacillus anthracis str. SVA11(1), Bacillus anthracis str. Sterne(2), Bacillus anthracis str. Turkey32(1), Bacillus anthracis str. V770-NP-1R(1), Bacillus anthracis str. Vollum(1)
■	RefSeq	235	26	Brucella abortus(18), Brucella abortus 104M(1), Brucella abortus 2308(1), Brucella abortus A13334(1), Brucella abortus S19(1), Brucella abortus bv. 1 str. 9-941(1), Brucella abortus bv. 2 str. 86/8/59(1), Bruc
■	RefSeq	36855	9	Brucella canis(6), Brucella canis ATCC 23365(1), Brucella canis HSK A52141(1), Brucella canis str. Oliveri(1)
■	RefSeq	120577	1	Brucella ceti TE10759-12(1)
■	RefSeq	29459	62	Brucella melitensis(54), Brucella melitensis ATCC 23457(1), Brucella melitensis M28(1), Brucella melitensis M5-90(1), Brucella melitensis NI(1), Brucella melitensis bv. 1 str. 16M(2), Brucella melitensis bv. 2 s
■	RefSeq	444163	1	Brucella microti CCM 4915(1)
■	RefSeq	236	1	Brucella ovis ATCC 25840(1)
■	RefSeq	120576	2	Brucella pinnipedialis(1), Brucella pinnipedialis B2/94(1)

RefDB-workflow I (externe DB)

Referenzgenome

Datenbank	Assemblierungslevel	<i>Brucella spp.</i>	<i>F. tularensis</i>	<i>B. anthracis</i>
RefSeq	Complete/ Chromosome Alle (inkl. draft)	124	39	54
		463	166	165
GenBank	Complete/ Chromosome Alle(inkl. draft)	139	39	60
		509 ¹	170	177
PATRIC	Complete Alle (inkl. WGS)	137	42	57
		650 ²	161	160
RefSeq/ PATRIC	Complete/ Chromosome Alle (inkl. draft/ WGS)	187	39	59
		710	226	178
GenBank/ PATRIC	Complete/ Chromosome All (inkl draft/ WGS)	187	39	62
		711 ³	230	179

Stand: 2.11.2017

¹*B. melitensis* (156), *B. abortus* (197), *B. suis* (58), *B. canis* (20), *B. ovis* (16), *B. ceti* (9),...

²*B. melitensis* (155), *B. abortus* (340), *B. suis* (63), *B. canis* (21), *B. ovis* (16), *B. ceti* (9),...

³*B. melitensis* (156+53), *B. abortus* (197+143), *B. suis* (58+6), *B. canis* (21+0), *B. ovis* (16+0), *B. ceti* (9+0),...

RefDB-workflow I (externe DB)

Funktionale Annotationen und STs

Datenbank	Sequenzen
Core VFDB	Gen- und Proteinsequenzen von experimentell verifizierten Virulenzfaktoren (2595)
Full VFDB	Gen- und Proteinsequenzen aller bekannten Virulenzfaktoren (26526)
CARD	Gen- und Proteinsequenzen von antimikrobiellen Resistenzgenen (2172)
PHAST	Proteinsequenzen von (Pro-)Phagen (257350)
pubMLST	Brucella MLST-21: 109 Sequenztypen in 21 Genen Francisella tularensis core MLST (<i>Antwerpen et al.</i>)

RefDB-workflow I (externe DBs)

- Genome/ Proteome ausgewählter Erreger
- Weitere Annotationen: Virulenzfaktoren, Resistenzen, Phagen, Sequenztypen
- Genom-Sequenzen und -Index aller bakteriellen Erreger aus NCBI oder PATRIC



Assembly-workflow

- Berechnung der Genomsequenzen und Annotationen der Referenzstämme



RefDB-workflow II (Ess-BAR-DB)

- Annotierte Genomsequenzen
- Proteinsequenzen mit Orthologen



Phylogenetische Analyse



Taxonomische Profilierung

- Identifikation der Spezies in einem Isolat
- Quantifizierung der Spezies in einer metagenomischen Probe



Speziesselektion

Analyse von Sub-Spezies und Stämmen

- Charakterisierung des Isolates z.B. anhand von Virulenzfaktoren, Resistenzen
- Intraspezifische Phylogeny (cgSNP)
- MLST-Analyse

Proteomische Analyse

LC/MS Analyse

- Identifikation von spezies-spezifischen Referenzpeptid-Profilen
- Charakterisierung der Proteome



Referenzmarker

Fingerprinting mit MRM/PRM-Methoden

- Identifizierung und Charakterisierung von Erreger-spezifischen Peptidmarkern

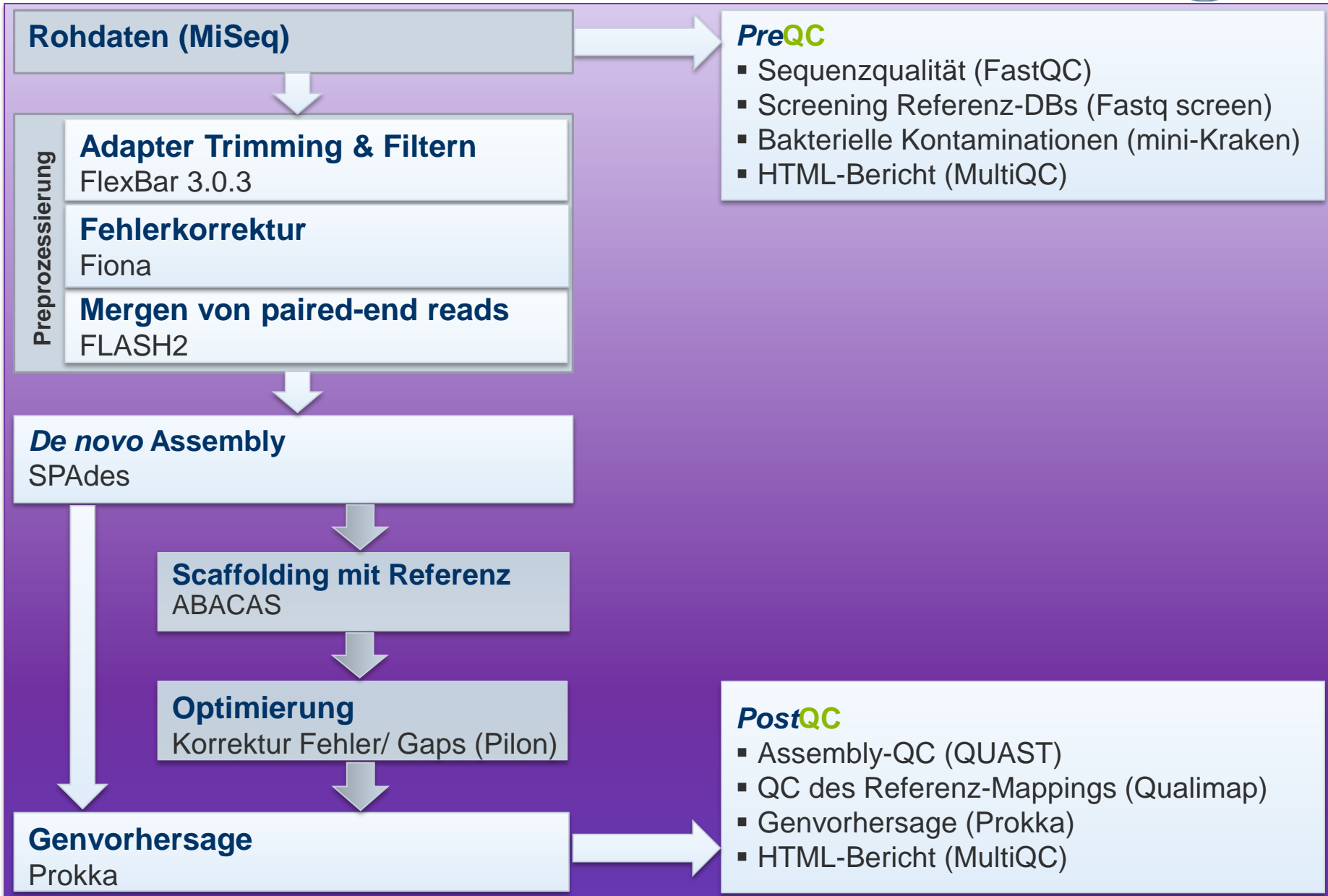
AP 4.1

AP 4.2

AP 4.3

AP 4.4

Assembly workflow

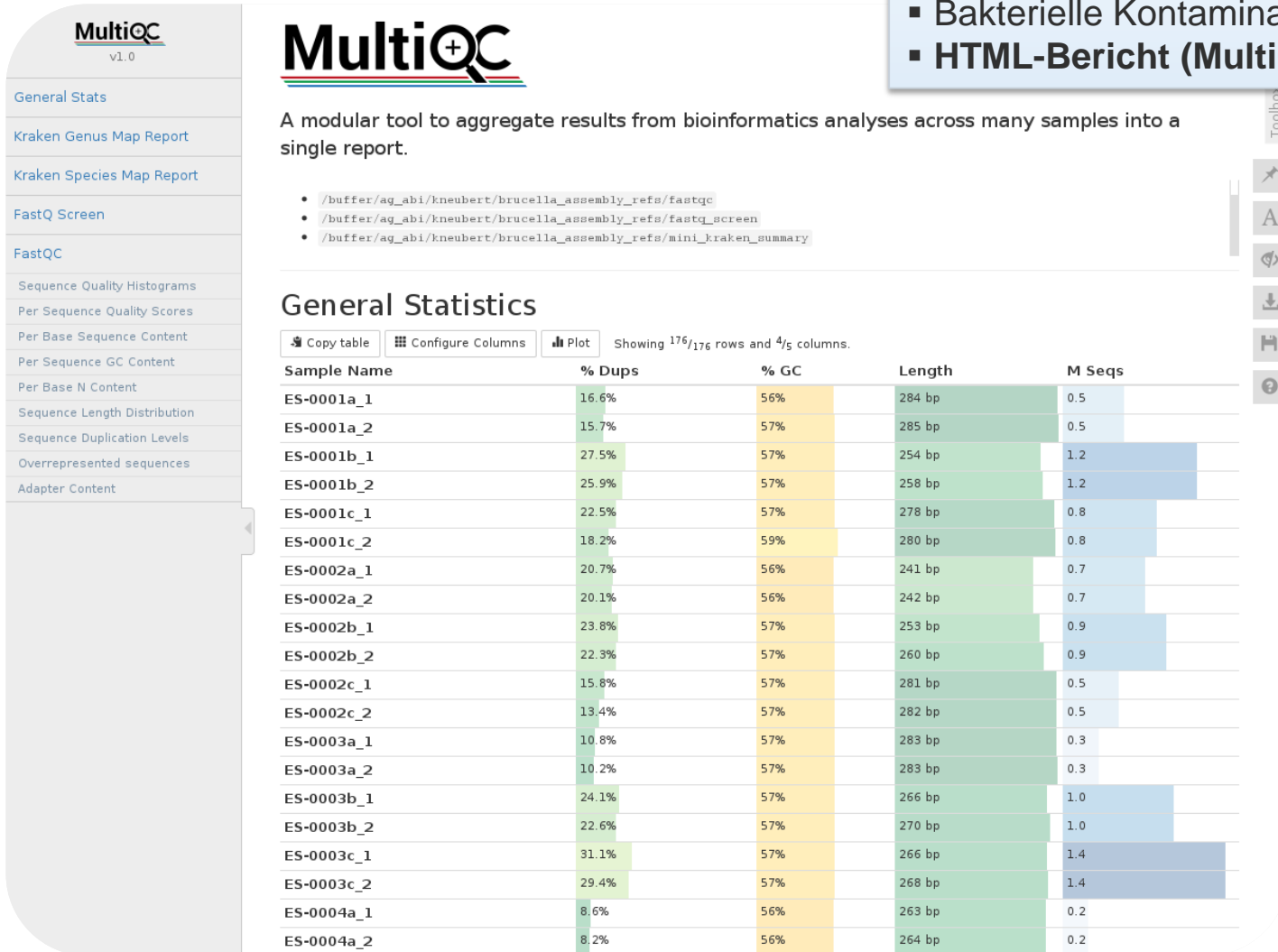


Assembly workflow

Rohdaten (MiSeq)

PreQC

- Sequenzqualität (FastQC)
- Screening Referenz-DBs (Fastq screen)
- Bakterielle Kontaminationen (mini-Kraken)
- **HTML-Bericht (MultiQC)**



Assembly workflow

Rohdaten (MiSeq)

PreQC

- Sequenzqualität (FastQC)
- Screening Referenz-DBs (Fastq screen)
- Bakterielle Kontaminationen (mini-Kraken)
- HTML-Bericht (MultiQC)

MultiQC
v1.0

- General Stats
- Kraken Genus Map Report
- Kraken Species Map Report
- FastQ Screen
- FastQC
- Sequence Quality Histograms
- Per Sequence Quality Scores
- Per Base Sequence Content
- Per Sequence GC Content
- Per Base N Content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Per Sequence Quality Scores 158

The number of reads with average quality scores. Shows if a subset of reads has poor quality. See the [FastQC help](#).

Flat image plot. Toolbox functions such as highlighting / hiding samples will not work (see the docs).



ES-0003c_2	29.4%	57%	268 bp	1.4
ES-0004a_1	8.6%	56%	263 bp	0.2
ES-0004a_2	8.2%	56%	264 bp	0.2

Assembly workflow

Rohdaten (MiSeq)

PreQC

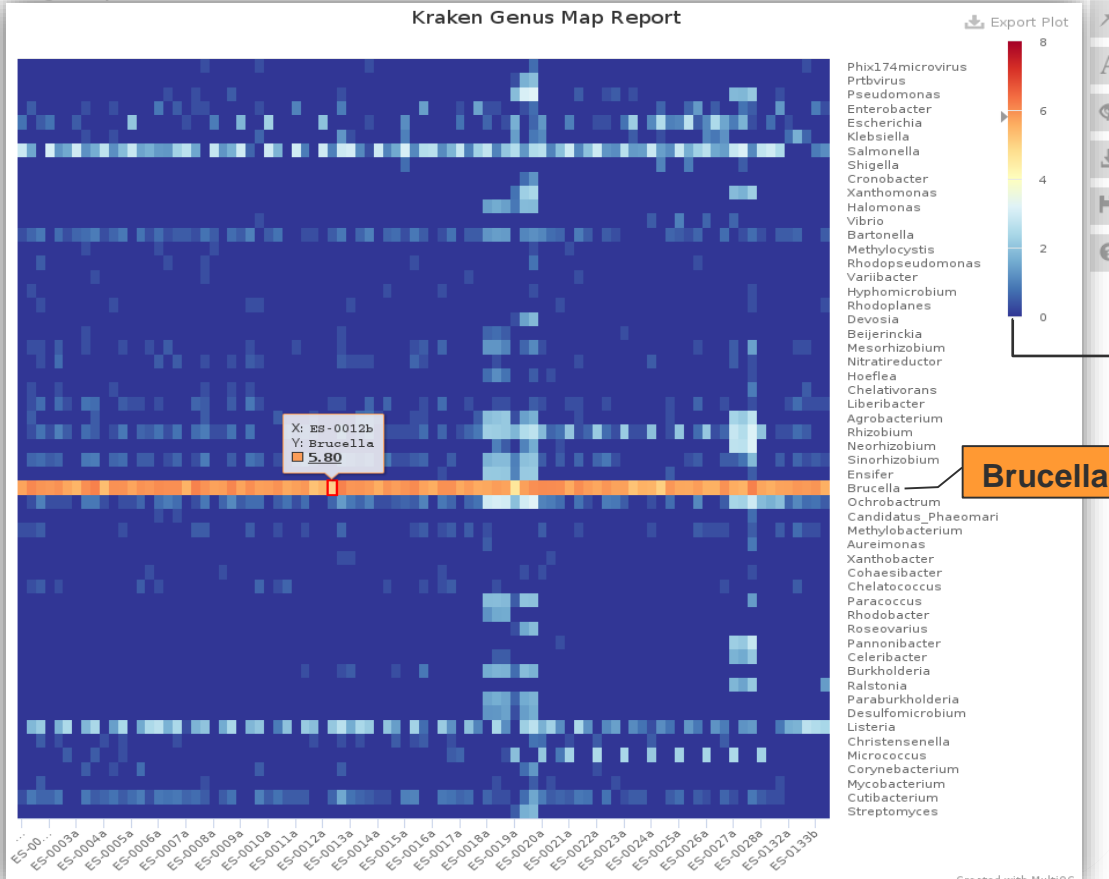
- Sequenzqualität (FastQC)
- Screening Referenz-DBs (Fastq screen)
- **Bakterielle Kontaminationen (mini-Kraken)**
- HTML-Bericht (MultiQC)

MultiQC
v1.0

- General Stats
- Kraken Genus Map Report
- Kraken Species Map Report
- FastQ Screen
- FastQC
- Sequence Quality Histograms
- Per Sequence Quality Scores
- Per Base Sequence Content
- Per Sequence GC Content
- Per Base N Content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.



Assembly workflow

Rohdaten (MiSeq)

PreQC

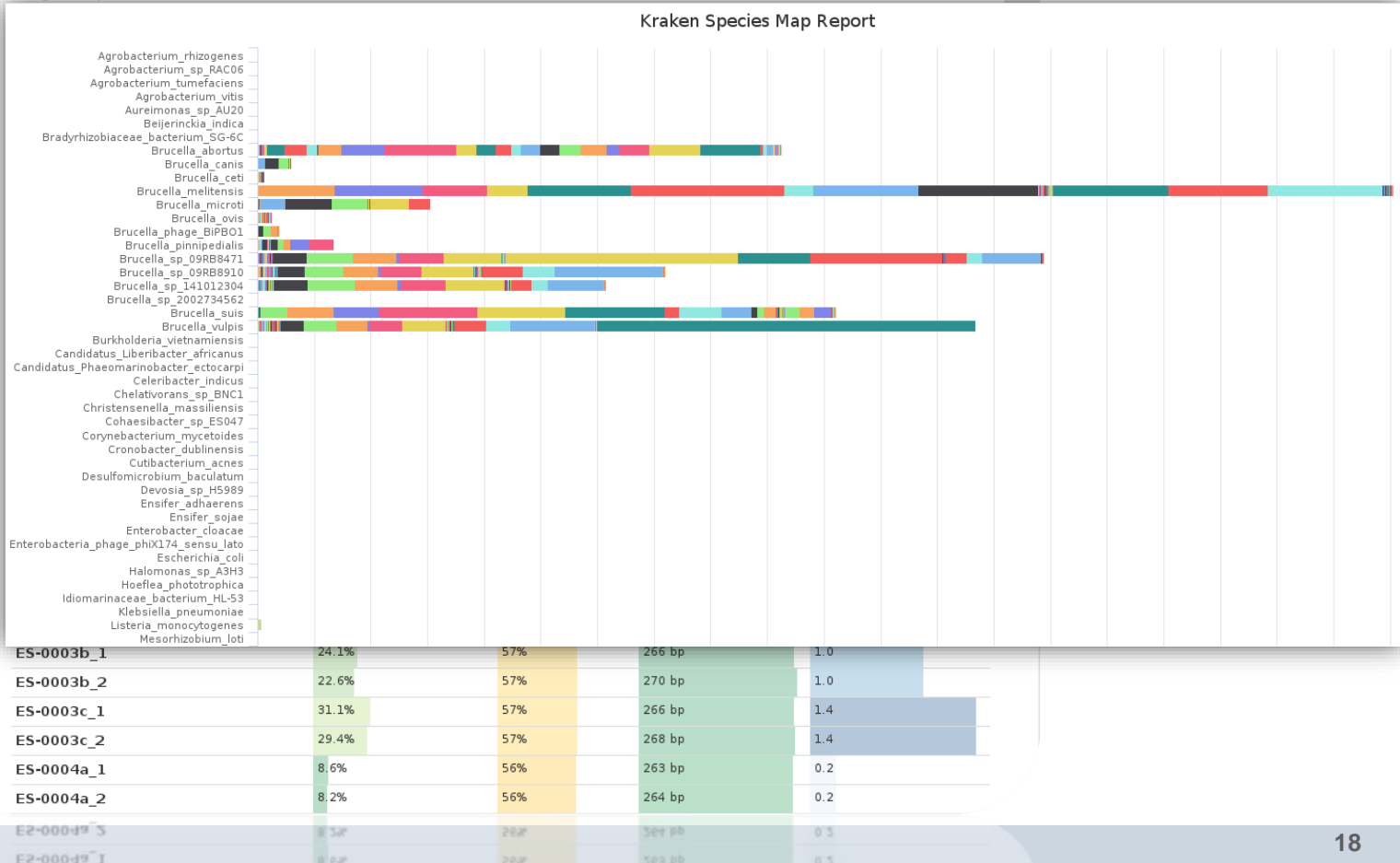
- Sequenzqualität (FastQC)
- Screening Referenz-DBs (Fastq screen)
- **Bakterielle Kontaminationen (mini-Kraken)**
- HTML-Bericht (MultiQC)

MultiQC
v1.0

- General Stats
- Kraken Genus Map Report
- Kraken Species Map Report
- FastQ Screen
- FastQC
- Sequence Quality Histograms
- Per Sequence Quality Scores
- Per Base Sequence Content
- Per Sequence GC Content
- Per Base N Content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.



MultiQC
v1.0

- General Stats
- QualiMap
- Coverage histogram
- Cumulative coverage genome fraction
- Insert size histogram
- GC content distribution
- QUAST
- Assembly Statistics
- Number of Contigs
- Prokka



A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

- /buffer/ag_abi/kneubert/brucella_assembly_refs/QUAST_denovo
- /buffer/ag_abi/kneubert/brucella_assembly_refs/qualimap
- /buffer/ag_abi/kneubert/brucella_assembly_refs/Prokka_contigs

Welcome! Not sure where to start? [Watch a tutorial video](#) (6:06) don't show again ✕

General Statistics

Copy table Configure Columns Plot Showing 90/90 rows and 10/17 columns.

Sample Name	Avg. GC	Insert Size	≥ 30X	Coverage	% Aligned	N50 (Kbp)	Length (Mbp)	Organism	Contigs	CDS
ES-0001a	58%	459bp	99.6%	90.0X	99.3%	250.8bp	3.3bp	ES-0132a	69.0	3141
ES-0001b	58%	376bp	99.8%	181.0X	99.0%	250.9bp	3.3bp	Brucella canis	32.0	3092
ES-0001c	59%	470bp	99.7%	136.0X	97.6%	367.7bp	3.3bp	Brucella canis	35.0	3089
ES-0002a	57%	453bp	99.8%	102.0X	99.1%	256.1bp	3.3bp	Brucella melitensis	26.0	3123
ES-0002b	58%	369bp	99.8%	137.0X	98.2%	221.8bp	3.3bp	Brucella melitensis	33.0	3122
ES-0002c	58%	541bp	99.4%	81.0X	98.2%	246.8bp	3.3bp	Brucella melitensis	24.0	3123
ES-0003a	58%	451bp	96.1%	52.0X	99.2%	189.7bp	3.3bp	Brucella melitensis	56.0	3141
ES-0003b	58%	390bp	99.8%	154.0X	99.0%	251.1bp	3.3bp	Brucella melitensis	33.0	3128
ES-0003c	58%	401bp	99.9%	225.0X	99.0%	251.1bp	3.3bp	Brucella melitensis	43.0	3130
ES-0004a	57%	493bp	66.2%	34.0X	98.8%	223.4bp	3.3bp	Brucella melitensis	36.0	3138
ES-0004b								Brucella melitensis	32.0	3143
ES-0004c	58%	477bp	99.7%	146.0X	98.8%	383.3bp	3.3bp	Brucella melitensis	30.0	3142
ES-0005a	57%	449bp	99.8%	75.0X	99.2%	207.8bp	3.3bp	Brucella suis	53.0	3073
ES-0005b	58%	399bp	99.8%	144.0X	99.1%	251.7bp	3.3bp	Brucella suis	38.0	3067
ES-0005c	58%	458bp	99.8%	127.0X	98.7%	334.6bp	3.3bp	Brucella suis	33.0	3070
ES-0006a	58%	386bp	99.9%	165.0X	98.6%	170.3bp	3.3bp	Brucella suis	37.0	3120
ES-0006b	58%	387bp	99.8%	148.0X	98.9%	140.1bp	3.3bp	Brucella suis	37.0	3127
E2-0000P	78%	381bp	99.8%	148.0X	98.8%	140.1bp	3.3bp	Brucella melitensis	31.0	3111
E2-0000Y	78%	386bp	99.8%	162.0X	98.8%	110.0bp	3.3bp	Brucella melitensis	31.0	3110
E2-0002C	78%	428bp	99.8%	171.0X	99.1%	134.0bp	3.3bp	Brucella melitensis	33.0	3010
E2-0002P	78%	388bp	99.8%	169.0X	98.8%	140.1bp	3.3bp	Brucella melitensis	36.0	3081
E2-0002Y	71%	412bp	99.8%	12.0X	98.8%	140.1bp	3.3bp	Brucella melitensis	29.0	3013
E2-0004C	78%	411bp	99.8%	148.0X	98.8%	140.1bp	3.3bp	Brucella melitensis	30.0	3141
E2-0004P	71%	411bp	99.8%	148.0X	98.8%	140.1bp	3.3bp	Brucella melitensis	30.0	3141

PostQC

- Assembly-QC (QUAST)
- QC des Referenz-Mappings (Qualimap)
- Genvorhersage (Prokka)
- **HTML-Bericht (MultiQC)**

MultiQC
v1.0

- General Stats
- QualiMap
- Coverage histogram
- Cumulative coverage genome fraction
- Insert size histogram
- GC content distribution
- QUAST**
- Assembly Statistics
- Number of Contigs
- Prokka

MultiQC

QUAST

QUAST is a quality assessment tool for genome assemblies, written by the Center for Algorithmic Biotechnology.

Assembly Statistics

Copy table | Configure Columns | Plot | Showing 87/87 rows and 7/10 columns.

Sample Name	N50 (Kbp)	Largest contig (Kbp)	Length (Mbp)	Misassemblies	Mismatches/100kbp	Indels/100kbp	Genome Fraction
ES-0001a	250.8bp	470.9bp	3.3bp	0.0	0.91	0.64	99.4%
ES-0001b	250.9bp	456.7bp	3.3bp	0.0	0.58	0.36	99.4%
ES-0001c	367.7bp	471.0bp	3.3bp	0.0	0.70	0.67	99.4%
ES-0002a	256.1bp	462.5bp	3.3bp	1.0	0.76	1.10	99.4%
ES-0002b	221.8bp	462.6bp	3.3bp	0.0	1.59	1.04	99.4%
ES-0002c	246.8bp	762.0bp	3.3bp	1.0	0.64	1.25	99.4%
ES-0003a	189.7bp	291.4bp	3.3bp	2.0	8.30	1.64	99.3%
ES-0003b	251.1bp	475.4bp	3.3bp	1.0	8.82	2.04	99.3%
ES-0003c	251.1bp	462.2bp	3.3bp	0.0	8.24	1.76	99.3%
ES-0004a	223.4bp	460.0bp	3.3bp	0.0	1.03	0.64	99.4%
ES-0004c	383.3bp	460.0bp	3.3bp	0.0	0.67	0.67	99.4%
ES-0005a	207.8bp	522.5bp	3.3bp	0.0	1.91	0.30	99.3%
ES-0005b	251.7bp	463.7bp	3.3bp	0.0	2.00	0.36	99.3%
ES-0005c	334.6bp	391.1bp	3.3bp	2.0	1.88	0.33	99.3%
ES-0006a	170.3bp	459.1bp	3.3bp	0.0	0.21	0.45	99.3%
ES-0006b	140.1bp	384.1bp	3.3bp				

PostQC

- Assembly-QC (QUAST)
- QC des Referenz-Mappings (Qualimap)
- Genvorhersage (Prokka)
- HTML-Bericht (MultiQC)

MultiQC
v1.0

General Stats

QualiMap

Coverage histogram

Cumulative coverage genome fraction

Insert size histogram

GC content distribution

QUAST

Assembly Statistics

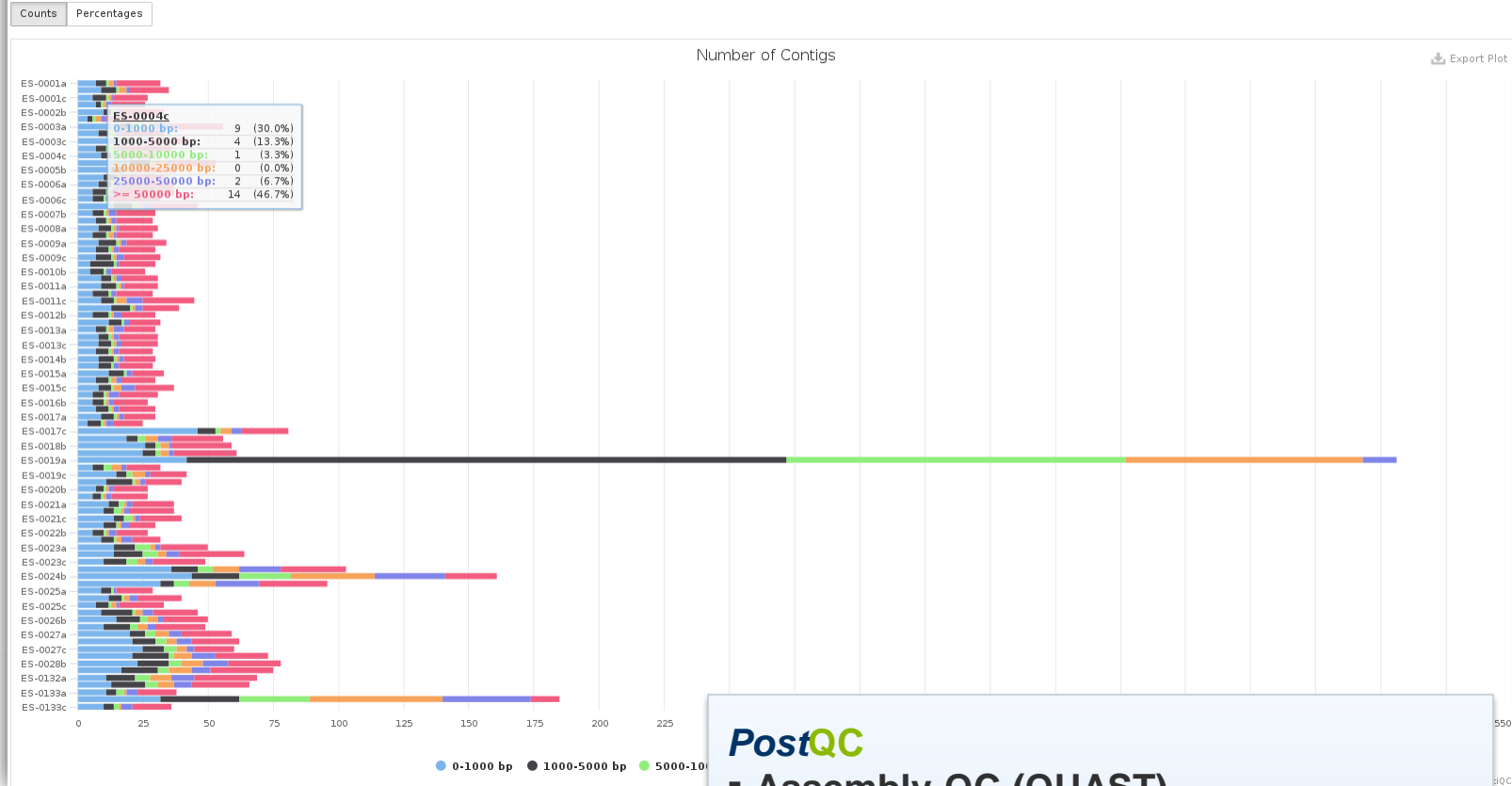
Number of Contigs

Prokka

MultiQC

Number of Contigs

This plot shows the number of contigs found for each assembly, broken down by length.



PostQC

- Assembly-QC (QUAST)
- QC des Referenz-Mappings (Qualimap)
- Genvorhersage (Prokka)
- HTML-Bericht (MultiQC)

MultiQC
v1.0

General Stats

QualiMap

Coverage histogram

Cumulative coverage genome fraction

Insert size histogram

GC content distribution

QUAST

Assembly Statistics

Number of Contigs

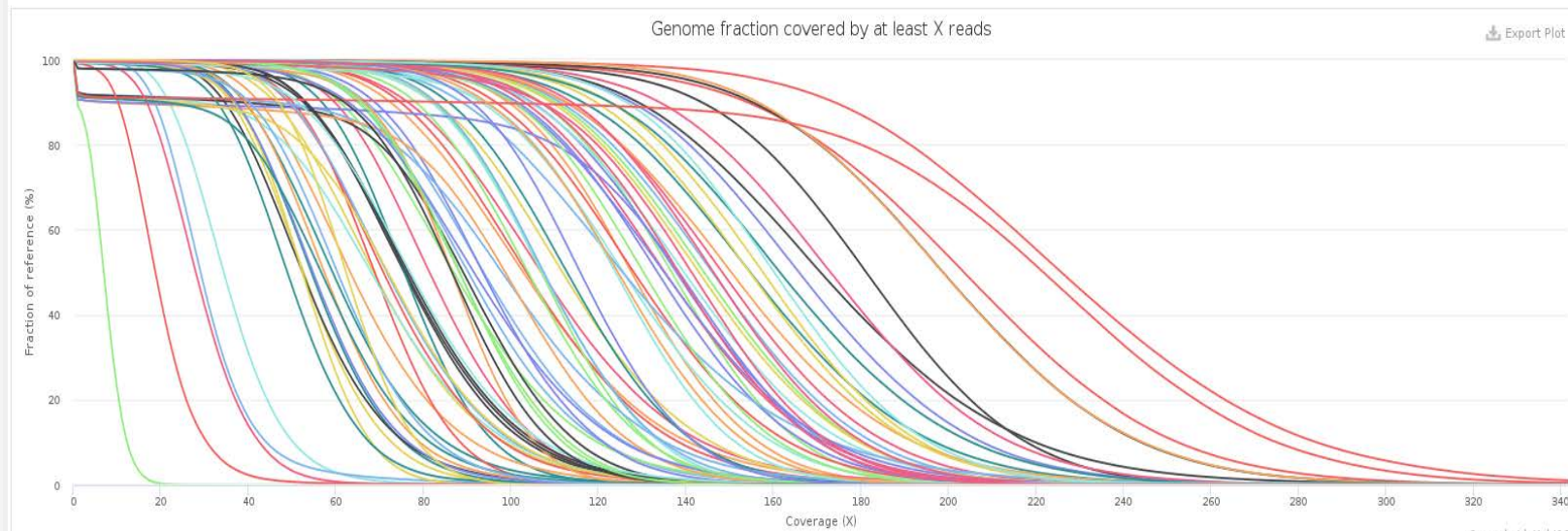
Prokka

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

- /buffer/ag_abi/kneubert/brucella_assembly_refs/QUAST_denovo
- /buffer/ag_abi/kneubert/brucella_assembly_refs/qualimap
- /buffer/ag_abi/kneubert/brucella_assembly_refs/Prokka_contigs

Cumulative coverage genome fraction



E2-0000P	75P	3812P	38.5P	762.0K	38.5P	240.22P	8.38P	Genome size	31.2	312.1
E2-0000Y	75P	3882P	38.5P	762.0K	38.5P	240.22P	8.38P	Genome size	31.2	312.1
E2-0002K	75P	4289P	38.5P	751.0K	38.5P	240.22P	8.38P	Genome size	31.2	312.1
E2-0002P	75P	3889P	38.5P	762.0K	38.5P	240.22P	8.38P	Genome size	31.2	312.1
E2-0002Y	71P	4120P	38.5P	762.0K	38.5P	240.22P	8.38P	Genome size	31.2	312.1
E2-0004K	75P	4110P	38.5P	762.0K	38.5P	240.22P	8.38P	Genome size	31.2	312.1
E2-0004P	75P	3889P	38.5P	762.0K	38.5P	240.22P	8.38P	Genome size	31.2	312.1
E2-0004Y	71P	4120P	38.5P	762.0K	38.5P	240.22P	8.38P	Genome size	31.2	312.1
E2-0005P	75P	3889P	38.5P	762.0K	38.5P	240.22P	8.38P	Genome size	31.2	312.1
E2-0005Y	71P	4120P	38.5P	762.0K	38.5P	240.22P	8.38P	Genome size	31.2	312.1

PostQC

- Assembly-QC (QUAST)
- QC des Referenz-Mappings (Qualimap)
- Genvorhersage (Prokka)
- HTML-Bericht (MultiQC)

MultiQC
v1.0

MultiQC

- General Stats
- QualiMap
- Coverage histogram
- Cumulative coverage genome fraction
- Insert size histogram
- GC content distribution
- QUAST
- Assembly Statistics
- Number of Contigs
- Prokka

Prokka

Prokka is a software tool for the rapid annotation of prokaryotic genomes.

Showing 85/85 rows and 5/5 columns.

Sample Name	Organism	# contigs	# bases	# CDS	# tRNA
ES-0001a	ES-0132a Brucella canis	69 32	3367472 3296796	3141 3092	54 54
ES-0001b	Brucella canis	35	3294348	3089	54
ES-0001c	Brucella canis	27	3294365	3092	54
ES-0002a	Brucella melitensis	26	3277810	3123	53
ES-0002b	Brucella melitensis	33	3277743	3122	53
ES-0002c	Brucella melitensis	24	3276496	3123	53
ES-0003a	Brucella melitensis	56	3293903	3141	54
ES-0003b	Brucella melitensis	33	3292219	3128	54
ES-0003c	Brucella melitensis	43	3292550	3130	54
ES-0004a	Brucella melitensis	36	3293076	3138	54
ES-0004b	Brucella melitensis	32	3292236	3143	54
ES-0004c	Brucella melitensis	30	3291984	3142	54
ES-0005a	Brucella suis	53	3307956	3073	54
ES-0005b	Brucella suis	38	3300252	3067	54
ES-0005c	Brucella suis	33	3301008	3070	54
ES-0006a	Brucella suis	37	3303141	3120	53
ES-0006b	Brucella suis	37			
ES-0006c	Brucella suis	32			
ES-0007a	Brucella suis	46			
ES-0007b	Brucella suis	30			
ES-0007c	Brucella suis	29			
ES-0008a	Brucella suis	31			
ES-0008b	Brucella suis	34			
ES-0008c	Brucella suis	29			

PostQC

- Assembly-QC (QUAST)
- QC des Referenz-Mappings (Qualimap)
- **Genvorhersage (Prokka)**
- HTML-Bericht (MultiQC)

MultiQC
v1.0

- General Stats
- QualiMap
- Coverage histogram
- Cumulative coverage genome fraction
- Insert size histogram
- GC content distribution
- QUAST
- Assembly Statistics
- Number of Contigs
- Prokka

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

- /buffer/ag_abi/kneubert/brucella_assembly_refs/QUAST_denovo
- /buffer/ag_abi/kneubert/brucella_assembly_refs/qualimap
- /buffer/ag_abi/kneubert/brucella_assembly_refs/Prokka_contigs

Welcome! Not sure where to start? [Watch a tutorial video](#) (6:06) don't show again ✕

General Statistics

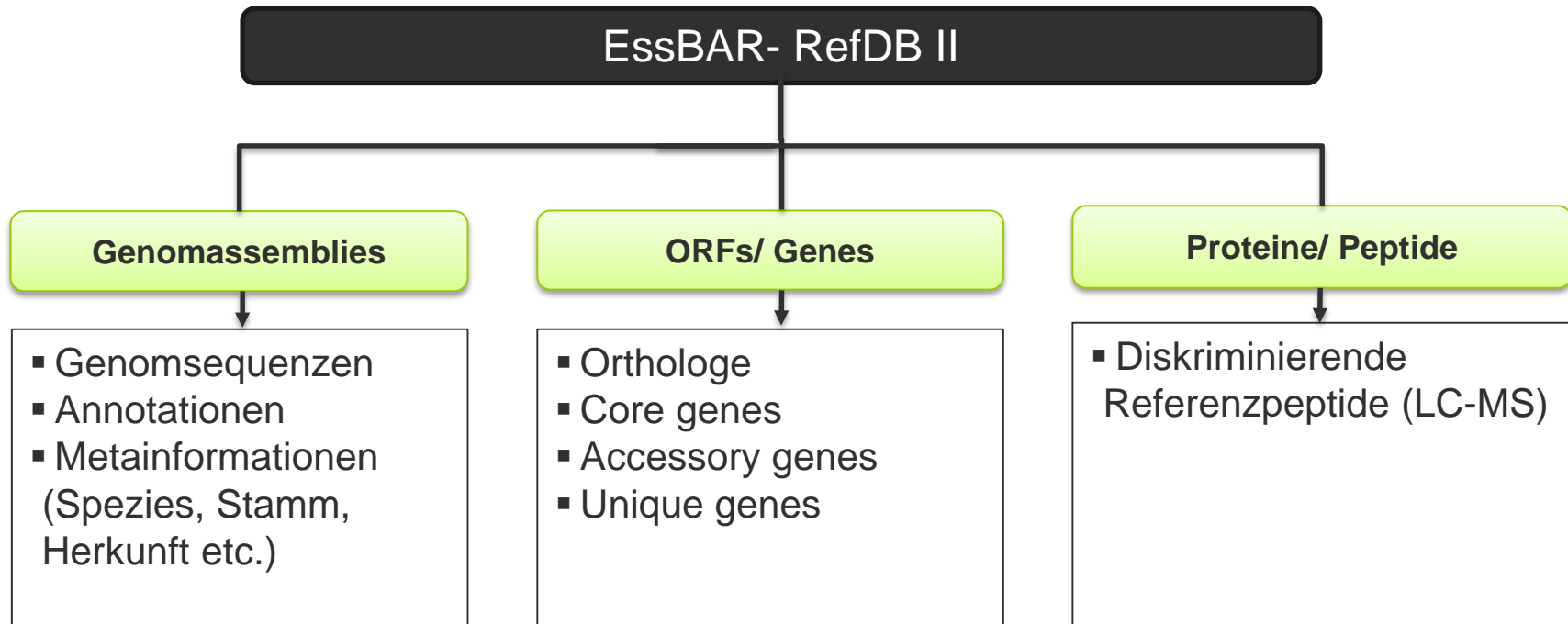
Copy table Configure Columns Plot Showing 90/90 rows and 10/17 columns.

Sample Name	Avg. GC	Insert Size	≥ 30X	Coverage	% Aligned	N50 (Kbp)	Length (Mbp)	Organism	Contigs	CDS
ES-0001a	58%	459bp	99.6%	90.0X	99.3%	250.8bp	3.3bp	ES-0132a	69.0	3141
ES-0001b	58%	376bp	99.8%	181.0X	99.0%	250.9bp	3.3bp	Brucella canis	32.0	3092
ES-0001c	59%	470bp	99.7%	136.0X	97.6%	367.7bp	3.3bp	Brucella canis	35.0	3089
ES-0002a	57%	453bp	99.8%	102.0X	99.1%	256.1bp	3.3bp	Brucella melitensis	26.0	3123
ES-0002b	58%	369bp	99.8%	137.0X	98.2%	221.8bp	3.3bp	Brucella melitensis	33.0	3122
ES-0002c	58%	541bp	99.4%	81.0X	98.2%	246.8bp	3.3bp	Brucella melitensis	24.0	3123
ES-0003a	58%	451bp	96.1%	52.0X	99.2%	189.7bp	3.3bp	Brucella melitensis	56.0	3141
ES-0003b	58%	390bp	99.8%	154.0X	99.0%	251.1bp	3.3bp	Brucella melitensis	33.0	3128
ES-0003c	58%	401bp	99.9%	225.0X	99.0%	251.1bp	3.3bp	Brucella melitensis	43.0	3130
ES-0004a	57%	493bp	66.2%	34.0X	98.8%	223.4bp	3.3bp	Brucella melitensis	36.0	3138
ES-0004b								Brucella melitensis	32.0	3143
ES-0004c	58%	477bp	99.7%	146.0X	98.8%	383.3bp	3.3bp	Brucella melitensis	30.0	3142
ES-0005a	57%	449bp	99.8%	75.0X	99.2%	207.8bp	3.3bp	Brucella suis	53.0	3073
ES-0005b	58%	399bp	99.8%	144.0X	99.1%	251.7bp	3.3bp	Brucella suis	38.0	3067
ES-0005c	58%	458bp	99.8%	127.0X	98.7%	334.6bp	3.3bp	Brucella suis	33.0	3070
ES-0006a	58%	386bp	99.9%	165.0X	98.6%	170.3bp	3.3bp	Brucella suis	37.0	3120
ES-0006b	58%	387bp	99.8%	148.0X	98.9%	140.1bp	3.3bp	Brucella suis	37.0	3127
E2-0008P	76%	381bp	99.8%	148.0X	98.8%	140.1bp	3.3bp	Brucella melitensis	31.0	3111
E2-0008Y	76%	386bp	99.8%	162.0X	98.8%	110.0bp	3.3bp	Brucella melitensis	31.0	3110
E2-0002C	76%	428bp	99.8%	171.0X	98.8%	134.0bp	3.3bp	Brucella melitensis	33.0	3010
E2-0002P	76%	388bp	99.8%	169.0X	98.8%	140.1bp	3.3bp	Brucella melitensis	30.0	3081
E2-0002Y	71%	412bp	99.8%	12.0X	98.8%	140.1bp	3.3bp	Brucella melitensis	29.0	3013
E2-0004C	76%	411bp	99.8%	148.0X	98.8%	140.1bp	3.3bp	Brucella melitensis	30.0	3141
E2-0004P	71%	411bp	99.8%	148.0X	98.8%	140.1bp	3.3bp	Brucella melitensis	30.0	3141
E2-0004Y	71%	411bp	99.8%	148.0X	98.8%	140.1bp	3.3bp	Brucella melitensis	30.0	3141

PostQC

- Assembly-QC (QUAST)
- QC des Referenz-Mappings (Qualimap)
- Genvorhersage (Prokka)
- HTML-Bericht (MultiQC)

RefDB II (interne DB)



RefDB II (interne DB)

Genomassembles

Quelle	Spezies	Daten	<i>De novo</i> Assemblies	Scaffold Assemblies
BfR	<i>Brucella spp.</i>	30 Referenzstämme (29 Triplikate, 1 Dublikat)	89/89	89/89
BfR	<i>Brucella spp.</i>	390 Feldstämme	81/390	81/390
FLI	<i>Francisella tularensis</i>	Referenzstämme	7/44	7/44








RefDB II (interne DB)

Diskriminierende Referenzpeptide

RefDB II (interne DB)

Diskriminierende Referenzpeptide

Zusammenfassung zum Stand der APs

AP	Inhalt	Stand
4.1 Referenz- datenbank	Zusammenstellung ausgewählter Referenzgenome- und proteome mit Hilfe des RefDB-workflows I (KNIME)	
	Genom- und Proteomsequenzen zu MLST-Analysen, Virulenzfaktoren, Resistenzen und Phagen	
	Berechnung von diskriminierenden Peptidsequenzen mit LC-MS	
4.2 Assembly- pipeline	Qualitätskontrolle und Fehlerbehandlung der Rohdaten (mit FastQC, FastQ Screen, Mini-Kraken und FlexBar)	
	De novo Assembly des Genoms (mit SPAdes)	
	Referenzbasiertes Scaffolding der Contigs (mit ABACAS)	
	Optimierung des Assemblies durch Schließung der Lücken mit einem Insert Assembly (PILON)	
	Berechnung der Genomassemblies und Proteinsequenzen für die Isolate (Übergabe an AP2 und AP5)	