Efficient Grouping Methods for the Annotation and Sorting of Single Cells

Disputation

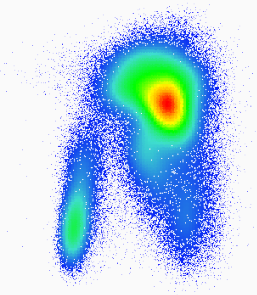
Markus Lux

April 24, 2018

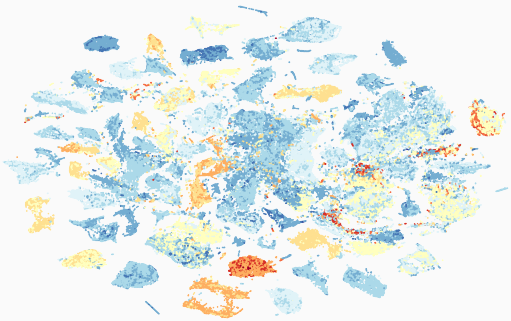
Overview

Efficient grouping methods for
the annotation and sorting of single cells

Flow cytometry



Metagenomics

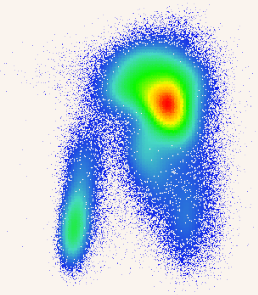


Single-cell genomics

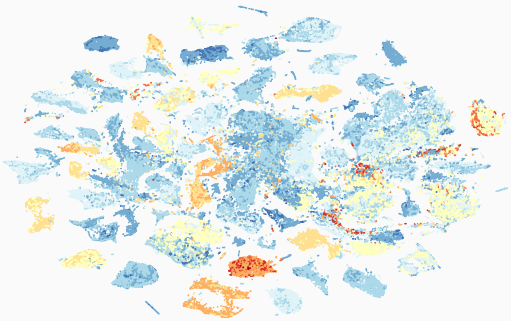
Overview

Efficient grouping methods for
the annotation and sorting of single cells

Flow cytometry



Metagenomics



Single-cell genomics

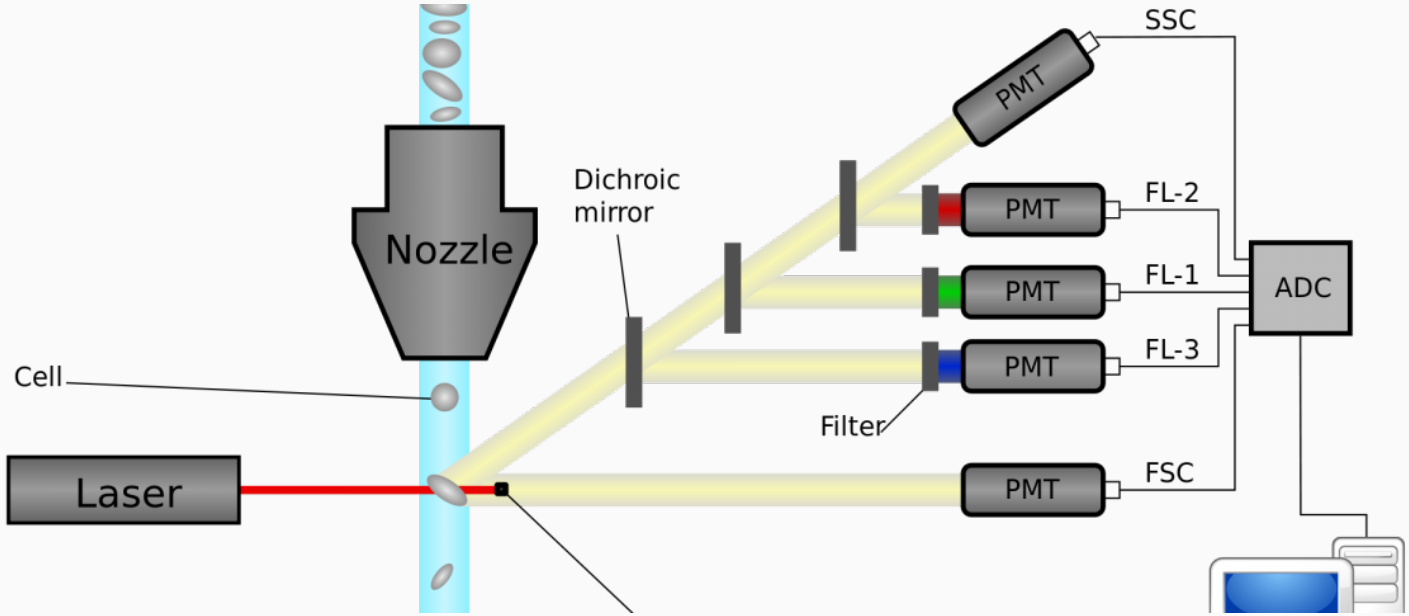
Flow Cytometry in a nutshell

- Use laser to measure cell parameters
- Cell counting and sorting using fluorescent markers
  - Viability
  - Antigens
  - Protein expression
  - DNA characteristics

Flow Cytometry in a nutshell

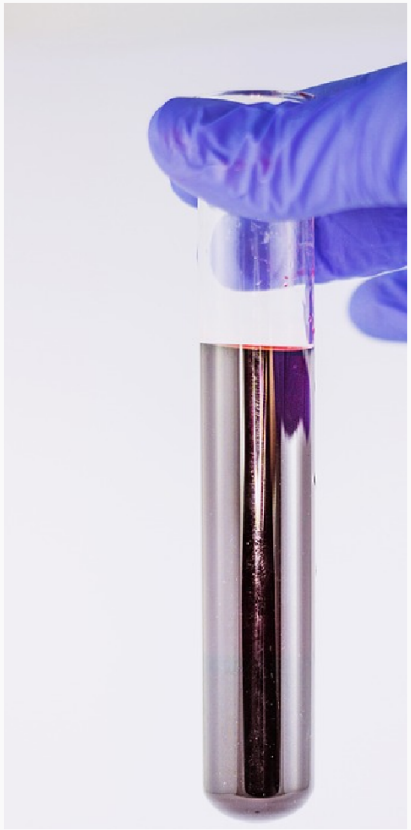
- Use laser to measure cell parameters
- Cell counting and sorting using fluorescent markers
  - Viability
  - Antigens
  - Protein expression
  - DNA characteristics

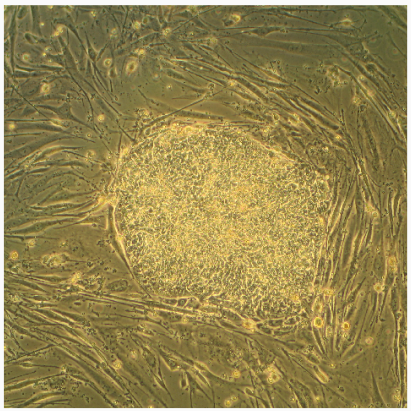
Flow Cytometry data generation



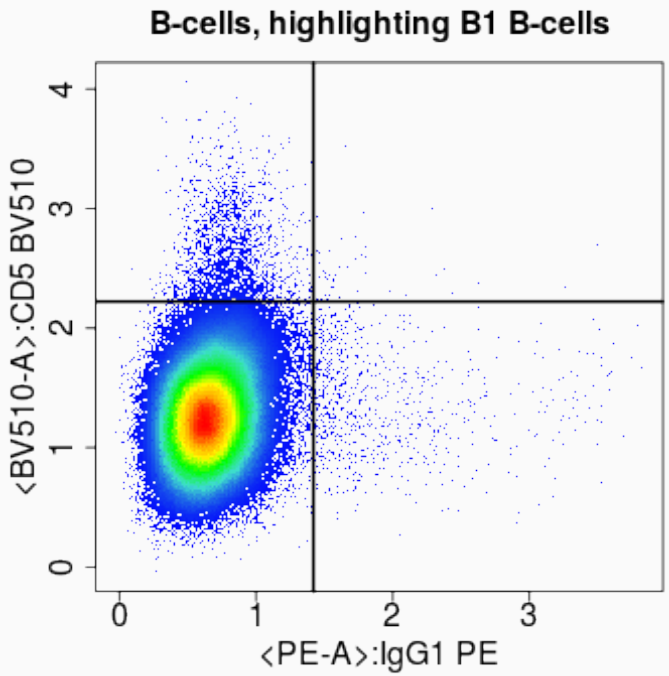
→ cells represented using
max. 20 features (channels/markers)

Flow Cytometry Applications

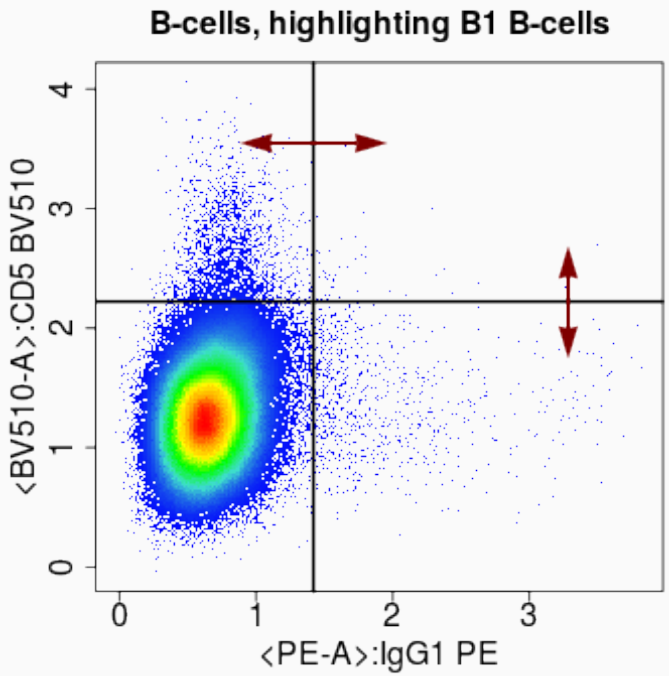




Gating


B-cells, highlighting B1 B-cells
<BV510-A>:CD5 BV510
<PE-A>:IgG1 PE
0
1
2
3
4

Gating



Measure of interest: **cell frequency**

Existing gating solutions

- **Manual**: resource-intensive, subjective
  Finak et al. (2016)
- **Clustering**: inaccurate for complex, rare populations
  Aghaeepour et al. (2011); Qiu et al. (2011); Van Gassen et al. (2015); Ge and Sealfon (2012); ...
- **Supervised**: difficult to tune parameters
  Finak et al. (2014); Malek et al. (2015); Li et al. (2017); ...
- **No gating**: missing acceptance, limited interpretability
  Mair et al. (2016)

Existing gating solutions

- **Manual**: resource-intensive, subjective
  Finak et al. (2016)

- **Clustering**: inaccurate for complex, rare populations
  Aghaeepour et al. (2011); Qiu et al. (2011); Van Gassen et al. (2015); Ge and Sealfon (2012); ...

- **Supervised**: difficult to tune parameters
  Finak et al. (2014); Malek et al. (2015); Li et al. (2017); ...

- **No gating**: missing acceptance, limited interpretability
  Mair et al. (2016)

Existing gating solutions

- **Manual**: resource-intensive, subjective
  Finak et al. (2016)

- **Clustering**: inaccurate for complex, rare populations
  Aghaeepour et al. (2011); Qiu et al. (2011); Van Gassen et al. (2015); Ge and Sealfon (2012); ...

- **Supervised**: difficult to tune parameters
  Finak et al. (2014); Malek et al. (2015); Li et al. (2017); ...

- **No gating**: missing acceptance, limited interpretability
  Mair et al. (2016)

Existing gating solutions

- **Manual**: resource-intensive, subjective
  Finak et al. (2016)
- **Clustering**: inaccurate for complex, rare populations
  Aghaeepour et al. (2011); Qiu et al. (2011); Van Gassen et al. (2015); Ge and Sealfon (2012); ...
- **Supervised**: difficult to tune parameters
  Finak et al. (2014); Malek et al. (2015); Li et al. (2017); ...
- **No gating**: missing acceptance, limited interpretability
  Mair et al. (2016)

Existing gating solutions

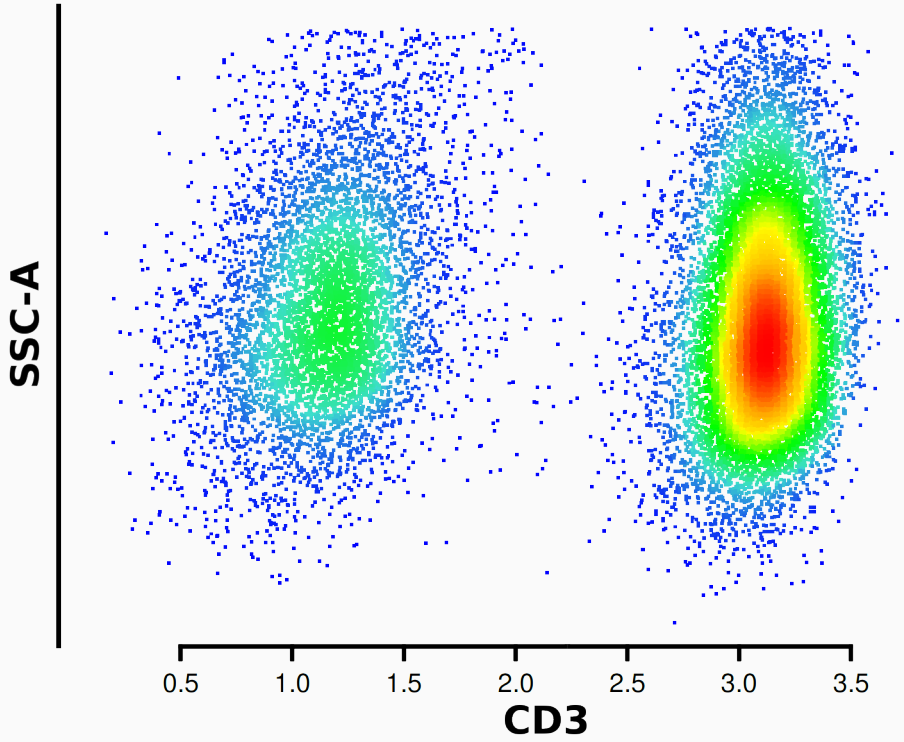
- **Manual**: resource-intensive, subjective
  Finak et al. (2016)
- **Clustering**: inaccurate for complex, rare populations
  Aghaeepour et al. (2011); Qiu et al. (2011); Van Gassen et al. (2015); Ge and Sealfon (2012); ...
- **Supervised**: difficult to tune parameters
  Finak et al. (2014); Malek et al. (2015); Li et al. (2017); ...
- **No gating**: missing acceptance, limited interpretability
  Mair et al. (2016)

$\rightarrow$ Semi-supervised approach:

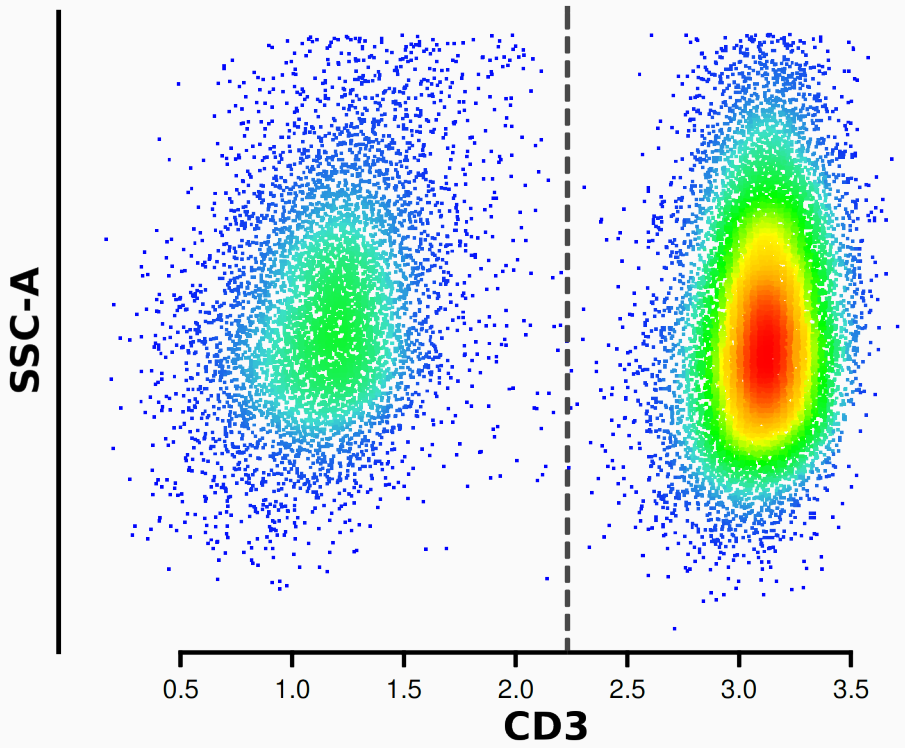
flowLearn

Lux et al. (2018)

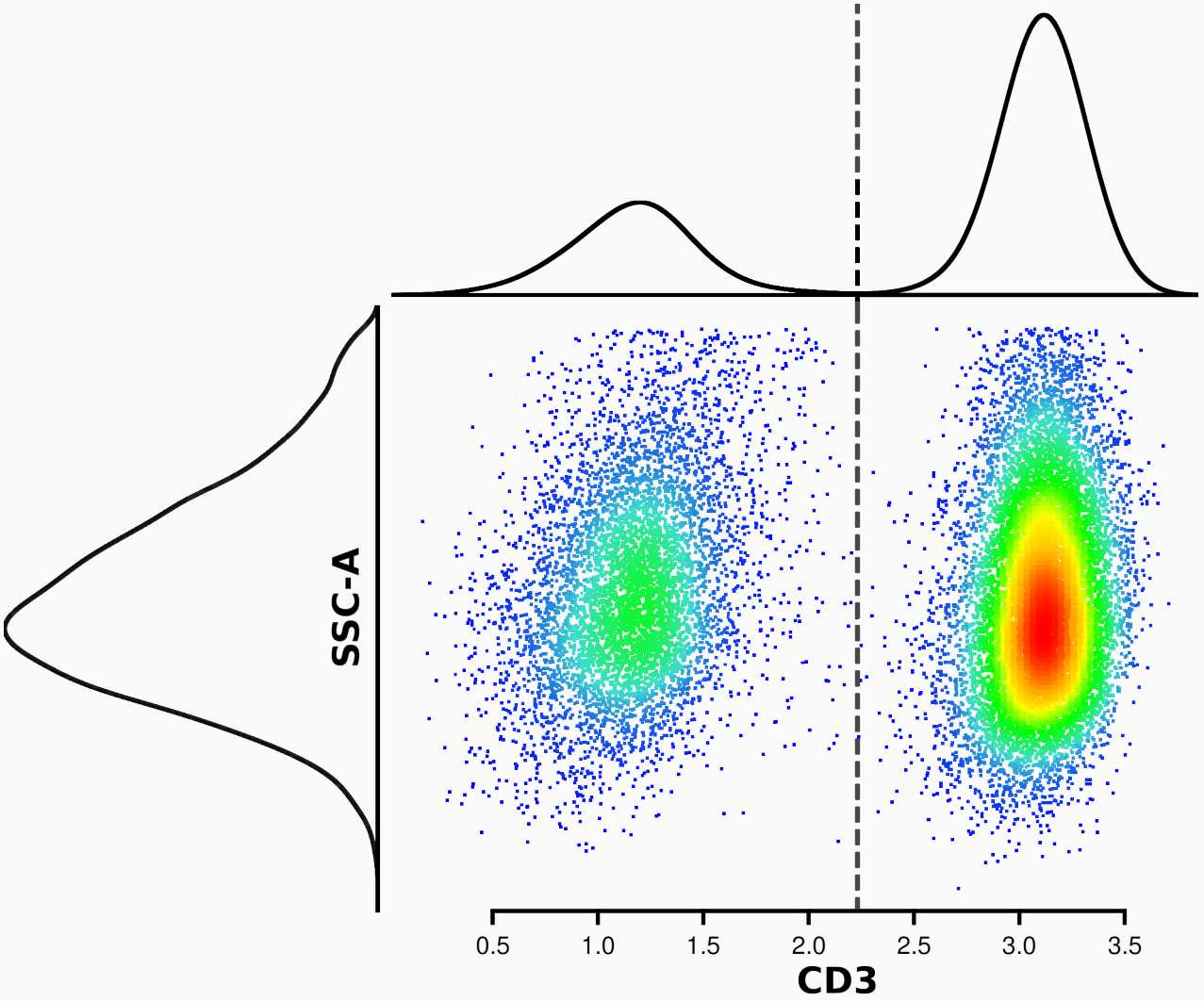
FlowLearn approach



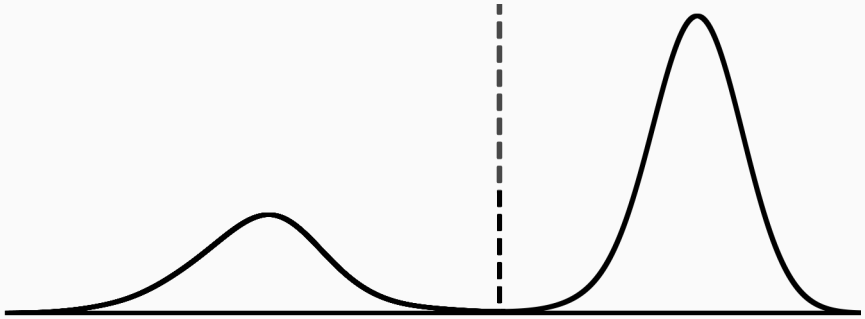
FlowLearn approach



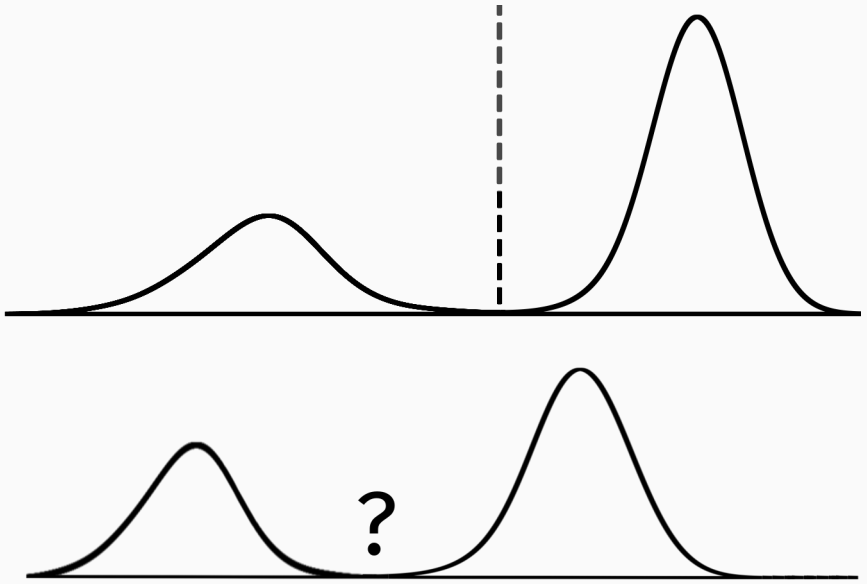
FlowLearn approach



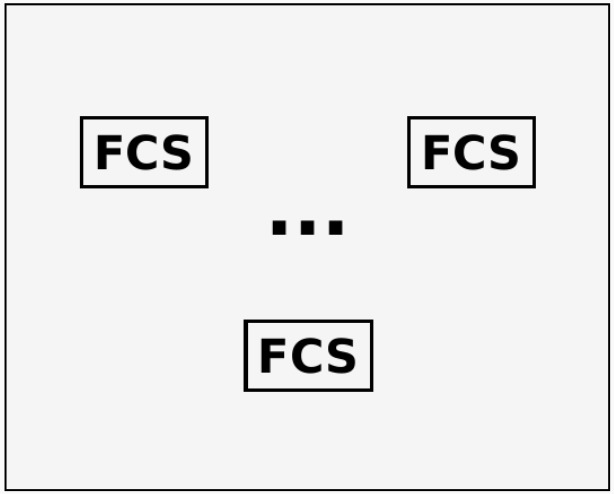
FlowLearn approach



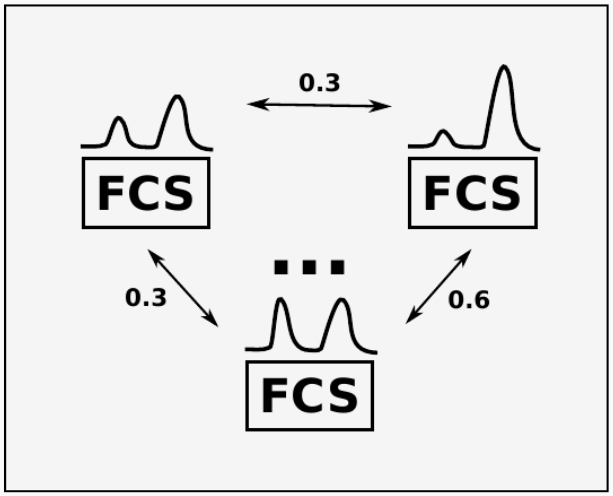
FlowLearn approach



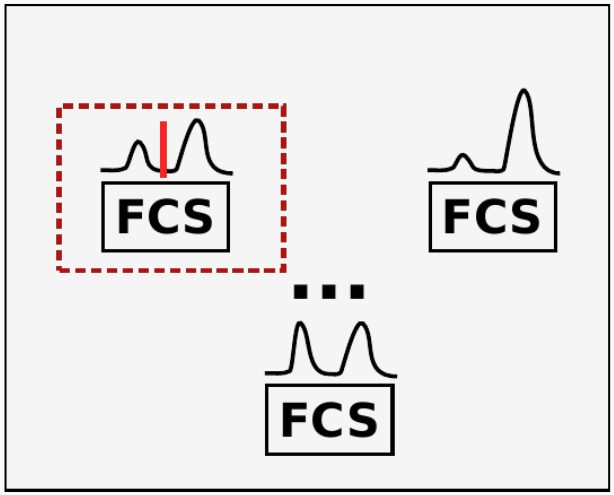
① Input FCM data



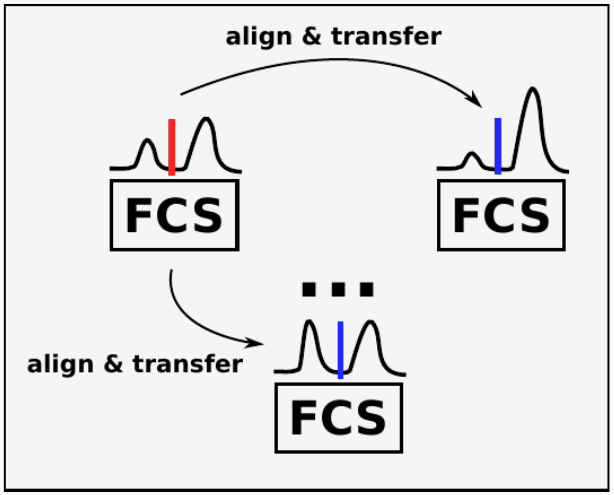
② Density estimation



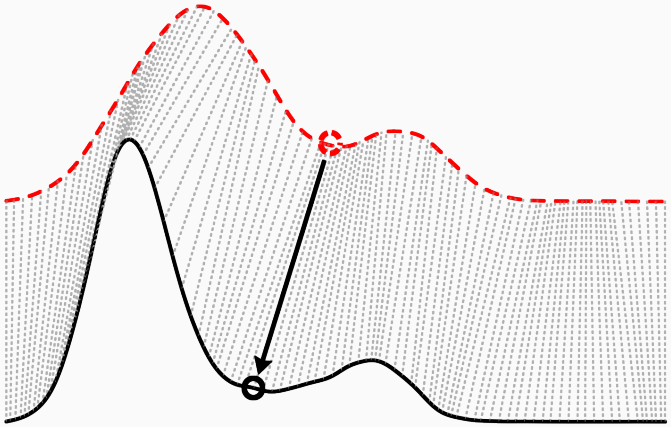
③ Example gates

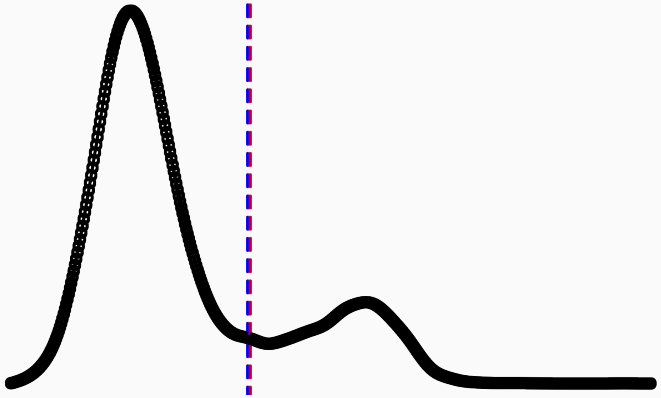


④ Alignment & transfer



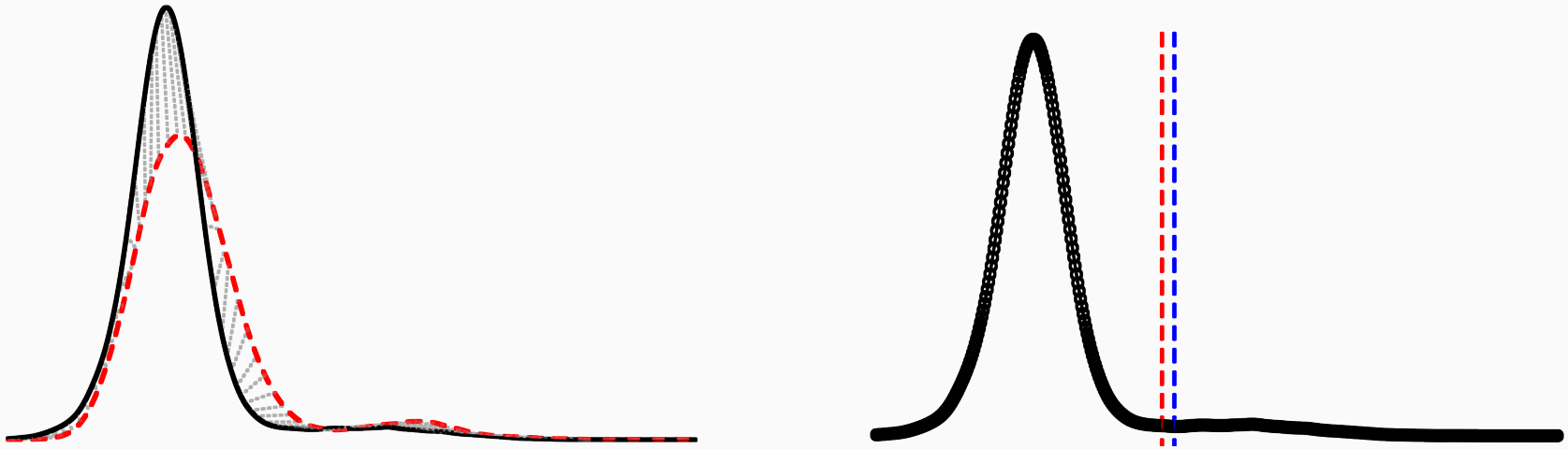
FlowLearn: Alignment and Prediction





1. Kernel density estimation & Smoothing splines
2. Derivative Dynamic Time Warping

Gating rare populations



Rare populations extend density tail.

Evaluation methodology

Benchmark data sets

- Mice (1 dataset, 2665 samples)
- FlowCAP (4 datasets, 63 samples, 7 centers)

Evaluation methodology

Benchmark data sets

- Mice (1 dataset, 2665 samples)
- FlowCAP (4 datasets, 63 samples, 7 centers)

Metrics

- per population
- F_1-score
- cell frequency error

Results

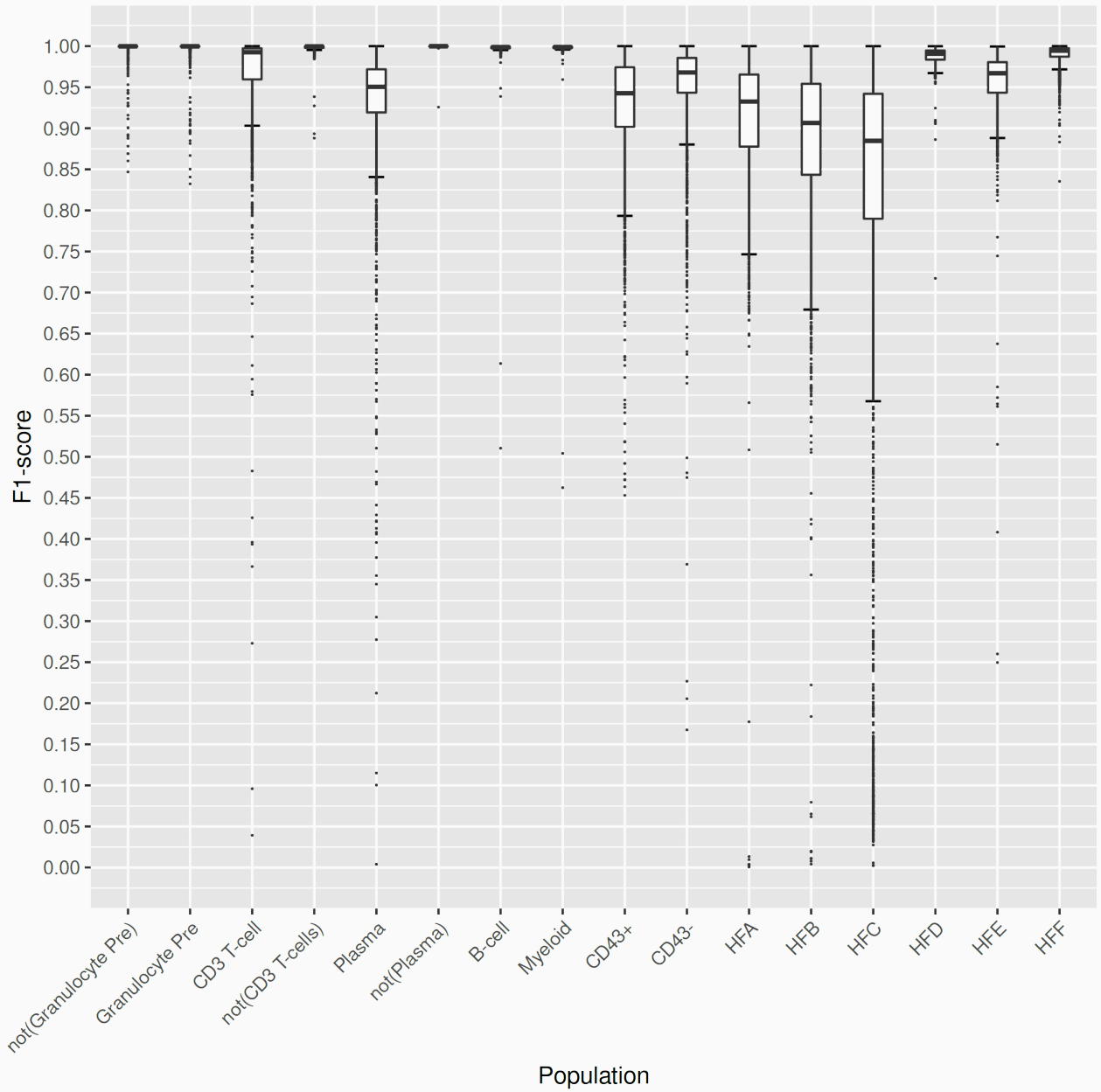
	# populations	$F_1 > 0.9$	$F_1 > 0.99$
Mice	16	15	9
FlowCap	48	36	11
Total	64	51	20

Results

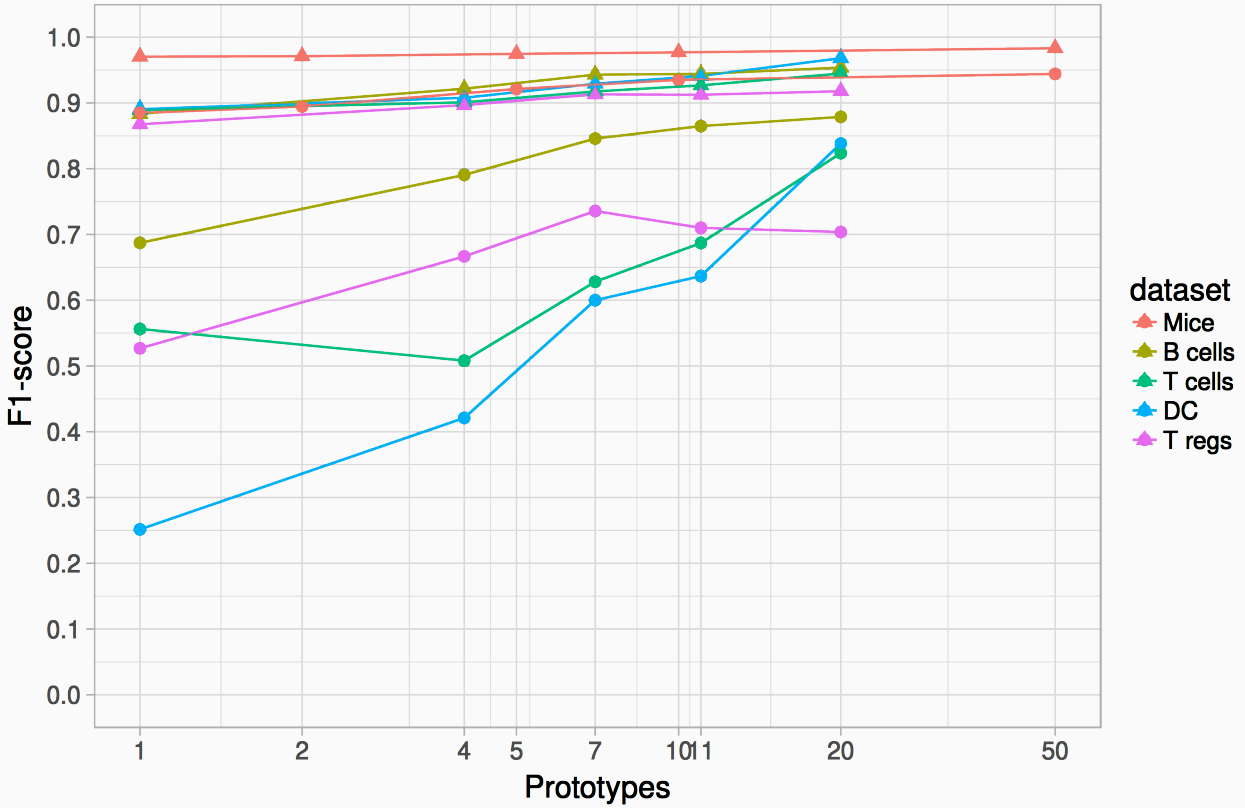
	# populations	$F_1 > 0.9$	$F_1 > 0.99$
Mice	16	15	9
FlowCap	48	36	11
Total	64	51	20

- Significantly better than current state-of-the-art methods
  - FlowSOM, Van Gassen et al. (2015)
  - DeepCyTOF, Li et al. (2017)

Mice, 1 prototype

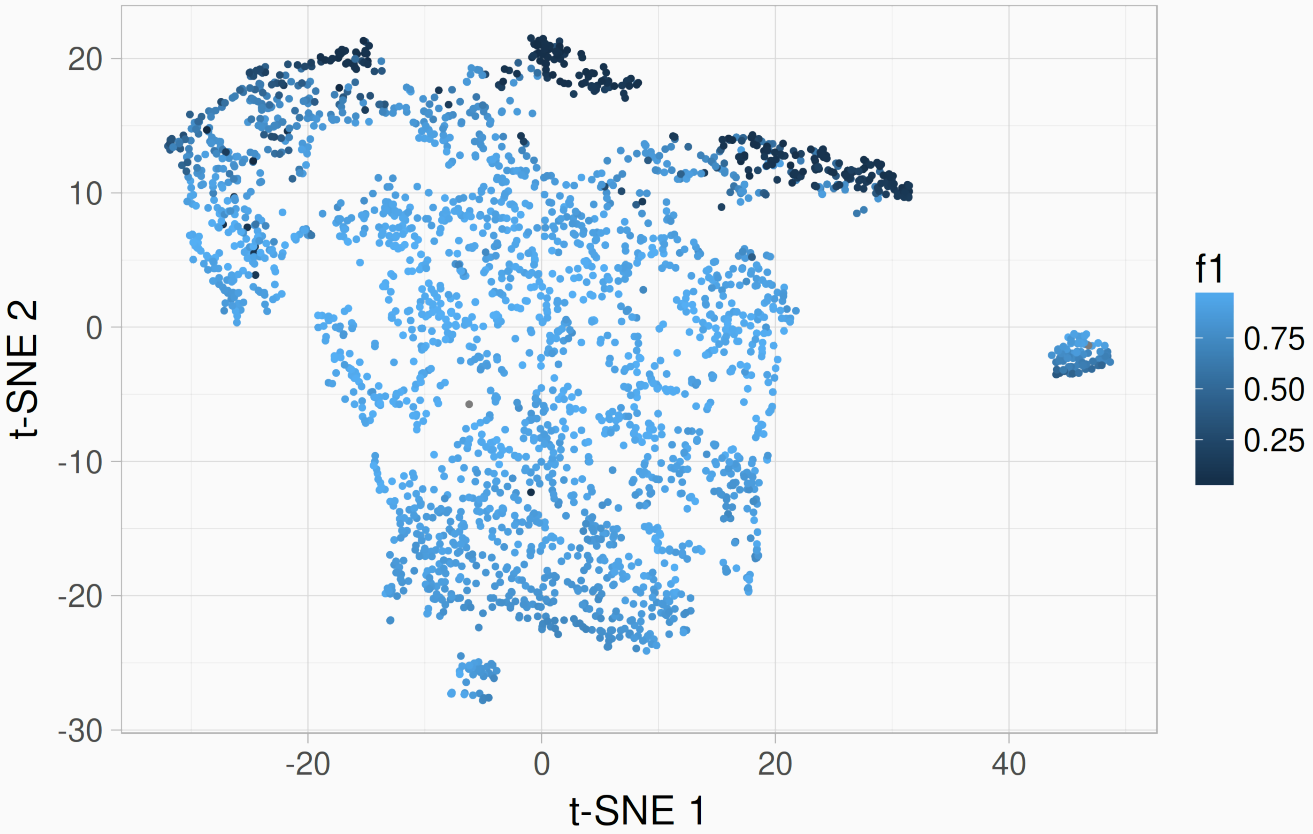


Influence of the number of prototypes



Quality checking

Use flowLearn for checking quality of existing gates.



Outlook

- Investigate other prototype choices
- Extend alignment to two dimensions
- Collaboration with clinical research

Markus Lux, Ryan Remy Brinkman, Cedric Chauve, Adam Laing, Anna Lorenc, Lucie Abeler-Dörner, Barbara Hammer

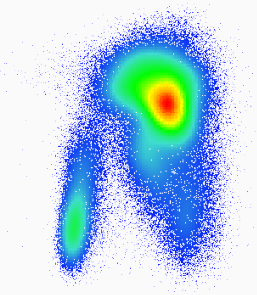
flowLearn: Fast and precise identification and quality checking of cell populations in flow cytometry

Bioinformatics, 2018

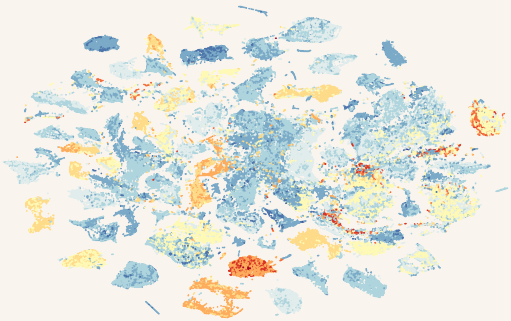
Overview

Efficient grouping methods for
the annotation and sorting of single cells

Flow cytometry

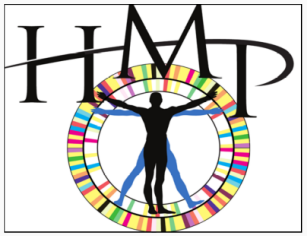


Metagenomics

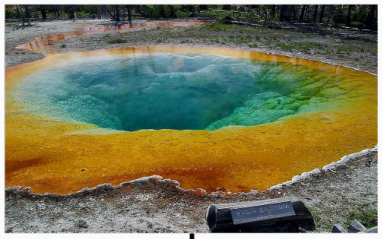


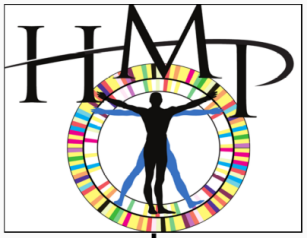
Single-cell genomics

Metagenomics in a nutshell



Metagenomics in a nutshell





Wet laboratory

Sampling

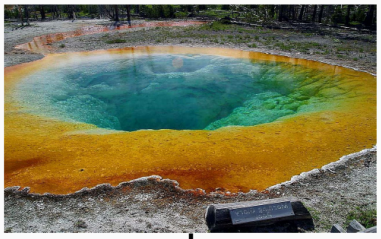
↓

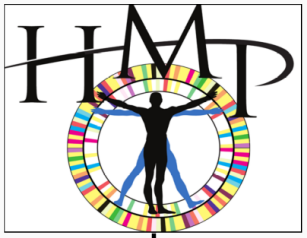
Sample fractionation

↓

DNA extraction

Metagenomics in a nutshell





Wet laboratory

Sampling

Sample fractionation

DNA extraction

DNA sequencing

Computational Methods

Assembly

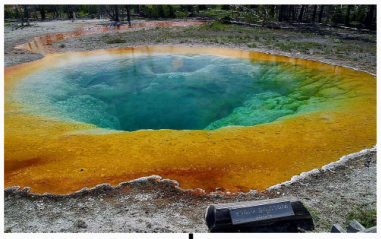
Annotation

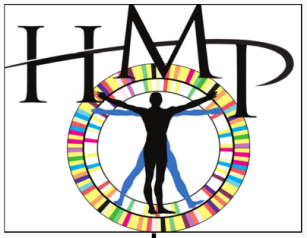
Statistical analysis

Data storage

Data sharing

Metagenomics in a nutshell





Wet laboratory

Sampling

Sample fractionation

DNA extraction

DNA sequencing

Computational Methods

Assembly

Binning

Annotation

Statistical analysis

Data storage

Data sharing

Binning:

What sequences belong to the same genome?

[1]Sedlar et al. (2017)

Problem statement

Binning:

What sequences belong to the same genome?

supervised

[1] Sedlar et al. (2017)

Problem statement

Binning:
What sequences belong to the same genome?

supervised

taxonomy-free[1]

abundance based

sequence composition based

[1] Sedlar et al. (2017)

Contribution

- Tools utilize sequence-composition based binning
  Laczny et al. (2014); Wu et al. (2015); Lin and Liao (2016), ...
- Integration of novel machine learning techniques
- Sensible w.r.t. involved techniques and parameters

Contribution

- Tools utilize sequence-composition based binning
  Laczny et al. (2014); Wu et al. (2015); Lin and Liao (2016), ...
- Integration of novel machine learning techniques
- Sensible w.r.t. involved techniques and parameters

→ **Thorough evaluation of pipeline ingredients,** Lux et al. (2015b)

→ Foundations for single-cell analysis

Binning pipeline evaluation

Ingredients:

- Vectorial representation
- Dimensionality reduction
- Clustering

Binning pipeline evaluation

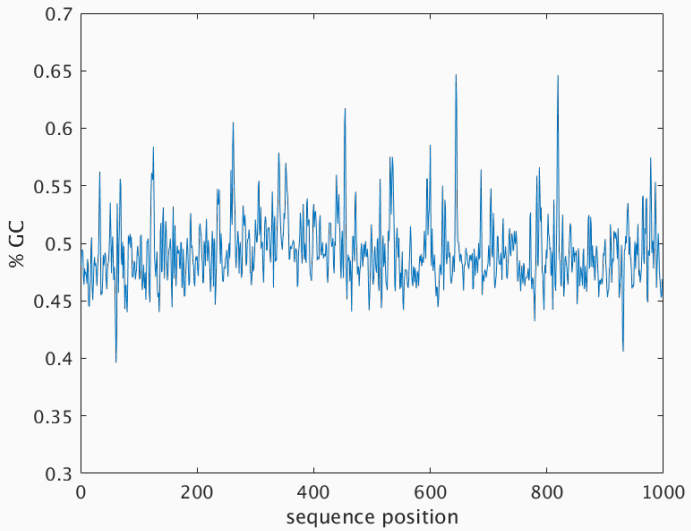
Ingredients:

- Vectorial representation
- Dimensionality reduction
- Clustering

Idea: 1 cluster $\leftrightarrow$ 1 genome

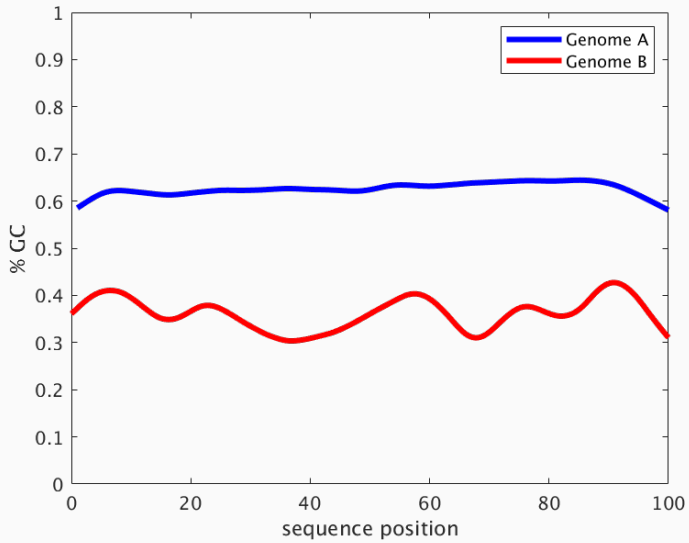
$\Rightarrow$ Resolve pronounced clusters

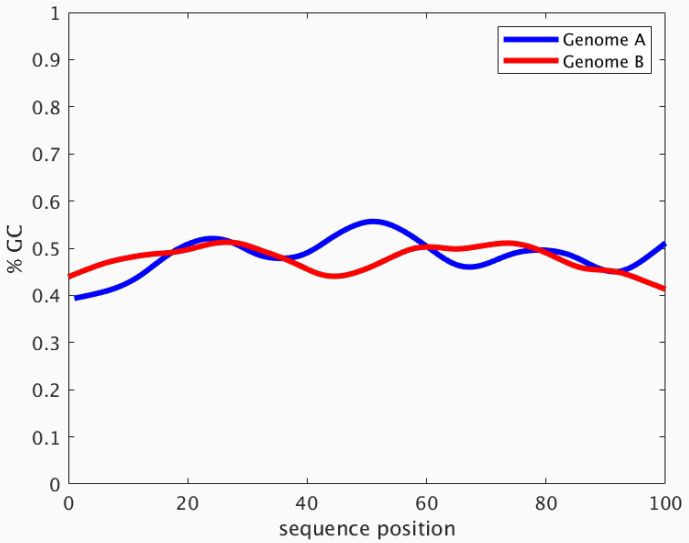
GC-content as a sequence signature



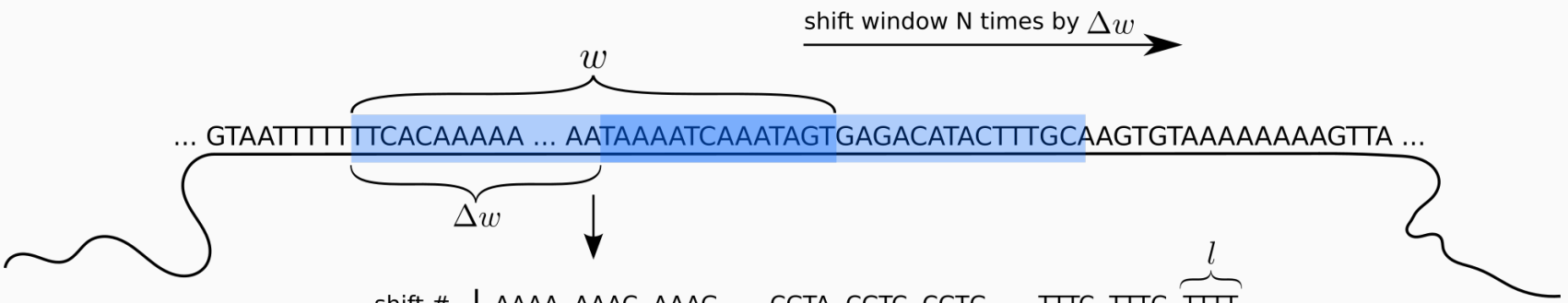
E. coli GC-content variation

GC-content as a sequence signature





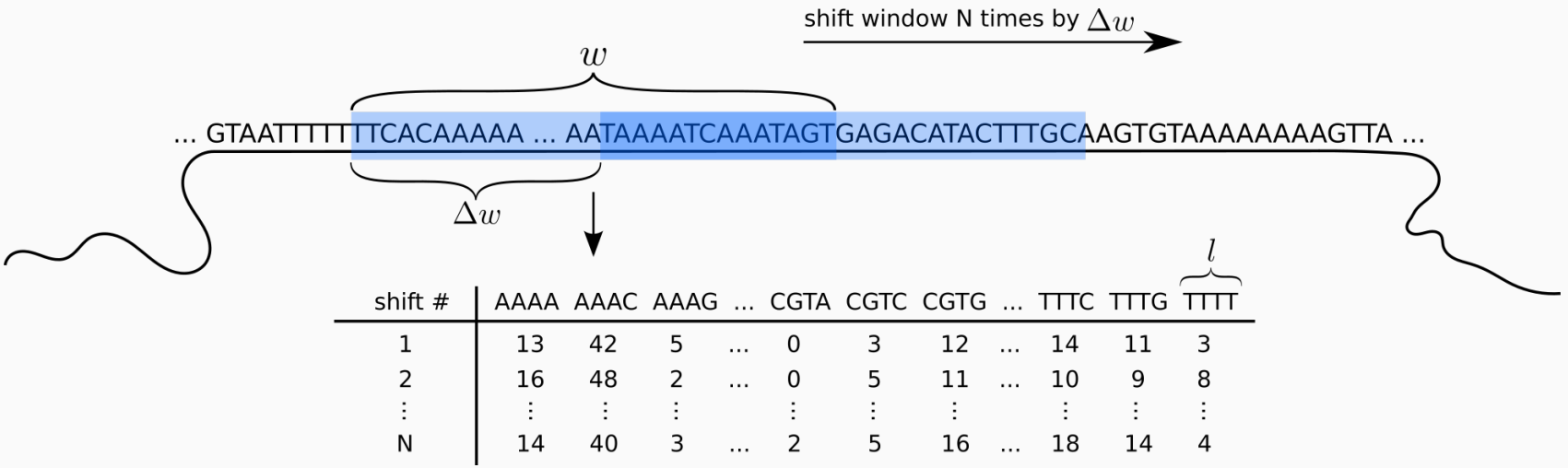
Calculation of l-mer frequencies



l

shift #	AAAA	AAAC	AAAG	...	CGTA	CGTC	CGTG	...	TTTC	TTTG	TTTT
1	13	42	5	...	0	3	12	...	14	11	3
2	16	48	2	...	0	5	11	...	10	9	8
⋮	⋮	⋮	⋮		⋮	⋮	⋮		⋮	⋮	⋮
N	14	40	3	...	2	5	16	...	18	14	4

Calculation of l-mer frequencies



shift #	AAAA	AAAC	AAAG	...	CGTA	CGTC	CGTG	...	TTTC	TTTG	TTTT
1	13	42	5	...	0	3	12	...	14	11	3
2	16	48	2	...	0	5	11	...	10	9	8
⋮	⋮	⋮	⋮		⋮	⋮	⋮		⋮	⋮	⋮
N	14	40	3	...	2	5	16	...	18	14	4

Observations:

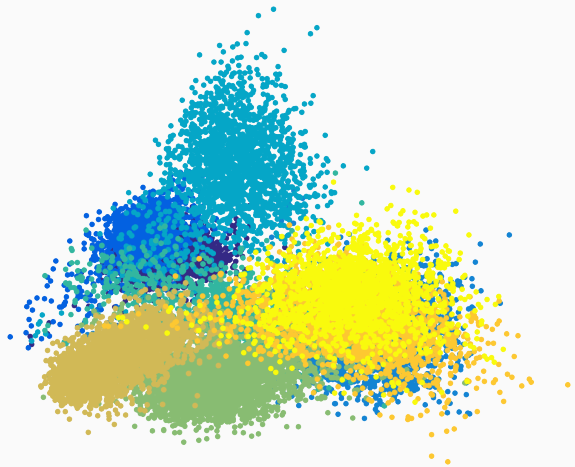
- N data points with dimensionality 4^l
- Taking $l = 1$ is equivalent to the GC-content
- Window parameters heavily influence representation

Evaluation questions

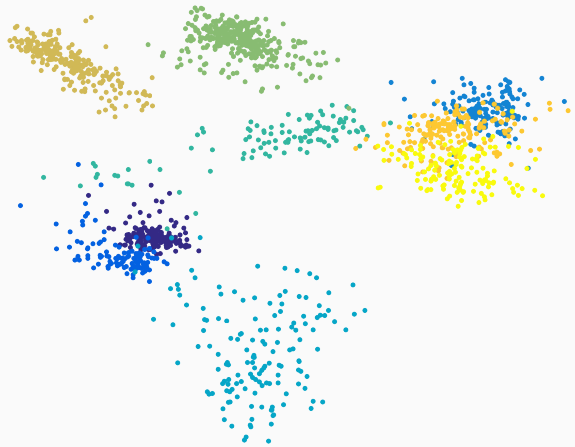
- What window parameters to choose?
- How to reduce dimensionality?
- How to cluster?

DR & window parameters, qualitative

small w



large w

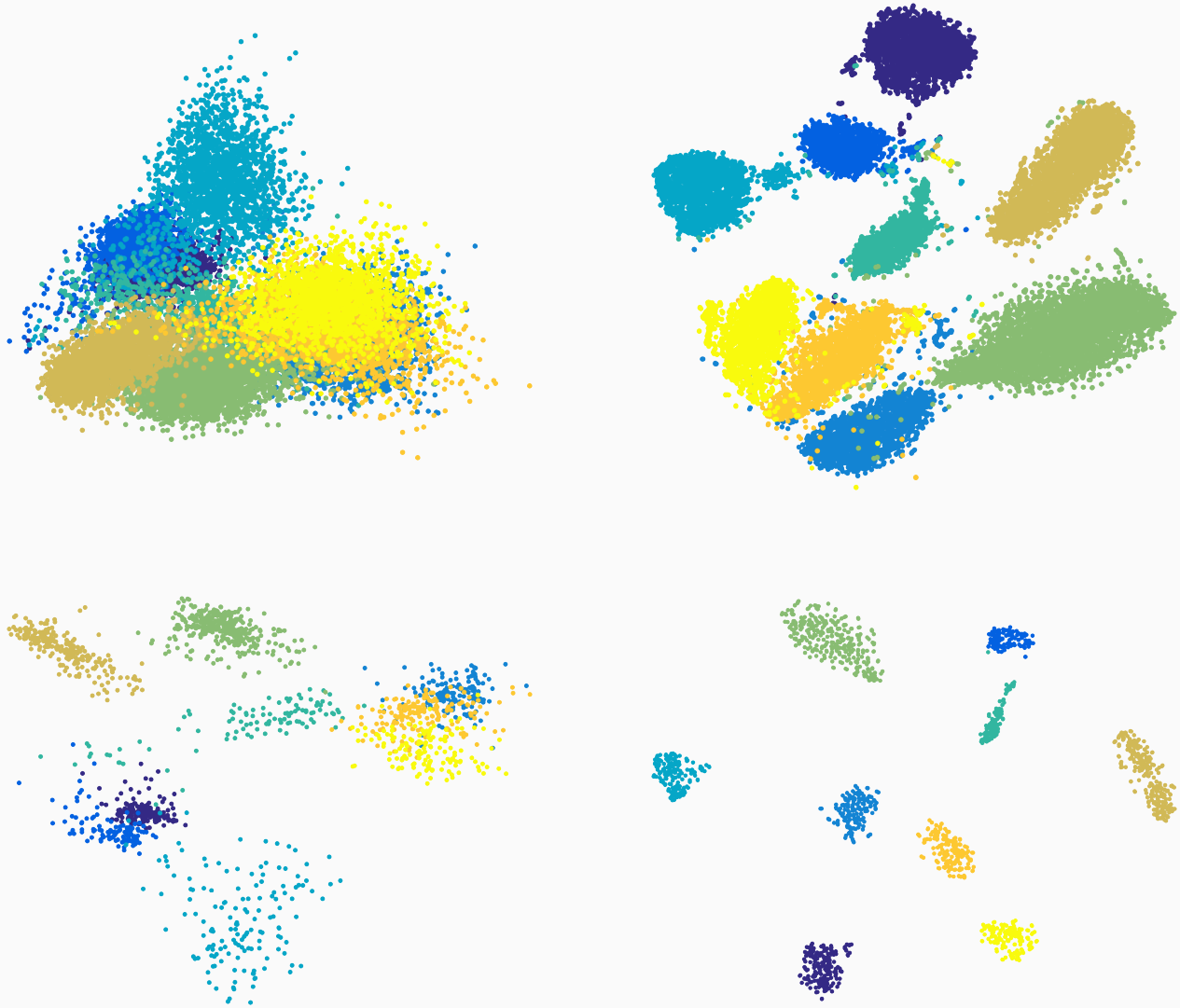


PCA

DR & window parameters, qualitative

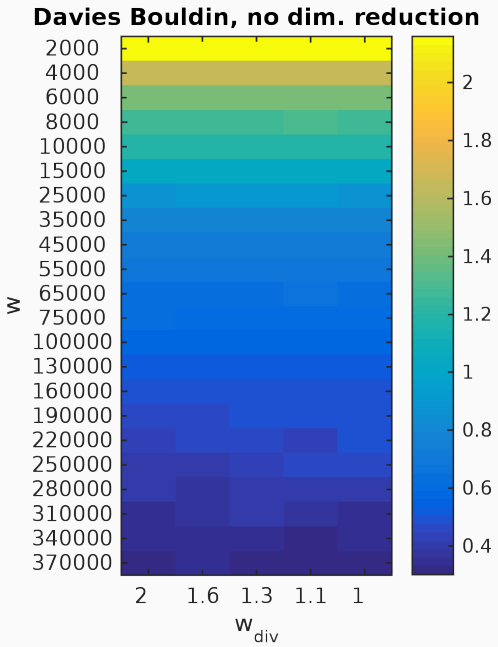
small w

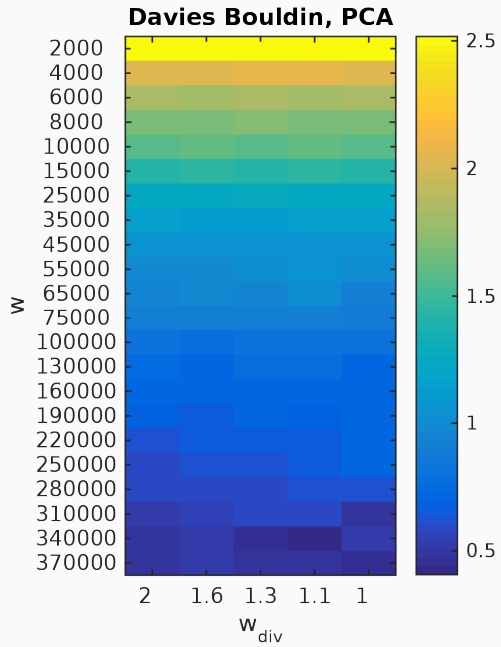
large w

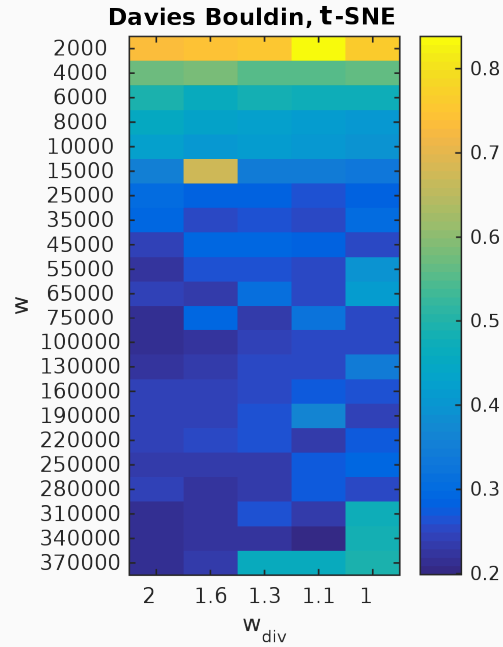


PCA t-SNE

DR & window parameters, quantitative







→ Significant advantage of non-linear DR

Clustering evaluation

Investigate suitable algorithms

How to determine number of clusters?

Clustering evaluation

THEO	DB K	DB J	DUNN K	DUNN J	XB K	XB J	GAP K	GAP J	SS K
	$k_{opt} = 5$								
KM	**5**	**1.00**	**5**	**1.00**	3	0.56	55	0.10	**5**
KM++	**5**	**1.00**	**5**	**1.00**	**5**	**1.00**	64	0.09	6
NG	**5**	**1.00**	**5**	**1.00**	**5**	**1.00**	**5**	**1.00**	**5**
HC	**5**	**1.00**	**5**	**1.00**	**5**	**1.00**	65	0.12	6
GMM	**5**	**1.00**	**5**	**1.00**	-	-	7	0.80	**5**
GMM++	**5**	**1.00**	**5**	**1.00**	-	-	12	0.47	7
SC	**5**	**1.00**	**5**	**1.00**	**5**	**1.00**	-	-	20
DBS	**K = 5, J = 1.00**								
SC-EIG	**K = 5, J = 1.00**								
	$k_{opt} = 10$								
KM	12	0.90	9	0.92	-	-	29	0.36	2
KM++	**10**	**1.00**	**10**	**1.00**	**10**	**1.00**	32	0.34	**10**
NG	9	0.93	**10**	**1.00**	11	**1.00**	9	0.93	2
HC	**10**	**1.00**	**10**	**1.00**	**10**	**1.00**	26	0.46	14
GMM	12	0.95	**10**	**1.00**	-	-	13	0.78	1
GMM++	**10**	**1.00**	**10**	**1.00**	-	-	15	0.71	11
SC	**10**	**1.00**	**10**	**1.00**	**10**	**1.00**	-	-	13
DBS	**K = 10, J = 1.00**								
SC-EIG	**K = 10, J = 1.00**								
	$k_{opt} = 20$								
KM	23	0.79	8	0.36	-	-	51	0.42	2
KM++	**20**	**1.00**	**20**	**1.00**	**20**	**1.00**	56	0.44	25
NG	**20**	0.89	15	0.74	18	0.86	18	0.80	5
HC	**20**	**1.00**	**20**	**1.00**	**20**	**1.00**	56	0.54	39
GMM	28	0.81	5	0.22	-	-	13	0.55	1
GMM++	**20**	**1.00**	**20**	**1.00**	**20**	**1.00**	40	0.58	23
SC	33	0.55	**20**	**1.00**	7	0.11	-	-	2
DBS	**K = 20, J = 1.00**								
SC-EIG	**K = 20**, J = 0.88								
	$k_{opt} = 30$								
KM	27	0.80	19	0.63	-	-	67	0.54	1
KM++	**30**	**1.00**	**30**	**1.00**	**30**	**1.00**	68	0.50	39
NG	26	0.86	34	0.94	32	0.94	24	0.83	3
HC	**30**	**1.00**	**30**	**1.00**	**30**	**1.00**	49	0.79	62
GMM	25	0.73	23	0.59	-	-	9	0.24	1
GMM++	**30**	**1.00**	**30**	**1.00**	**30**	**1.00**	44	0.80	35
SC	33	0.93	31	0.97	31	0.97	-	-	31
DBS	**K = 30, J = 1.00**								
SC-EIG	K = 31, J = 0.97								
	$k_{opt} = 40$								
KM	39	0.66	21	0.49	-	-	68	0.65	1
KM++	39	0.96	35	0.77	35	0.77	**40**	**1.00**	43
NG	34	0.82	29	0.64	45	0.96	26	0.58	3
HC	39	0.96	35	0.77	35	0.77	**40**	**1.00**	61
GMM	44	0.62	9	0.21	-	-	6	0.15	3
GMM++	39	0.96	35	0.77	35	0.77	**40**	**1.00**	44
SC	35	0.56	3	0.03	2	0.28	-	-	41
DBS	K = 35, J = 0.77								
SC-EIG	K = **40**, J = 0.88								
	$k_{opt} = 50$								
KM	29	0.52	16	0.27	-	-	63	0.63	3
KM++	**50**	**1.00**	**50**	**1.00**	**50**	**1.00**	**50**	**1.00**	54
NG	35	0.63	51	0.83	43	0.74	24	0.45	11
HC	**50**	**1.00**	**50**	**1.00**	**50**	**1.00**	**50**	**1.00**	70
GMM	49	0.59	9	0.15	-	-	3	0.06	2
GMM++	**50**	**1.00**	**50**	**1.00**	**50**	**1.00**	**50**	**1.00**	55
SC	58	0.67	2	0.02	4	0.02	-	-	13
DBS	K = 36, J = 0.48								
SC-EIG	K = 51, J = 0.65								

NCBI	DB K	DB J	DUNN K	DUNN J	XB K	XB J	GAP K	GAP J	SS K
	$k_{opt} = 5$								
KM	4	0.87	4	0.87	4	0.87	12	0.38	**5**
KM++	**5**	**1.00**	4	0.87	4	0.87	11	0.41	7
NG	**5**	**1.00**	4	0.87	4	0.87	10	0.45	8
HC	**5**	**1.00**	4	0.87	4	0.87	6	0.80	6
GMM	**5**	0.98	4	0.87	-	-	7	0.69	**5**
GMM++	**5**	0.98	4	0.87	4	0.87	-	-	8
SC	4	0.73	4	0.73	4	0.73	-	-	10
DBS	K = 4, J = 0.87								
SC-EIG	K = **5**, J = 0.67								
	$k_{opt} = 10$								
KM	**10**	**0.98**	6	0.63	9	0.94	68	0.13	4
KM++	**10**	**0.98**	**10**	**0.98**	**10**	**0.98**	53	0.16	11
NG	**10**	**0.98**	6	0.63	8	0.89	17	0.50	12
HC	**10**	**0.98**	**10**	**0.98**	**10**	**0.98**	18	0.50	11
GMM	9	0.79	8	0.71	-	-	13	0.89	9
GMM++	**10**	**0.98**	**10**	**0.98**	**10**	**0.98**	50	0.20	11
SC	2	0.13	2	0.13	2	0.13	-	-	12
DBS	**K = 10, J = 0.98**								
SC-EIG	K = 12 J = 0.79								
	$k_{opt} = 20$								
KM	22	0.85	5	0.32	-	-	61	0.26	3
KM++	22	0.73	16	0.91	16	0.90	63	0.24	**20**
NG	**20**	0.74	4	0.27	13	0.76	30	0.53	18
HC	19	**0.97**	17	0.90	18	0.96	39	0.44	19
GMM	18	0.91	16	0.85	-	-	16	0.88	1
GMM++	16	0.87	17	0.94	17	0.92	34	0.48	**20**
SC	14	0.61	12	0.46	11	0.44	-	-	21
DBS	K = 18, J = 0.96								
SC-EIG	K = 24, J = 0.58								
	$k_{opt} = 30$								
KM	23	0.78	10	0.40	-	-	60	0.45	4
KM++	**30**	**0.93**	12	0.49	17	0.63	51	0.46	**30**
NG	27	0.77	18	0.71	26	0.92	28	0.80	16
HC	27	0.91	13	0.53	26	0.89	65	0.47	36
GMM	23	0.75	52	0.67	-	-	17	0.65	1
GMM++	31	0.84	19	0.70	21	0.76	55	0.49	**30**
SC	2	0.44	9	0.14	2	0.04	-	-	14
DBS	K = 28, J = 0.91								
SC-EIG	K = 39, J = 0.22								
	$k_{opt} = 40$								
KM	**39**	0.75	17	0.53	-	-	70	0.53	4
KM++	46	0.75	25	0.70	24	0.66	68	0.49	4
NG	37	0.81	17	0.53	16	0.51	31	0.78	14
HC	**41**	0.92	11	0.35	36	**0.95**	53	0.69	48
GMM	29	0.78	17	0.49	4	0.12	17	0.45	1
GMM++	43	0.81	29	0.80	4	0.12	-	-	1
SC	31	0.6	4	0.04	7	0.06	-	-	-
DBS	K = 36, J = 0.91								
SC-EIG	K = 49, J = 0.25								
	$k_{opt} = 50$								
KM	40	0.76	34	0.71	-	-	66	0.70	1
KM++	**50**	0.81	68	0.64	31	0.68	69	0.63	4
NG	42	0.80	37	0.80	34	0.76	33	0.73	1
HC	47	**0.89**	15	0.33	44	0.88	46	**0.89**	66
GMM	65	0.82	65	0.79	-	-	6	0.12	1
GMM++	49	0.83	2	0.05	28	0.62	-	-	1
SC	2	0.03	2	0.03	2	0.03	-	-	10
DBS	K = 45, J = 0.82								
SC-EIG	K = 68, J = 0.18								

Clustering evaluation – main findings

- Majority of clustering methods work well on theoretical data
- Few methods deliver accurate results on simulated data
- Promising candidates:
  - Hierarchical Clustering
  - Connected Components Clustering
  - Davies-Bouldin index
- More complex metagenomes are more difficult

Clustering evaluation – main findings

- Majority of clustering methods work well on theoretical data
- Few methods deliver accurate results on simulated data
- Promising candidates:
  - Hierarchical Clustering
  - Connected Components Clustering
  - Davies-Bouldin index
- More complex metagenomes are more difficult

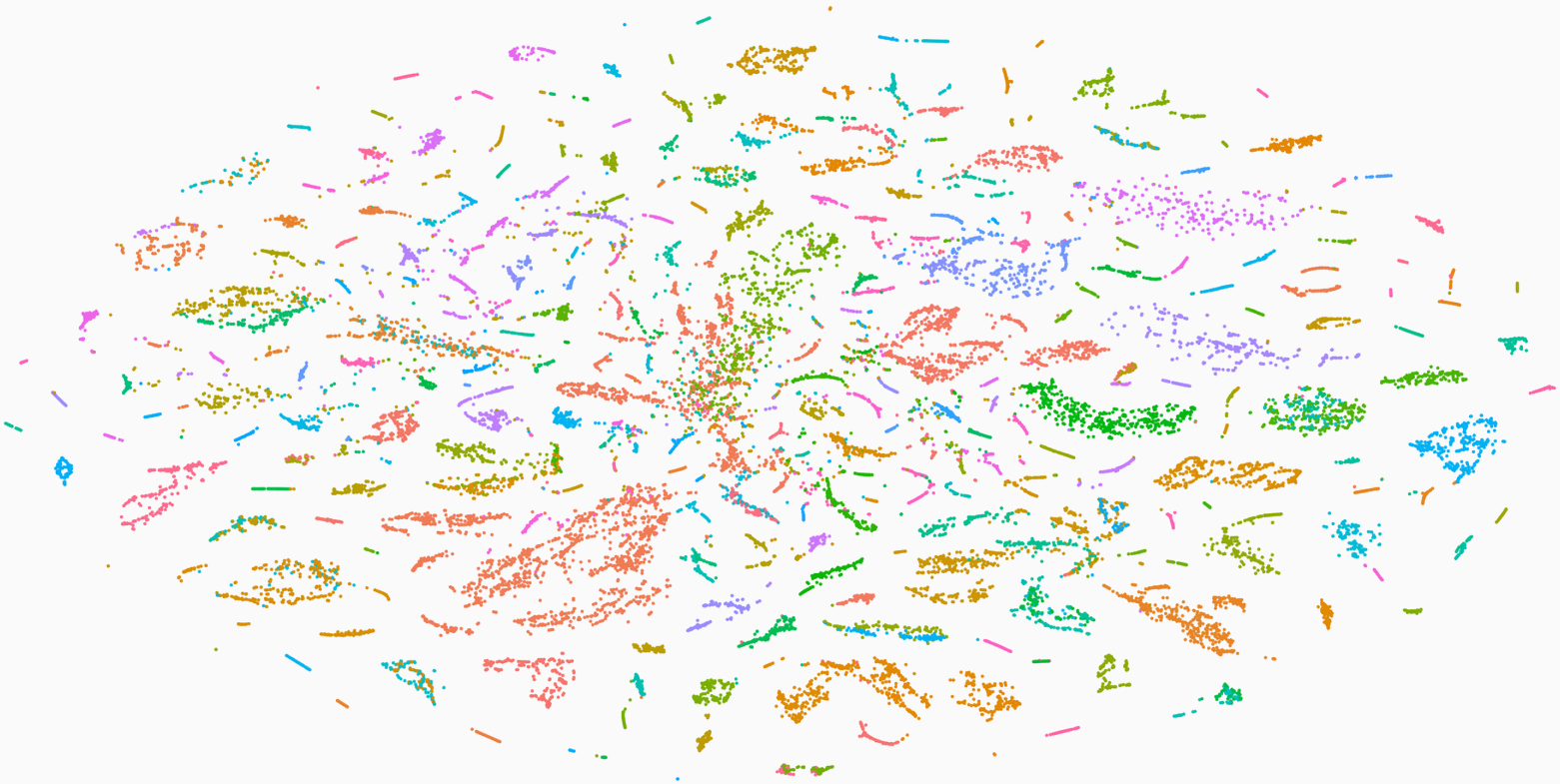
Clustering evaluation – main findings

- Majority of clustering methods work well on theoretical data
- Few methods deliver accurate results on simulated data
- Promising candidates:
  - Hierarchical Clustering
  - Connected Components Clustering
  - Davies-Bouldin index
- More complex metagenomes are more difficult

Clustering evaluation – main findings

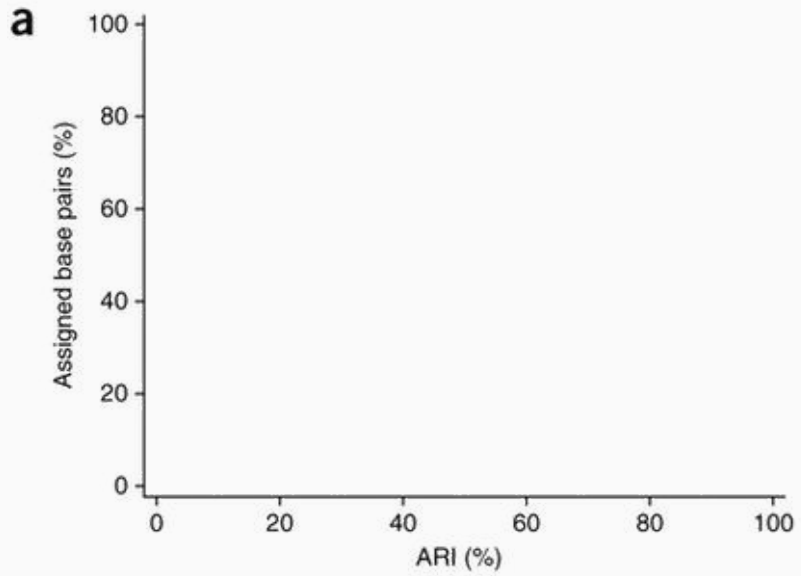
- Majority of clustering methods work well on theoretical data
- Few methods deliver accurate results on simulated data
- Promising candidates:
  - Hierarchical Clustering
  - Connected Components Clustering
  - Davies-Bouldin index
- More complex metagenomes are more difficult

Application to complex metagenomes



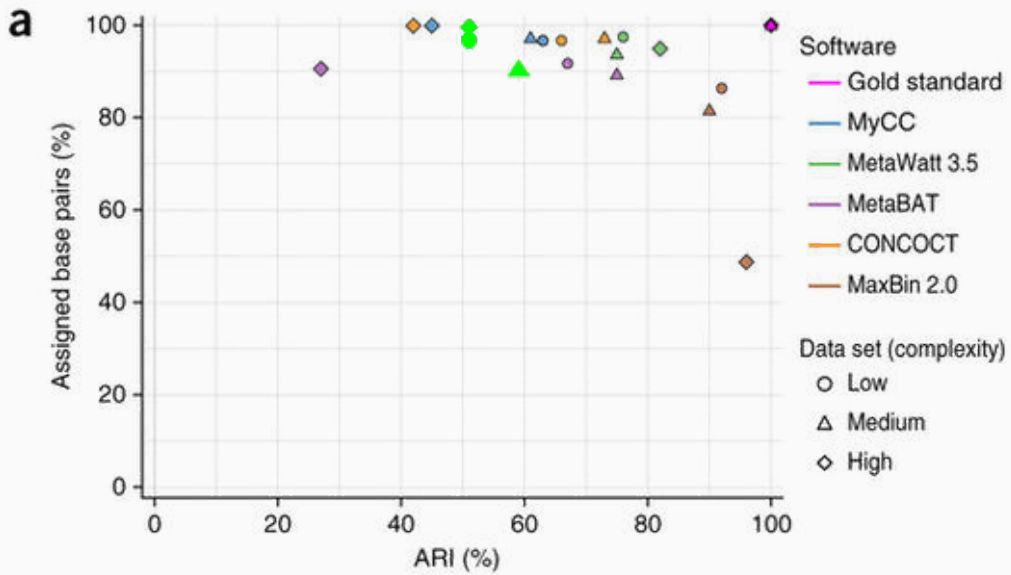
High complexity CAMI metagenome t-SNE representation, colored by gold standard

Application to complex metagenomes



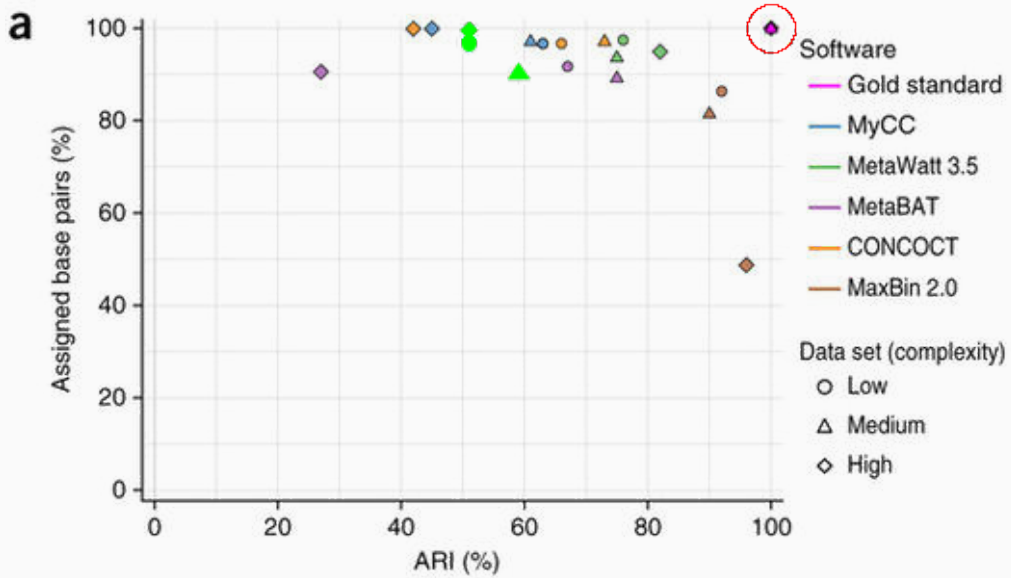
Comparison to other CAMI contestants

Application to complex metagenomes



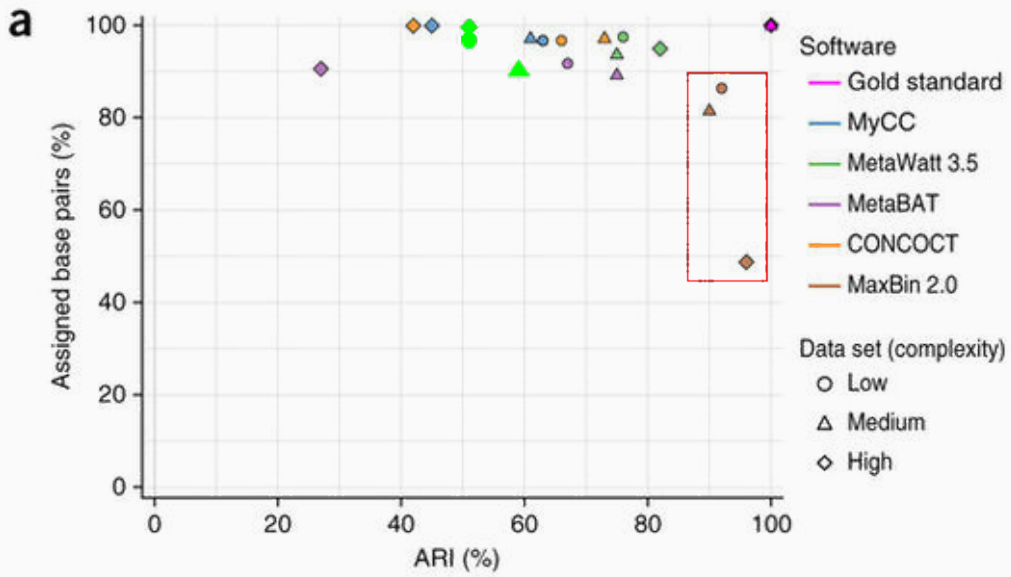
Comparison to other CAMI contestants

Application to complex metagenomes



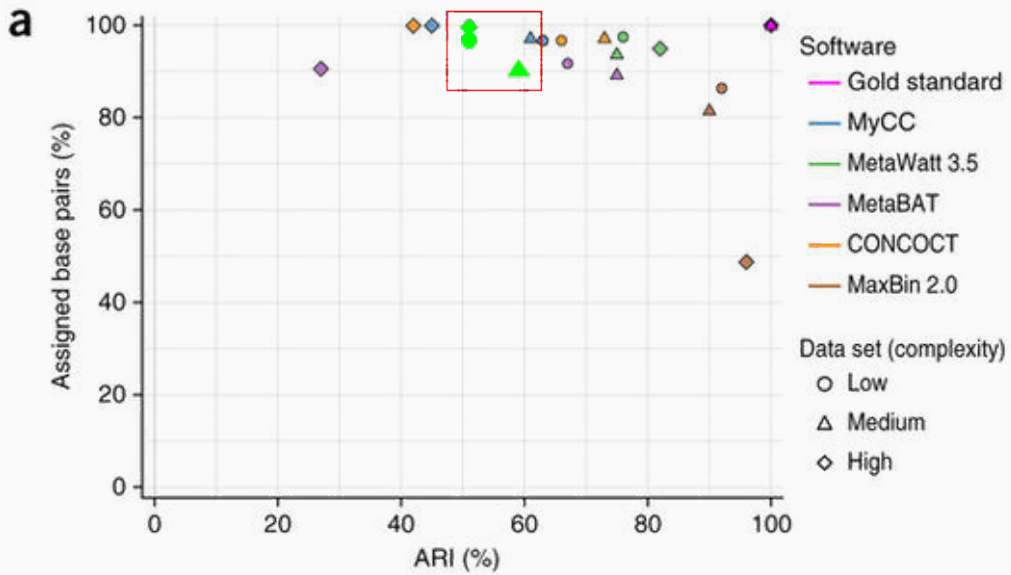
Comparison to other CAMI contestants

Application to complex metagenomes



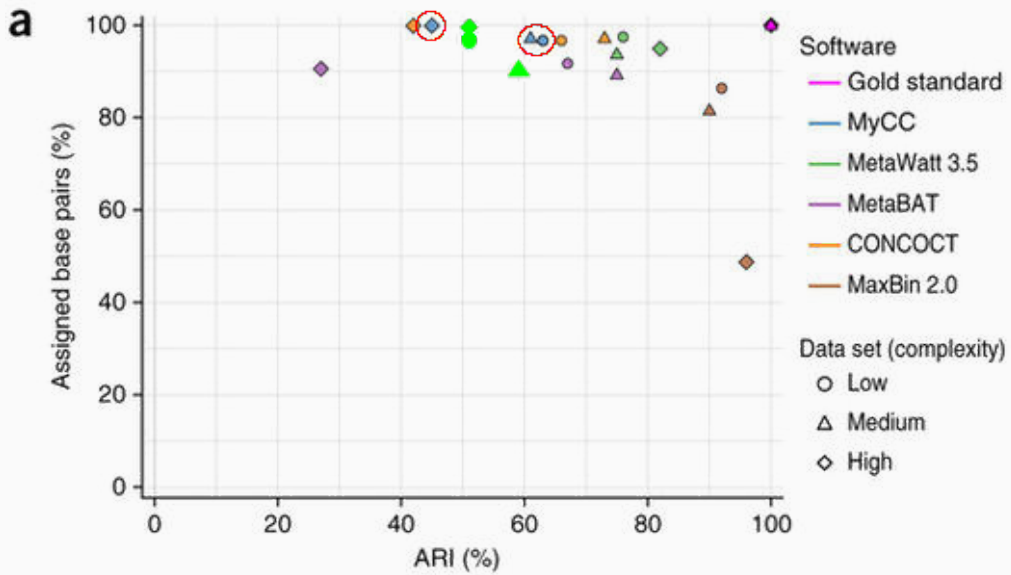
Comparison to other CAMI contestants

Application to complex metagenomes



Comparison to other CAMI contestants

Application to complex metagenomes



Comparison to other CAMI contestants

Guidelines

- Choose large, half-overlapping windows for a fixed number of data points.
- Reduce dimensionality using cluster-focused t-SNE
- Hierarchical clustering + Davies-Bouldin index

Outlook

- Crucial: utilize auxiliary information
- Incorporate findings to extant tools
- Iterative clustering process
- Investigate recent DR alternatives (UMAP[2])

[2] McInnes and Healy (2018)

Markus Lux, Alexander Sczyrba, Barbara Hammer

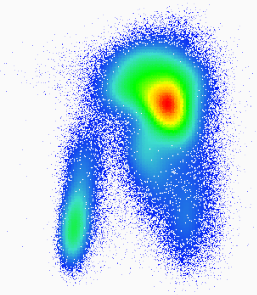
Automatic discovery of metagenomic structure

2015 International Joint Conference on Neural Networks (IJCNN)

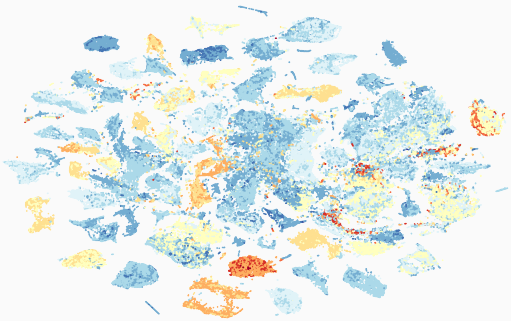
Overview

Efficient grouping methods for
the annotation and sorting of single cells

Flow cytometry

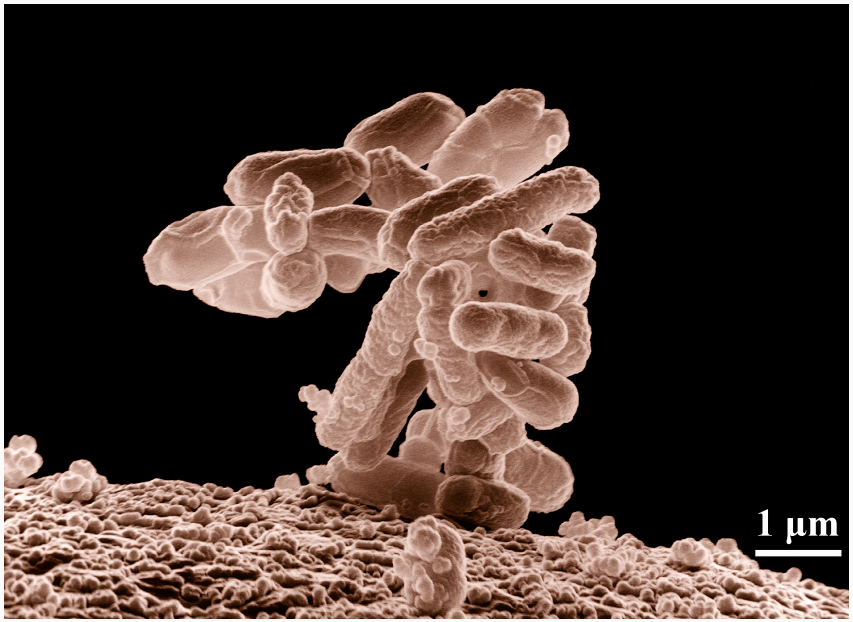


Metagenomics



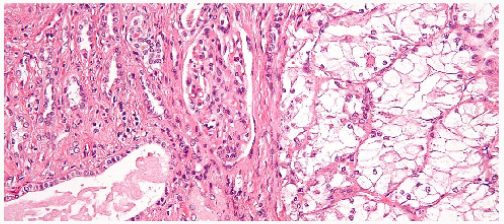
Single-cell genomics

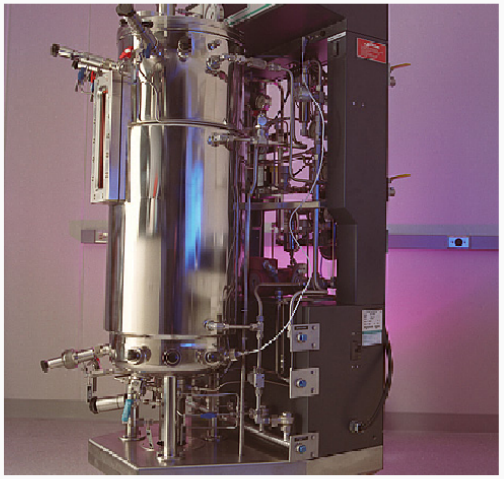
Single-cell sequencing

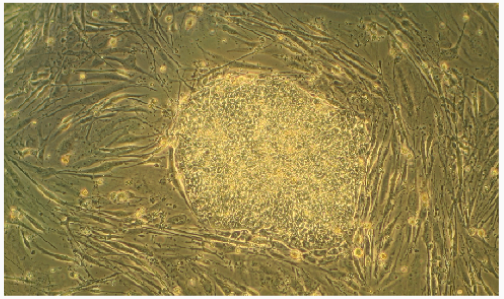


- Eliminates need for cultivation
- Amplification of one DNA molecule
- Focus on specifics of individual single cells

Single-cell genomics applications



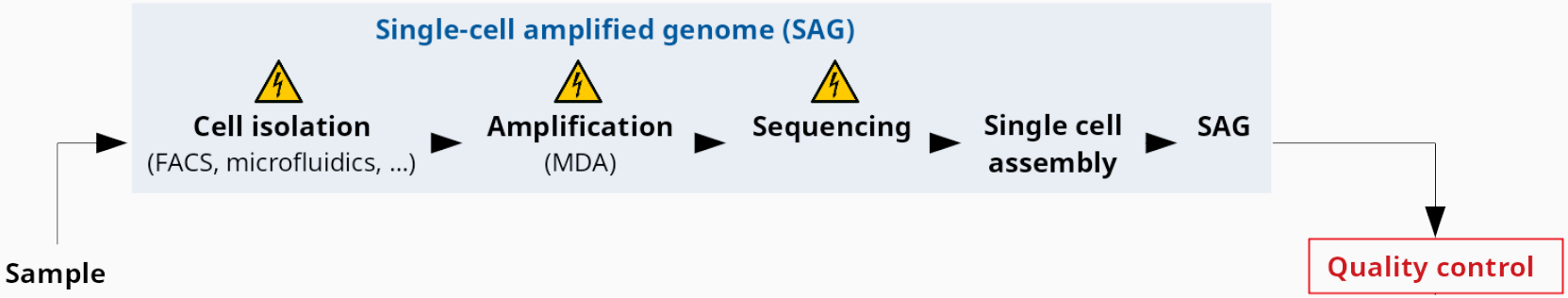




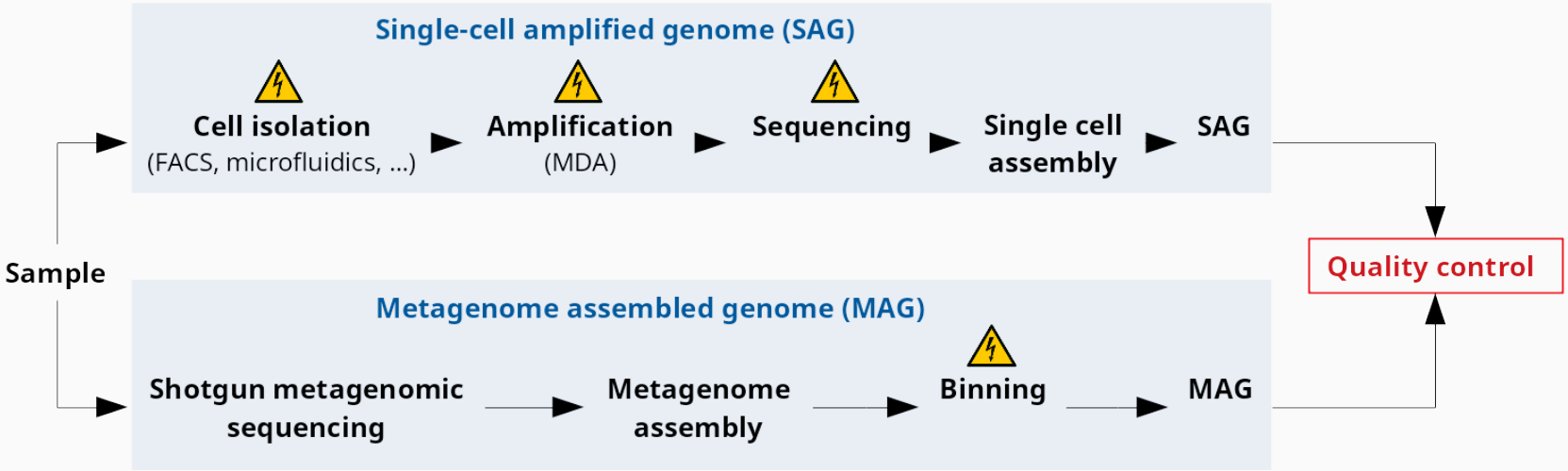
What if:

All these applications are based on wrong data?

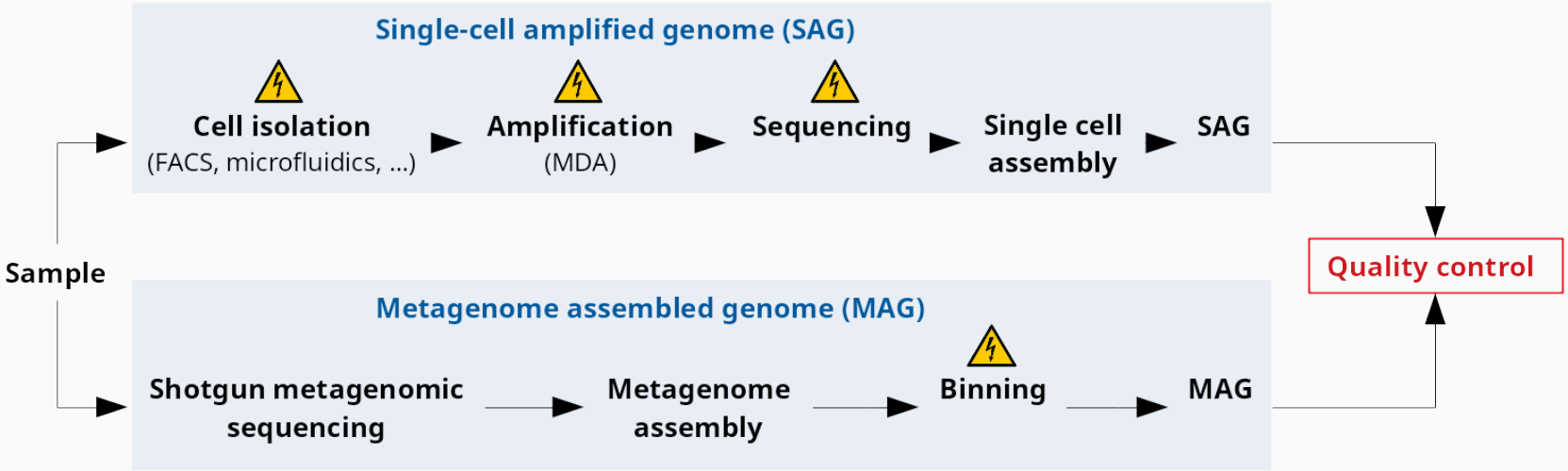
Problem: Contaminated samples



Problem: Contaminated samples



Problem: Contaminated samples



- Sources of contamination:
  - In-sample
  - Unclean laboratory environment
  - Library preparation kits, reagents, ...

- Prevent introduction to public databases!

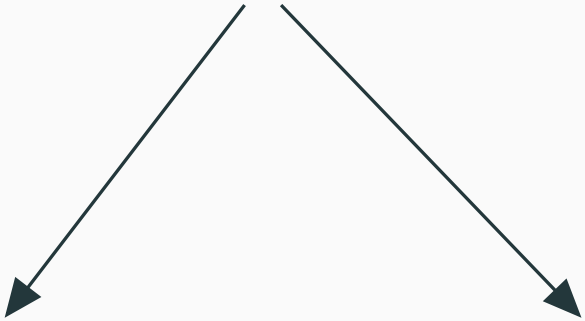
Quality standards for reporting SAGs and MAGs

Contamination rates[3]:

$< 10\%$	for low-quality draft genomes
$< 5\%$	for high-quality draft genomes

[3] Bowers et al. (2017)

Detecting contamination



supervised

de novo

ProDeGe, Tennessen et al. (2015)

Contribution

Interactive detection and
cleansing without re-sequencing

Contribution

Interactive detection and
cleansing without re-sequencing

acdc – **a**utomated **c**ontamination **d**etection and **c**onfidence estimation for single-cell genome data

Lux et al. (2015a, 2016)

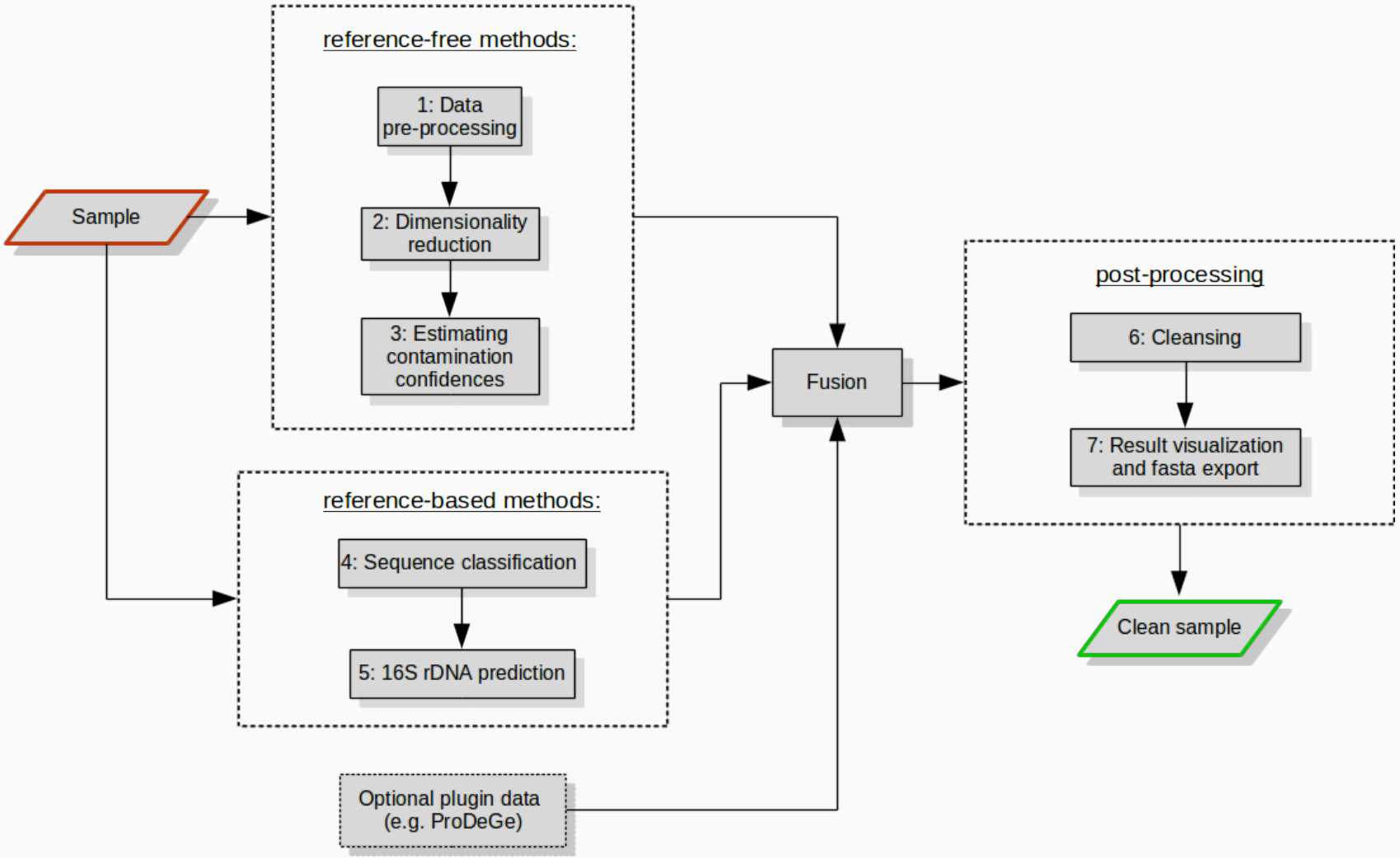
Approach

Unsupervised machine learning and clustering techniques.

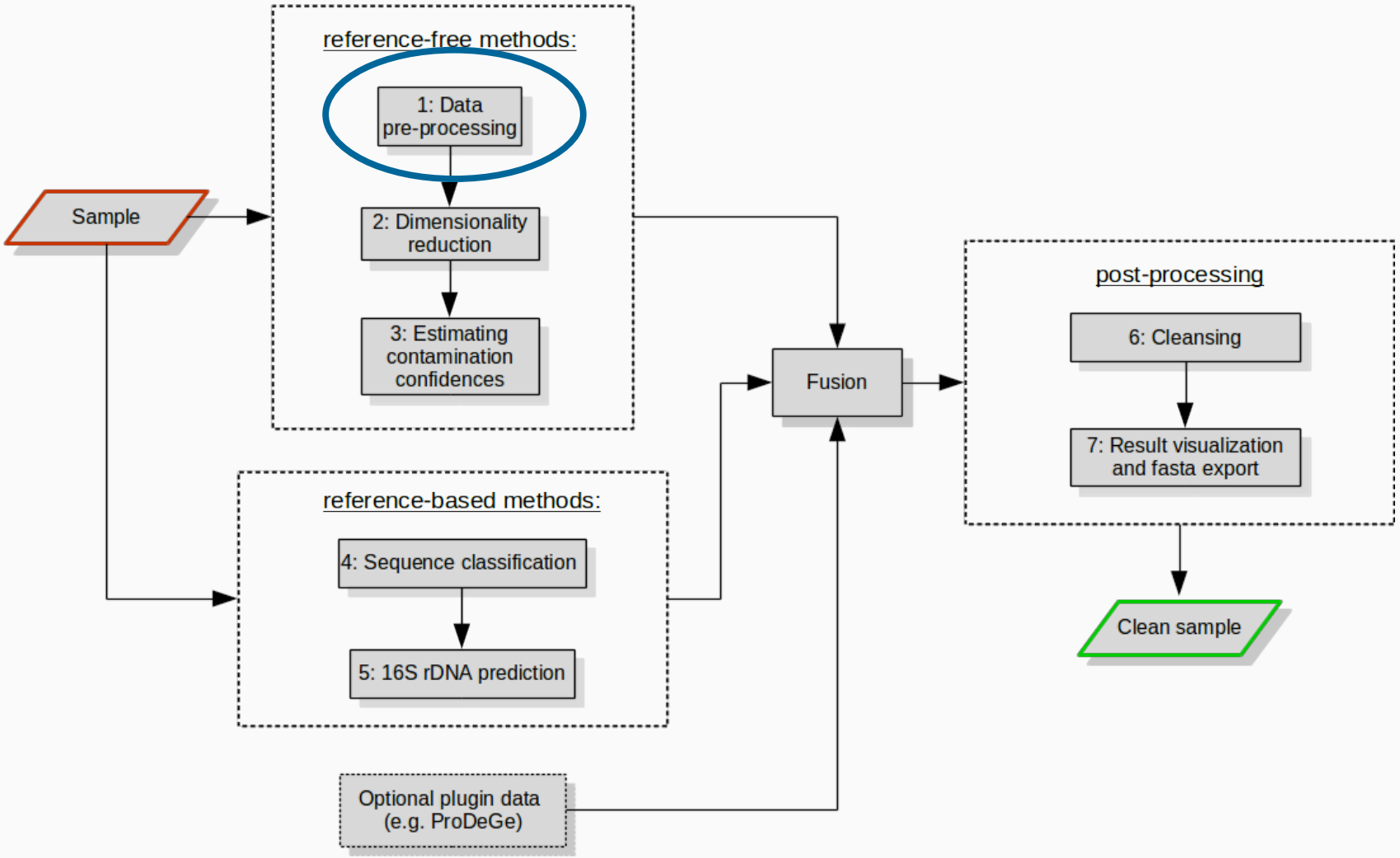
Idea: one cluster $\Leftrightarrow$ one genome

$\rightarrow$ detect number of clusters

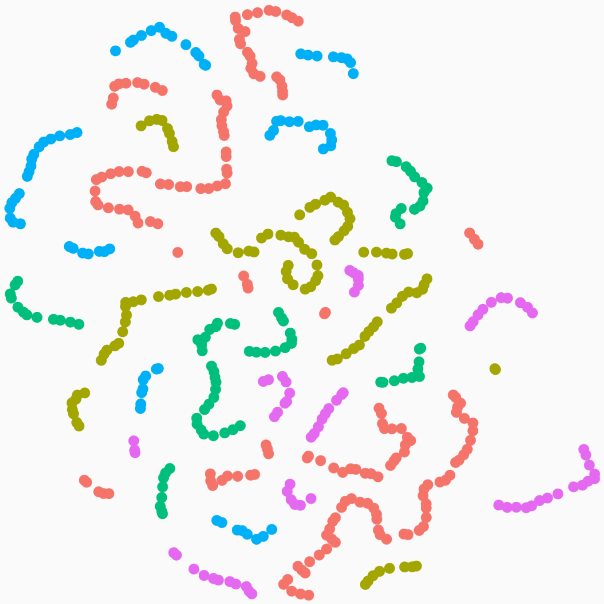
Contamination detection and cleansing

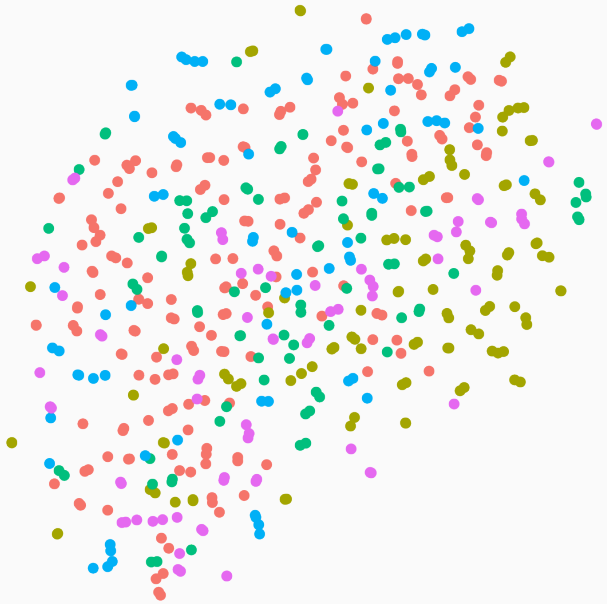


Contamination detection and cleansing



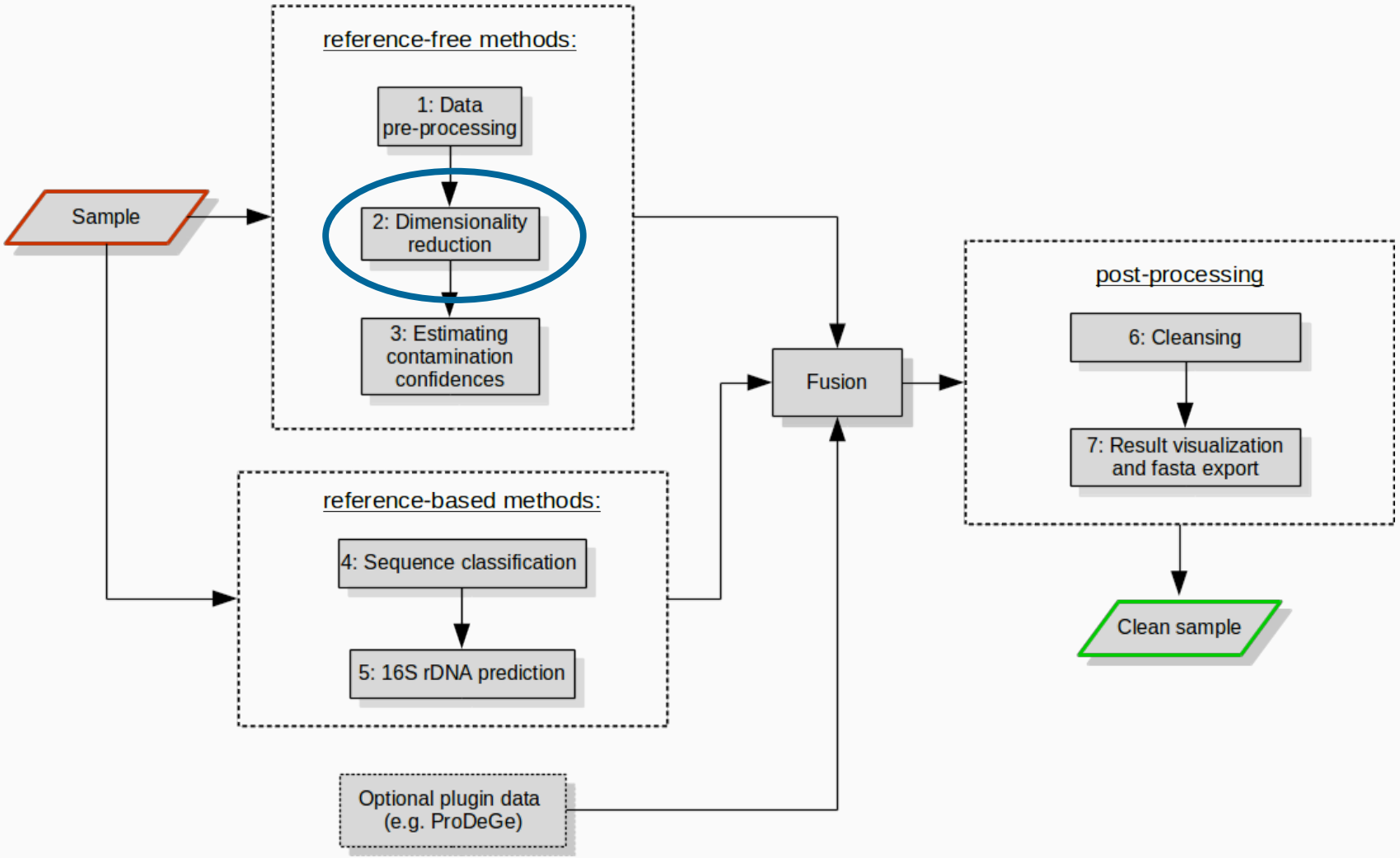
Window parameters



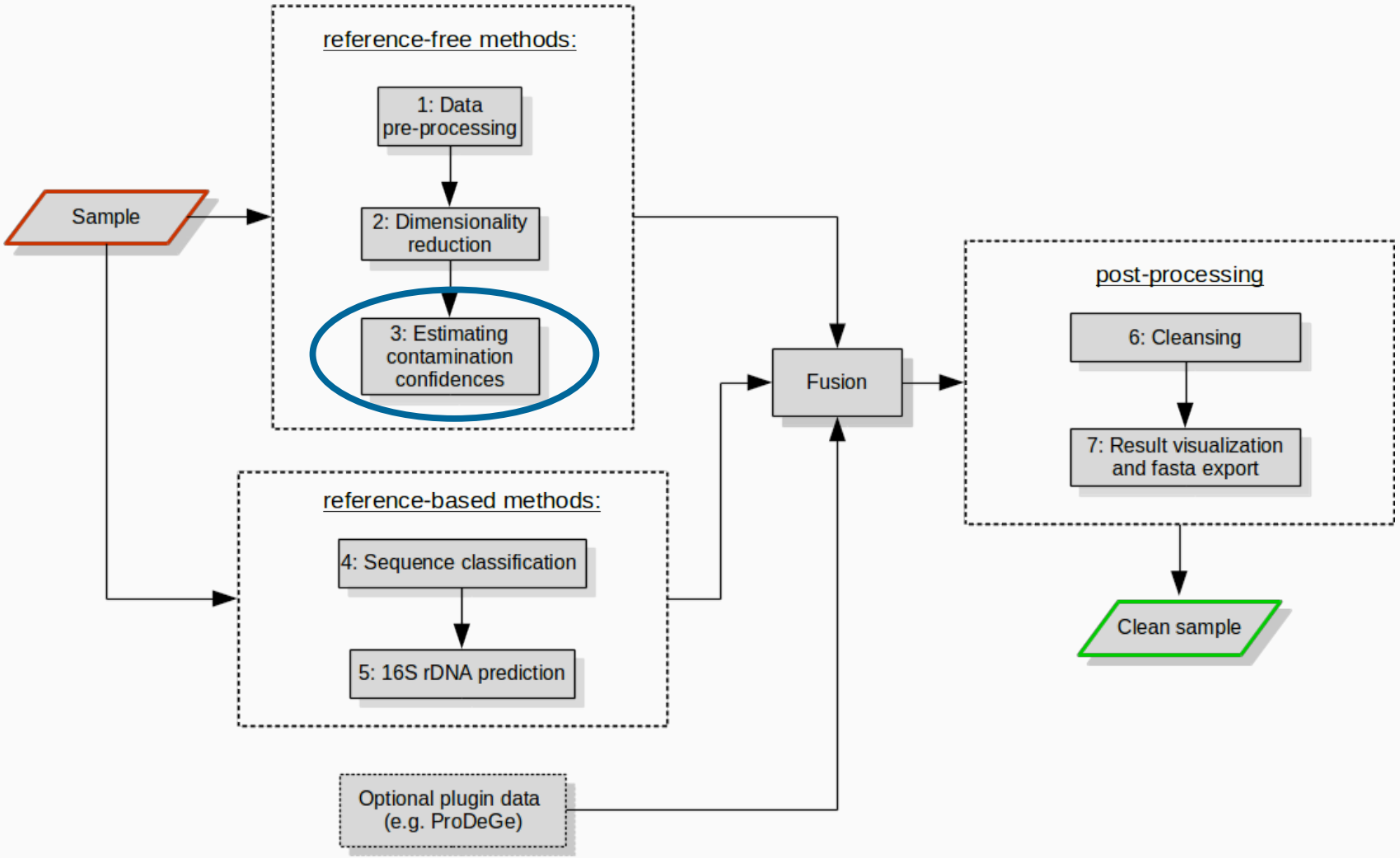


$$\Delta w = \lceil \frac{n_{bp}}{n_{target}} \rceil, w = 2 \cdot \Delta w$$

Contamination detection and cleansing



Contamination detection and cleansing



Contamination yes/no?

- Task: Determine whether number of clusters
  - $k = 1$ (clean)
  - $k > 1$ (contaminated)
- Many known algorithms not applicable[4]

[4] Lux et al. (2015a)

Contamination yes/no?

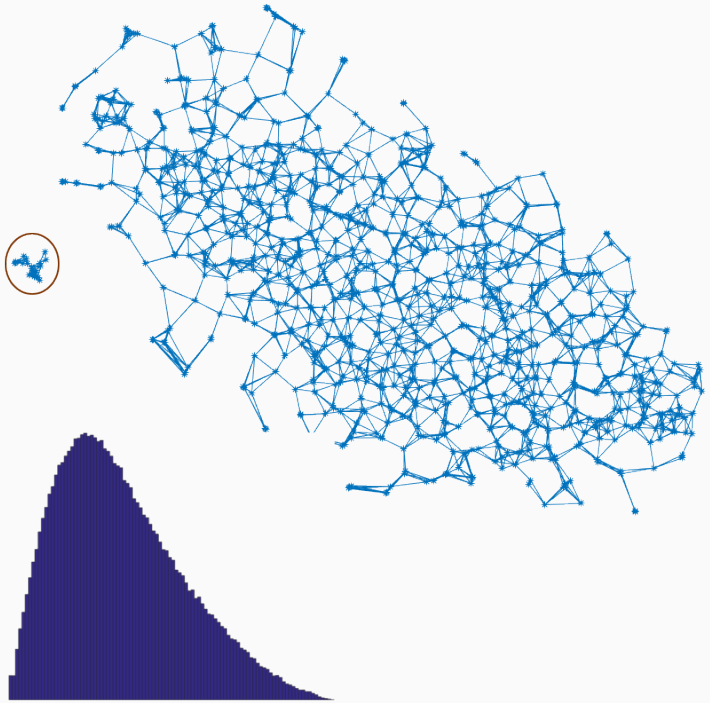
- Task: Determine whether number of clusters
  - $k = 1$ (clean)
  - $k > 1$ (contaminated)
- Many known algorithms not applicable[4]

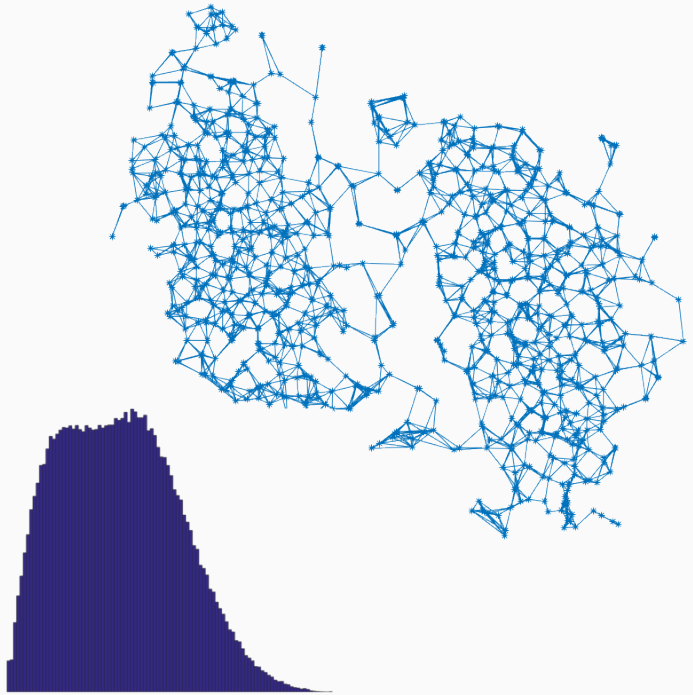
Contamination detection and confidence estimation:

1. The dip-statistic test for multimodality
2. Connected components in a neighborhood graph
3. Estimate confidences using bootstrapping

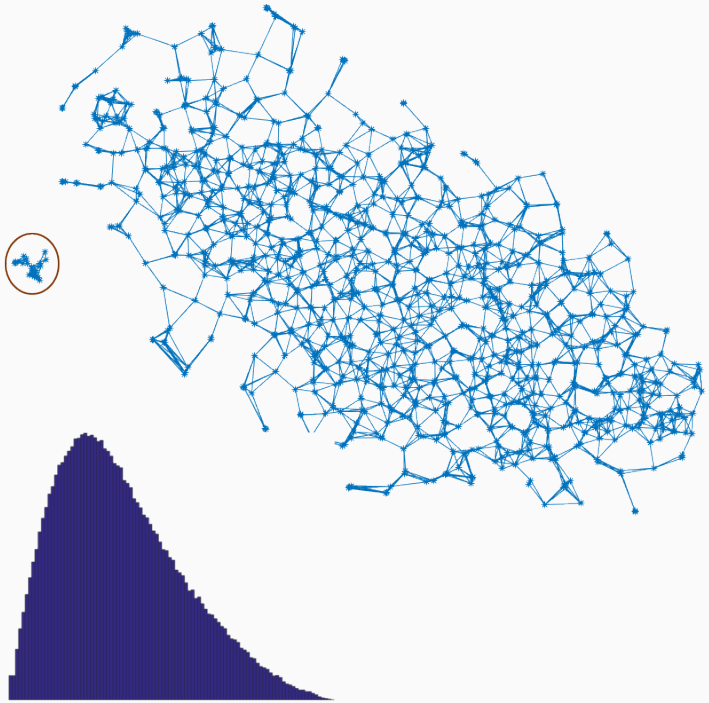
[4] Lux et al. (2015a)

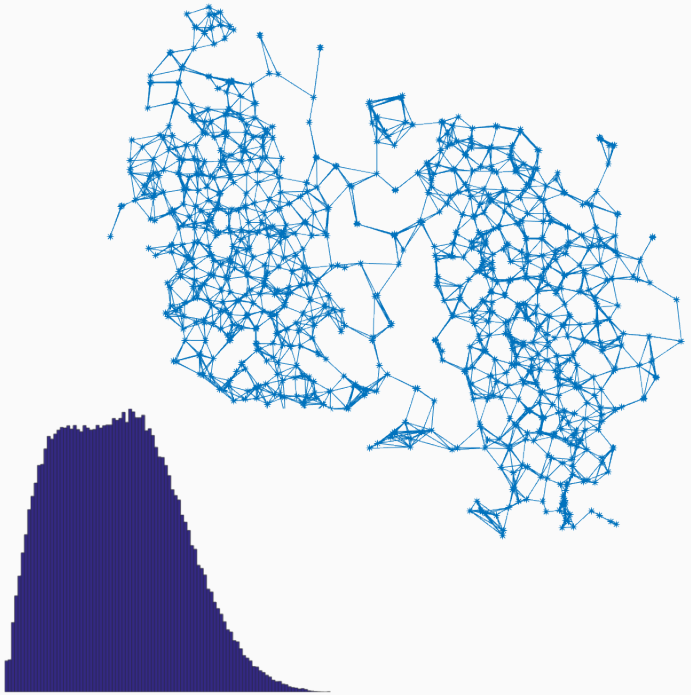
Contamination yes/no?





Contamination yes/no?

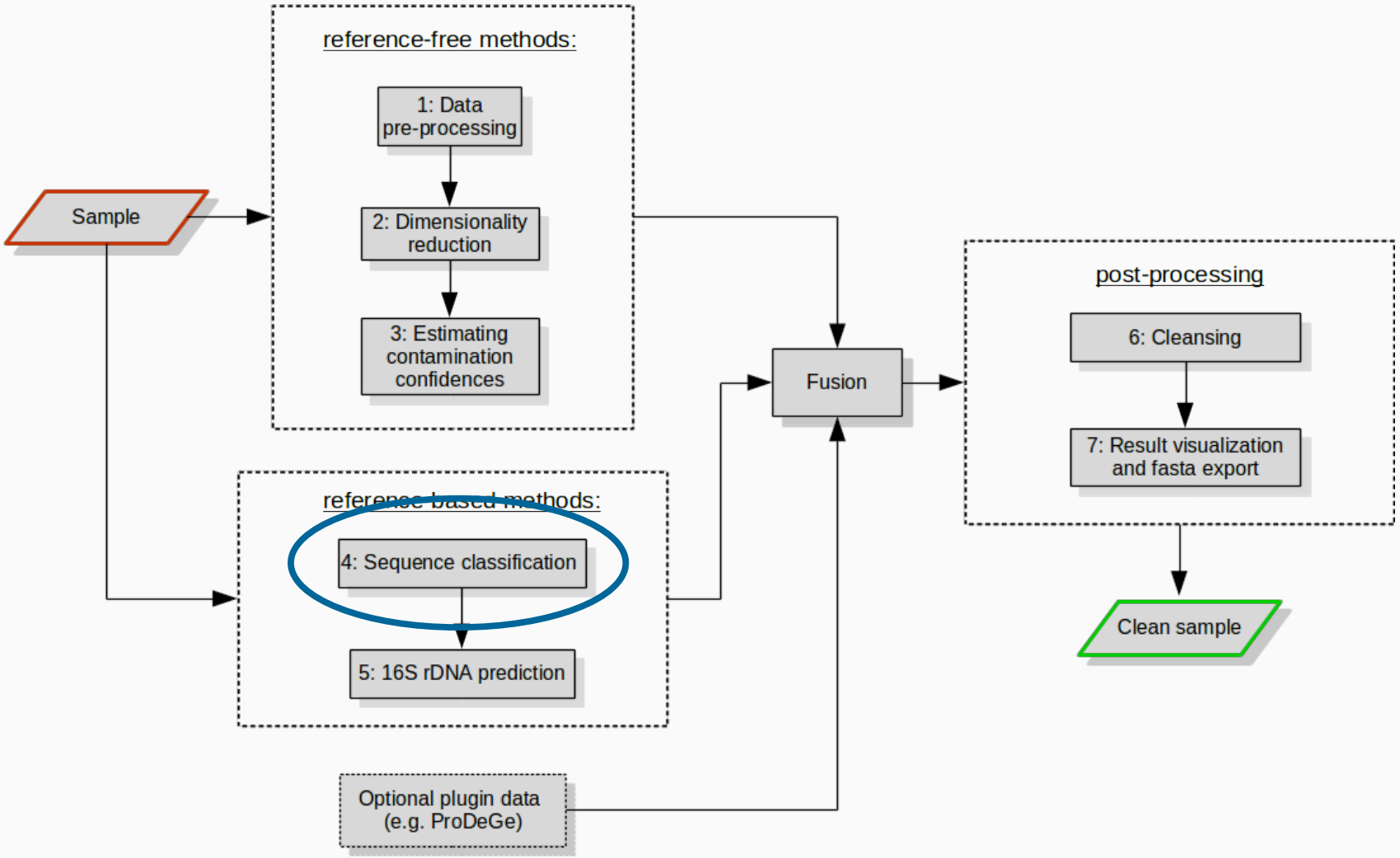




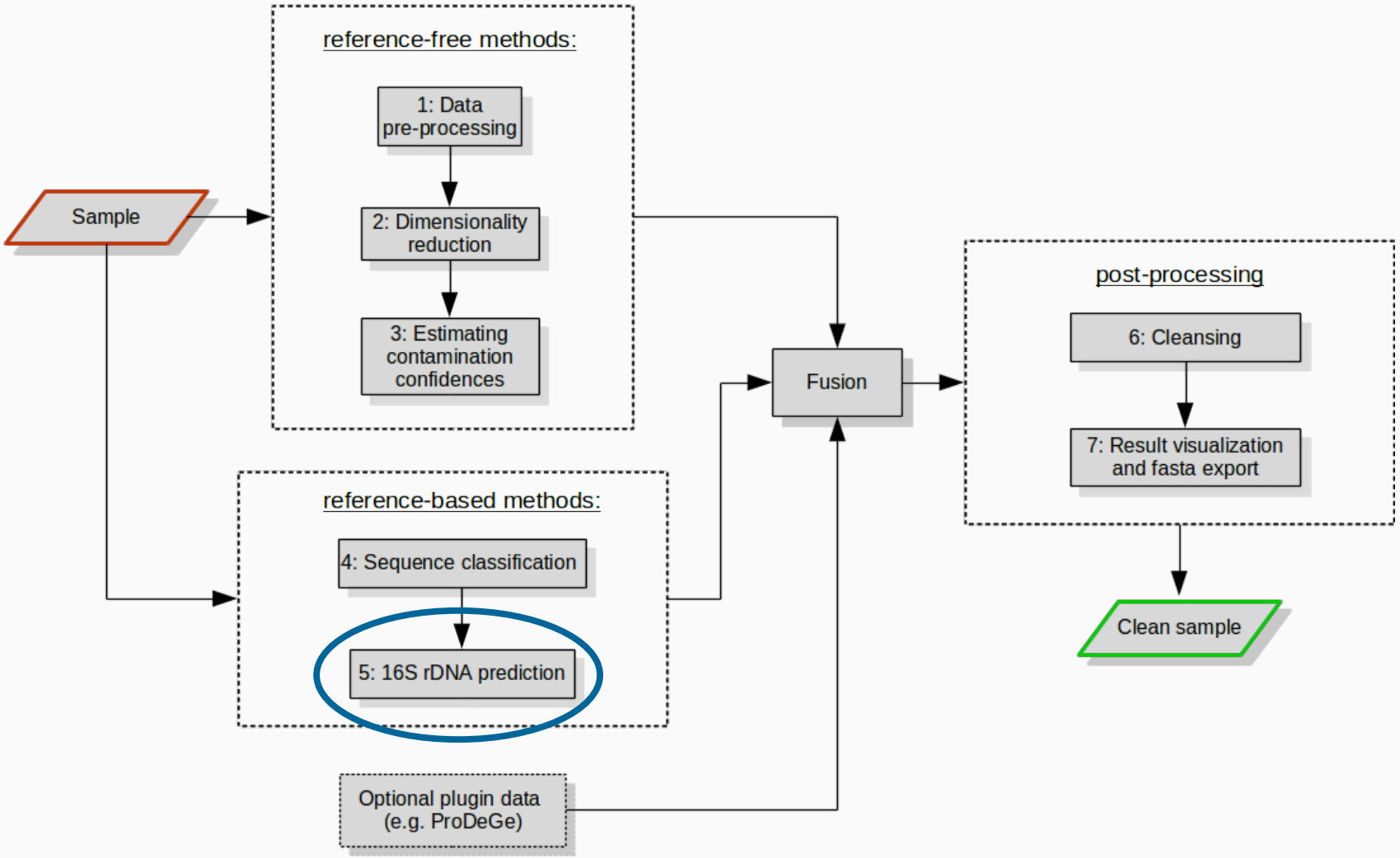
Confidence value:

$$\nu_{dip|cc} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{k_{dip}+k_{cc}>2}$$

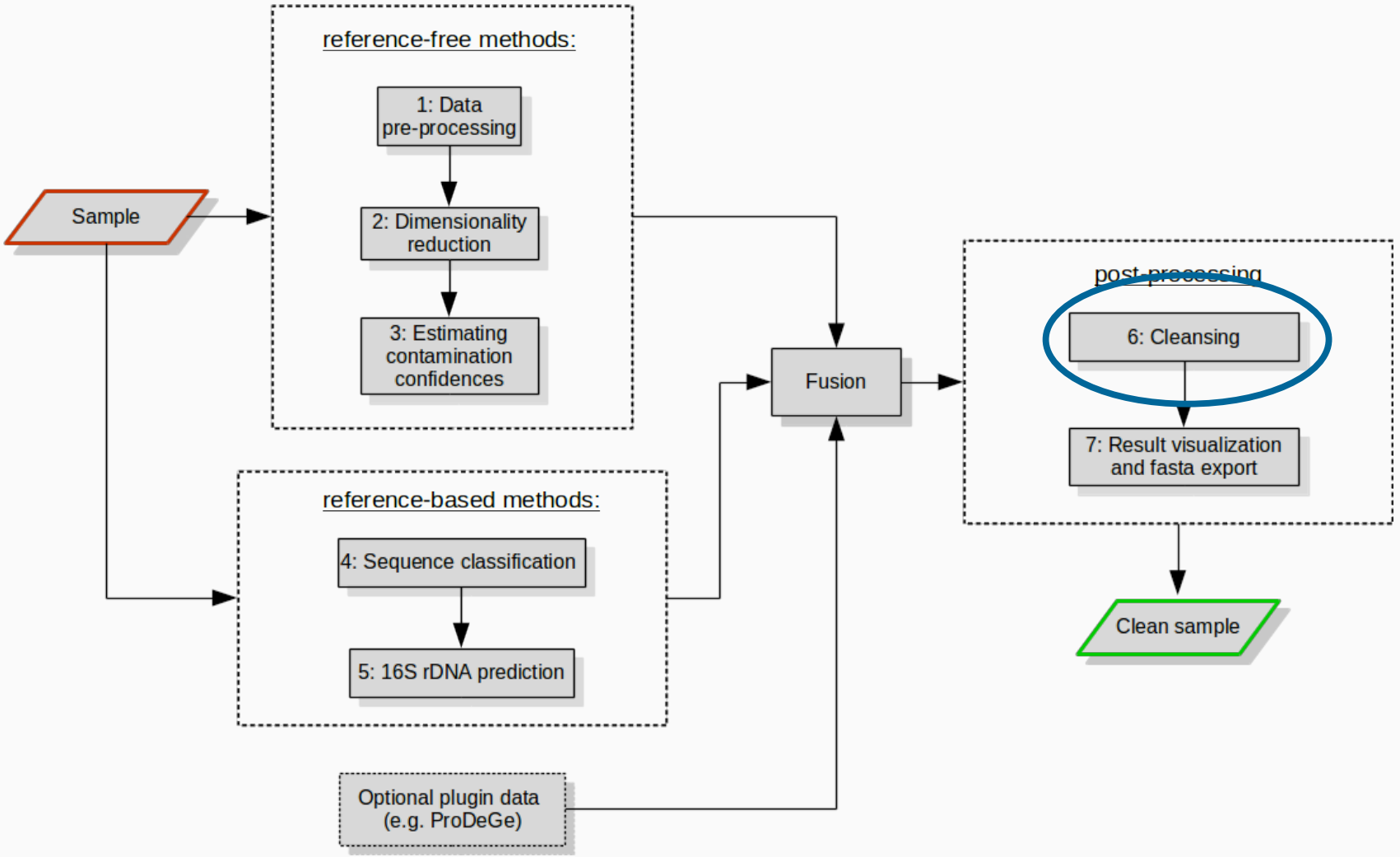
Contamination detection and cleansing



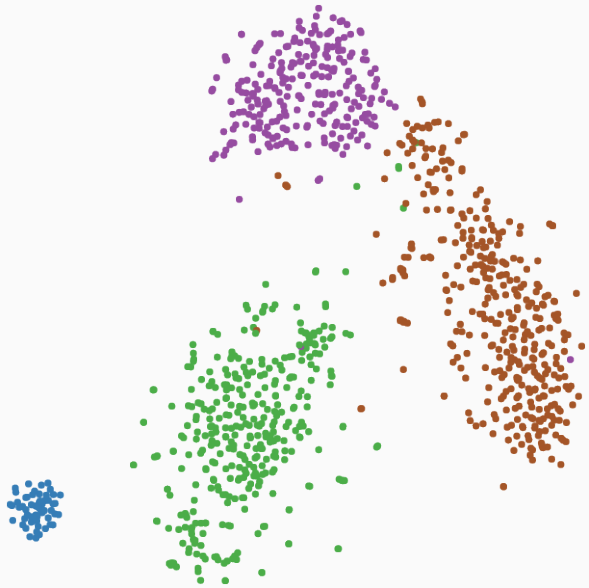
Contamination detection and cleansing



Contamination detection and cleansing



Cleansing

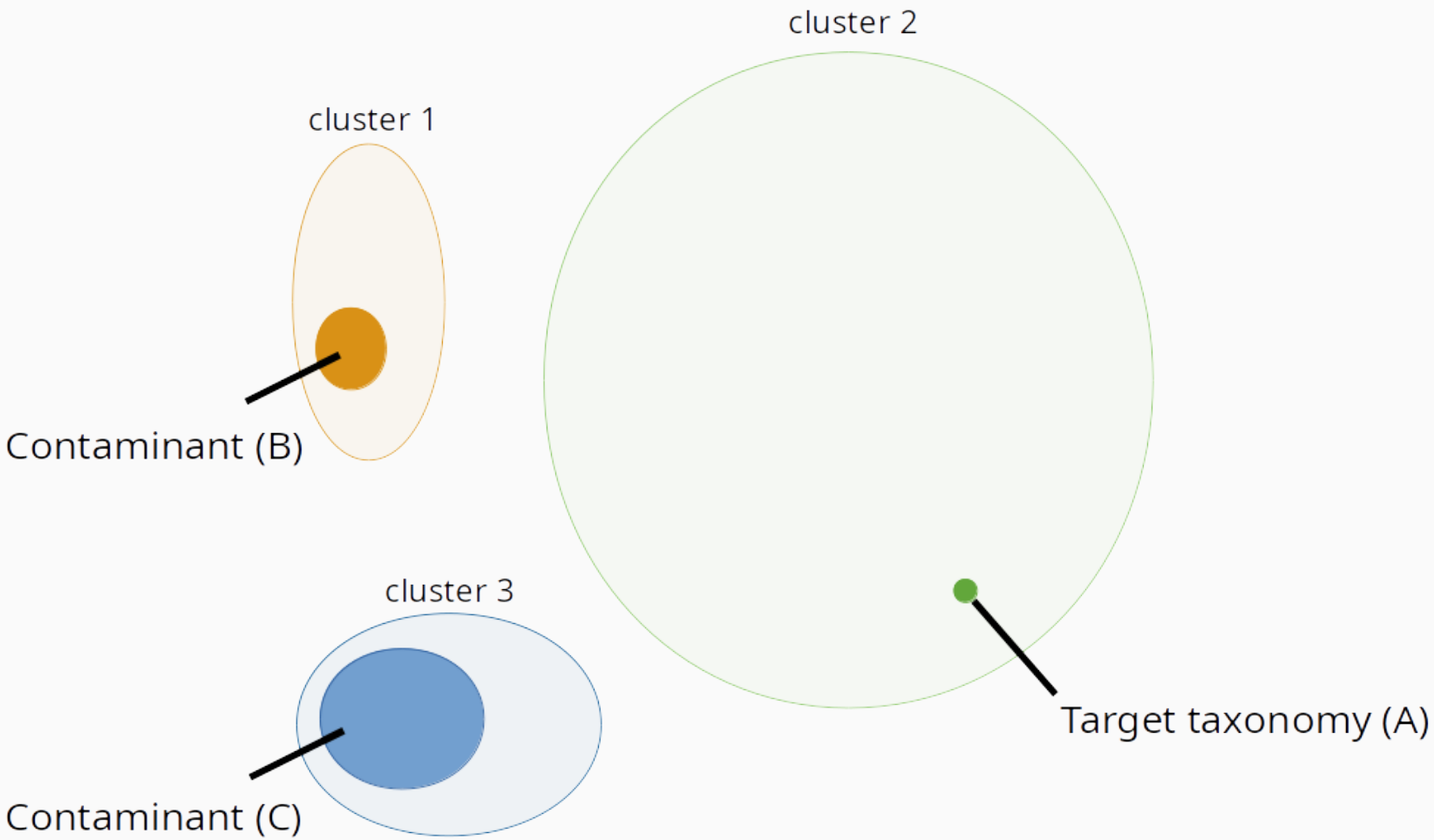


Find optimal clustering for a given range of k

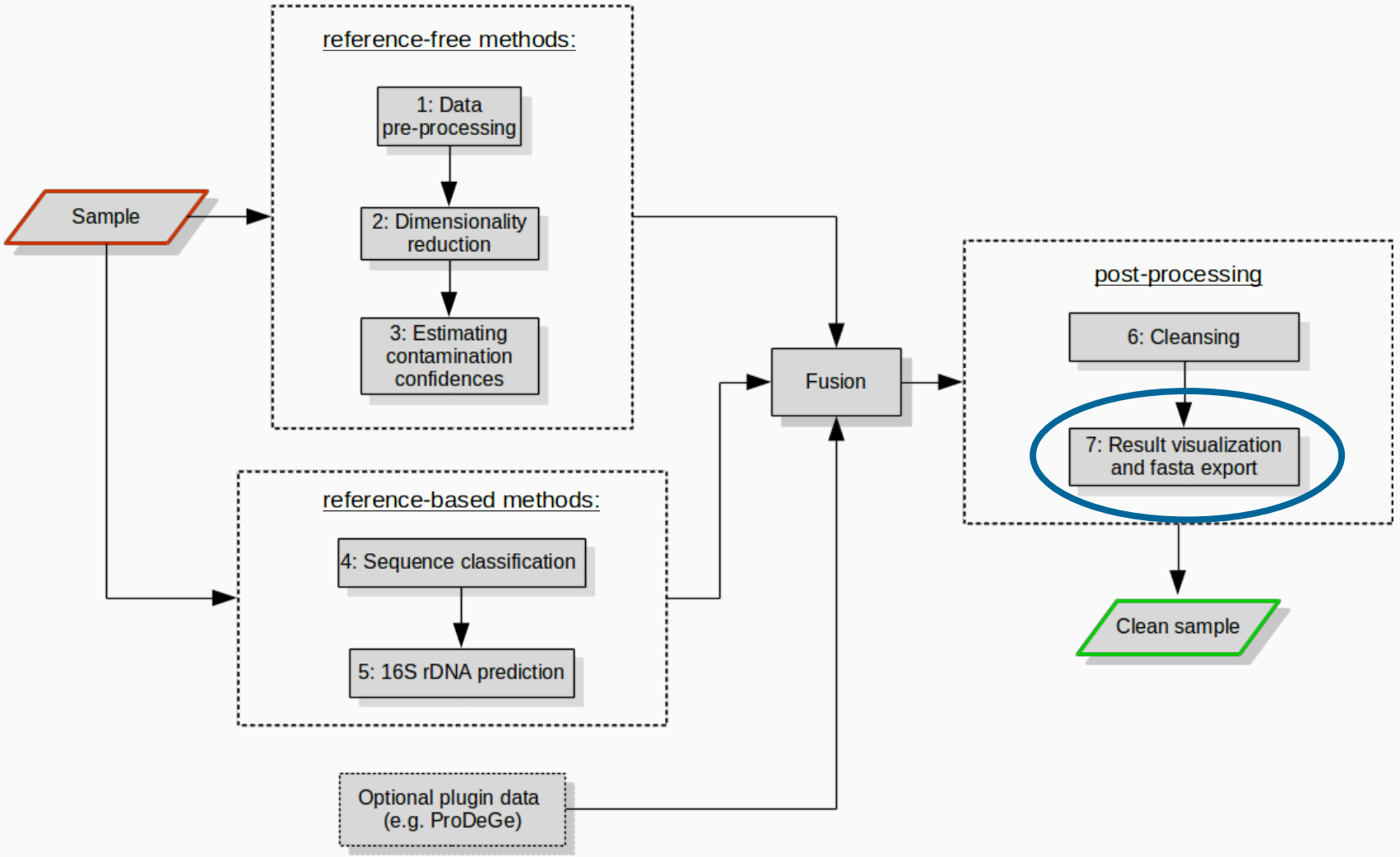
$\rightarrow$ export individual clusters

Integration of external knowledge: Taxonomy inference

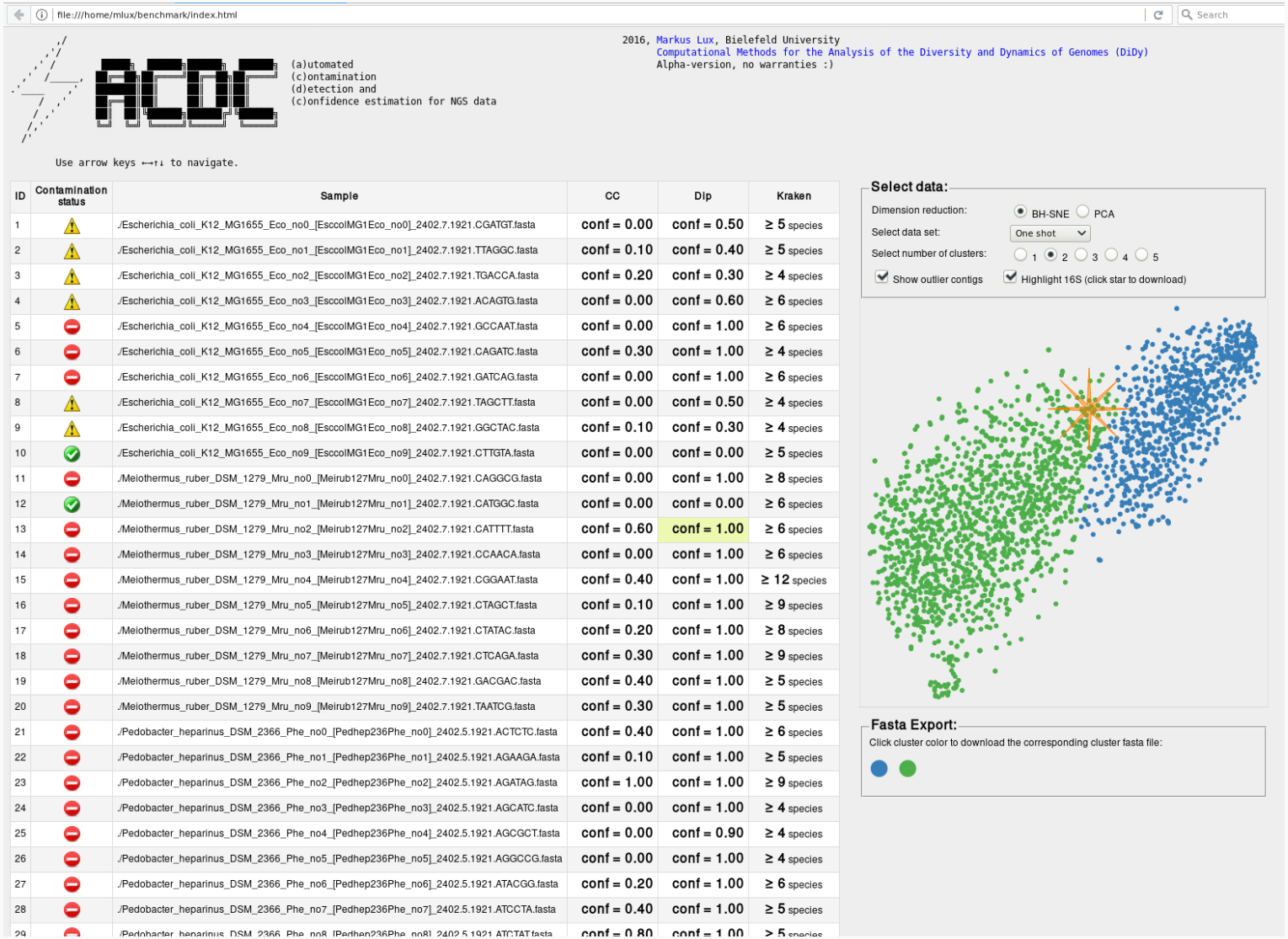
Integration of external knowledge: Taxonomy inference



Contamination detection and cleansing



Example 1



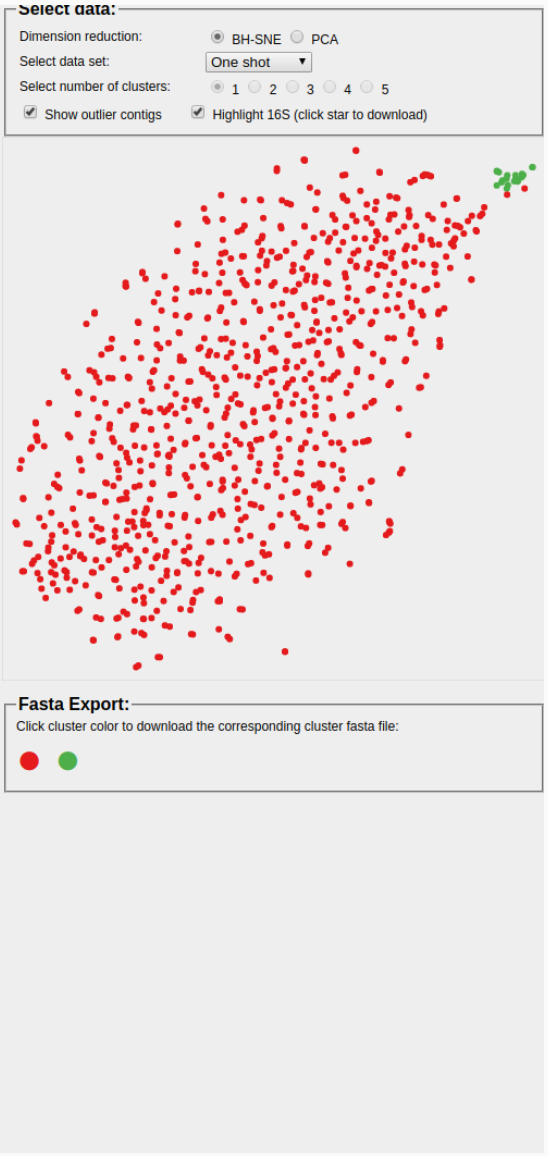
ID	Contamination status	Sample	CC	Dip	Kraken
1	⚠	./Escherichia_coli_K12_MG1655_Eco_no0_[EsccolMG1Eco_no0]_2402.7.1921.CGATGT.fasta	conf = 0.00	conf = 0.50	≥ 5 species
2	⚠	./Escherichia_coli_K12_MG1655_Eco_no1_[EsccolMG1Eco_no1]_2402.7.1921.TTAGGC.fasta	conf = 0.10	conf = 0.40	≥ 5 species
3	⚠	./Escherichia_coli_K12_MG1655_Eco_no2_[EsccolMG1Eco_no2]_2402.7.1921.TGACCA.fasta	conf = 0.20	conf = 0.30	≥ 4 species
4	⚠	./Escherichia_coli_K12_MG1655_Eco_no3_[EsccolMG1Eco_no3]_2402.7.1921.ACAGTG.fasta	conf = 0.00	conf = 0.60	≥ 6 species
5	⛔	./Escherichia_coli_K12_MG1655_Eco_no4_[EsccolMG1Eco_no4]_2402.7.1921.GCCAAT.fasta	conf = 0.00	conf = 1.00	≥ 6 species
6	⛔	./Escherichia_coli_K12_MG1655_Eco_no5_[EsccolMG1Eco_no5]_2402.7.1921.CAGATC.fasta	conf = 0.30	conf = 1.00	≥ 4 species
7	⛔	./Escherichia_coli_K12_MG1655_Eco_no6_[EsccolMG1Eco_no6]_2402.7.1921.GATCAG.fasta	conf = 0.00	conf = 1.00	≥ 6 species
8	⚠	./Escherichia_coli_K12_MG1655_Eco_no7_[EsccolMG1Eco_no7]_2402.7.1921.TAGCTT.fasta	conf = 0.00	conf = 0.50	≥ 4 species
9	⚠	./Escherichia_coli_K12_MG1655_Eco_no8_[EsccolMG1Eco_no8]_2402.7.1921.GGCTAC.fasta	conf = 0.10	conf = 0.30	≥ 4 species
10	✅	./Escherichia_coli_K12_MG1655_Eco_no9_[EsccolMG1Eco_no9]_2402.7.1921.CTTGTA.fasta	conf = 0.00	conf = 0.00	≥ 5 species
11	⛔	./Meiothermus_ruber_DSM_1279_Mru_no0_[Meirub127Mru_no0]_2402.7.1921.CAGGCG.fasta	conf = 0.00	conf = 1.00	≥ 8 species
12	✅	./Meiothermus_ruber_DSM_1279_Mru_no1_[Meirub127Mru_no1]_2402.7.1921.CATGGC.fasta	conf = 0.00	conf = 0.00	≥ 6 species
13	⛔	./Meiothermus_ruber_DSM_1279_Mru_no2_[Meirub127Mru_no2]_2402.7.1921.CATTTT.fasta	conf = 0.60	conf = 1.00	≥ 6 species
14	⛔	./Meiothermus_ruber_DSM_1279_Mru_no3_[Meirub127Mru_no3]_2402.7.1921.CCAACA.fasta	conf = 0.00	conf = 1.00	≥ 6 species
15	⛔	./Meiothermus_ruber_DSM_1279_Mru_no4_[Meirub127Mru_no4]_2402.7.1921.CGGAAT.fasta	conf = 0.40	conf = 1.00	≥ 12 species
16	⛔	./Meiothermus_ruber_DSM_1279_Mru_no5_[Meirub127Mru_no5]_2402.7.1921.CTAGCT.fasta	conf = 0.10	conf = 1.00	≥ 9 species
17	⛔	./Meiothermus_ruber_DSM_1279_Mru_no6_[Meirub127Mru_no6]_2402.7.1921.CTATAC.fasta	conf = 0.20	conf = 1.00	≥ 8 species
18	⛔	./Meiothermus_ruber_DSM_1279_Mru_no7_[Meirub127Mru_no7]_2402.7.1921.CTCAGA.fasta	conf = 0.30	conf = 1.00	≥ 9 species
19	⛔	./Meiothermus_ruber_DSM_1279_Mru_no8_[Meirub127Mru_no8]_2402.7.1921.GACGAC.fasta	conf = 0.40	conf = 1.00	≥ 5 species
20	⛔	./Meiothermus_ruber_DSM_1279_Mru_no9_[Meirub127Mru_no9]_2402.7.1921.TAATCG.fasta	conf = 0.30	conf = 1.00	≥ 5 species
21	⛔	./Pedobacter_heparinus_DSM_2366_Phe_no0_[Pedhep236Phe_no0]_2402.5.1921.ACTCTC.fasta	conf = 0.40	conf = 1.00	≥ 6 species
22	⛔	./Pedobacter_heparinus_DSM_2366_Phe_no1_[Pedhep236Phe_no1]_2402.5.1921.AGAAGA.fasta	conf = 0.10	conf = 1.00	≥ 5 species
23	⛔	./Pedobacter_heparinus_DSM_2366_Phe_no2_[Pedhep236Phe_no2]_2402.5.1921.AGATAG.fasta	conf = 1.00	conf = 1.00	≥ 9 species
24	⛔	./Pedobacter_heparinus_DSM_2366_Phe_no3_[Pedhep236Phe_no3]_2402.5.1921.AGCATC.fasta	conf = 0.00	conf = 1.00	≥ 4 species
25	⛔	./Pedobacter_heparinus_DSM_2366_Phe_no4_[Pedhep236Phe_no4]_2402.5.1921.AGCGCT.fasta	conf = 0.00	conf = 0.90	≥ 4 species
26	⛔	./Pedobacter_heparinus_DSM_2366_Phe_no5_[Pedhep236Phe_no5]_2402.5.1921.AGGCCG.fasta	conf = 0.00	conf = 1.00	≥ 4 species
27	⛔	./Pedobacter_heparinus_DSM_2366_Phe_no6_[Pedhep236Phe_no6]_2402.5.1921.ATACGG.fasta	conf = 0.20	conf = 1.00	≥ 6 species
28	⛔	./Pedobacter_heparinus_DSM_2366_Phe_no7_[Pedhep236Phe_no7]_2402.5.1921.ATCCTA.fasta	conf = 0.40	conf = 1.00	≥ 5 species
29	⛔	./Pedobacter_heparinus_DSM_2366_Phe_no8_[Pedhep236Phe_no8]_2402.5.1921.ATCTAT.fasta	conf = 0.80	conf = 1.00	≥ 5 species

Example 2

8	✓	./2264867056.fna	conf = 0.00	conf = 0.00	≥ 1 species
9	⚠	./2264867057.fna	conf = 0.00	conf = 0.30	≥ 1 species
10	✓	./2264867058.fna	conf = 0.00	conf = 0.00	≥ 1 species
11	✓	./2264867059.fna	conf = 0.00	conf = 0.00	≥ 1 species
12	✓	./2264867060.fna	conf = 0.00	conf = 0.00	≥ 1 species
13	✓	./2264867061.fna	conf = 0.00	conf = 0.20	≥ 1 species
14	✓	./2264867062.fna	conf = 0.10	conf = 0.00	≥ 1 species
15	✓	./2264867063.fna	conf = 0.00	conf = 0.00	≥ 1 species
16	✓	./2264867064.fna	conf = 0.00	conf = 0.00	≥ 1 species
17	✓	./2264867065.fna	conf = 0.00	conf = 0.00	≥ 1 species
18	✓	./2264867066.fna	conf = 0.00	conf = 0.10	≥ 1 species
19	✓	./2264867067.fna	conf = 0.10	conf = 0.00	≥ 1 species
20	✓	./2264867069.fna	conf = 0.00	conf = 0.00	≥ 1 species
21	⚠	./2264867070.fna	conf = 0.30	conf = 0.00	≥ 1 species
22	✓	./2264867071.fna	conf = 0.00	conf = 0.00	≥ 1 species
23	✓	./2264867072.fna	conf = 0.00	conf = 0.00	≥ 1 species
24	✓	./2264867073.fna	conf = 0.00	conf = 0.00	≥ 1 species
25	⚠	./2264867074.fna	conf = 0.40	conf = 0.00	≥ 1 species
26	✓	./2264867075.fna	conf = 0.10	conf = 0.00	≥ 1 species
27	⚠	./2264867076.fna	conf = 0.50	conf = 0.00	≥ 1 species
28	✓	./2264867077.fna	conf = 0.00	conf = 0.00	≥ 1 species
29	⛔	./2264867079.fna	conf = 0.90	conf = 0.00	≥ 1 species
30	✓	./2264867080.fna	conf = 0.00	conf = 0.00	≥ 1 species
31	⛔	./2264867081.fna	conf = 0.80	conf = 0.00	≥ 1 species
32	✓	./2264867082.fna	conf = 0.00	conf = 0.00	≥ 1 species
33	✓	./2264867083.fna	conf = 0.00	conf = 0.00	≥ 1 species
34	⚠	./2264867084.fna	conf = 0.60	conf = 0.00	≥ 1 species
35	✓	./2264867085.fna	conf = 0.00	conf = 0.00	≥ 1 species
36	✓	./2264867086.fna	conf = 0.00	conf = 0.00	≥ 1 species
37	✓	./2264867087.fna	conf = 0.00	conf = 0.00	≥ 1 species
38	✓	./2264867088.fna	conf = 0.20	conf = 0.00	≥ 1 species
39	✓	./2264867089.fna	conf = 0.00	conf = 0.00	≥ 1 species
40	✓	./2264867090.fna	conf = 0.00	conf = 0.00	≥ 1 species
41	✓	./2264867092.fna	conf = 0.00	conf = 0.00	≥ 1 species

Clean samples actually contaminated

8	✓	./2264867056.fna	conf = 0.00	conf = 0.00	≥ 1 species
9	⚠	./2264867057.fna	conf = 0.00	conf = 0.30	≥ 1 species
10	✓	./2264867058.fna	conf = 0.00	conf = 0.00	≥ 1 species
11	✓	./2264867059.fna	conf = 0.00	conf = 0.00	≥ 1 species
12	✓	./2264867060.fna	conf = 0.00	conf = 0.00	≥ 1 species
13	✓	./2264867061.fna	conf = 0.00	conf = 0.20	≥ 1 species
14	✓	./2264867062.fna	conf = 0.10	conf = 0.00	≥ 1 species
15	✓	./2264867063.fna	conf = 0.00	conf = 0.00	≥ 1 species
16	✓	./2264867064.fna	conf = 0.00	conf = 0.00	≥ 1 species
17	✓	./2264867065.fna	conf = 0.00	conf = 0.00	≥ 1 species
18	✓	./2264867066.fna	conf = 0.00	conf = 0.10	≥ 1 species
19	✓	./2264867067.fna	conf = 0.10	conf = 0.00	≥ 1 species
20	✓	./2264867069.fna	conf = 0.00	conf = 0.00	≥ 1 species
21	⚠	./2264867070.fna	conf = 0.30	conf = 0.00	≥ 1 species
22	✓	./2264867071.fna	conf = 0.00	conf = 0.00	≥ 1 species
23	✓	./2264867072.fna	conf = 0.00	conf = 0.00	≥ 1 species
24	✓	./2264867073.fna	conf = 0.00	conf = 0.00	≥ 1 species
25	⚠	./2264867074.fna	conf = 0.40	conf = 0.00	≥ 1 species
26	✓	./2264867075.fna	conf = 0.10	conf = 0.00	≥ 1 species
27	⚠	./2264867076.fna	conf = 0.50	conf = 0.00	≥ 1 species
28	✓	./2264867077.fna	conf = 0.00	conf = 0.00	≥ 1 species
29	⛔	./2264867079.fna	conf = 0.90	conf = 0.00	≥ 1 species
30	✓	./2264867080.fna	conf = 0.00	conf = 0.00	≥ 1 species
31	⛔	./2264867081.fna	conf = 0.80	conf = 0.00	≥ 1 species
32	✓	./2264867082.fna	conf = 0.00	conf = 0.00	≥ 1 species
33	✓	./2264867083.fna	conf = 0.00	conf = 0.00	≥ 1 species
34	⚠	./2264867084.fna	conf = 0.60	conf = 0.00	≥ 1 species
35	✓	./2264867085.fna	conf = 0.00	conf = 0.00	≥ 1 species
36	✓	./2264867086.fna	conf = 0.00	conf = 0.00	≥ 1 species
37	✓	./2264867087.fna	conf = 0.00	conf = 0.00	≥ 1 species
38	✓	./2264867088.fna	conf = 0.20	conf = 0.00	≥ 1 species
39	✓	./2264867089.fna	conf = 0.00	conf = 0.00	≥ 1 species
40	✓	./2264867090.fna	conf = 0.00	conf = 0.00	≥ 1 species
41	✓	./2264867092.fna	conf = 0.00	conf = 0.00	≥ 1 species



Results

- Correctly detected contamination state on large set of SAGs
- Contaminants of same genus
- Removed contaminants with high precision and recall
- Good prediction of pure and impure CAMI bins
- Quadratic complexity but low practical requirements

Results

- Correctly detected contamination state on large set of SAGs
- Contaminants of same genus
- Removed contaminants with high precision and recall
- Good prediction of pure and impure CAMI bins
- Quadratic complexity but low practical requirements

Results

- Correctly detected contamination state on large set of SAGs
- Contaminants of same genus
- Removed contaminants with high precision and recall
- Good prediction of pure and impure CAMI bins
- Quadratic complexity but low practical requirements

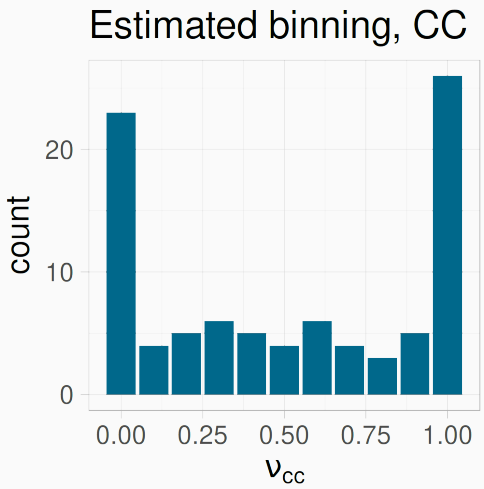
Results

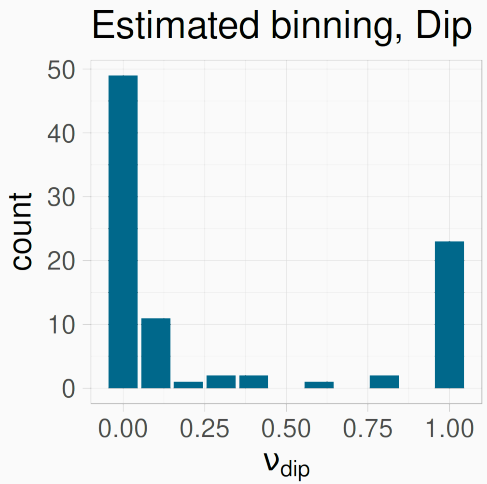
- Correctly detected contamination state on large set of SAGs
- Contaminants of same genus
- Removed contaminants with high precision and recall
- Good prediction of pure and impure CAMI bins
- Quadratic complexity but low practical requirements

Results

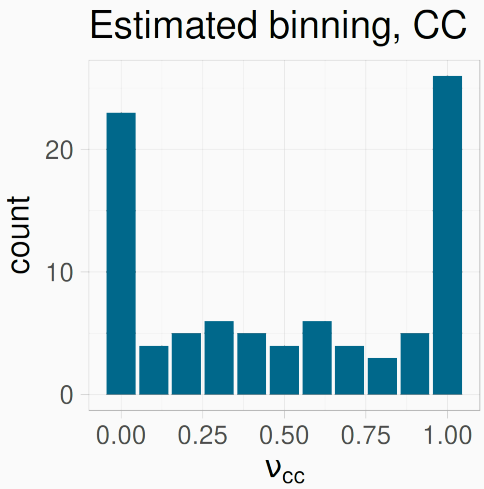
- Correctly detected contamination state on large set of SAGs
- Contaminants of same genus
- Removed contaminants with high precision and recall
- Good prediction of pure and impure CAMI bins
- Quadratic complexity but low practical requirements

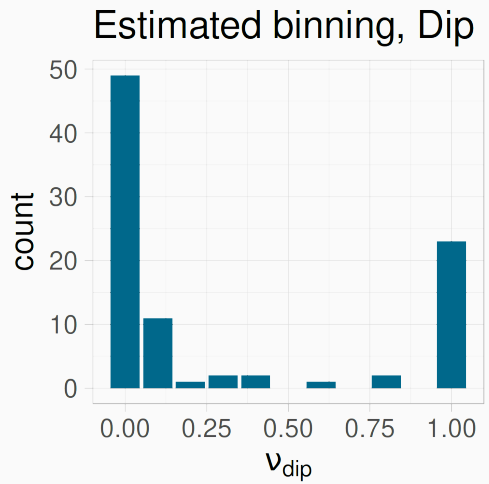
Early results on MAG data

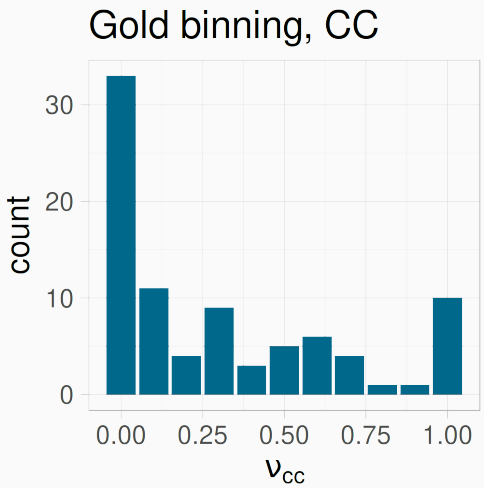


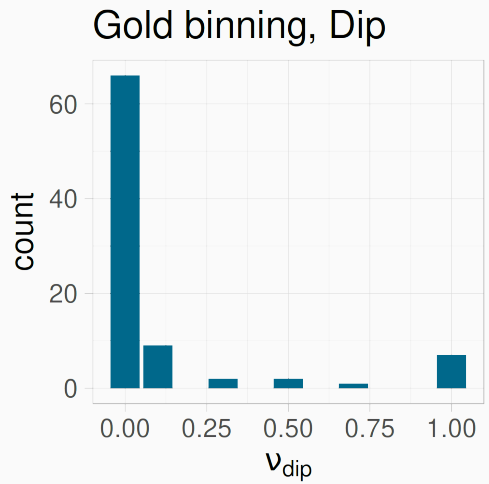


Early results on MAG data









Outlook

- Collaboration with Joint Genome Institute (JGI), USA
- Research other types of data representation to overcome species-rank limitation
- Further investigation of MAG capabilities
- Investigate horizontal gene transfer detection

Outlook

- Collaboration with Joint Genome Institute (JGI), USA
- Research other types of data representation to overcome species-rank limitation
- Further investigation of MAG capabilities
- Investigate horizontal gene transfer detection

Outlook

- Collaboration with Joint Genome Institute (JGI), USA
- Research other types of data representation to overcome species-rank limitation
- Further investigation of MAG capabilities
- Investigate horizontal gene transfer detection

Outlook

- Collaboration with Joint Genome Institute (JGI), USA
- Research other types of data representation to overcome species-rank limitation
- Further investigation of MAG capabilities
- Investigate horizontal gene transfer detection

Markus Lux, Jan Krüger, Christian Rinke, Irena Maus, Andreas Schlueter, Tanja Woyke, Alexander Sczyrba, Barbara Hammer

acdc – Automated Contamination Detection and Confidence estimation for single-cell genome data

BMC Bioinformatics, 2016

Markus Lux, Barbara Hammer, Alexander Sczyrba

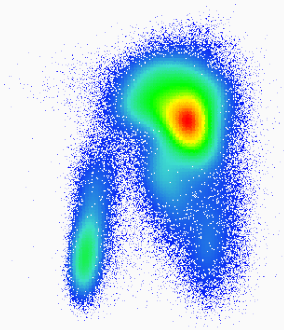
Automated Contamination Detection in Single-Cell Sequencing

bioRxiv, 2015

Poster @ Third Microbial Single Cell Genomics Workshop

Bigelow Laboratory for Ocean Sciences, 2015

Overall Summary

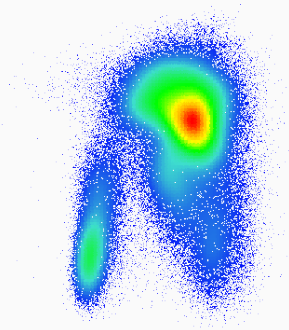


→ Semi-supervised approach for flow cytometry gating

→ Gate thousands of samples from one example, outperforming state-of-the-art methods

Lux et al. (2018)

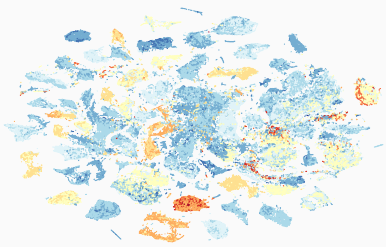
Overall Summary



→ Semi-supervised approach for flow cytometry gating

→ Gate thousands of samples from one example, outperforming state-of-the-art methods

Lux et al. (2018)

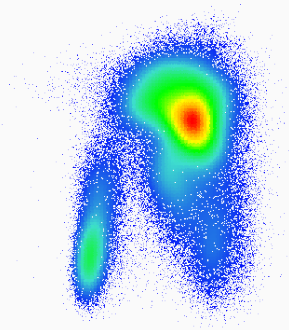


→ Evaluation of taxonomy-independent, sequence-composition based binning

→ Guidelines for data representation, dimensionality reduction and clustering

Lux et al. (2015b)

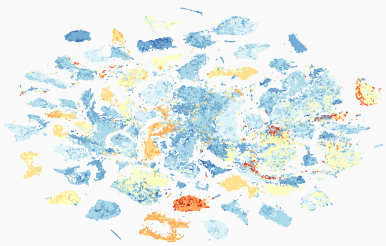
Overall Summary



→ Semi-supervised approach for flow cytometry gating

→ Gate thousands of samples from one example, outperforming state-of-the-art methods

Lux et al. (2018)



→ Evaluation of taxonomy-independent, sequence-composition based binning

→ Guidelines for data representation, dimensionality reduction and clustering

Lux et al. (2015b)

→ Automated contamination detection for single-cell genome data

→ Currently adapted for production by leading single-cell institute

Lux et al. (2015a, 2016)

References

Nima Aghaeepour, Radina Nikolic, Holger H Hoos, and Ryan R Brinkman. Rapid cell population identification in flow cytometry data. *Cytometry Part A*, 79(1):6–13, 2011.

Robert M Bowers, Nikos C Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, TBK Reddy, Frederik Schulz, Jessica Jarett, Adam R Rivers, Emiley A Eloe-Fadrosh, et al. Minimum information about a single amplified genome (misag) and a metagenome-assembled genome (mimag) of bacteria and archaea. *Nature biotechnology*, 35(8), 2017.

Greg Finak, Jacob Frelinger, Wenxin Jiang, Evan W Newell, John Ramey, Mark M Davis, Spyros A Kalams, Stephen C De Rosa, and Raphael Gottardo. Opencyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS computational biology*, 10(8):e1003806, 2014.

Greg Finak, Marc Langweiler, Maria Jaimes, Mehrnoush Malek, Jafar Taghiyar, Yael Korin, Khadir Raddassi, Lesley Devine, Gerlinde Obermoser, Marcin L Pekalski, et al. Standardizing flow cytometry immunophenotyping analysis from the human immunophenotyping consortium. *Scientific reports*, 6, 2016.

Yongchao Ge and Stuart C Sealfon. flowpeaks: a fast unsupervised clustering for flow cytometry data via k-means and density peak finding. *Bioinformatics*, 28(15):2052–2058, 2012.

Cedric C Laczny, Nicolás Pinel, Nikos Vlassis, and Paul Wilmes. Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Scientific reports*, 4, 2014.

Huamin Li, Uri Shaham, Kelly P. Stanton, Yi Yao, Ruth Montgomery, and Yuval Kluger. Gating mass cytometry data by deep learning. *Bioinformatics*, 2017.

Hsin-Hung Lin and Yu-Chieh Liao. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Scientific reports*, 6:24175, 2016.

Markus Lux, Barbara Hammer, and Alexander Sczyrba. Automated contamination detection in single-cell sequencing. *bioRxiv*, page 020859, 2015a.

Markus Lux, Alexander Sczyrba, and Barbara Hammer. Automatic discovery of metagenomic structure. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–8. IEEE, 2015b.

Markus Lux, Jan Krüger, Christian Rinke, Irena Maus, Andreas Schlüter, Tanja Woyke, Alexander Sczyrba, and Barbara Hammer. acdc–automated contamination detection and confidence estimation for single-cell genome data. *BMC bioinformatics*, 17(1):543, 2016.

Markus Lux, Ryan Remy Brinkman, Cedric Chauve, Adam Laing, Anna Lorenc, Lucie Abeler-Dörner, Barbara Hammer, and Jonathan Wren. flowlearn: Fast and precise identification and quality checking of cell populations in flow cytometry. *Bioinformatics*, 1:9, 2018.

Florian Mair, Felix J Hartmann, Dunja Mrdjen, Vinko Tosevski, Carsten Krieg, and Burkhard Becher. The end of gating? an introduction to automated analysis of high dimensional cytometry data. *European journal of immunology*, 46(1):34–43, 2016.

Mehrnoush Malek, Mohammad Jafar Taghiyar, Lauren Chong, Greg Finak, Raphael Gottardo, and Ryan R Brinkman. flowdensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinformatics*, 31(4):606–607, 2015.

Leland McInnes and John Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Peng Qiu, Erin F Simonds, Sean C Bendall, Kenneth D Gibbs Jr, Robert V Bruggner, Michael D Linderman, Karen Sachs, Garry P Nolan, and Sylvia K Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nature biotechnology*, 29(10):886–891, 2011.

Karel Sedlar, Kristyna Kupkova, and Ivo Provaznik. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Computational and structural biotechnology journal*, 15: 48–55, 2017.

Kristin Tennessen, Evan Andersen, Scott Clingenpeel, Christian Rinke, Derek S Lundberg, James Han, Jeff L Dangl, Natalia Ivanova, Tanja Woyke, Nikos Kyrpides, et al. Prodege: a computational protocol for fully automated decontamination of genomes. *The ISME journal*, 2015.

Sofie Van Gassen, Britt Callebaut, Mary J Van Helden, Bart N Lambrecht, Piet Demeester, Tom Dhaene, and Yvan Saeys. Flowsom: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, 87(7):636–645, 2015.

Yu-Wei Wu, Blake A Simmons, and Steven W Singer. Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–607, 2015.