TADA: phylogenetic augmentation of microbiome samples enhances phenotype classification

Erfan Sayyari[1], Ban Kawas [2], Siavash Mirarab[1]

[1] University of California, San Diego,
[2] IBM, Research—Almaden

Erfan Sayyari

Ban Kawas

Many studies report successful application of supervised learning to microbiome data

Gut microbiome development along the colorectal adenoma–carcinoma sequence

Qiang Feng, Suisha Liang […] Jun Wang

Nature Communications **6**, Article number: 6528 (2015) | Download Citation

Gastrointestinal Microbiome Signatures of Pediatric Patients With Irritable Bowel Syndrome

Delphine M. Saulnier[*,†,§], Kevin Riehle[‡,¶], Toni–Ann Mistretta[*,‡], Maria–Alejandra Diaz[*,‡], Debasmita Mandal[#], Sabeen Raza[*,‡], Erica M. Weidler[**,‡‡], Xiang Qin[§§], Cristian Coarfa[‡,¶], Aleksandar Milosavljevic[‡,¶], Joseph F. Petrosino[§§,ıı,¶¶], Sarah Highlander[§§,ıı], Richard Gibbs[§§], Susan V. Lynch[#], Robert J. Shulman[**,‡‡], James Versalovic[*,†,‡,**,ıı,*]

A Metagenomic Approach to Characterization of the Vaginal Microbiome Signature in Pregnancy

Kjersti Aagaard, Kevin Riehle, Jun Ma, Nicola Segata, Toni-Ann Mistretta, Cristian Coarfa, Sabeen Raza, Sean Rosenbaum, Ignatia Van den Veyver, Aleksandar Milosavljevic, Dirk Gevers, Curtis Huttenhower, Joseph Petrosino, James Versalovic

Published: June 13, 2012 • https://doi.org/10.1371/journal.pone.0036466

Machine Learning Techniques Accurately Classify Microbial Communities by Bacterial Vaginosis Characteristics

Daniel Beck, James A. Foster

Published: February 3, 2014 • https://doi.org/10.1371/journal.pone.0087830

Article | OPEN | Published: 10 October 2017

The gut microbiome in atherosclerotic cardiovascular disease

Zhuye Jie, Huihua Xia, […] Karsten Kristiansen

Nature Communications **8**, Article number: 845 (2017) | Download Citation

Article

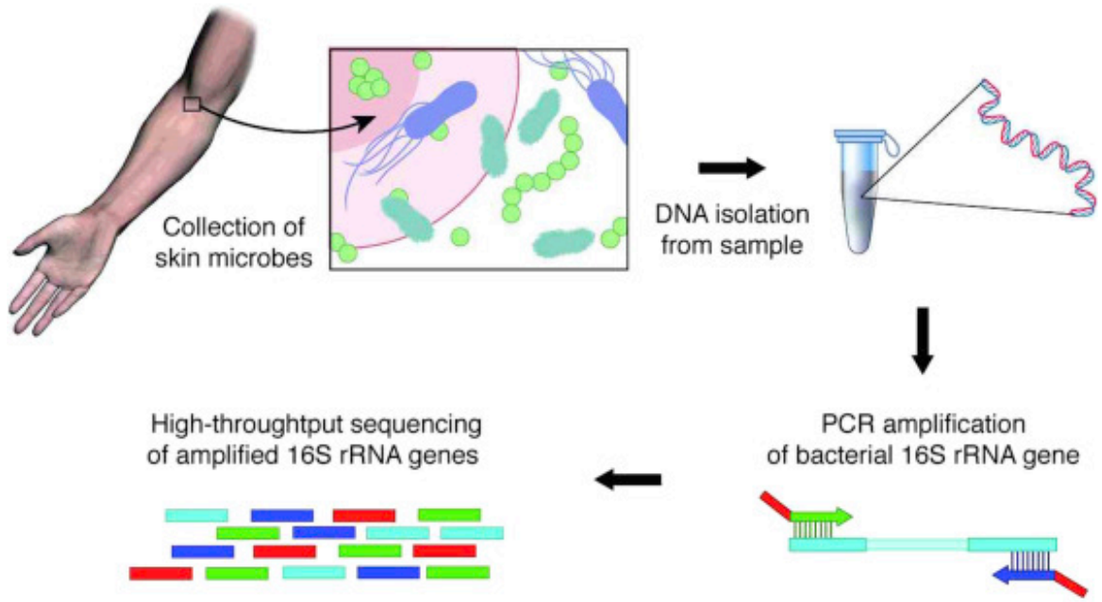
Prediction of Early Childhood Caries via Spatial-Temporal Variations of Oral Microbiota

Fei Teng [1, 2, 5], Fang Yang [3, 5], Shi Huang [2, 5], Cunpei Bo [2], Zhenjiang Zech Xu [4], Amnon Amir [4], Rob Knight [4], Junqi Ling [1], Jian Xu [2]

Show more

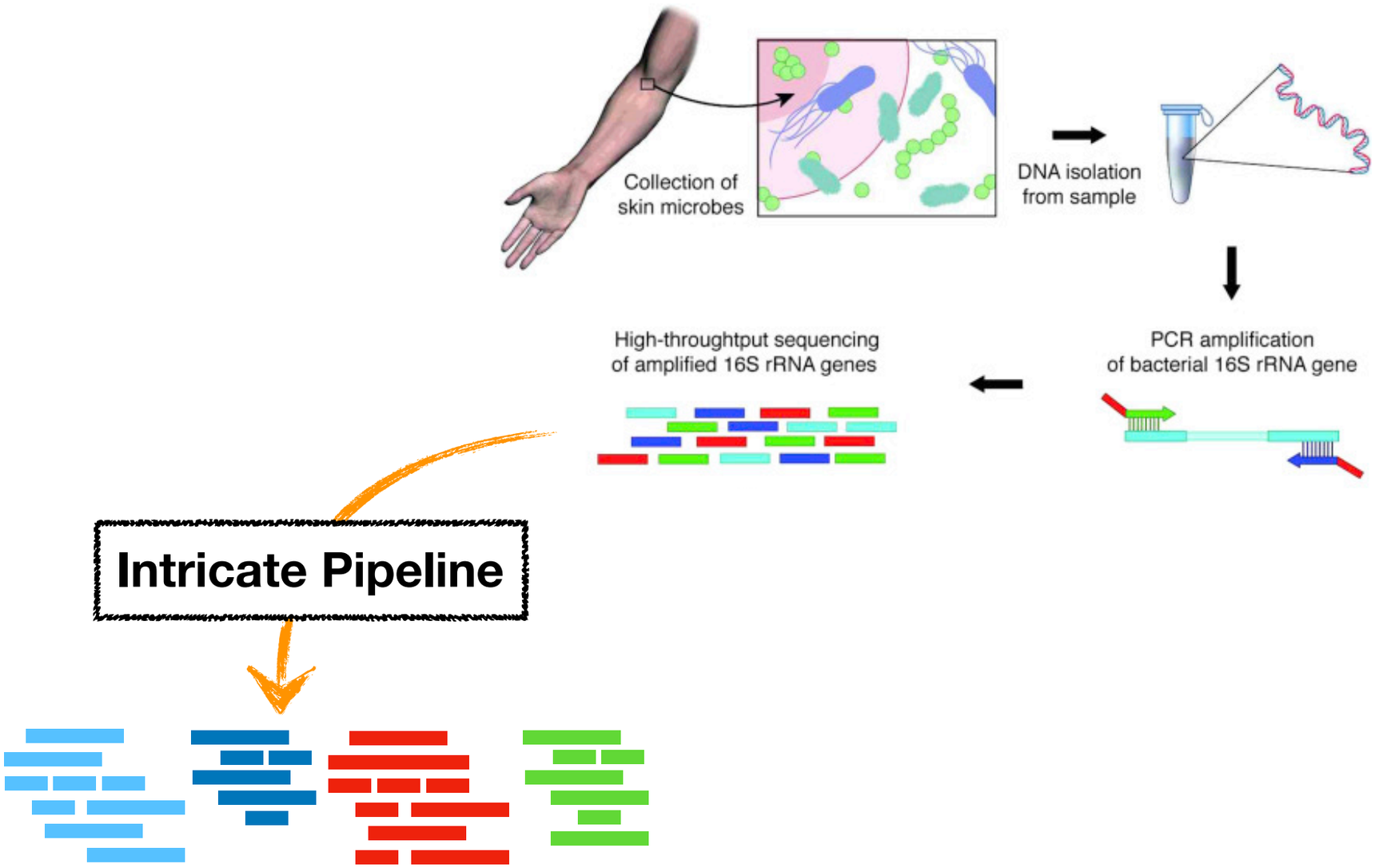
https://doi.org/10.1016/j.chom.2015.08.005

Sequencing microbiome (16S)



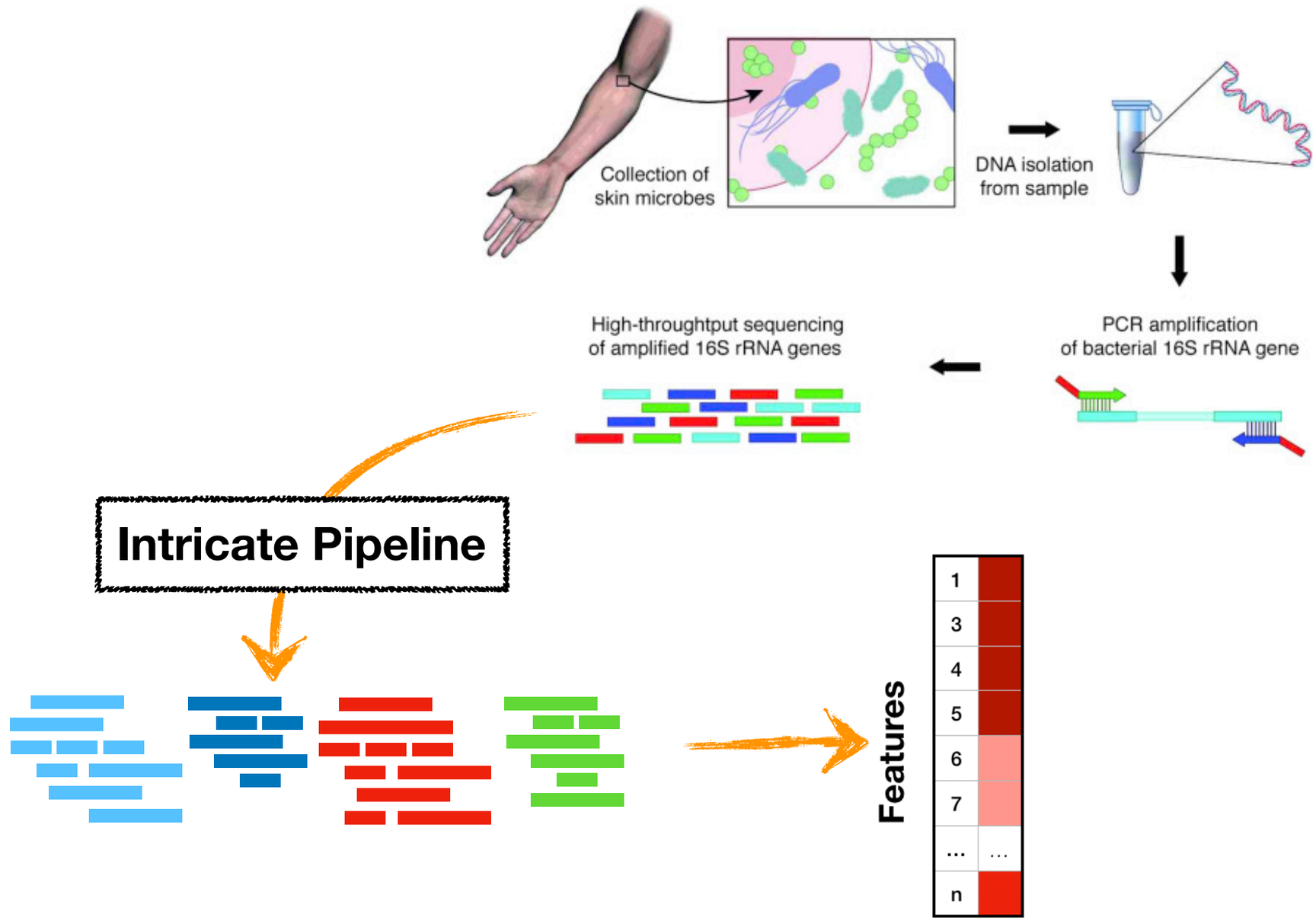
Jo et al., Journal of Investigative Dermatology, 2016

Sequencing microbiome (16S)



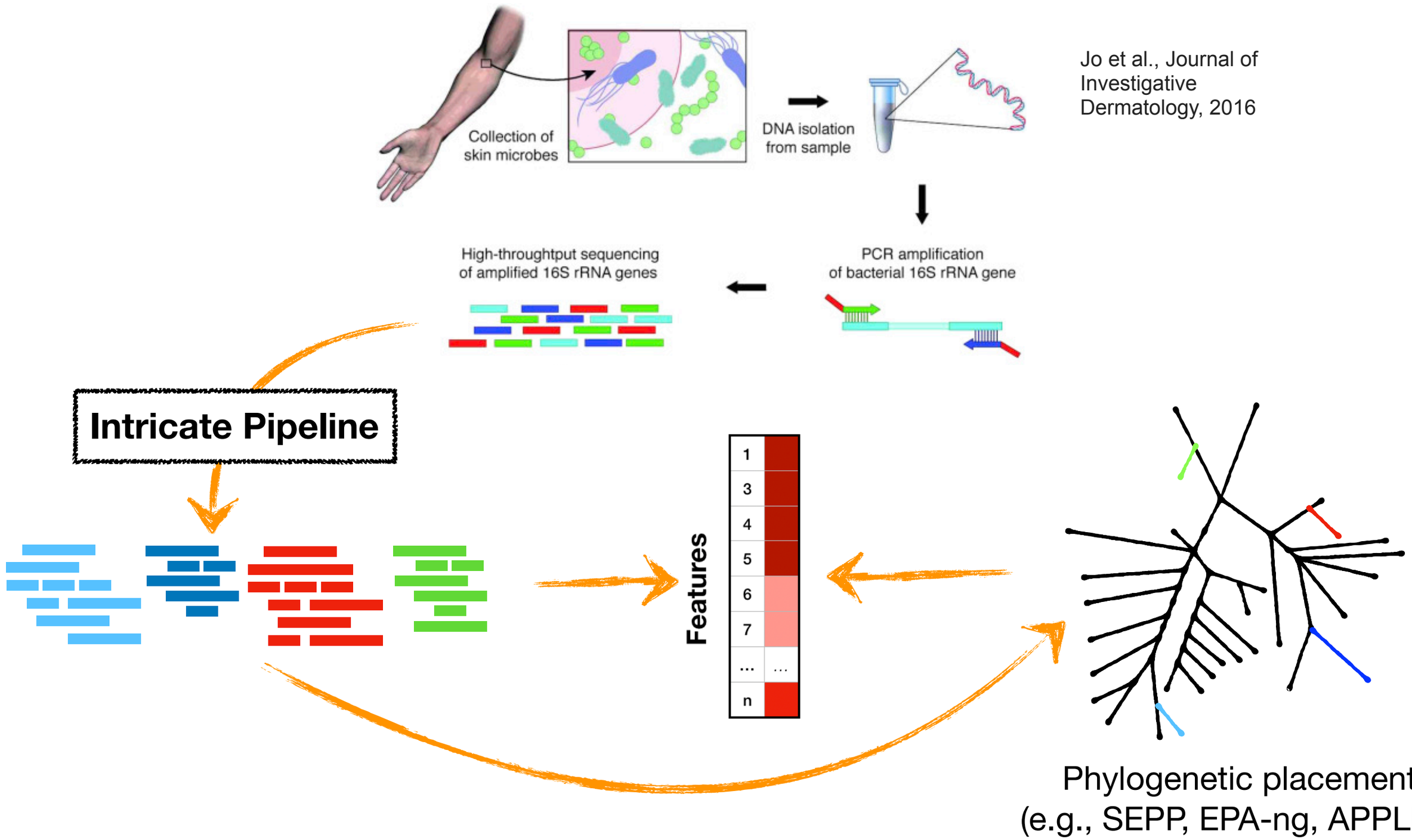
Jo et al., Journal of Investigative Dermatology, 2016

Sequencing microbiome (16S)



Jo et al., Journal of Investigative Dermatology, 2016

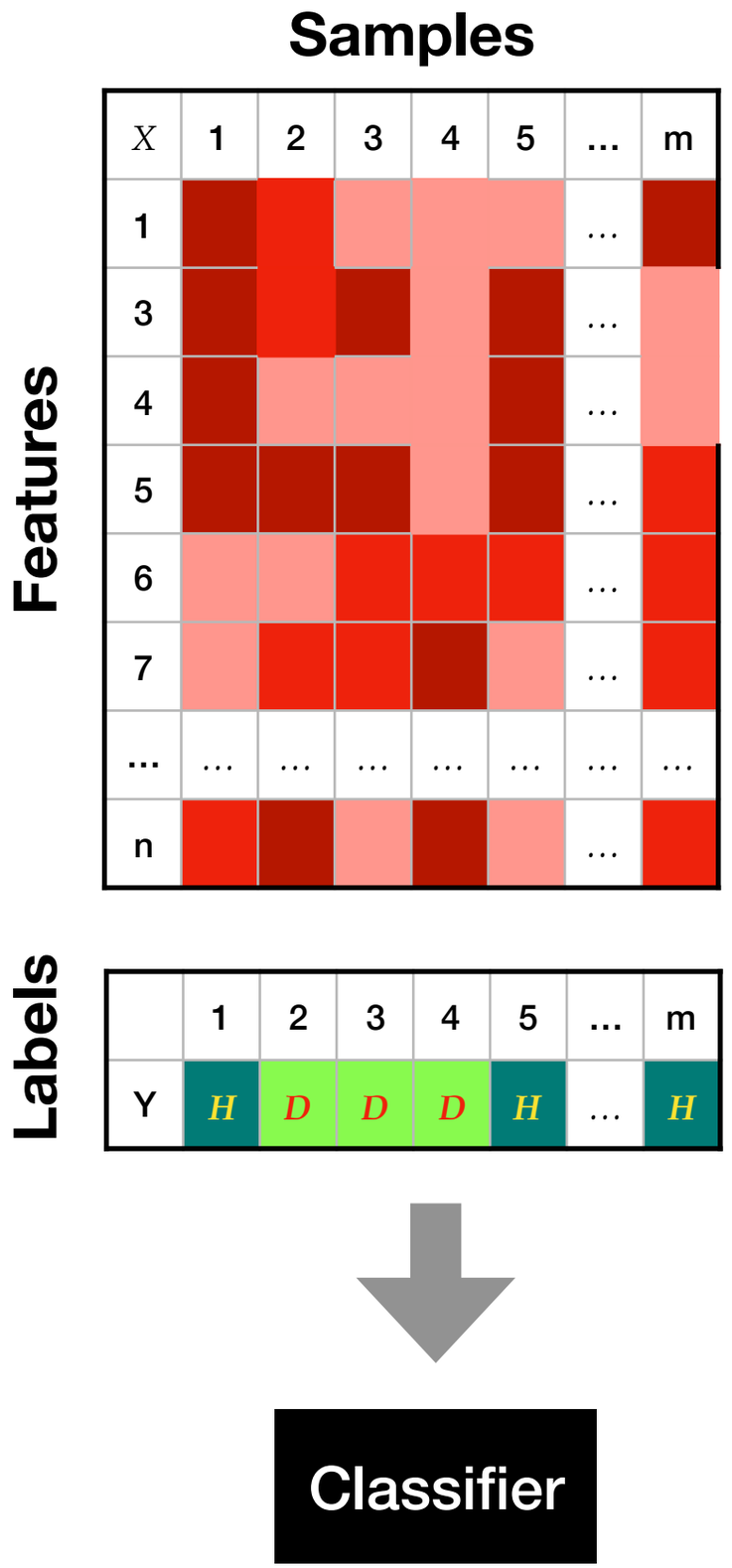
Sequencing microbiome (16S)



Jo et al., Journal of Investigative Dermatology, 2016

Supervised learning from microbiome data

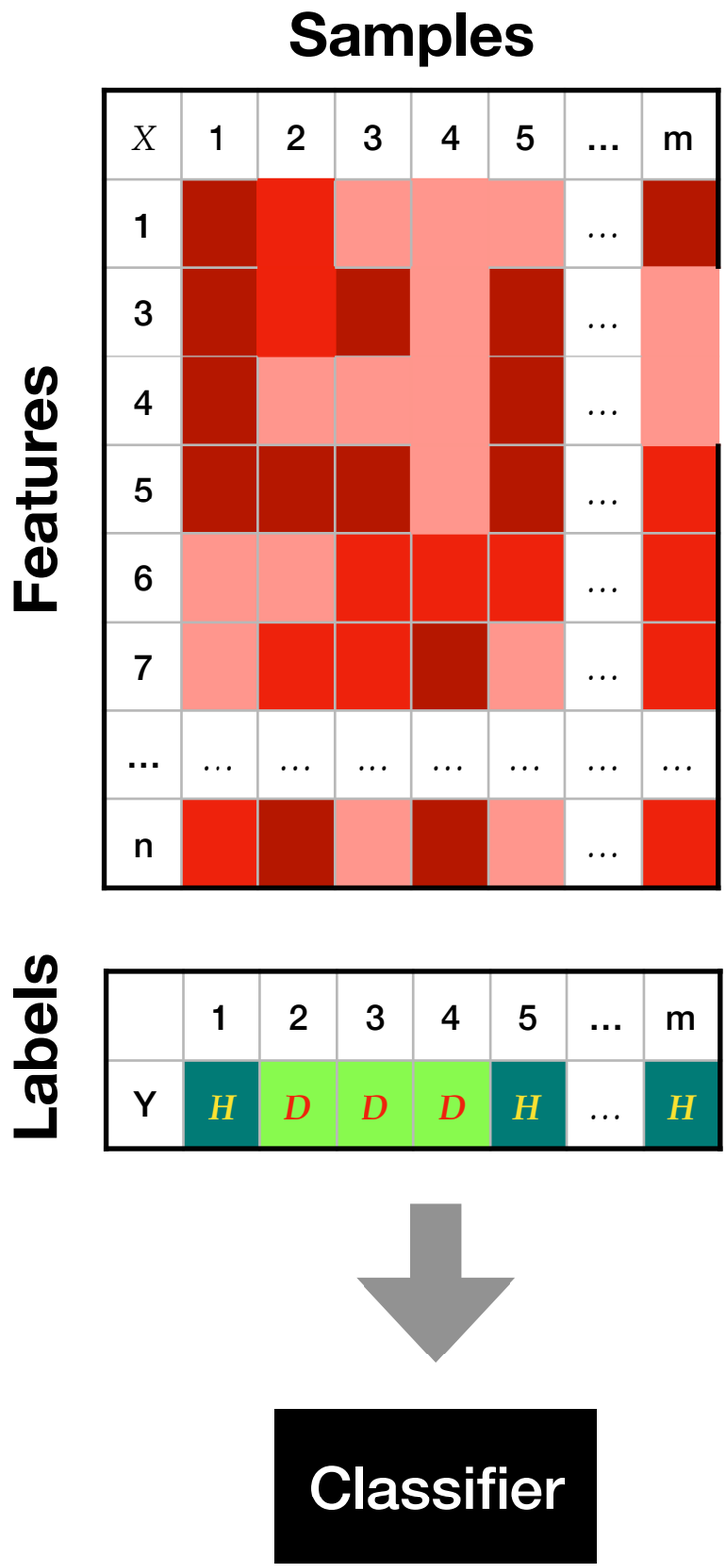
- **Goal**: train a classifier to predict the labels (**phenotype**) **from** the feature matrix (**genotype**)



Supervised learning from microbiome data

- **Goal**: train a classifier to predict the labels (**phenotype**) **from** the feature matrix (**genotype**)

- Random Forest has outperformed other methods, including Neural Networks
  [Statnikov et al., Microbiome, 2013]



Overfitting seems prevalent

Sze and Schloss, Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome, mBio, 2016

Dave et. al., The human gut microbiome: current Knowledge, challenges, and future directions, Translational Research, 2012

- Large number of features, but small number of samples
- High (biological) variation (confounders)
- Highly noisy and incomplete data

Overfitting seems prevalent

Sze and Schloss, Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome, mBio, 2016

Dave et. al., The human gut microbiome: current Knowledge, challenges, and future directions, Translational Research, 2012

- Large number of features, but small number of samples
- High (biological) variation (confounders)
- Highly noisy and incomplete data
- **Unbalanced and biased representation** of labels
  — Example: control group often is underrepresented

Benchmarking dataset

- Inflammatory Bowel Disease (IBD) study [Gevers *et al., 2014*]
- 1,359 samples, reduced after filtering to
  - 647 diseased and 243 healthy samples
- Around 10,000 features
  - Roughly: denoised unique sequences (defined using deblur)

[Amir et al., 2017]

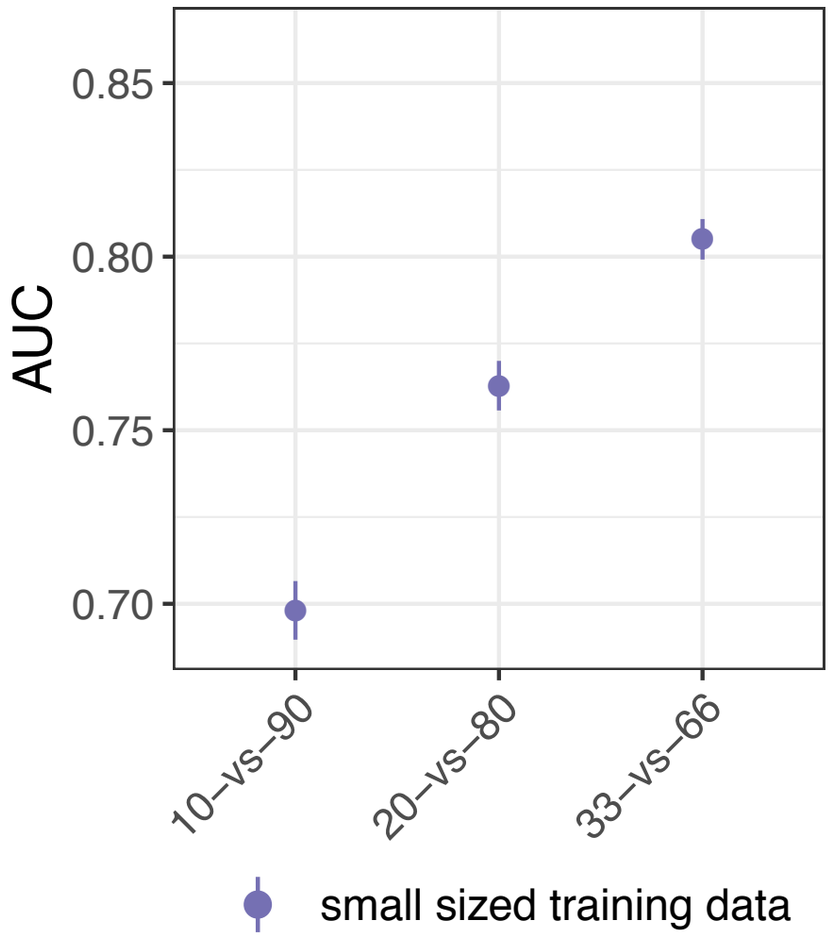
Cell Host & Microbe
Resource

The Treatment-Naive Microbiome in New-Onset Crohn's Disease

Dirk Gevers,[1] Subra Kugathasan,[4,24] Lee A. Denson,[5,24] Yoshiki Vázquez-Baeza,[6] Will Van Treuren,[7] Boyu Ren,[8] Emma Schwager,[8] Dan Knights,[9,10] Se Jin Song,[7] Moran Yassour,[1] Xochitl C. Morgan,[8] Aleksandar D. Kostic,[1] Chengwei Luo,[1] Antonio González,[7] Daniel McDonald,[7] Yael Haberman,[5] Thomas Walters,[11] Susan Baker,[12] Joel Rosh,[13] Michael Stephens,[14] Melvin Heyman,[15] James Markowitz,[16] Robert Baldassano,[17] Anne Griffiths,[18] Francisco Sylvester,[19] David Mack,[20] Sandra Kim,[21] Wallace Crandall,[21] Jeffrey Hyams,[19] Curtis Huttenhower,[1,8] Rob Knight,[7,22,23] and Ramnik J. Xavier[1,2,3,*]

Are **unbalanced classes** problematic?

- Fixed the size of the training dataset (243 samples in total)
- Control the proportion of healthy-vs-diseased



Data augmentation

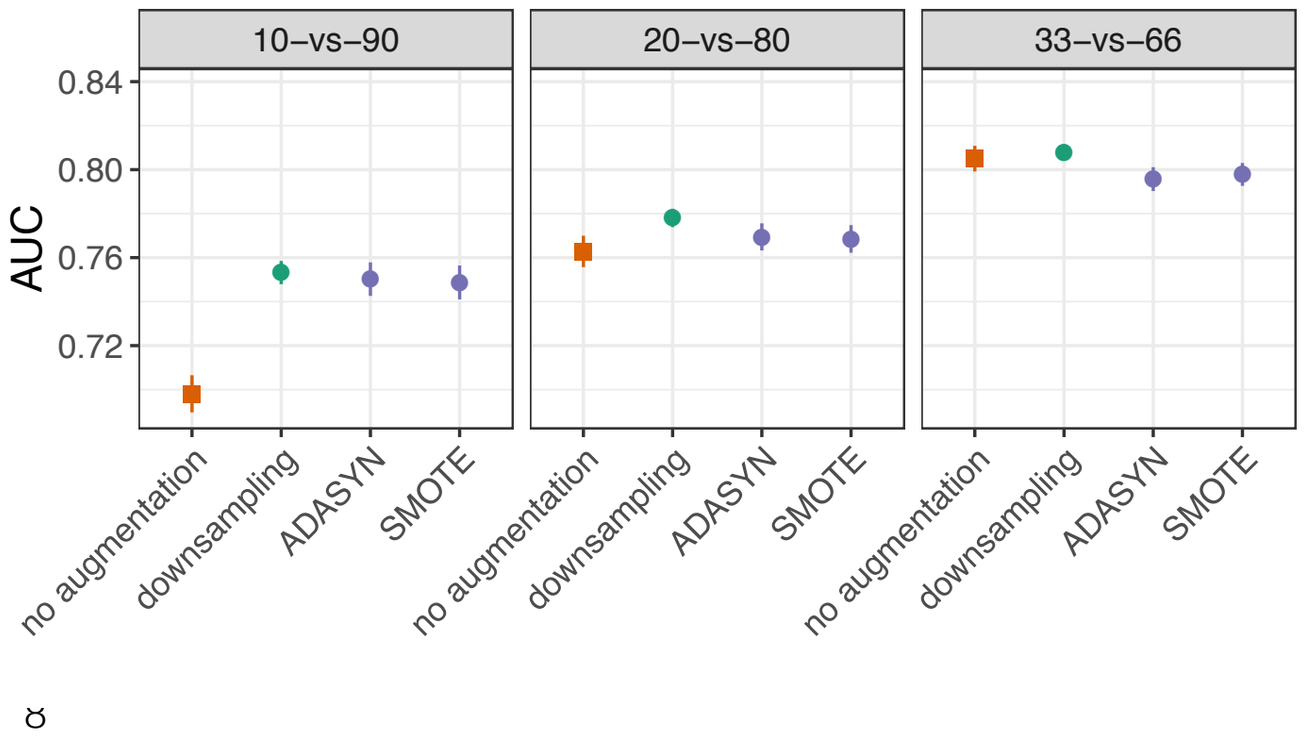
- Downsample the overrepresented class

Data augmentation

- Downsample the overrepresented class
- Upsample: generate synthetic data to augment the underrepresented class until classes have similar sizes
  - Synthetic data should resemble real data (could have been but are not seen)
  - SMOTE [Chawla *et al*., 2002]
    ADASYN [He *et al*.,2008]
    (kNN+linear combination)

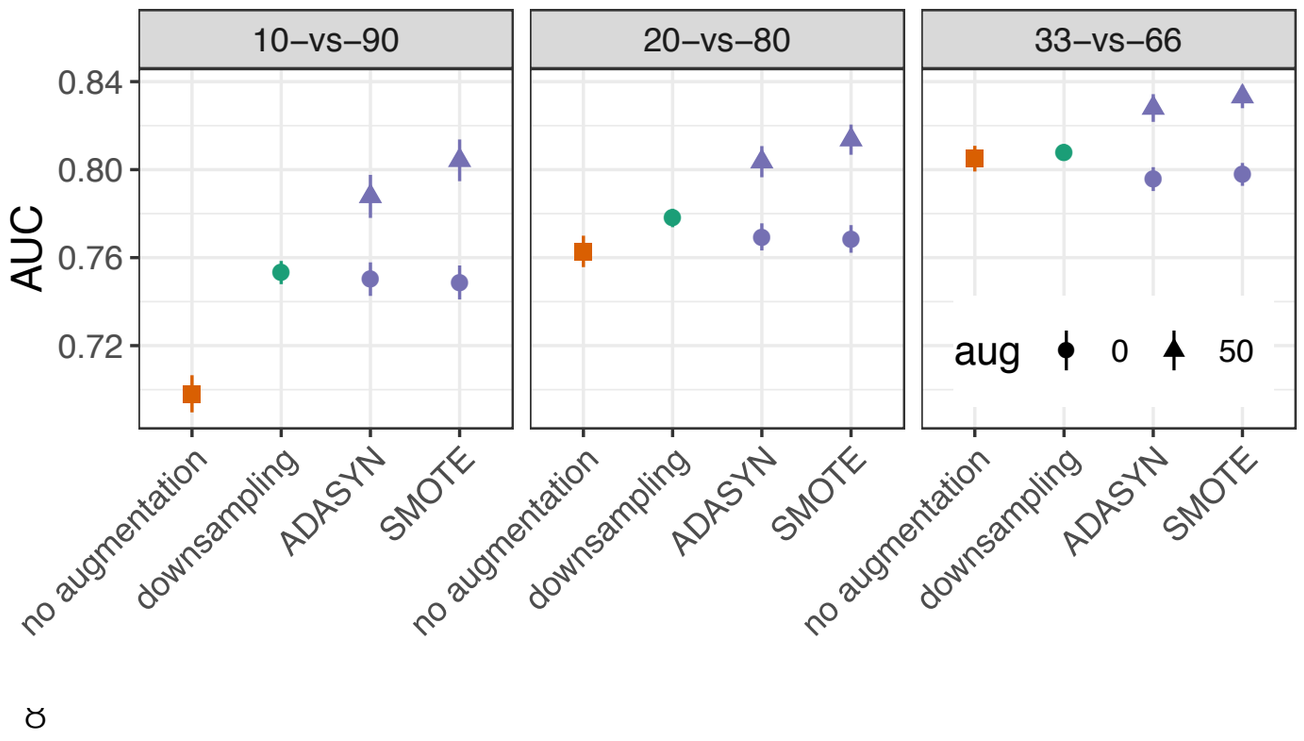
Data augmentation

- Downsample the overrepresented class
- Upsample: generate synthetic data to augment the underrepresented class until classes have similar sizes
- Synthetic data should resemble real data (could have been but are not seen)
- SMOTE [Chawla *et al.*, 2002]
  ADASYN [He *et al.*,2008]
  (kNN+linear combination)



Data augmentation

- Downsample the overrepresented class
- Upsample: generate synthetic data to augment the underrepresented class until classes have similar sizes
  - Synthetic data should resemble real data (could have been but are not seen)
  - SMOTE [Chawla *et al.*, 2002]
    ADASYN [He *et al.*,2008]
    (kNN+linear combination)
- Keep augmenting (50X) *after* balancing labels (▲)



Combining domain knowledge with data augmentation

- Microbes are related through a phylogeny

Combining domain knowledge with data augmentation

- Microbes are related through a phylogeny

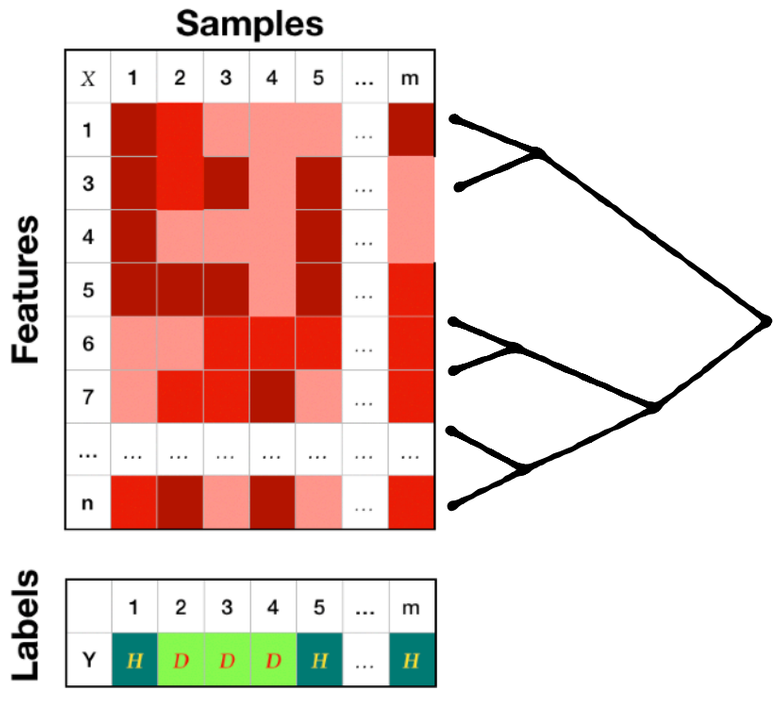
- **Hypothesis:** Variations in the feature vectors across samples are governed by the phylogeny

Combining domain knowledge with data augmentation

- Microbes are related through a phylogeny

- **Hypothesis:** Variations in the feature vectors across samples are governed by the phylogeny

- Thus, data augmentation should take the phylogeny into account



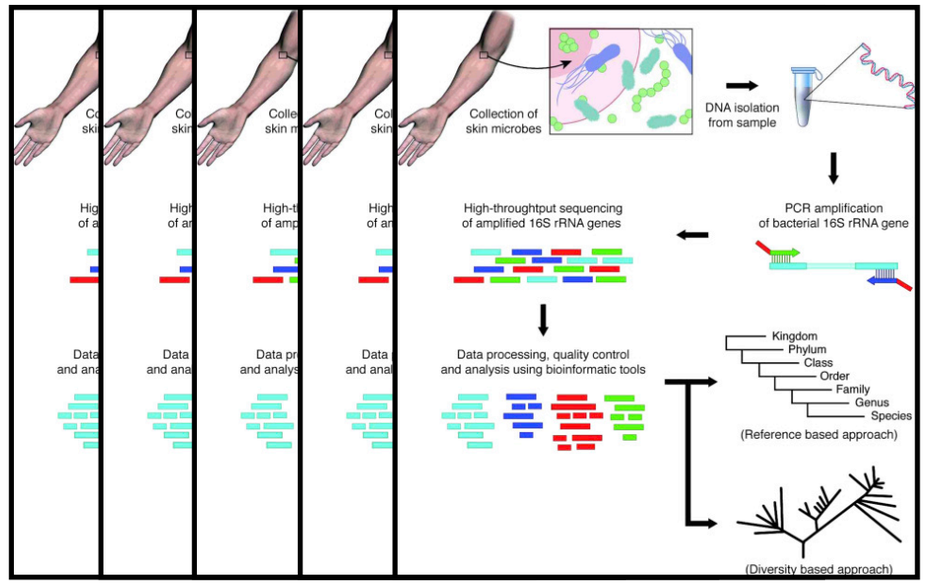
TADA: Tree-based Associative Data Augmentation

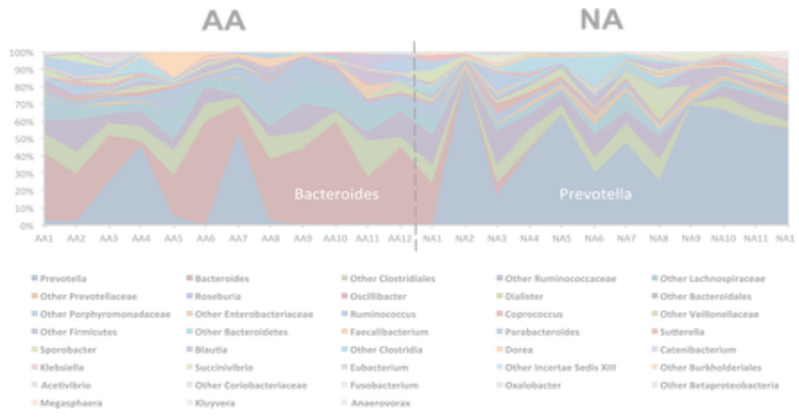
- Design a phylogeny-aware generative model with parameters that can be learnt from (subsets of) training data

- Use the generative model to generate synthetic samples and add them to the training data

Sources of variation

- **Sampling (sequencing) variation**
- Biological variation
  - Confounding factors: ethnicity, age, gender, diet, lifestyle
- Temporal variations
- Not fully understood

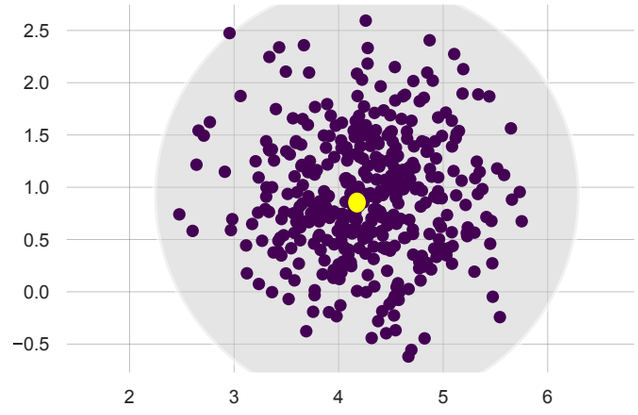




Ou et al, Am J Clin Nutr 98:111, 2013

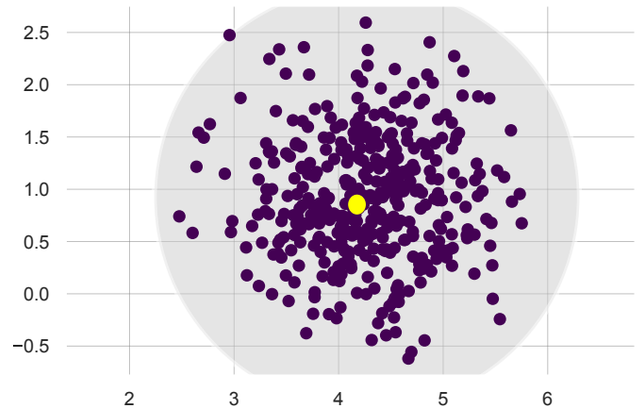
Sampling variation (SV)

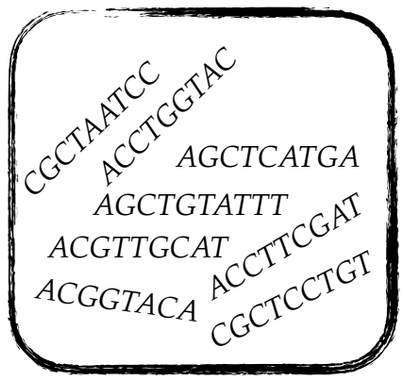
- If another round of sequencing is performed, what could we observe?



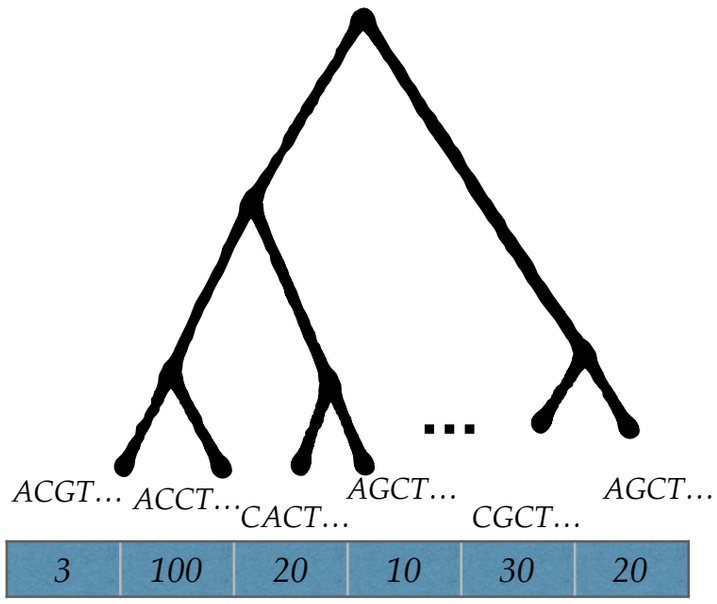
Sampling variation (SV)

- If another round of sequencing is performed, what could we observe?



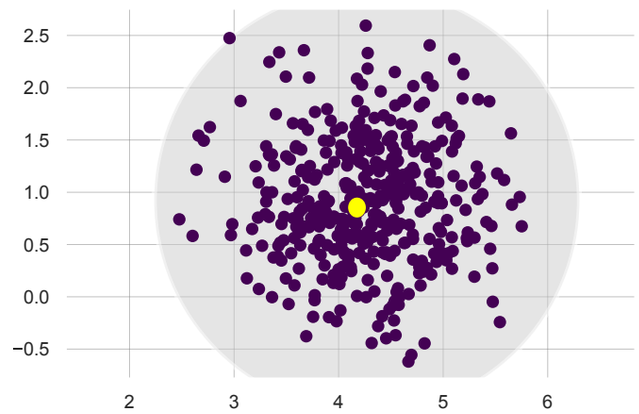


ACGT…	3
ACCT…	100
CACT…	20
AGCT…	10
…	…
CGCT…	30
AGCT…	20

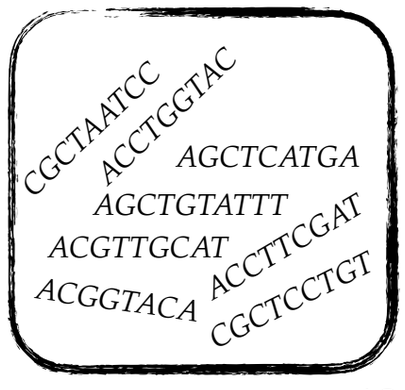


Sampling variation (SV)

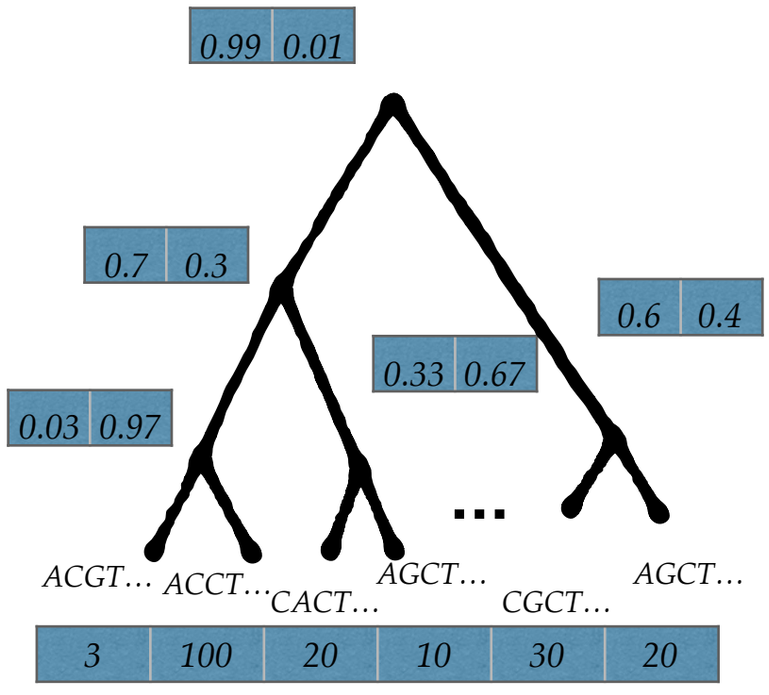
- If another round of sequencing is performed, what could we observe?



- Map normalized abundances to nodes

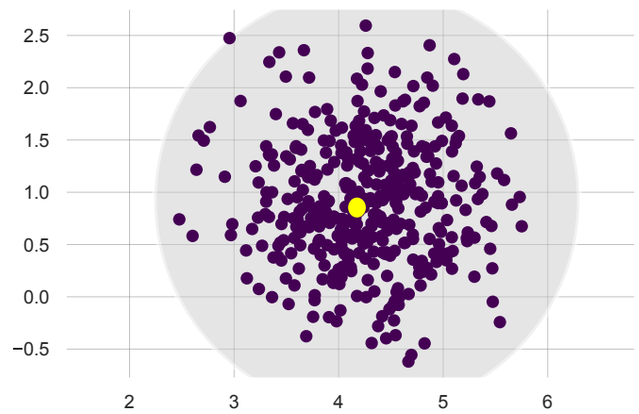


ACGT…	3
ACCT…	100
CACT…	20
AGCT…	10
…	…
CGCT…	30
AGCT…	20



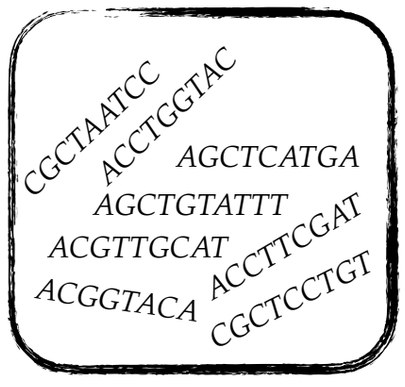
Sampling variation (SV)

- If another round of sequencing is performed, what could we observe?

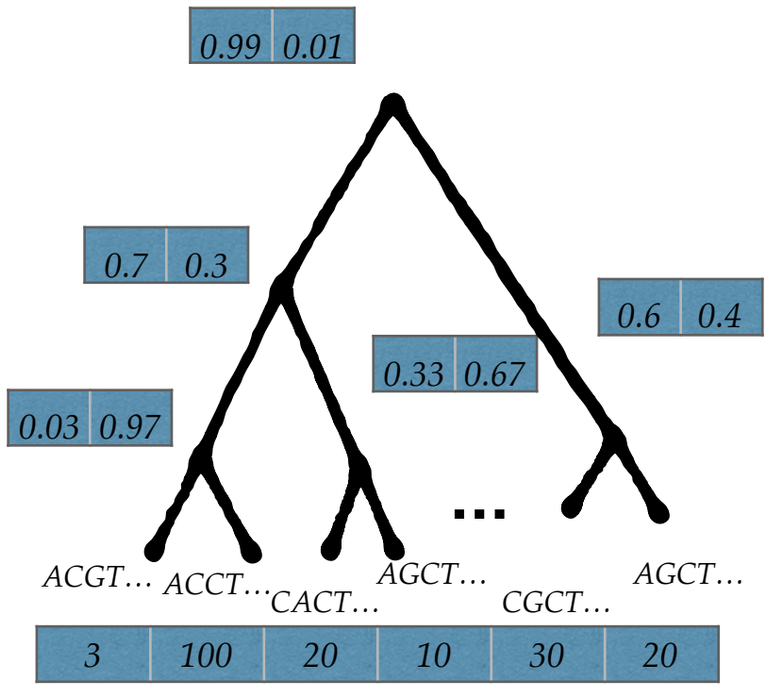


- Map normalized abundances to nodes

- Gives a hierarchy of binomial distributions

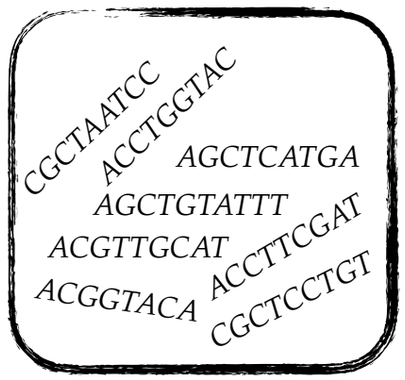


ACGT…	3
ACCT…	100
CACT…	20
AGCT…	10
…	…
CGCT…	30
AGCT…	20

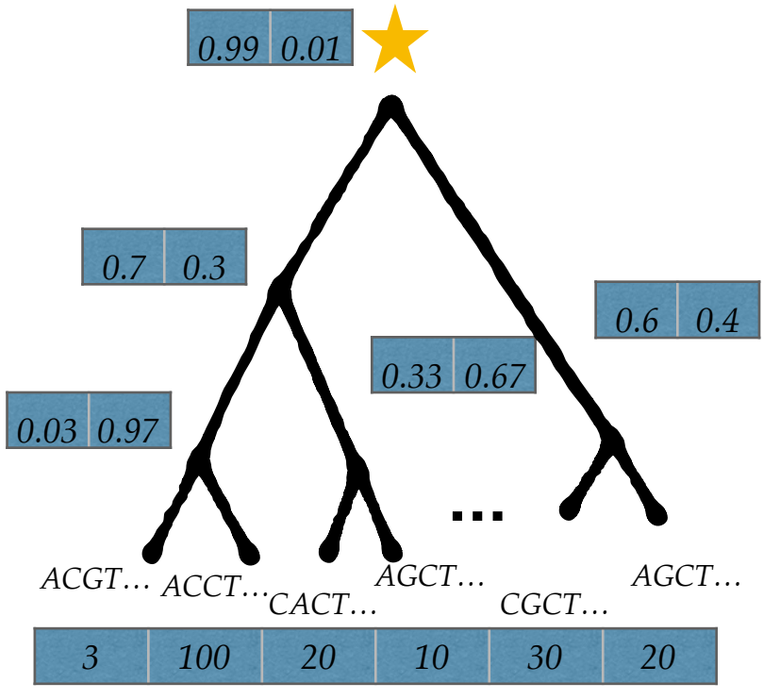


TADA-SV

Repeat:



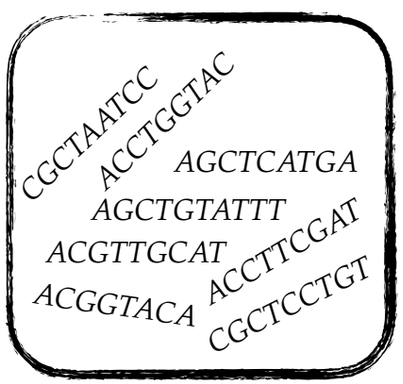
ACGT…	3
ACCT…	100
CACT…	20
AGCT…	10
…	…
CGCT…	30
AGCT…	20



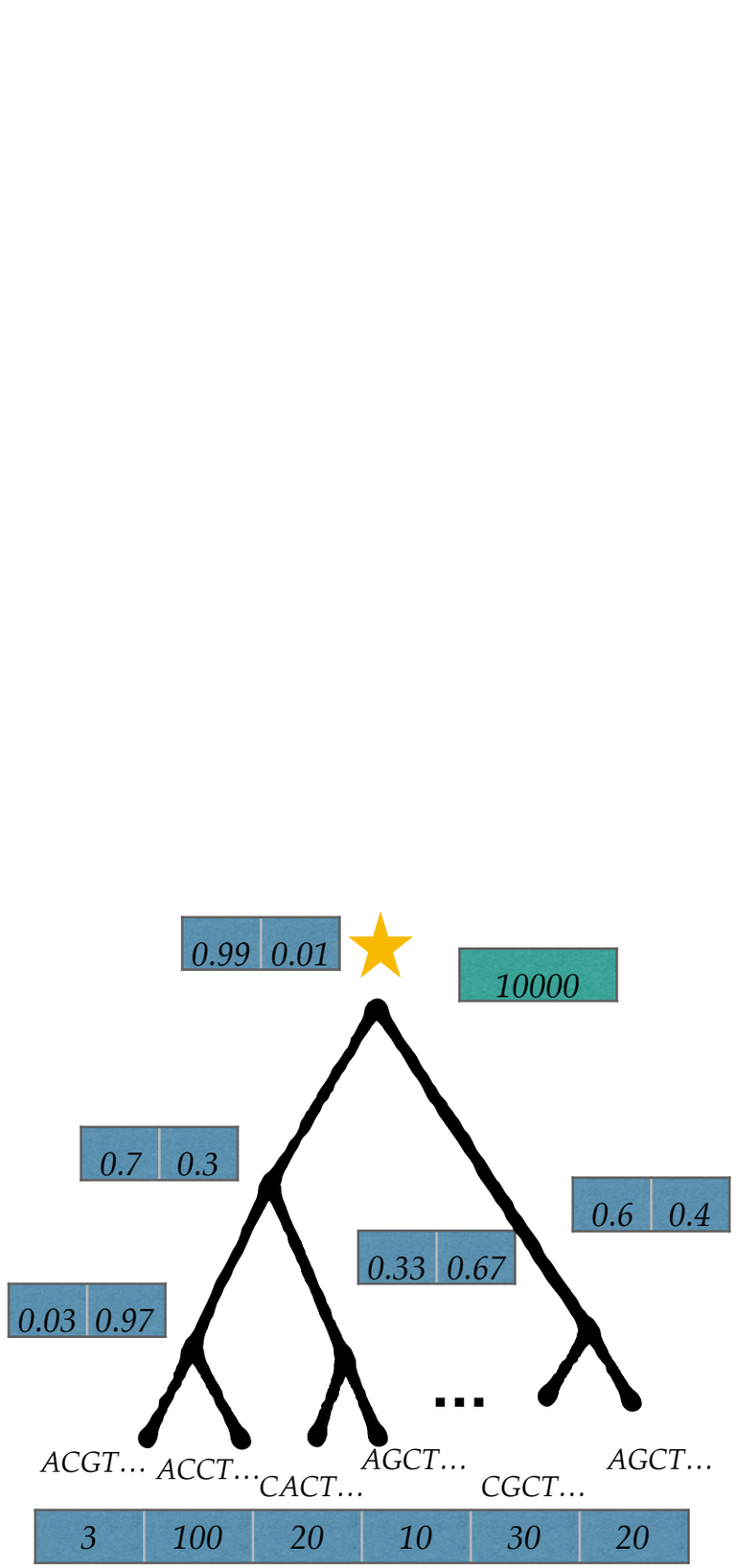
TADA-SV

Repeat:

- Start from the root
    - Assume some total counts over the root



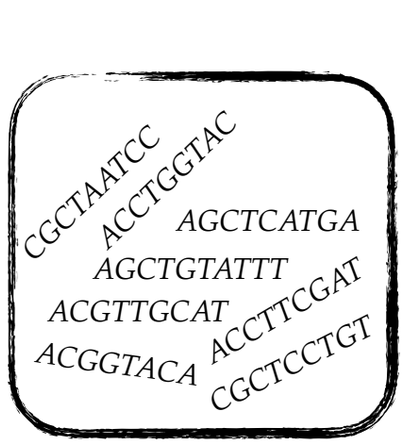
ACGT…	3
ACCT…	100
CACT…	20
AGCT…	10
…	…
CGCT…	30
AGCT…	20



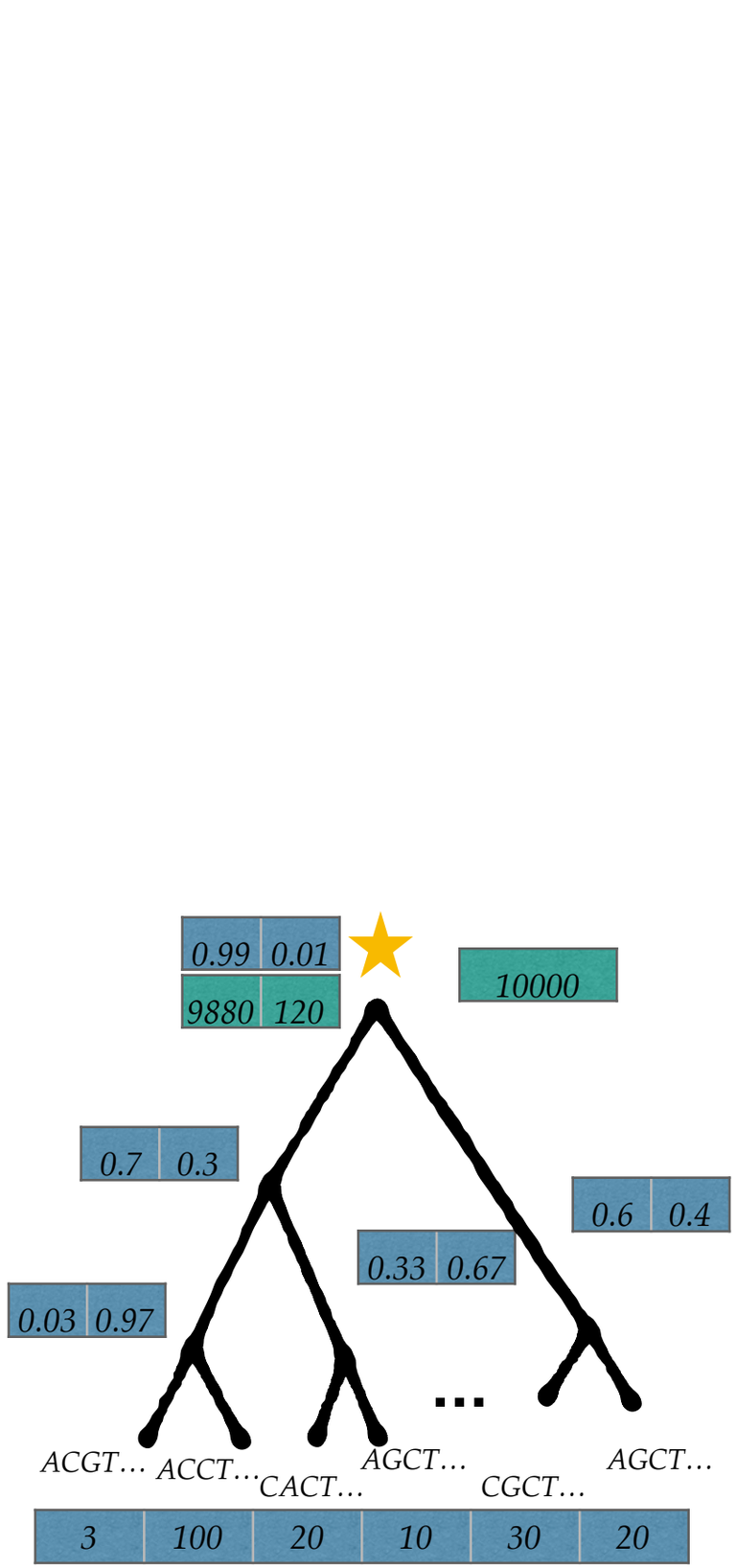
TADA-SV

Repeat:

- Start from the root
    - Assume some total counts over the root
- Draw a number from Binomial distribution



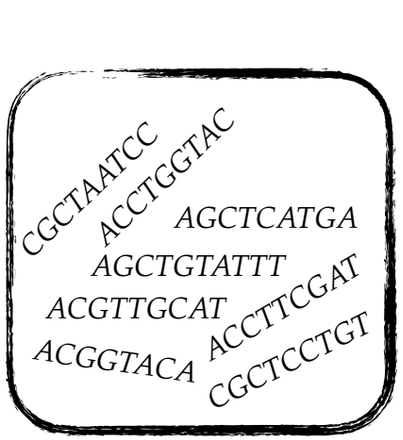
ACGT…	3
ACCT…	100
CACT…	20
AGCT…	10
…	…
CGCT…	30
AGCT…	20



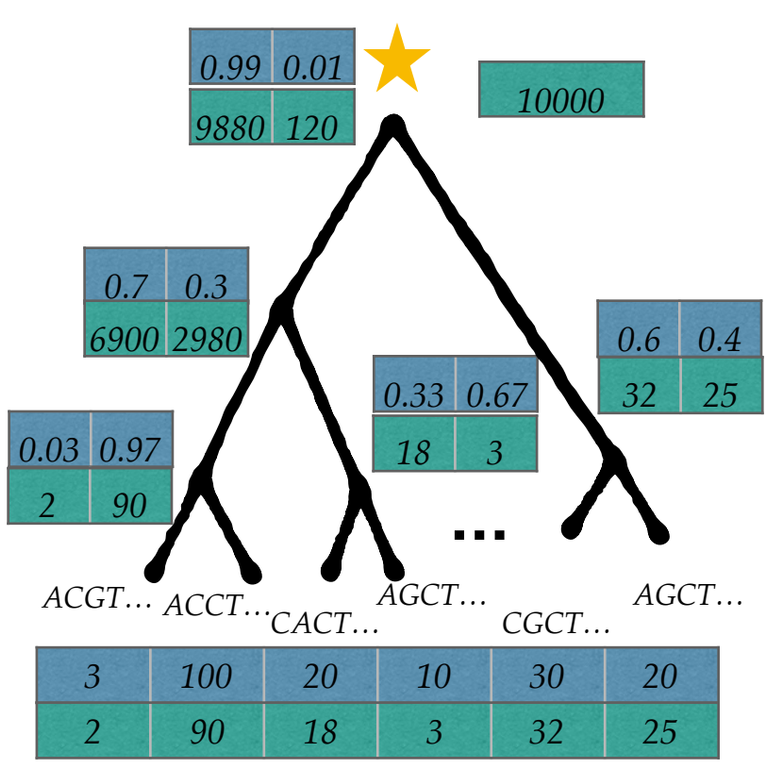
TADA-SV

Repeat:

- Start from the root
    - Assume some total counts over the root
- Draw a number from Binomial distribution
- Recurse on children



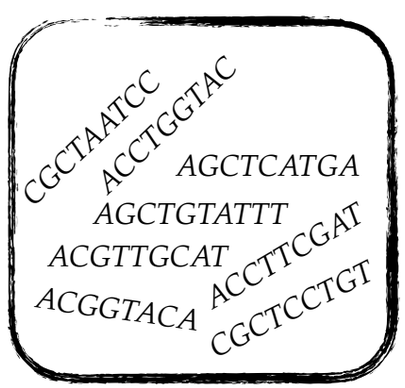
ACGT…	3	2
ACCT…	100	90
CACT…	20	18
AGCT…	10	3
…	…	…
CGCT…	30	32
AGCT…	20	25

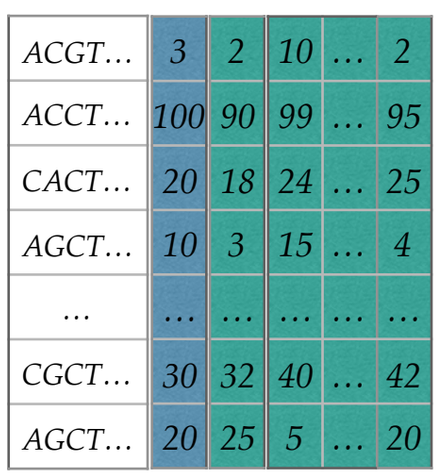


TADA-SV

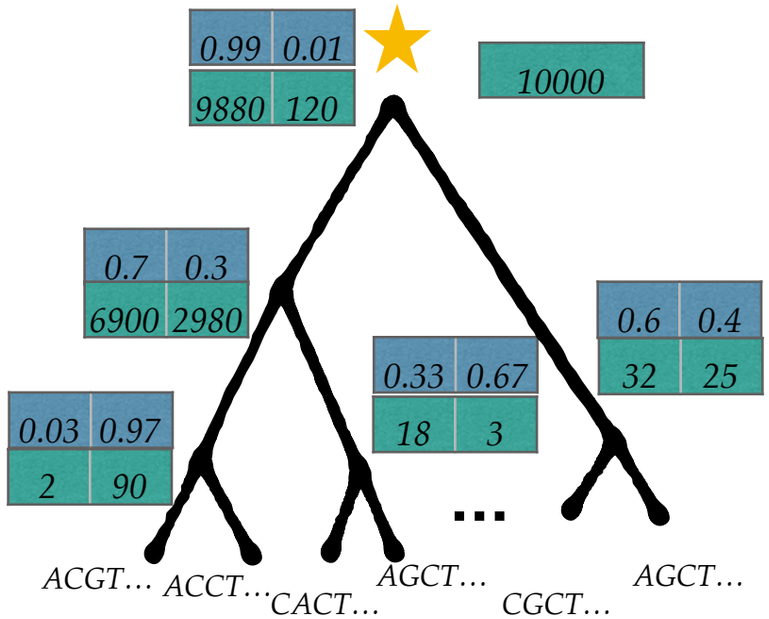
Repeat:

- Start from the root
    - Assume some total counts over the root
- Draw a number from Binomial distribution
- Recurse on children





ACGT…	3	2	10	…	2
ACCT…	100	90	99	…	95
CACT…	20	18	24	…	25
AGCT…	10	3	15	…	4
…	…	…	…	…	…
CGCT…	30	32	40	…	42
AGCT…	20	25	5	…	20

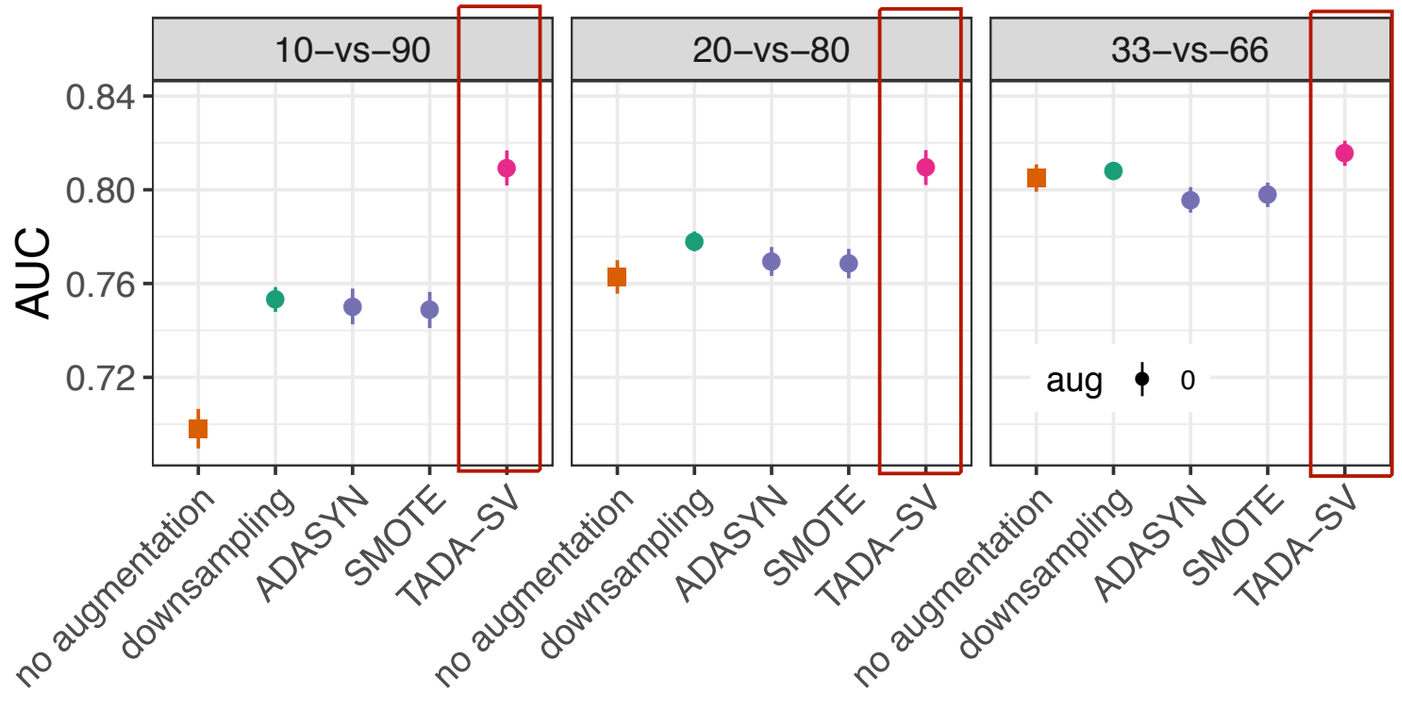


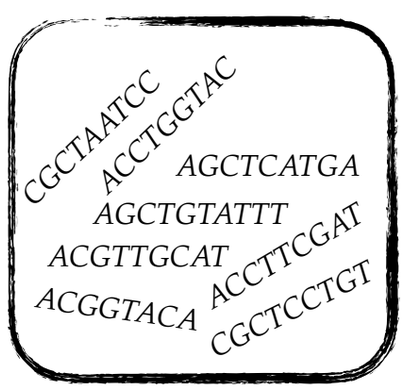
3	100	20	10	30	20
2	90	18	3	32	25
10	99	24	15	40	5
…	…	…	…	…	…
2	95	25	4	42	25

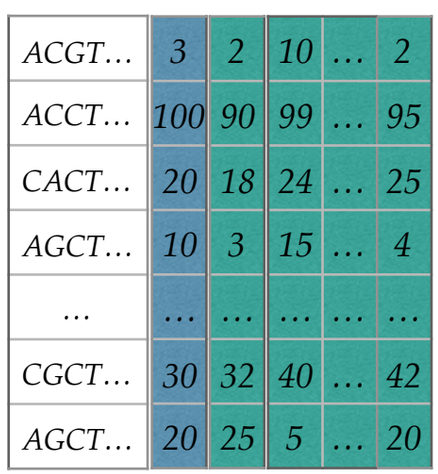
TADA-SV

Repeat:

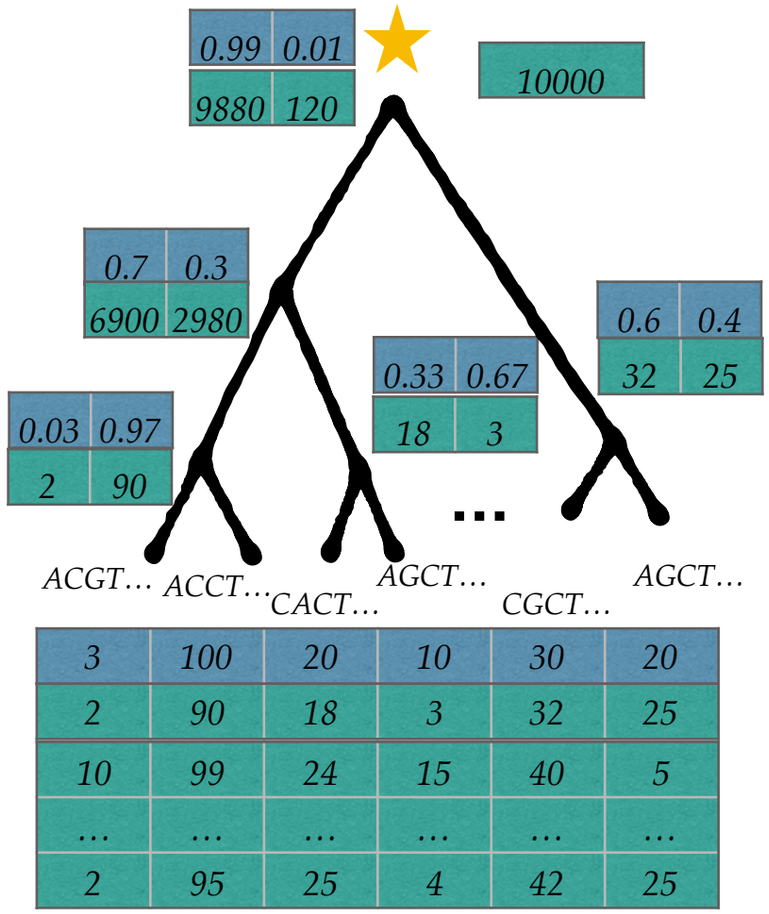
- Start from the root
  - Assume some total counts over the root
- Draw a number from Binomial distribution
- Recurse on children







ACGT…	3	2	10	…	2
ACCT…	100	90	99	…	95
CACT…	20	18	24	…	25
AGCT…	10	3	15	…	4
…	…	…	…	…	…
CGCT…	30	32	40	…	42
AGCT…	20	25	5	…	20

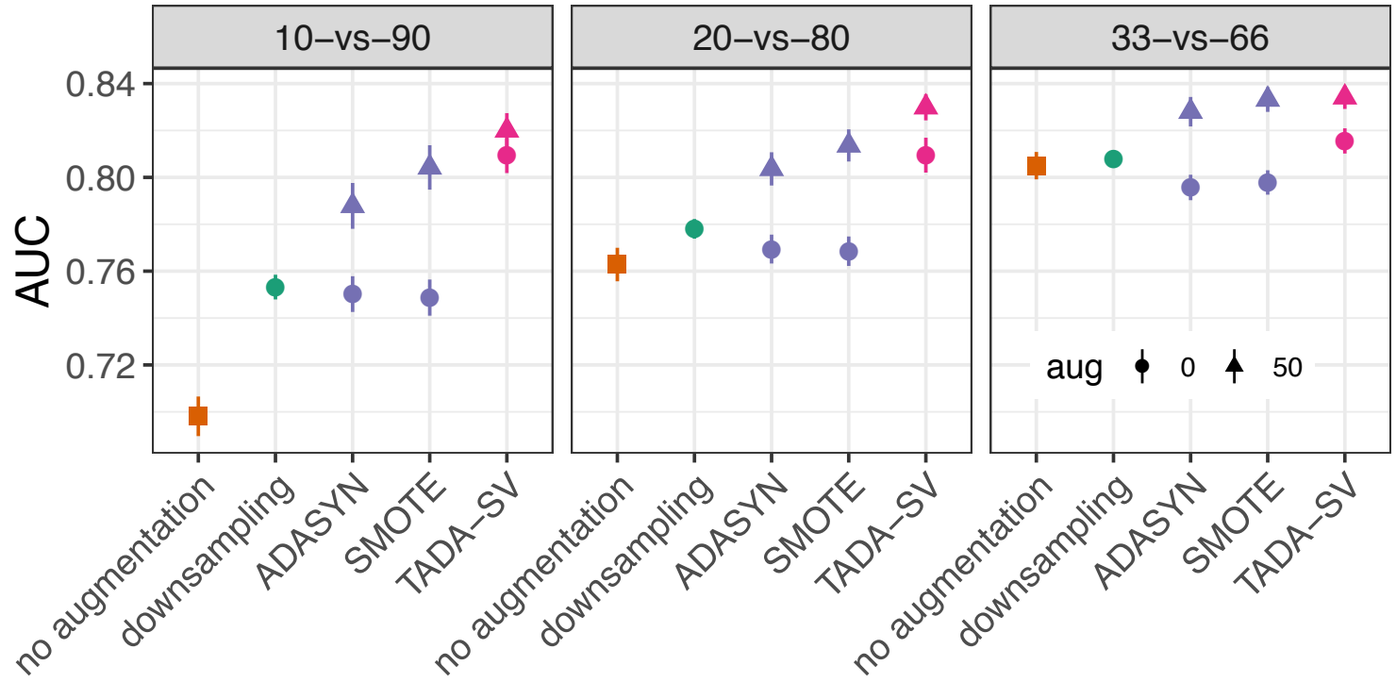


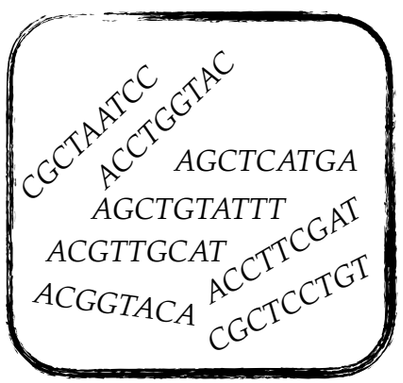
3	100	20	10	30	20
2	90	18	3	32	25
10	99	24	15	40	5
…	…	…	…	…	…
2	95	25	4	42	25

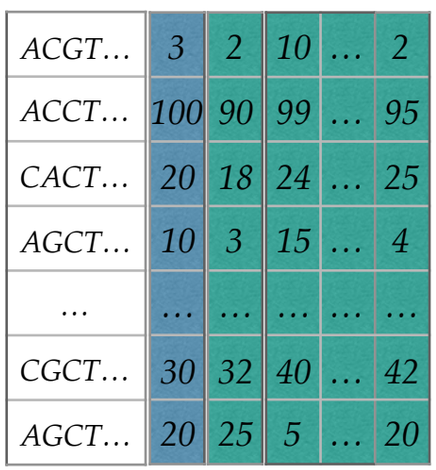
TADA-SV

Repeat:

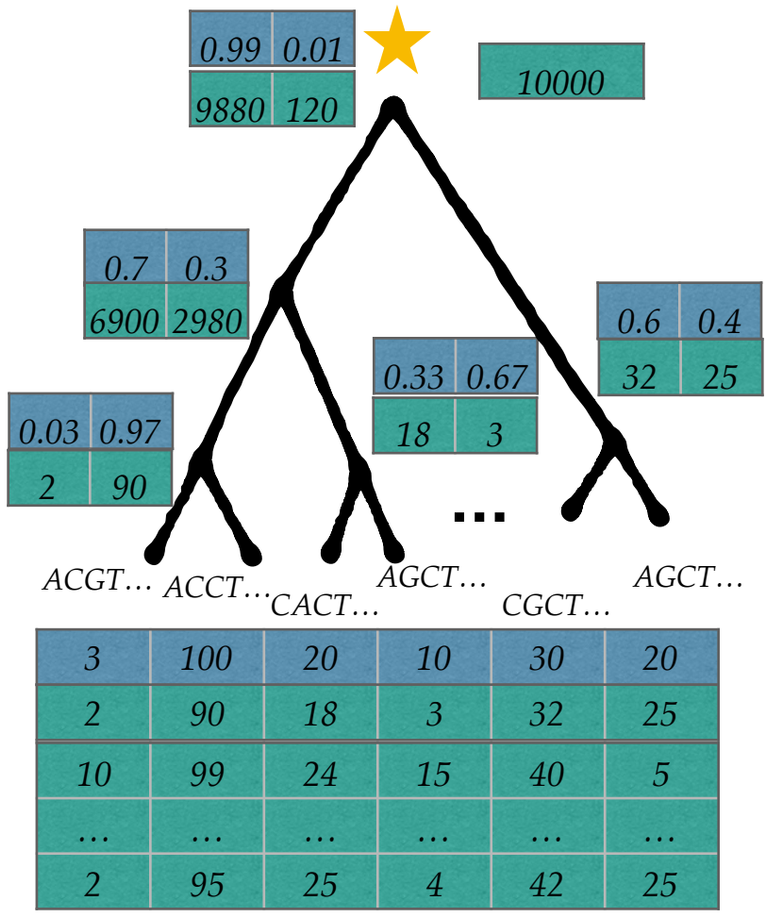
- Start from the root
  - Assume some total counts over the root
- Draw a number from Binomial distribution
- Recurse on children







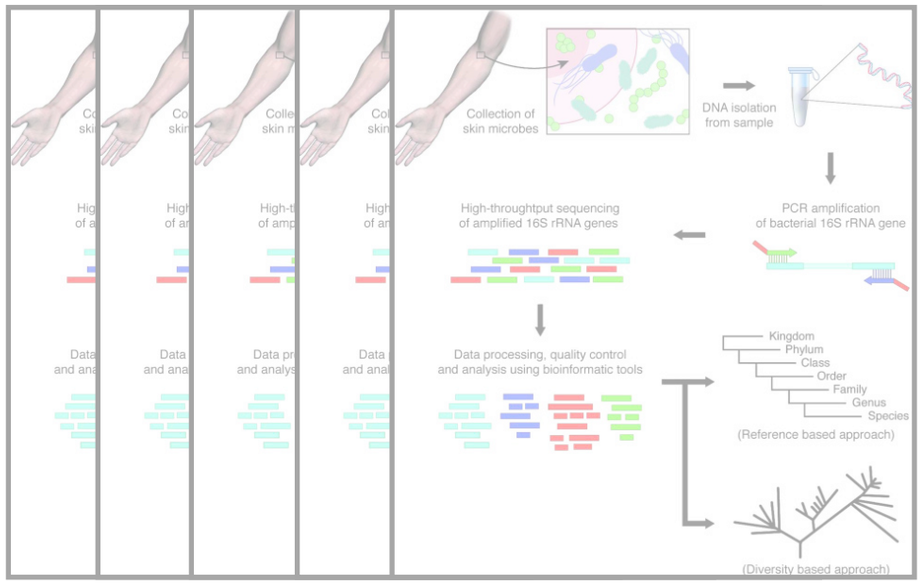
ACGT…	3	2	10	…	2
ACCT…	100	90	99	…	95
CACT…	20	18	24	…	25
AGCT…	10	3	15	…	4
…	…	…	…	…	…
CGCT…	30	32	40	…	42
AGCT…	20	25	5	…	20

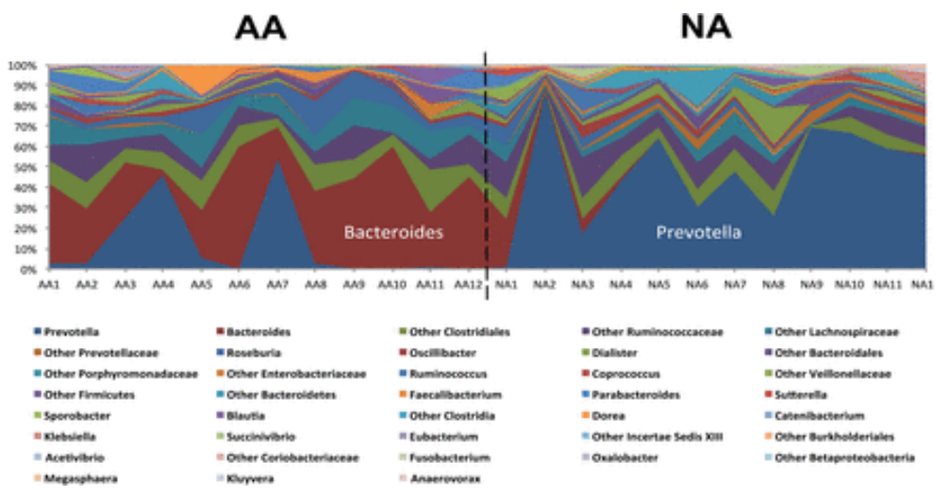


3	100	20	10	30	20
2	90	18	3	32	25
10	99	24	15	40	5
…	…	…	…	…	…
2	95	25	4	42	25

Sources of variation

- Sampling (sequencing) variation
- **Biological variation**
  - Confounding factors: ethnicity, age, gender, diet, lifestyle
  - Temporal variations
  - Not fully understood

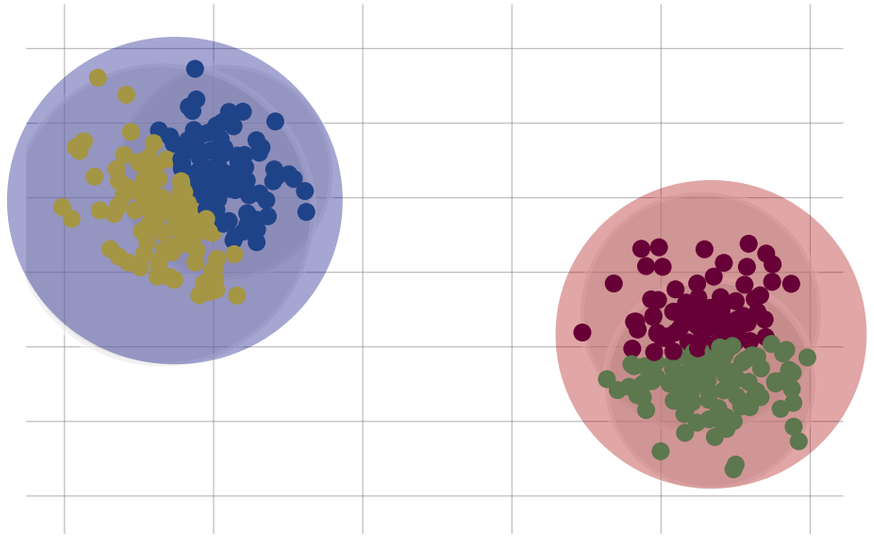


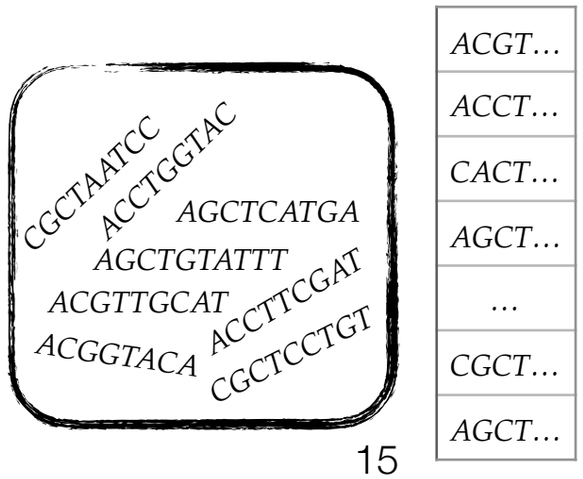


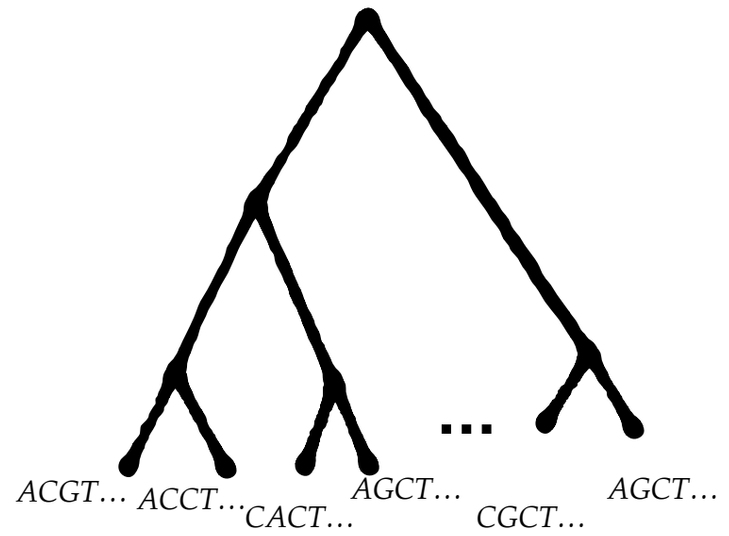
Ou et al, Am J Clin Nutr 98:111, 2013

Biological variation (BV)

- Considering variations between groups of samples

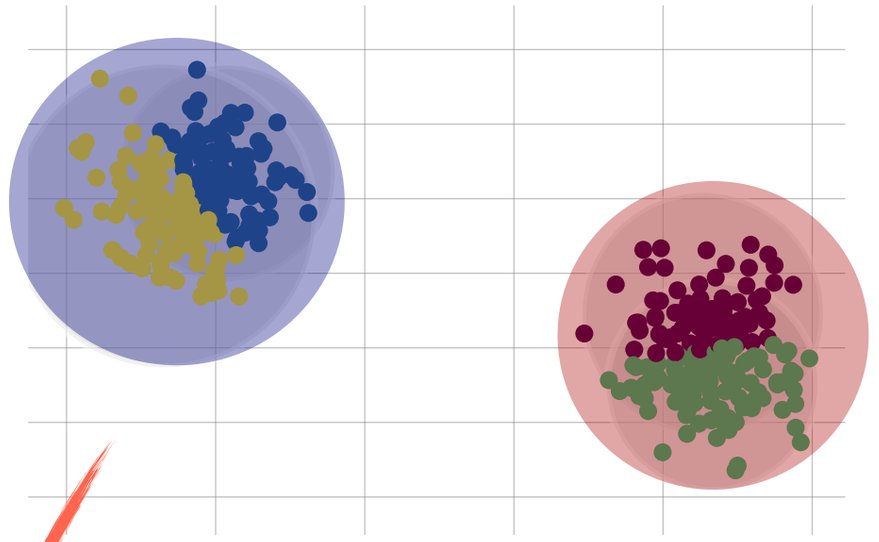


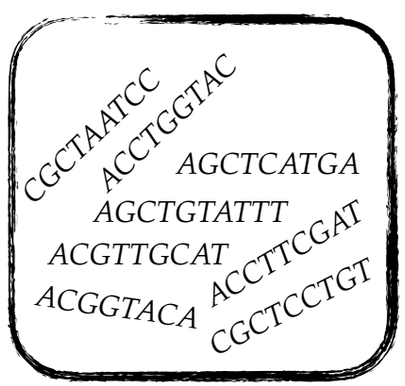




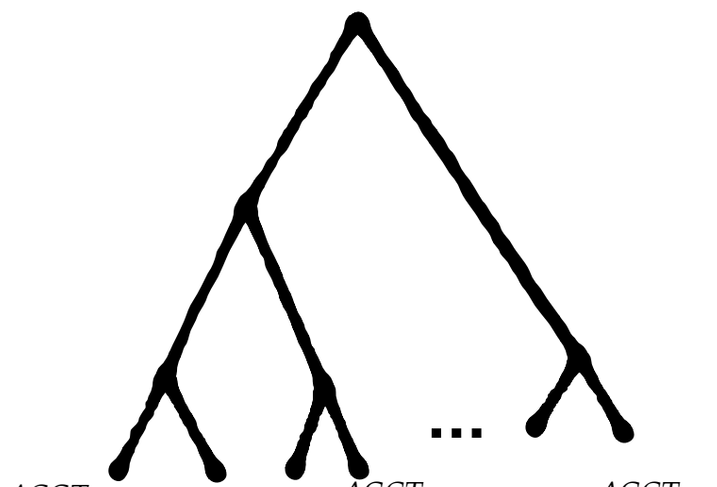
Biological variation (BV)

- Considering variations between groups of samples





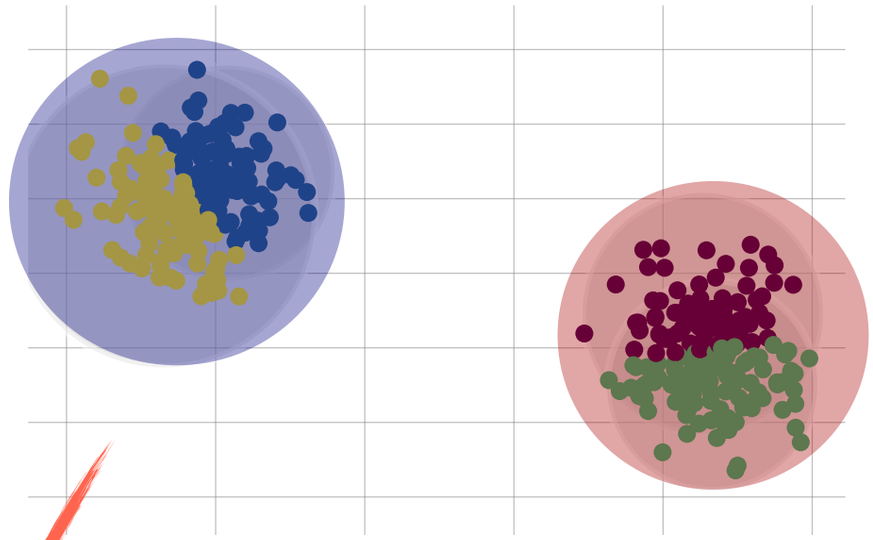
ACGT…	4	3	…	5
ACCT…	98	100	…	100
CACT…	23	20	…	25
AGCT…	23	10	…	1
…		…	…	…
CGCT…	20	30	…	10
AGCT…	19	20	…	20

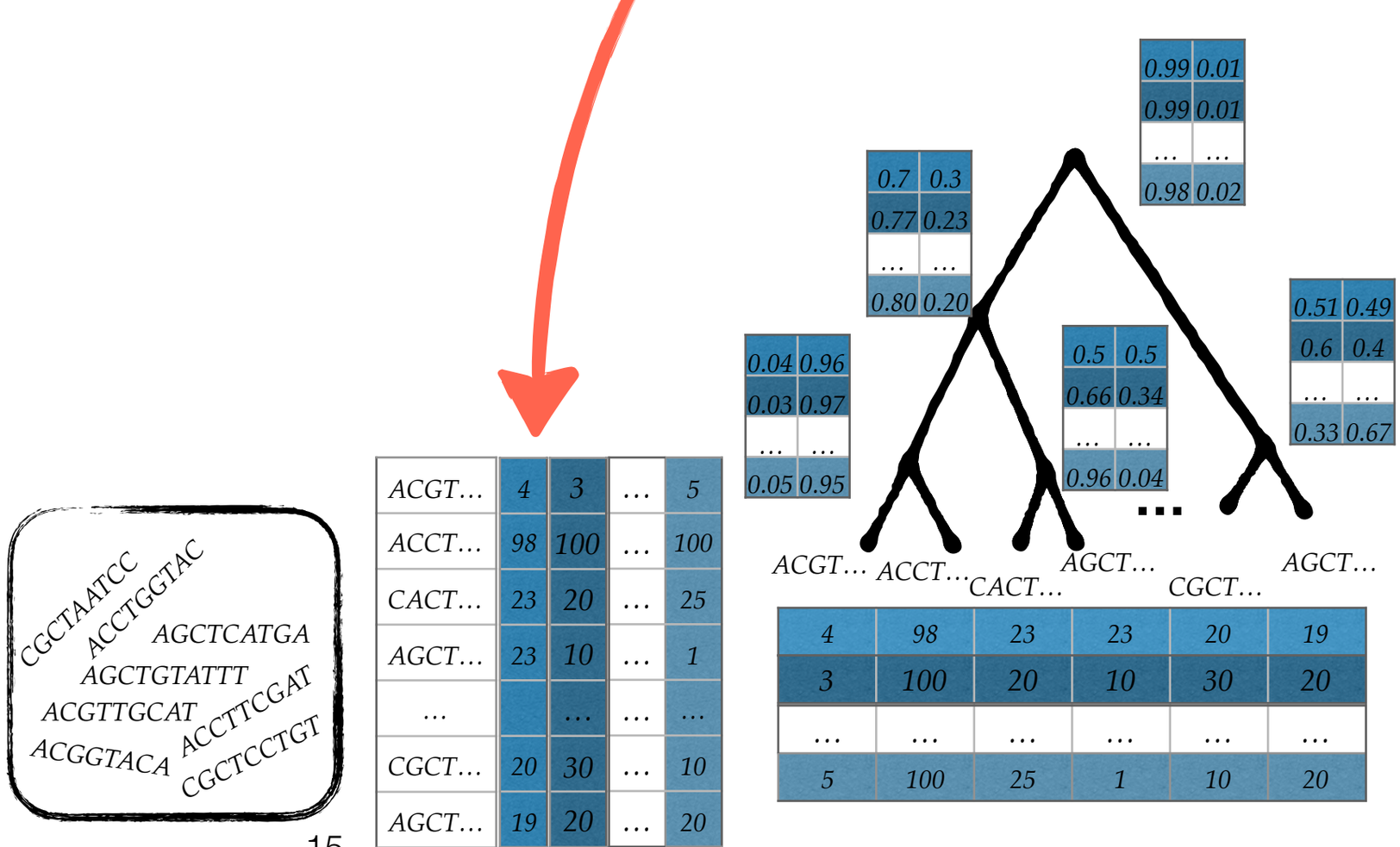


4	98	23	23	20	19
3	100	20	10	30	20
…	…	…	…	…	…
5	100	25	1	10	20

Biological variation (BV)

- Considering variations between groups of samples

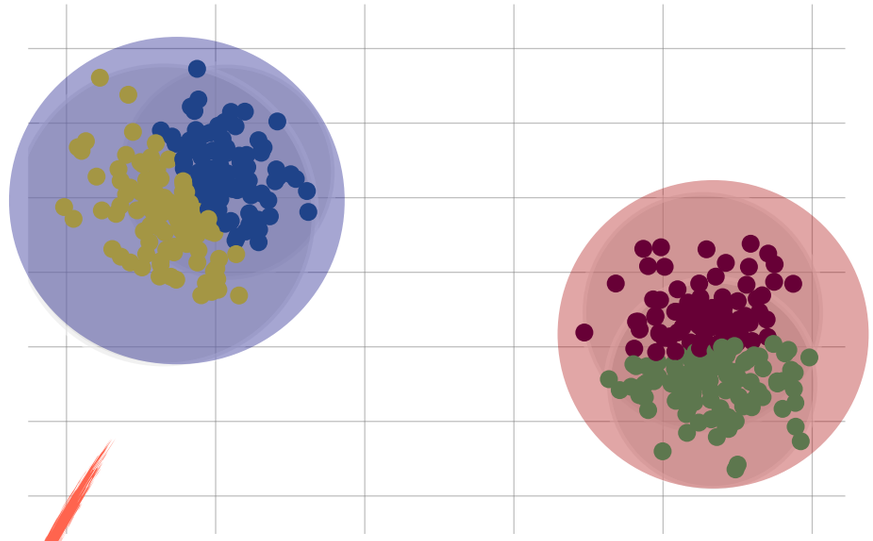


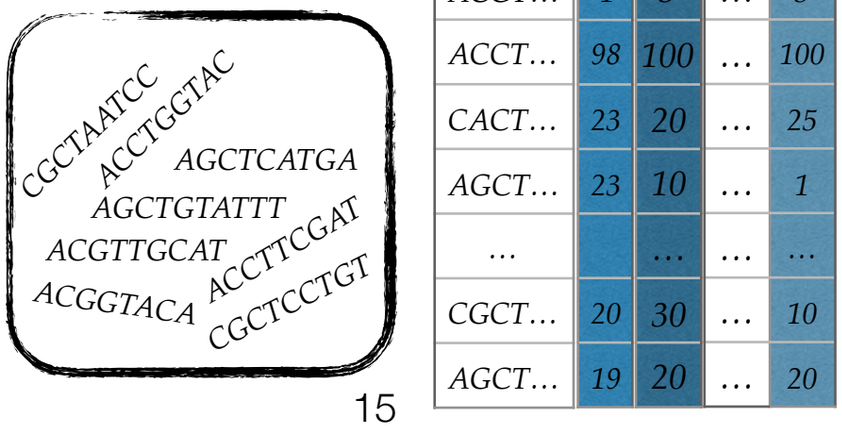


Biological variation (BV)

- Considering variations between groups of samples

- BV: Beta distributions



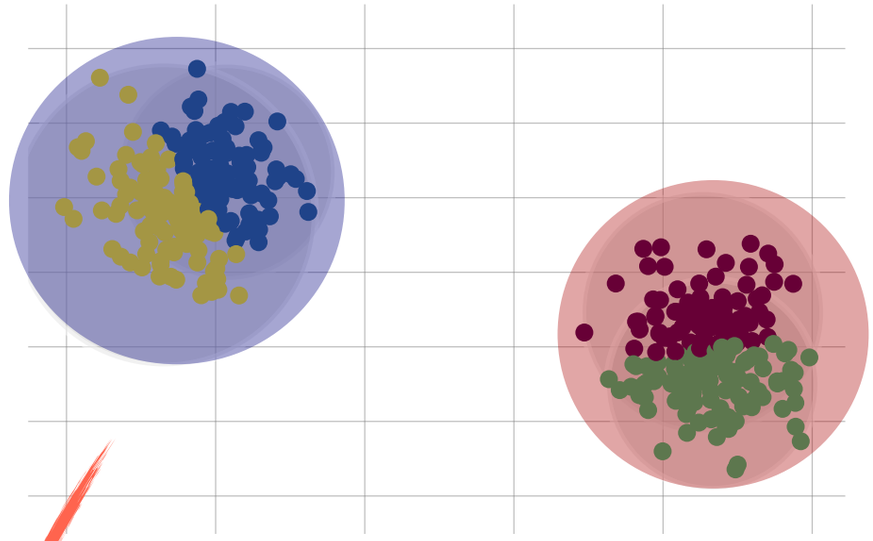


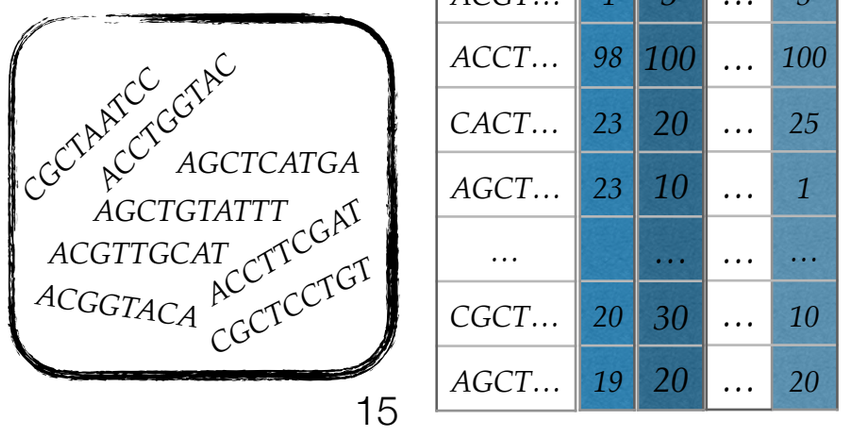
Biological variation (BV)

- Considering variations between groups of samples

- BV: Beta distributions

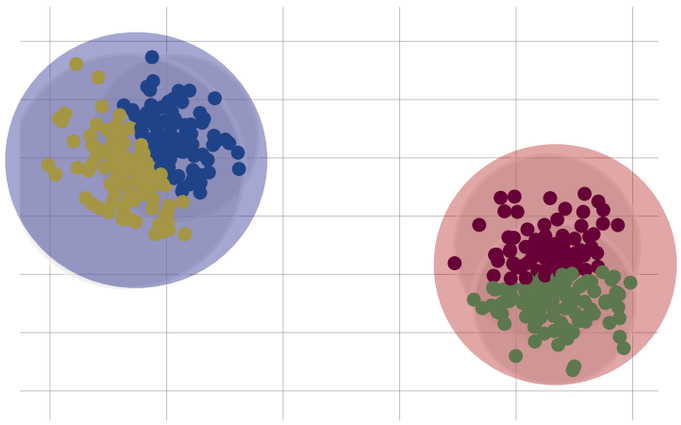
- SV: Binomial distributions

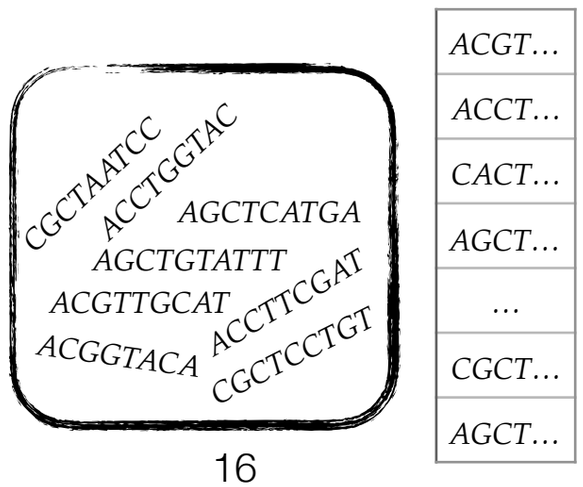


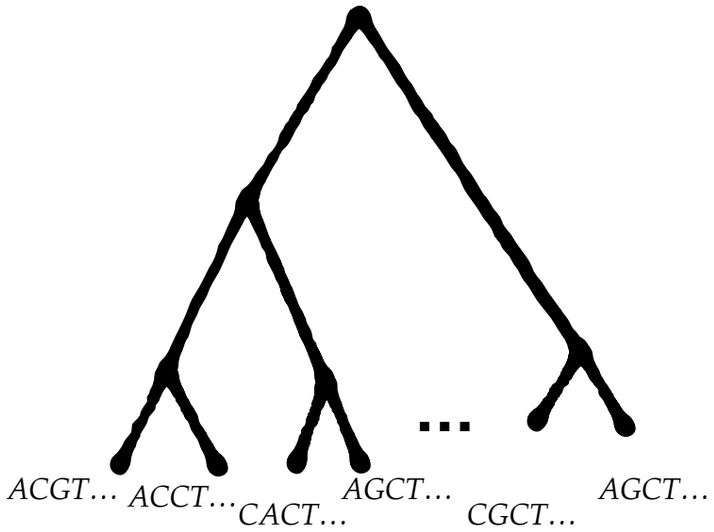


TADA-BV+SV

1. Group samples — Use class labels
   — Further cluster samples in each class





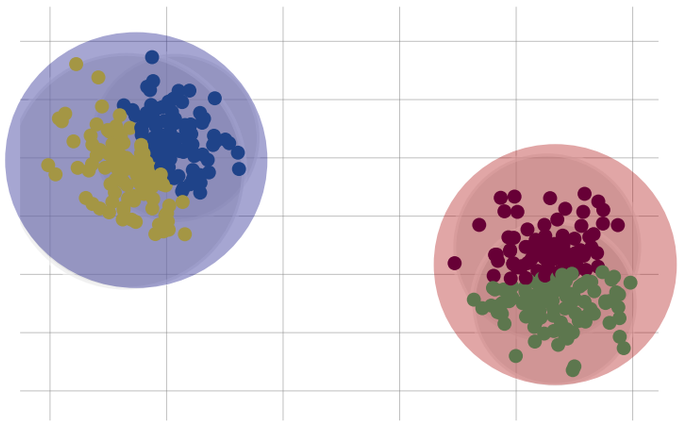


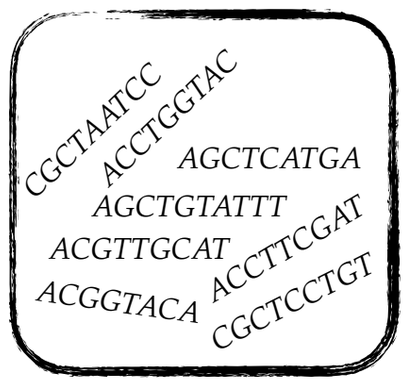
TADA-BV+SV

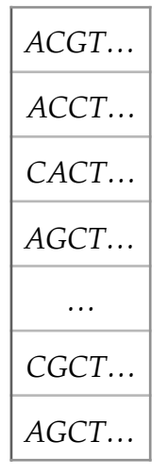
1. Group samples — Use class labels
   — Further cluster samples in each class

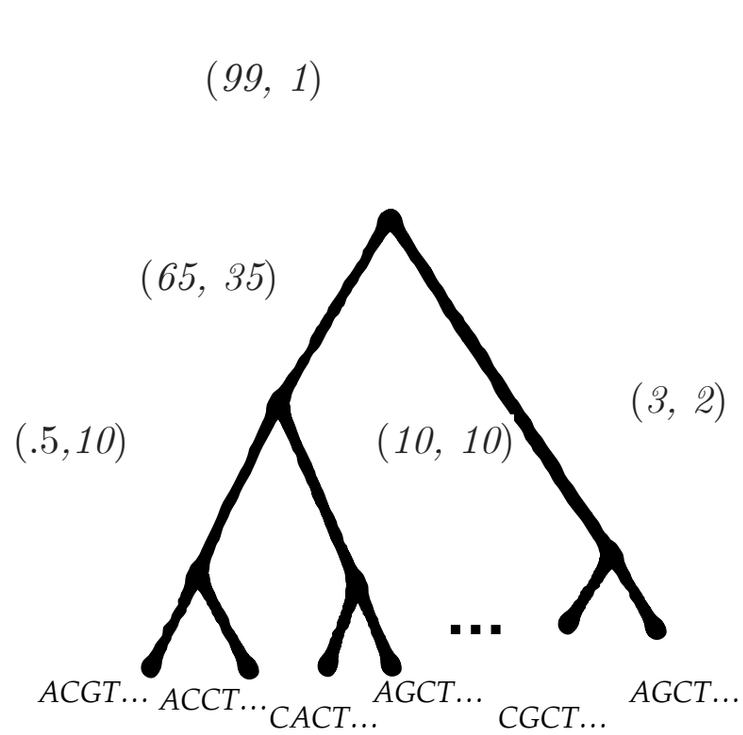
2. Per each **group**:

   - Learn parameters of Beta for all nodes
     (use tree branch lengths for controlling variance)









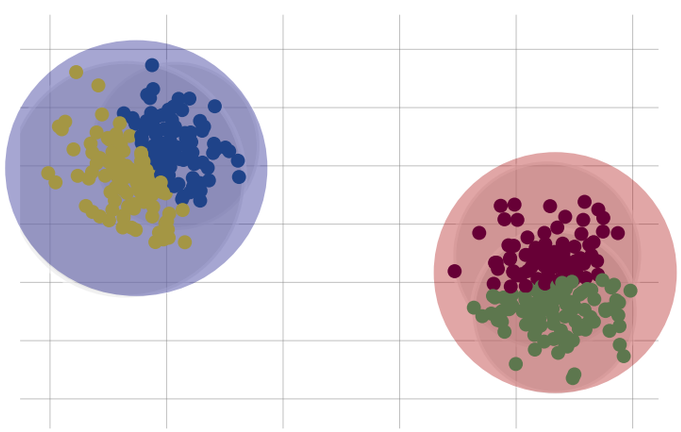
TADA-BV+SV

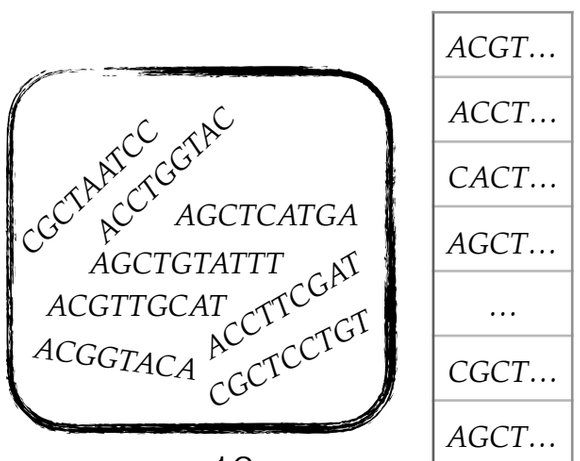
1. Group samples — Use class labels
   — Further cluster samples in each class

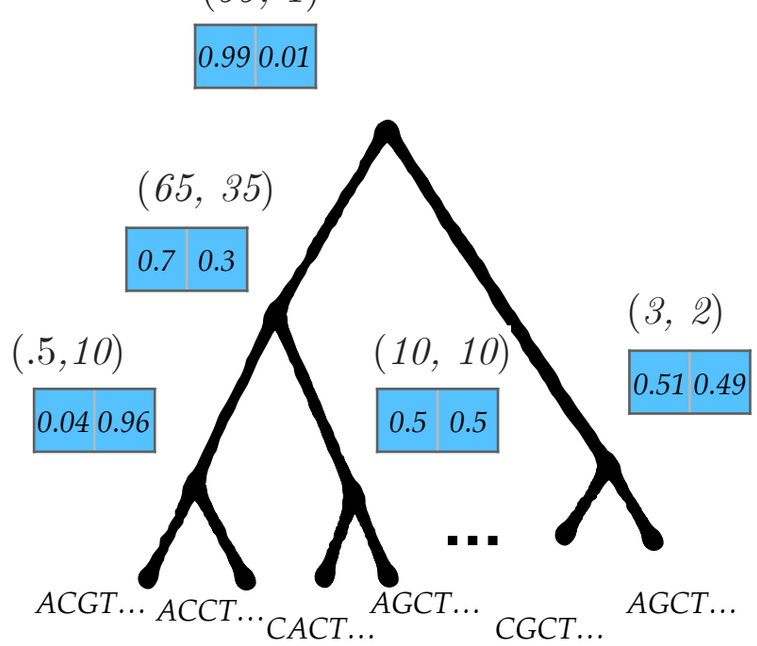
2. Per each **group**:

   - Learn parameters of Beta for all nodes
     (use tree branch lengths for controlling variance)

   - Repeat: Draw left/right probabilities from Beta for all nodes







TADA-BV+SV

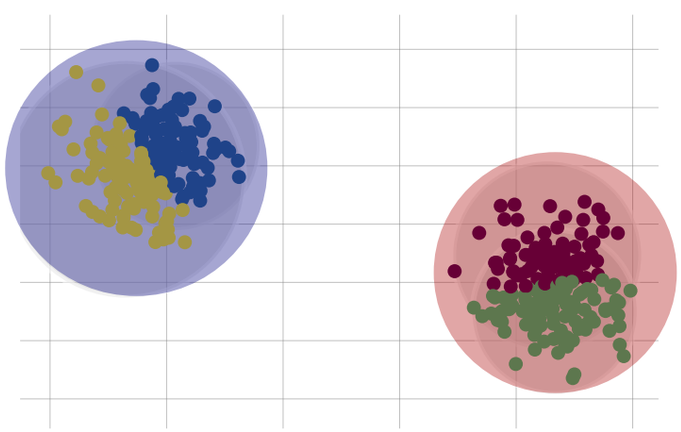
1. Group samples — Use class labels
   — Further cluster samples in each class

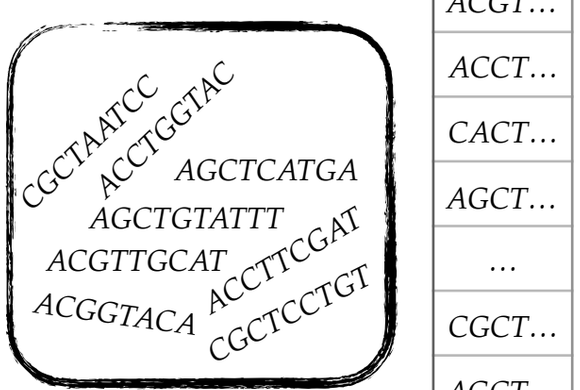
2. Per each **group**:

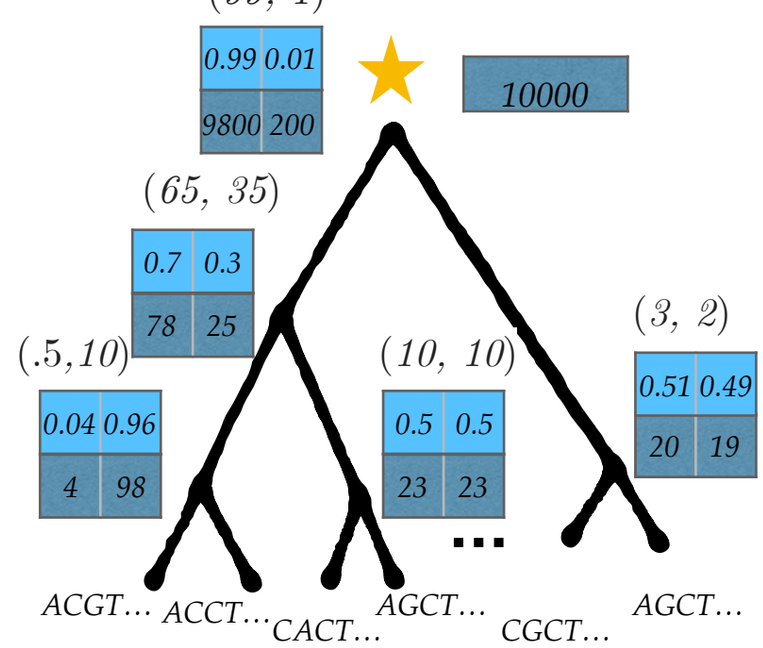
   - Learn parameters of Beta for all nodes
     (use tree branch lengths for controlling variance)

   - Repeat: Draw left/right probabilities from Beta for all nodes

     - Repeat: Draw counts using left/right probabilities from Binomial distribution







TADA-BV+SV

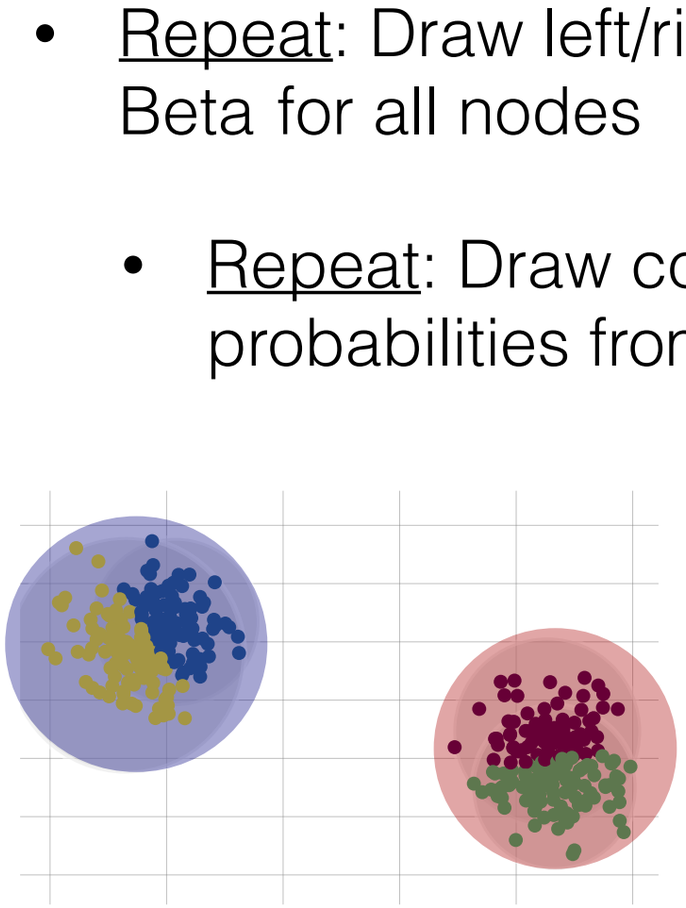
1. Group samples — Use class labels
   — Further cluster samples in each class

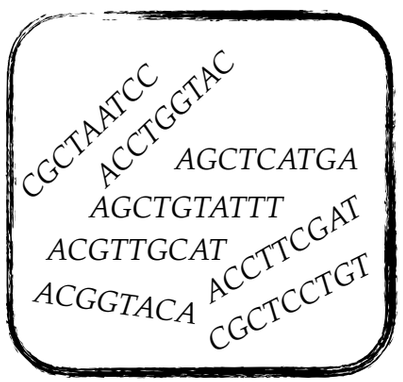
2. Per each **group**:

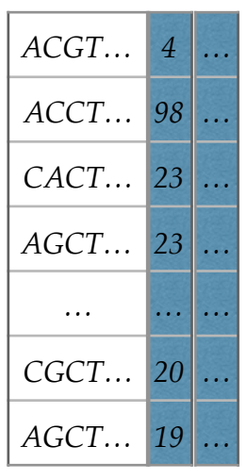
   - Learn parameters of Beta for all nodes
     (use tree branch lengths for controlling variance)

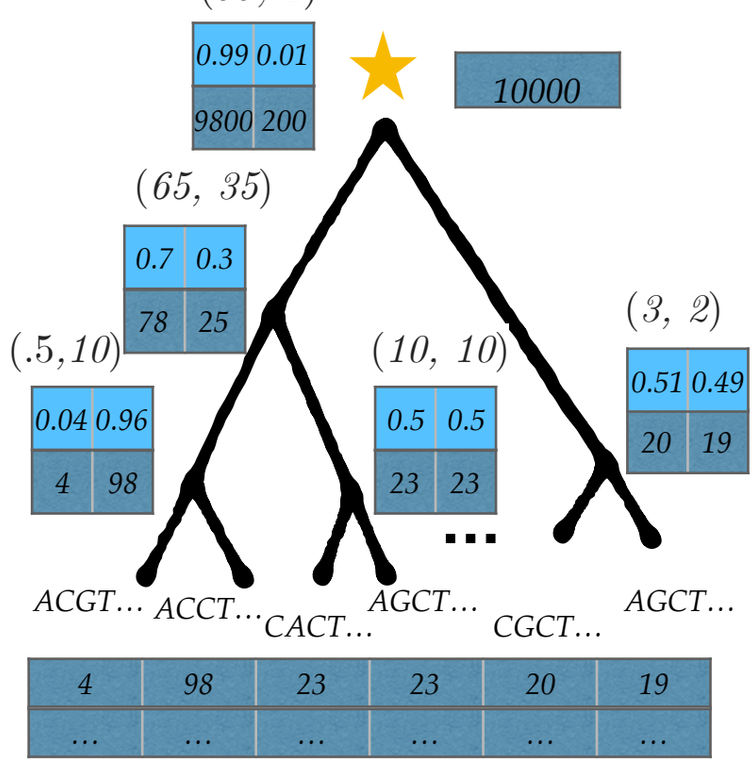
   - Repeat: Draw left/right probabilities from Beta for all nodes

     - Repeat: Draw counts using left/right probabilities from Binomial distribution









TADA-BV+SV

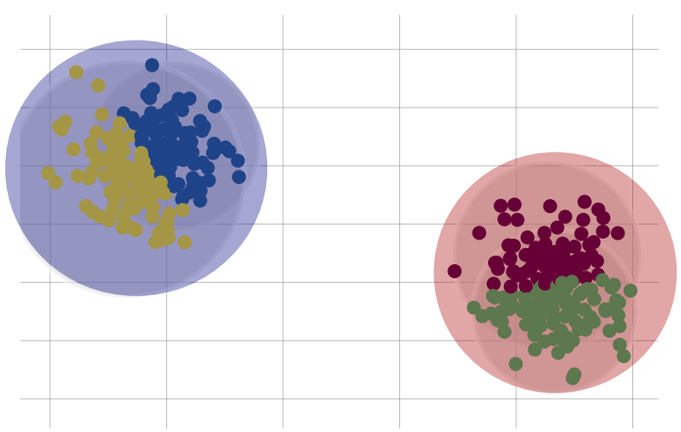
1. Group samples — Use class labels
   — Further cluster samples in each class

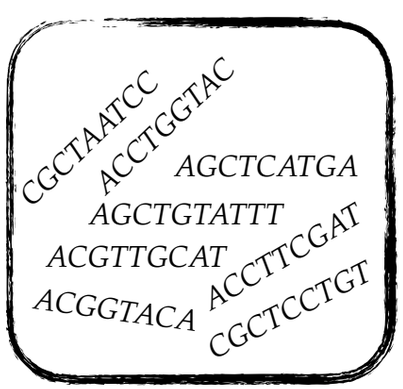
2. Per each **group**:

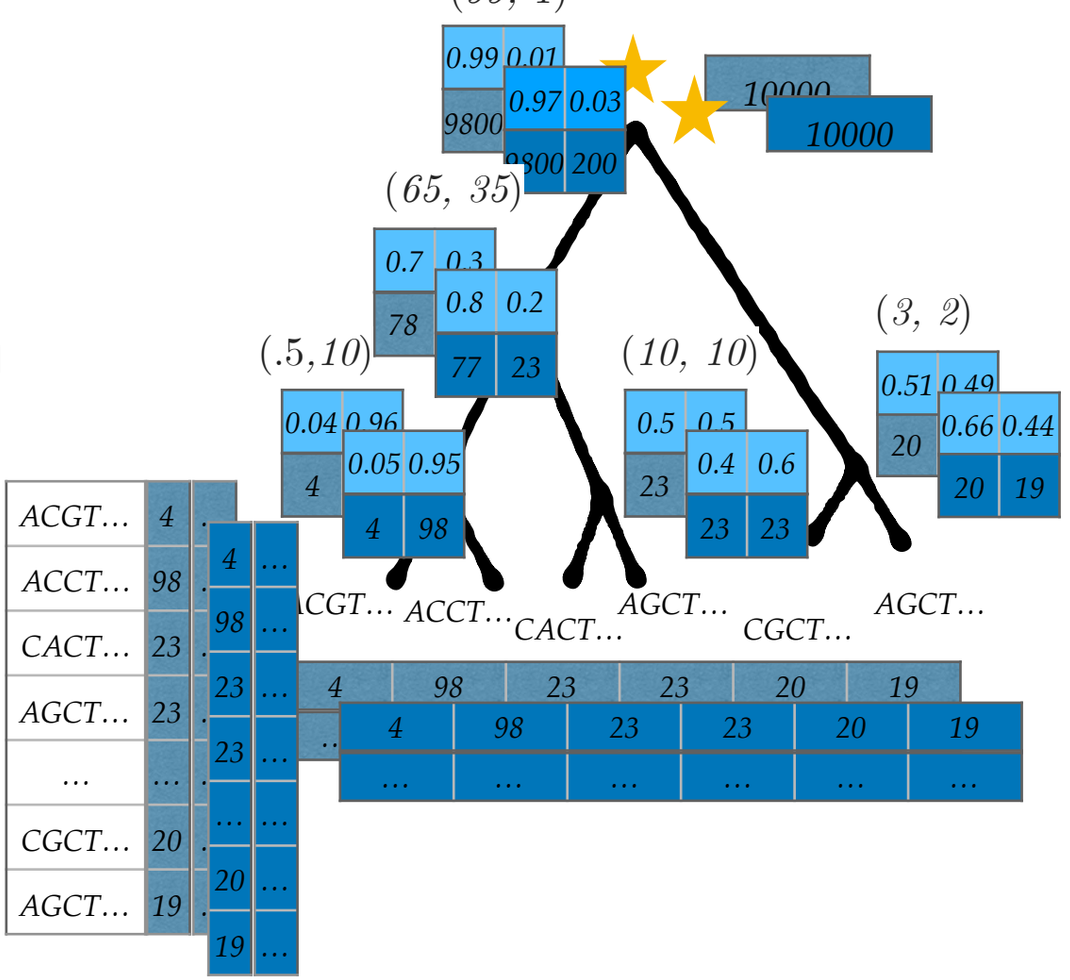
   - Learn parameters of Beta for all nodes
     (use tree branch lengths for controlling variance)

   - Repeat: Draw left/right probabilities from Beta for all nodes

     - Repeat: Draw counts using left/right probabilities from Binomial distribution







TADA-BV+SV

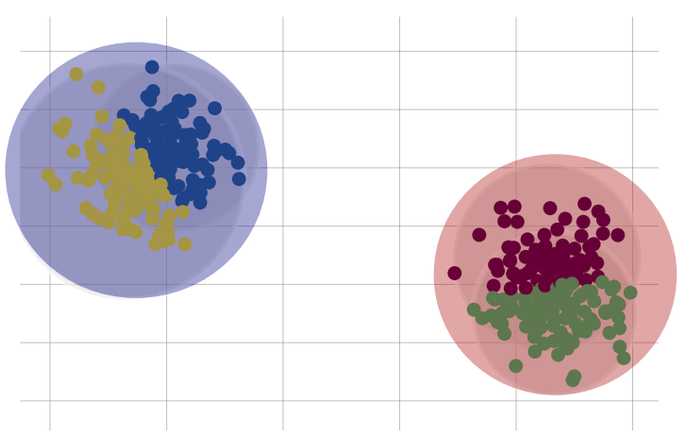
1. Group samples — Use class labels
   — Further cluster samples in each class

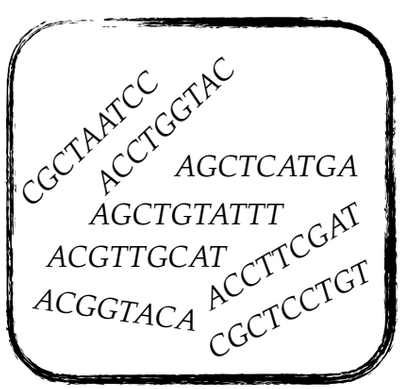
2. Per each **group**:

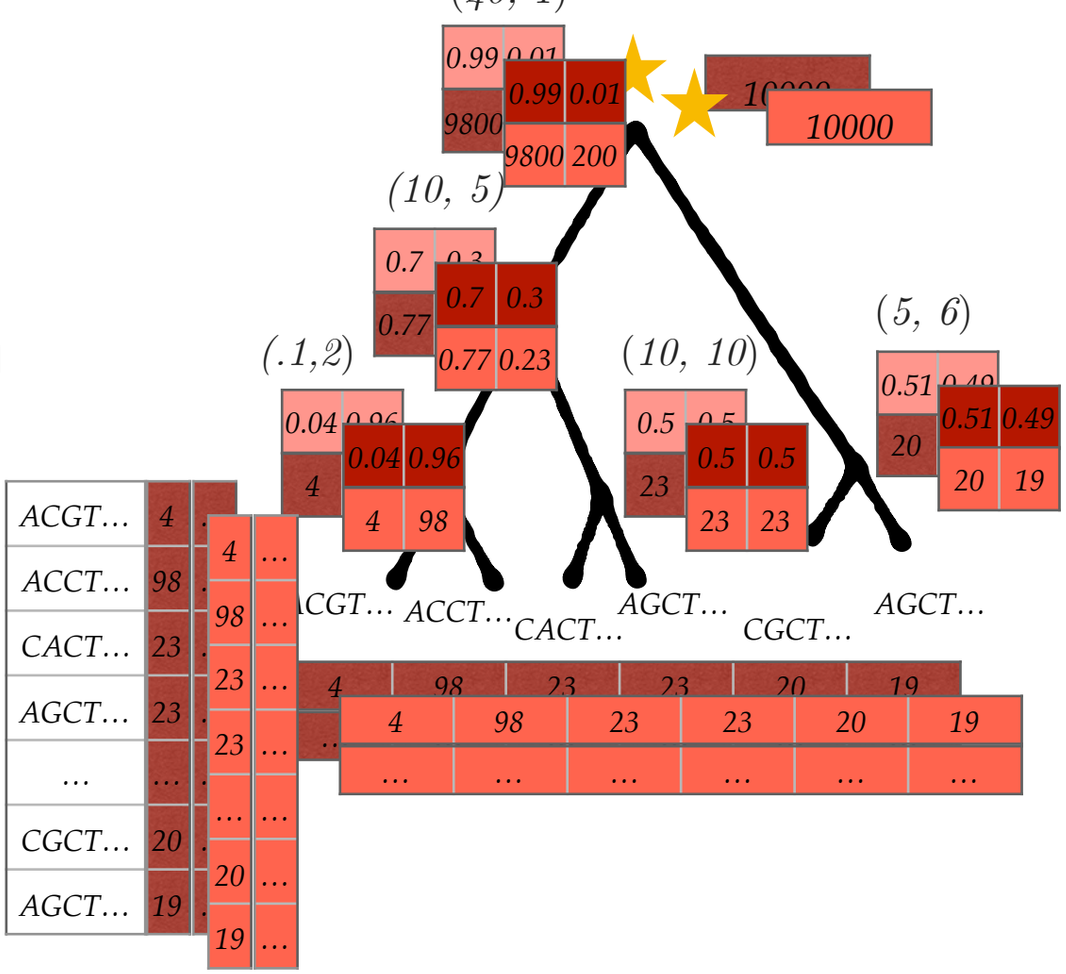
   - Learn parameters of Beta for all nodes
     (use tree branch lengths for controlling variance)

   - Repeat: Draw left/right probabilities from Beta for all nodes

     - Repeat: Draw counts using left/right probabilities from Binomial distribution





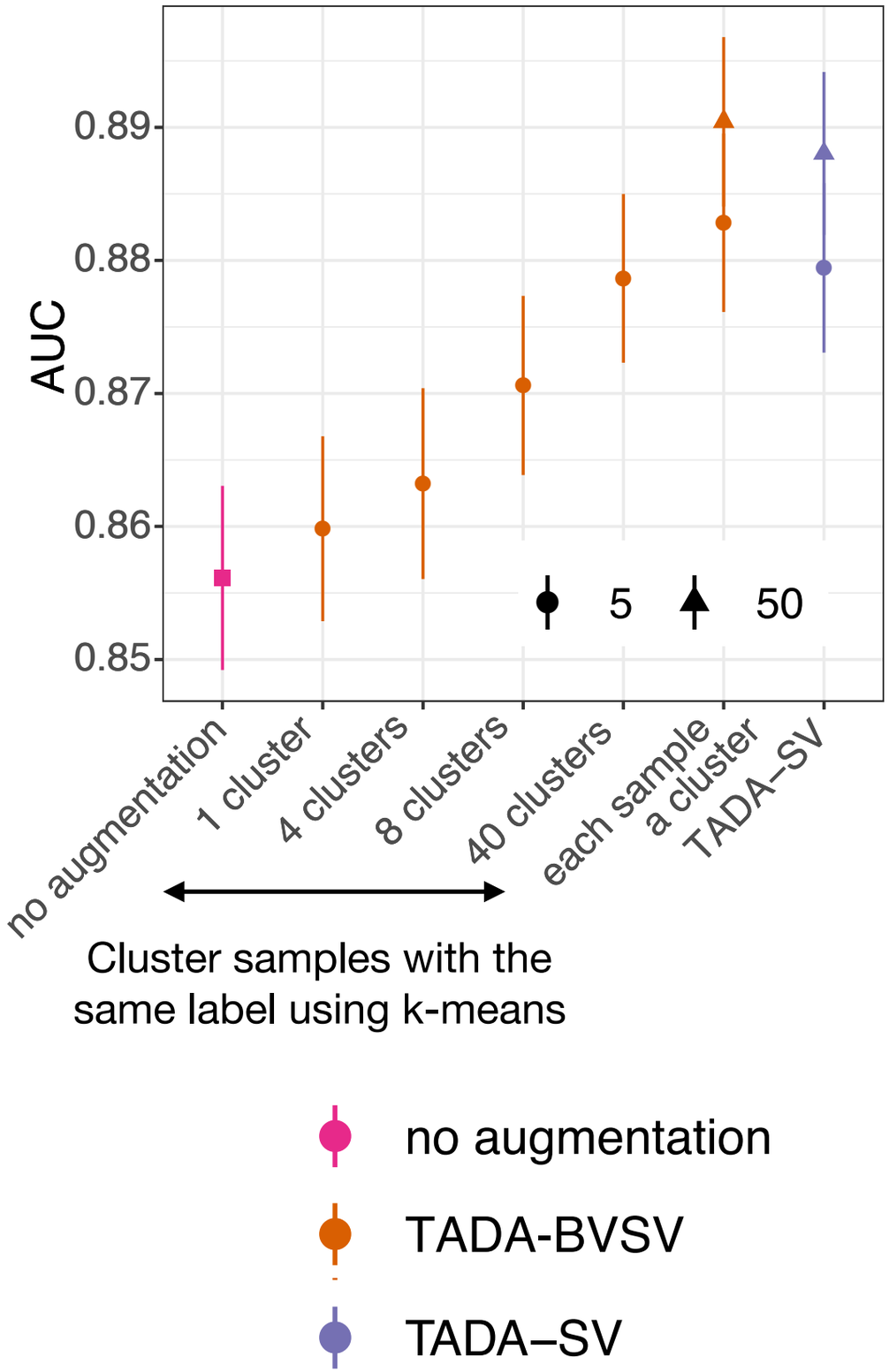


TADA BVSV+clustering

- 33% healthy vs 66% IBD
- Medium sized training dataset (712 samples)

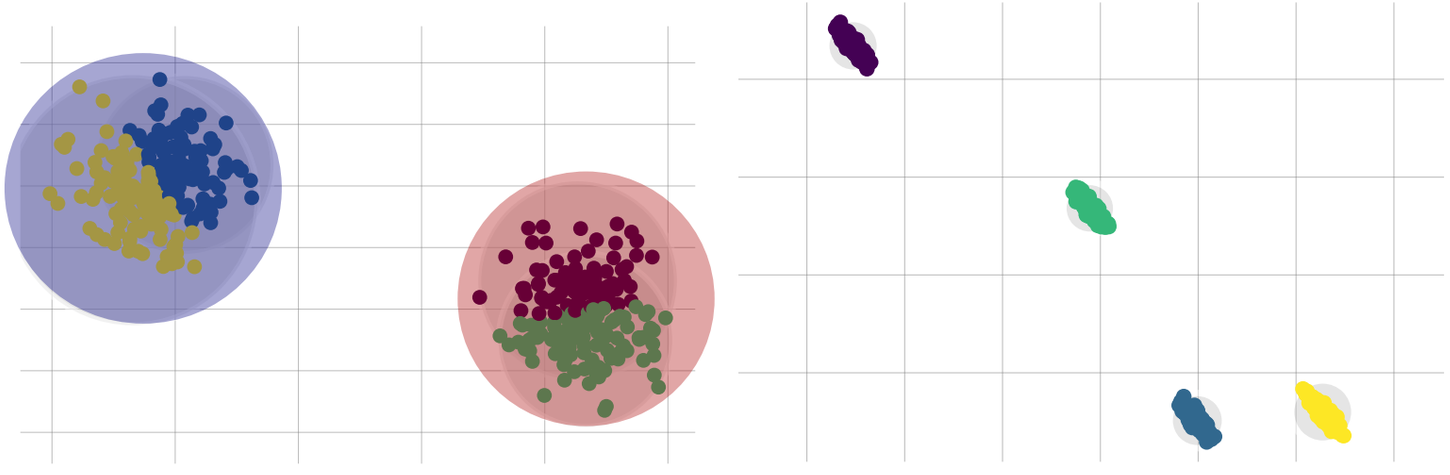
TADA BVSV+clustering

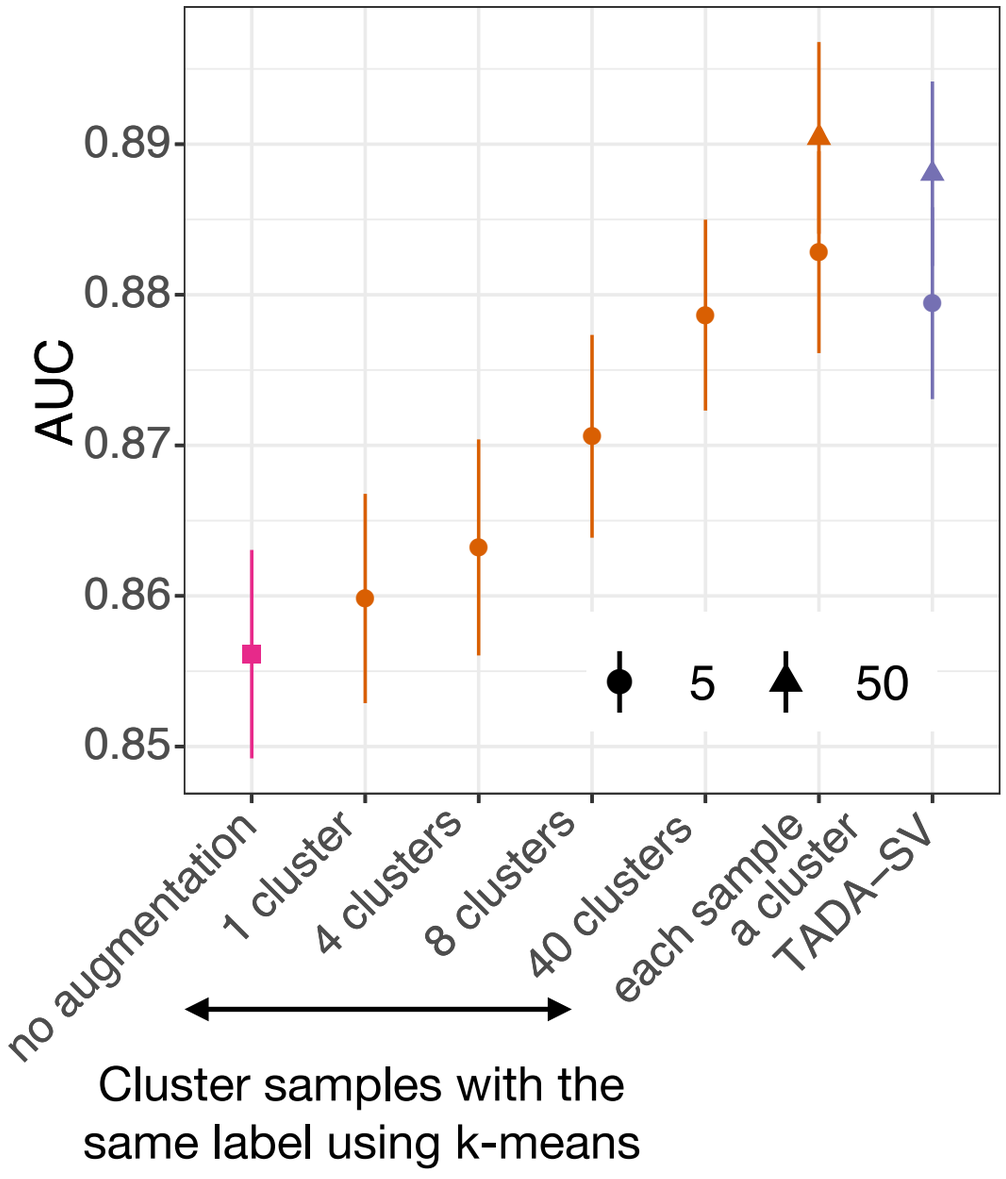
- 33% healthy vs 66% IBD
- Medium sized training dataset (712 samples)



TADA BVSV+clustering

- 33% healthy vs 66% IBD
- Medium sized training dataset (712 samples)
- What if there are as many clusters as there are samples?





See the paper for …

Sayyari E., Kawas B., Mirarab S. 2019. TADA: phylogenetic augmentation of microbiome samples enhances phenotype classification. Bioinformatics. 35:i31–i40.

- Details of the TADA algorithm
- Many more experiments on the IBD data
  - Biased (but balanced) sampling
- Results using NN
- Results on a second dataset (American Gut) for a different trait (BMI)

Code and data

- They will be made available in a couple of days (hopefully) on GitHub

- https://github.com/tada-alg/TADA

- Apologies the code is not uploaded yet!!!

Acknowledgments

Prof. Rob Knights

Dr. Austin Swafford

Dr. Se Jin Song

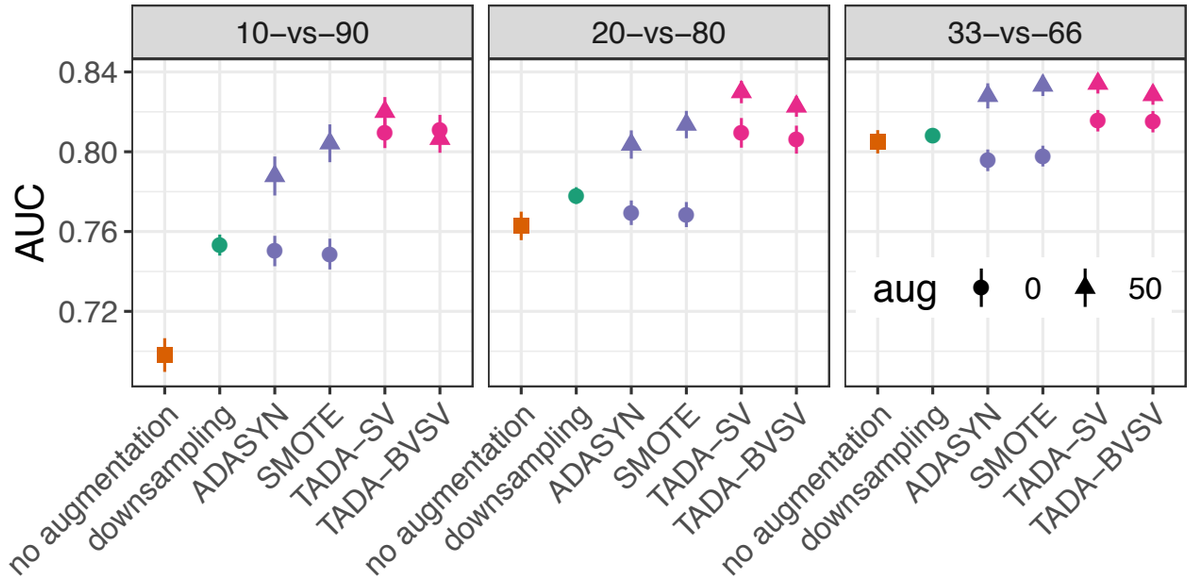
Dr. Sandrine Miller

Dr. Ho-Cheol Kim

Thank you

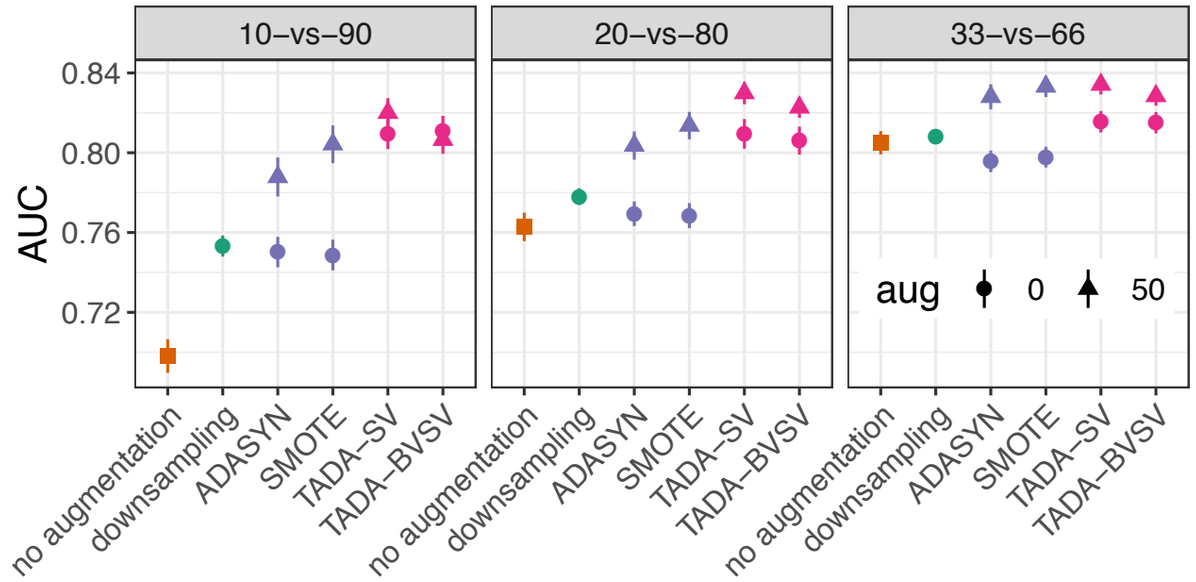
Why TADA-BVSV doesn't beat TADA-SV?

- BV might be sufficiently captured



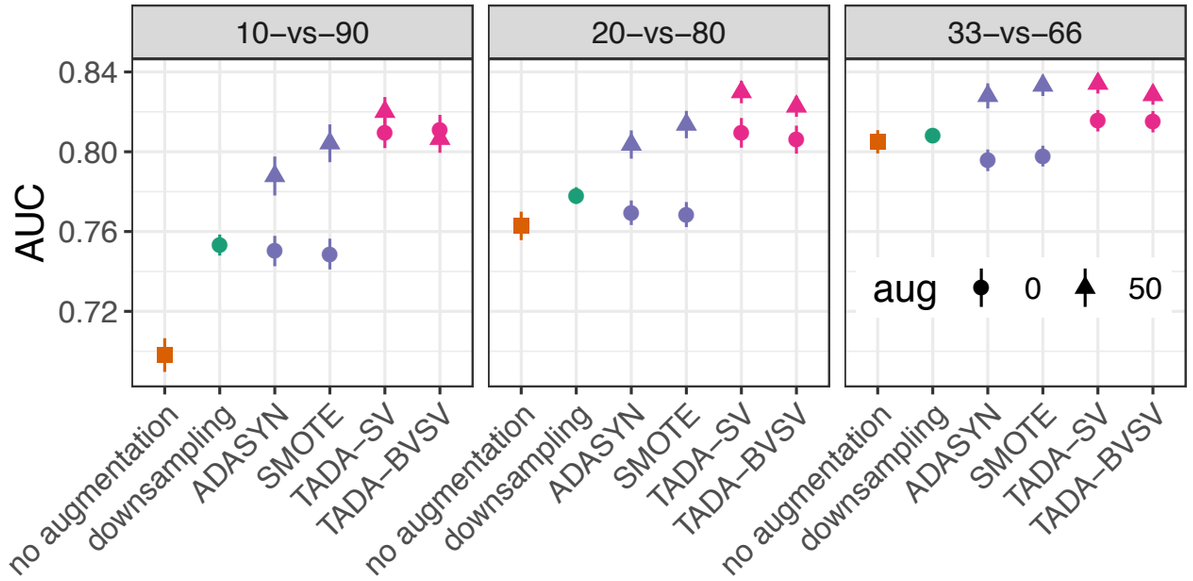
Why TADA-BVSV doesn't beat TADA-SV?

- BV might be sufficiently captured
- Beta distribution might not be the correct choice



Why TADA-BVSV doesn't beat TADA-SV?

- BV might be sufficiently captured
- Beta distribution might not be the correct choice
- Conditional independence assumptions on the phylogenetic tree might not match the biology
    - Example: Horizontal gene transfer



Parameters of Beta distribution

- Beta distribution has two parameters
  - Mean and variance
- Learn them using the samples of the same cluster/class
  - Method of moments
  - Did not work well
- With few samples, learning all parameters from data is not feasible
  - We want a procedure to consider each point as the center of its own cluster

Heuristic solution

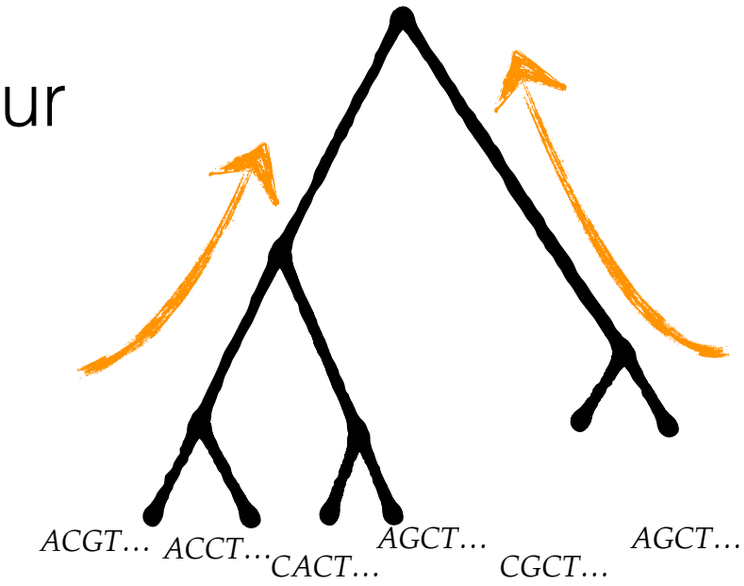
- If two nodes are far away from each other: our count estimates are robust, less variation

Heuristic solution

- If two nodes are far away from each other: our count estimates are robust, less variation
- If two nodes are closer to each other: more variation

Heuristic solution

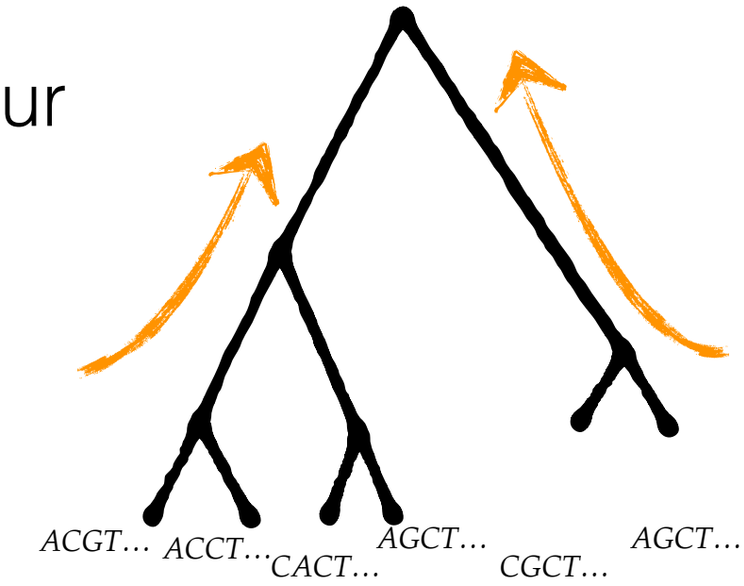
- If two nodes are far away from each other: our count estimates are robust, less variation
- If two nodes are closer to each other: more variation



Heuristic solution

- If two nodes are far away from each other: our count estimates are robust, less variation

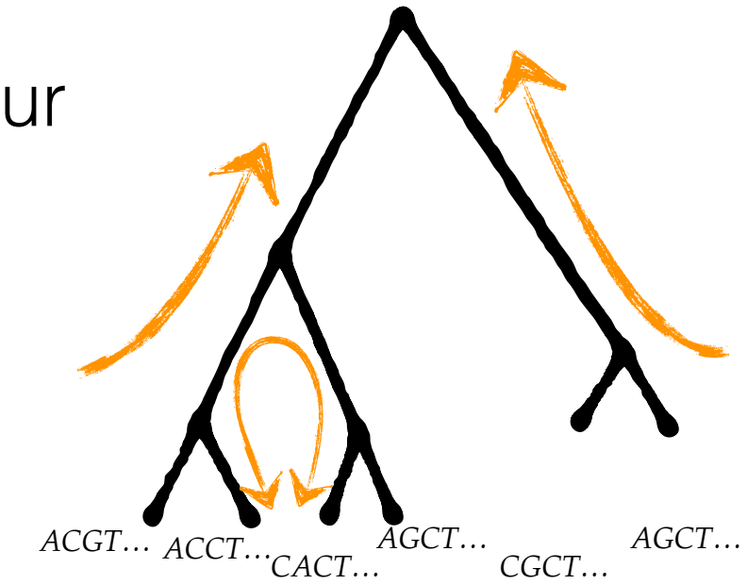
- If two nodes are closer to each other: more variation



- d: average pairwise distances of leaves contained under each node

Heuristic solution

- If two nodes are far away from each other: our count estimates are robust, less variation
- If two nodes are closer to each other: more variation

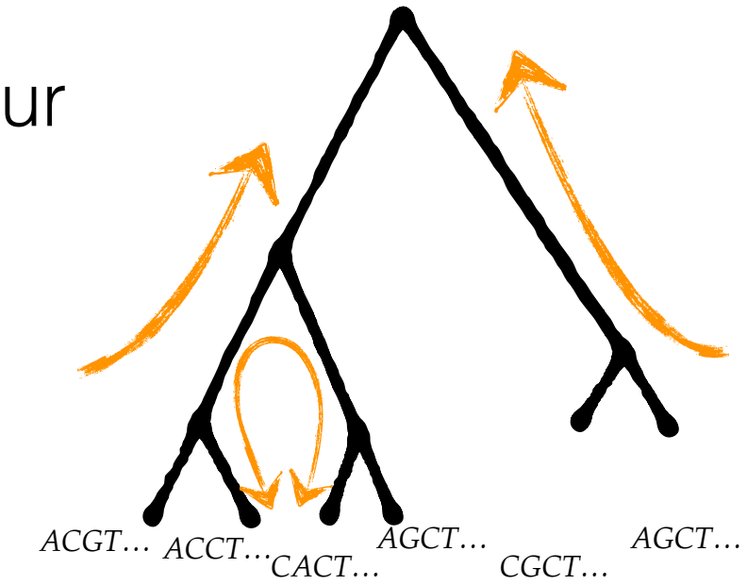


- d: average pairwise distances of leaves contained under each node

Heuristic solution

- If two nodes are far away from each other: our count estimates are robust, less variation

- If two nodes are closer to each other: more variation



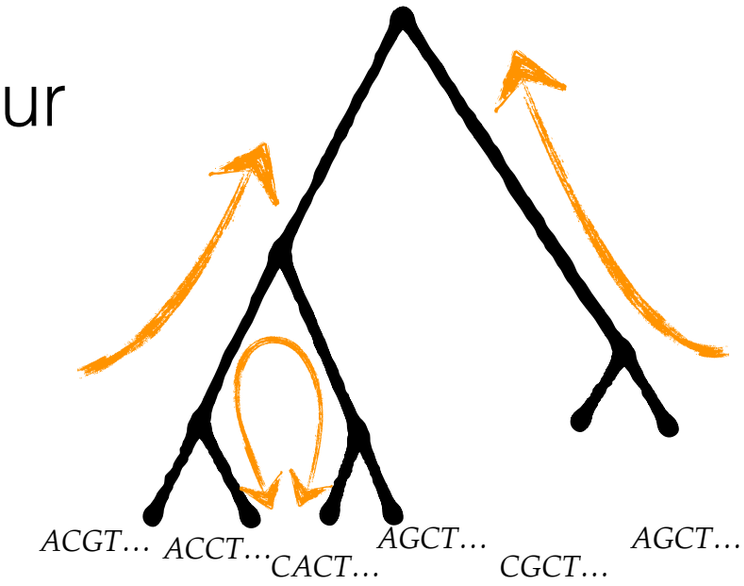
- d: average pairwise distances of leaves contained under each node

- Variance $\propto f(d)=d^{-r}$ (e.g., $r=0.5$)

Heuristic solution

- If two nodes are far away from each other: our count estimates are robust, less variation

- If two nodes are closer to each other: more variation



- d: average pairwise distances of leaves contained under each node

- Variance $\propto f(d)=d^{-r}$ (e.g., $r=0.5$)

- Mean is average proportions of right and left

Applications

Applications

- Reduce unbalancedness with respect to class labels
  - e.g. having 20 samples for class 1 and 200 samples for class 2

Applications

- Reduce unbalancedness with respect to class labels
  - e.g. having 20 samples for class 1 and 200 samples for class 2
- Reduce overfitting in machine learning analyses
  - e.g. having only 50 samples for each class

Applications

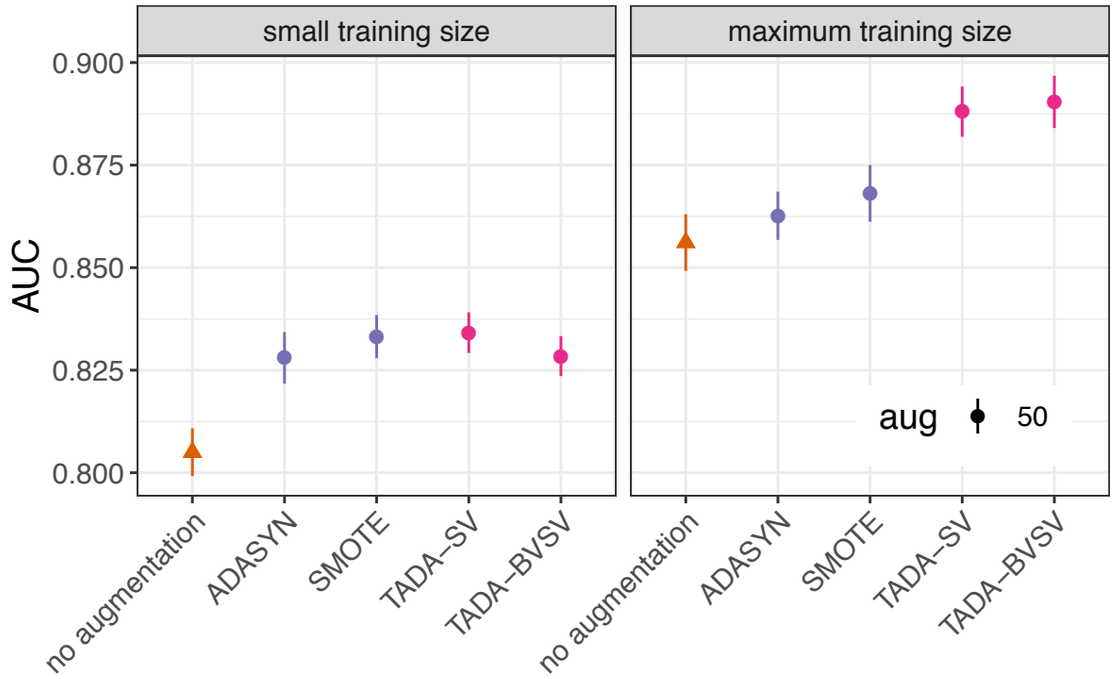
- Reduce unbalancedness with respect to class labels
  - e.g. having 20 samples for class 1 and 200 samples for class 2
- **Reduce overfitting in machine learning analyses**
  - e.g. having only 50 samples for each class

Augmentation to reduce overfitting

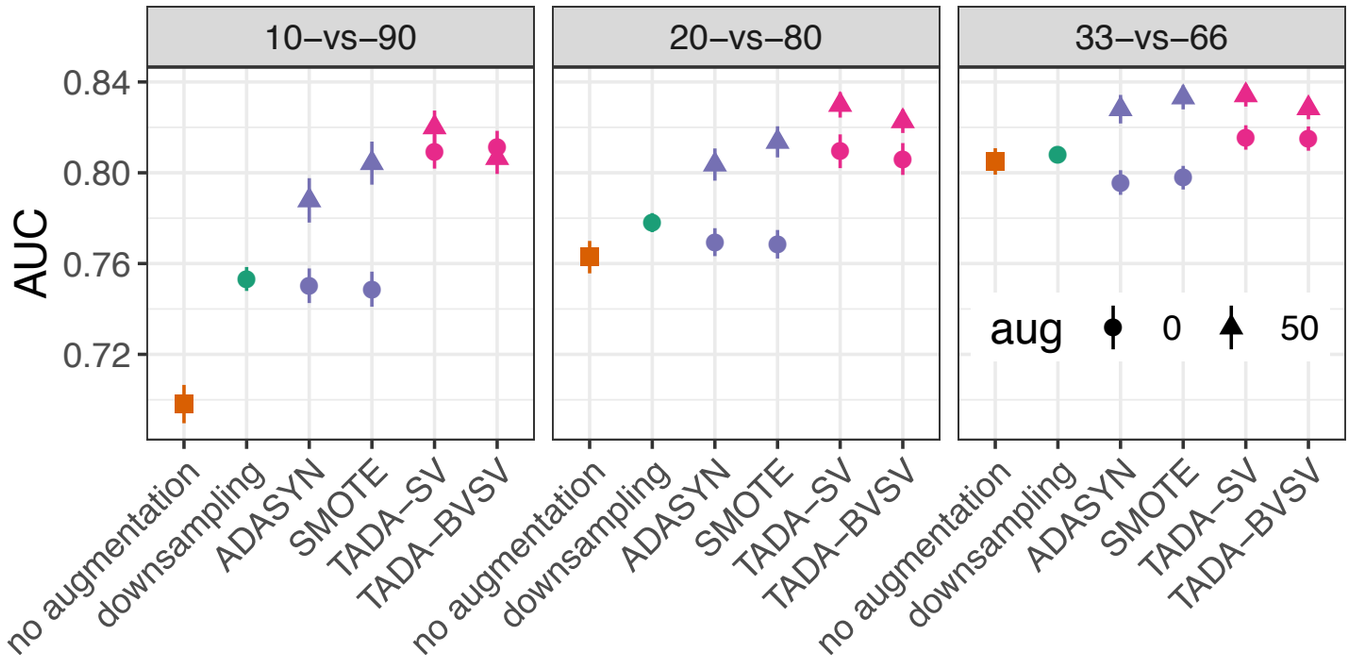
- Small sized training data
  - 243 training points
- Medium sized training data
  - 712 training points
- 33% healthy vs 66% IBD
- Augmentation level: 50x

Augmentation to reduce overfitting

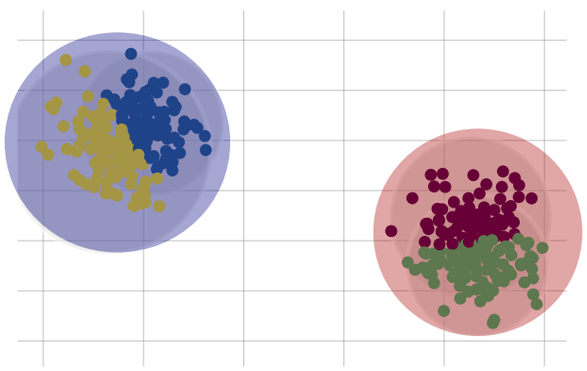
- Small sized training data
  - 243 training points
- Medium sized training data
  - 712 training points
- 33% healthy vs 66% IBD
- Augmentation level: 50x



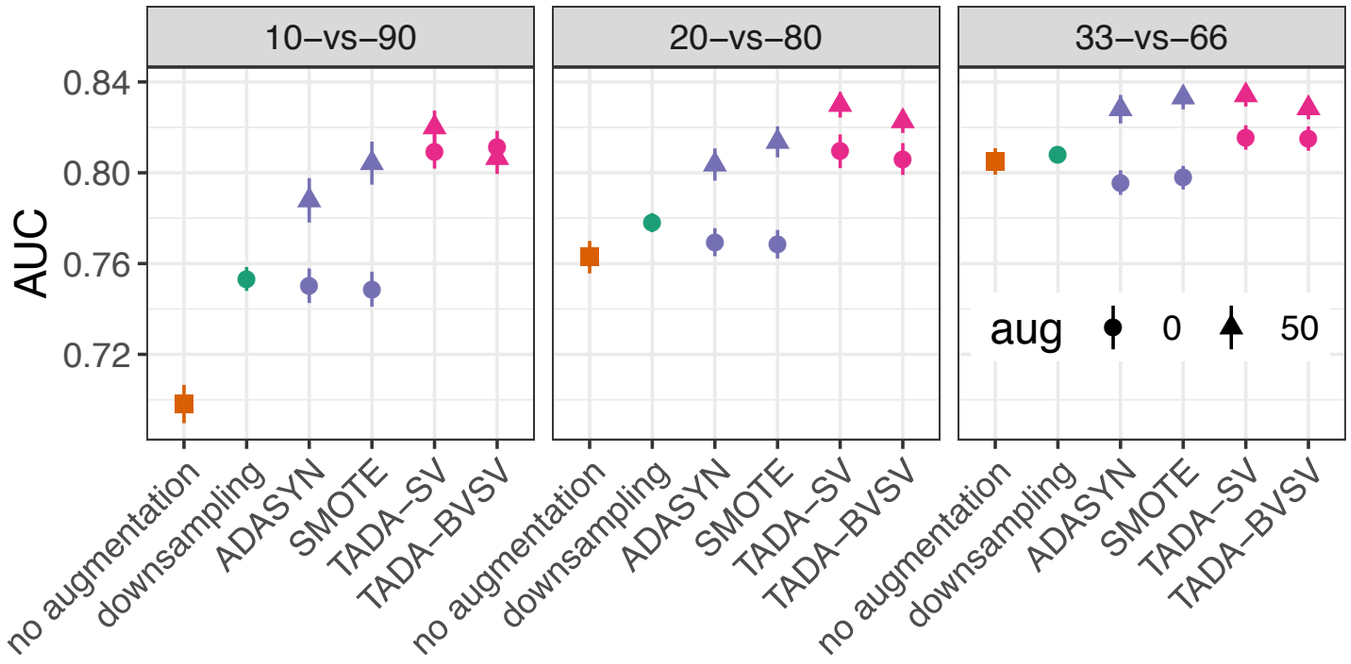
Why not helping?



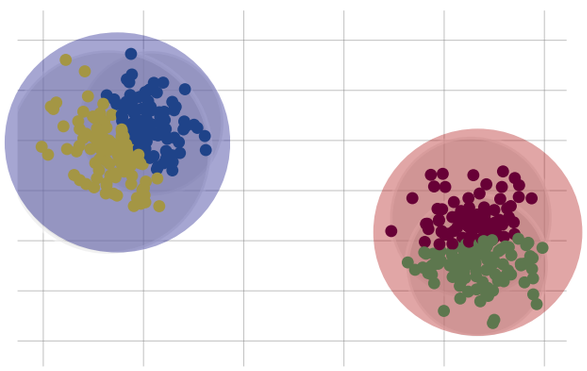
- What if there are as many clusters as there are samples?

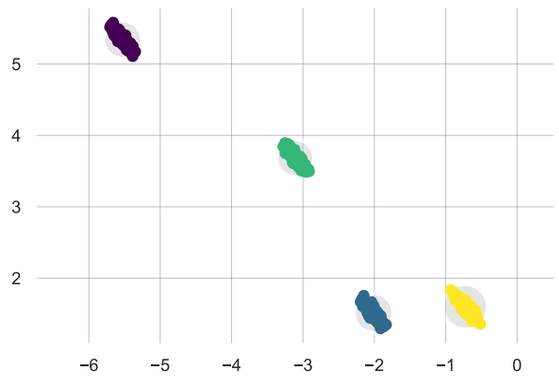


Why not helping?



- What if there are as many clusters as there are samples?



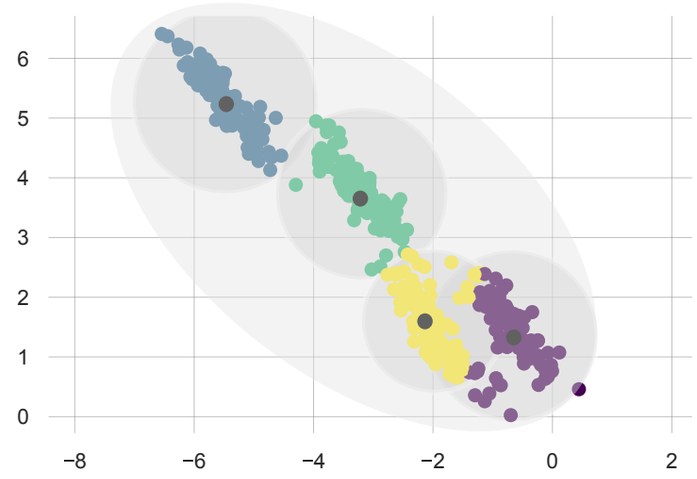


Biological variation (BV)

- Considering variations between samples

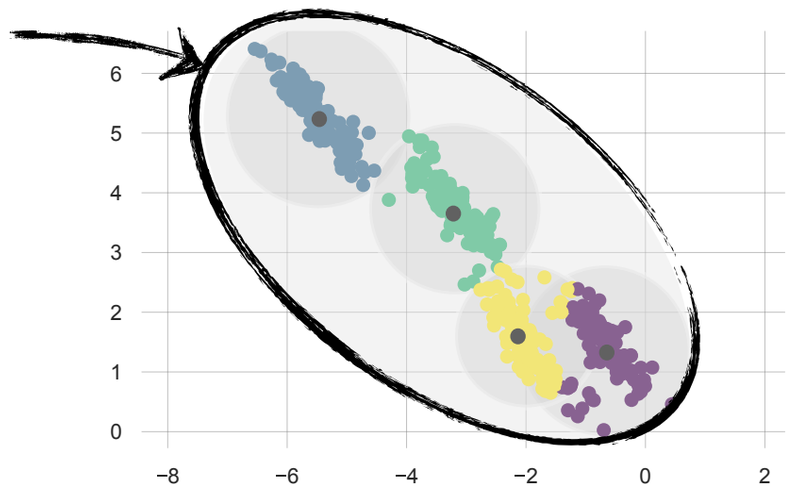
Biological variation (BV)

- Considering variations between samples



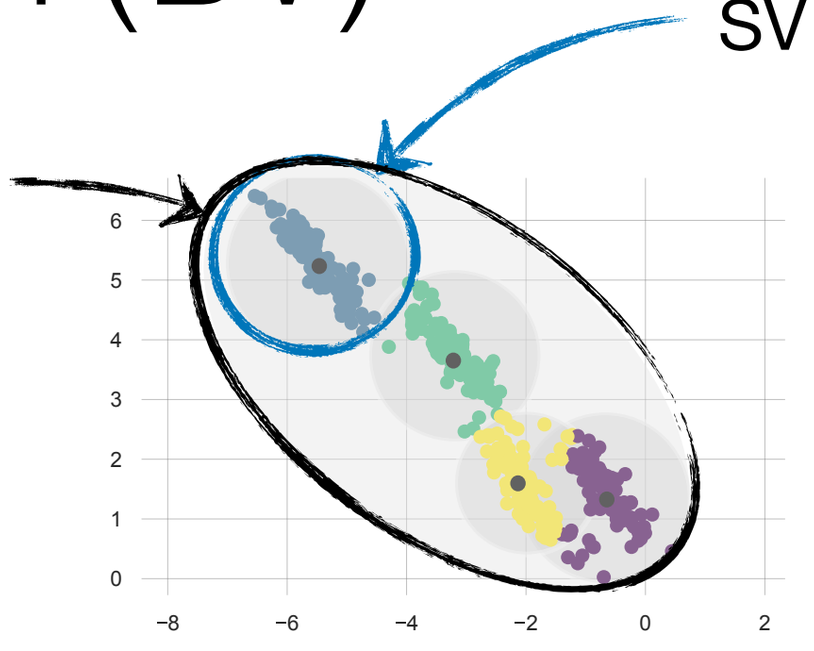
Biological variation (BV)

- Considering variations between samples

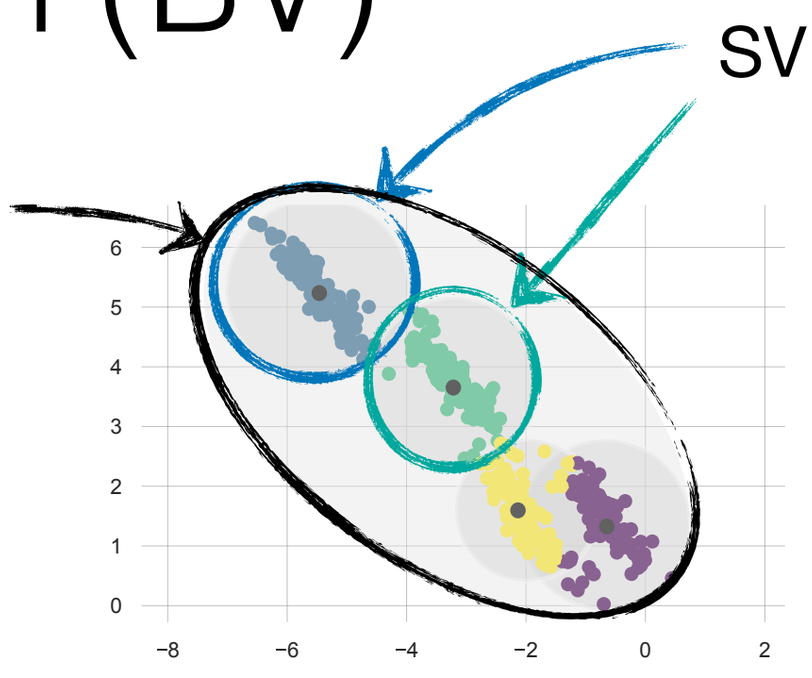


Biological variation (BV)

- Considering variations between samples

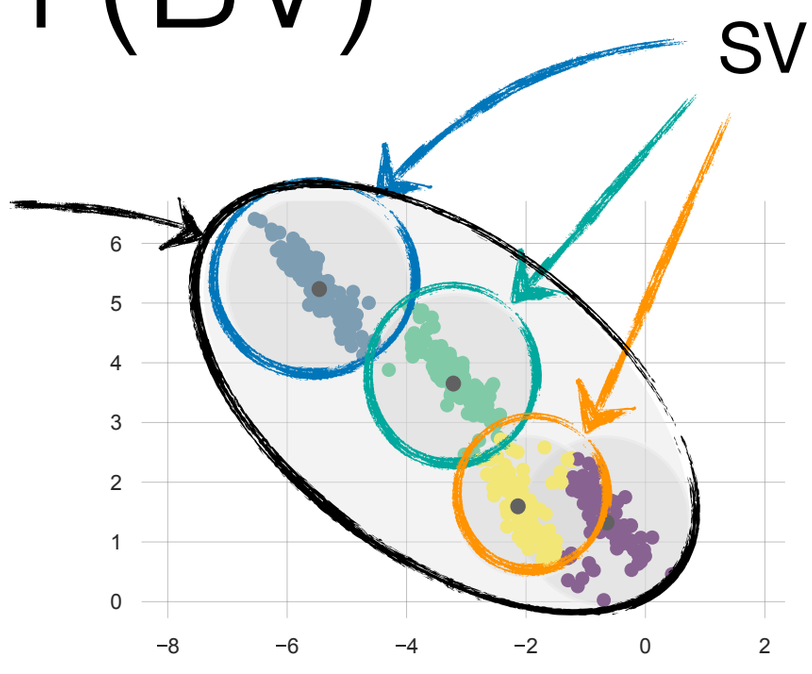


Biological variation (BV)



- Considering variations between samples

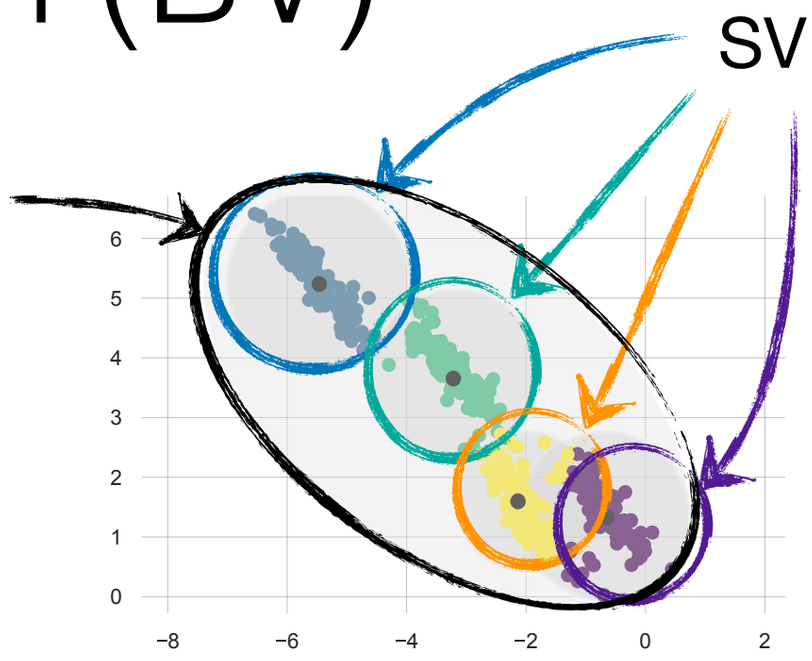
Biological variation (BV)



- Considering variations between samples

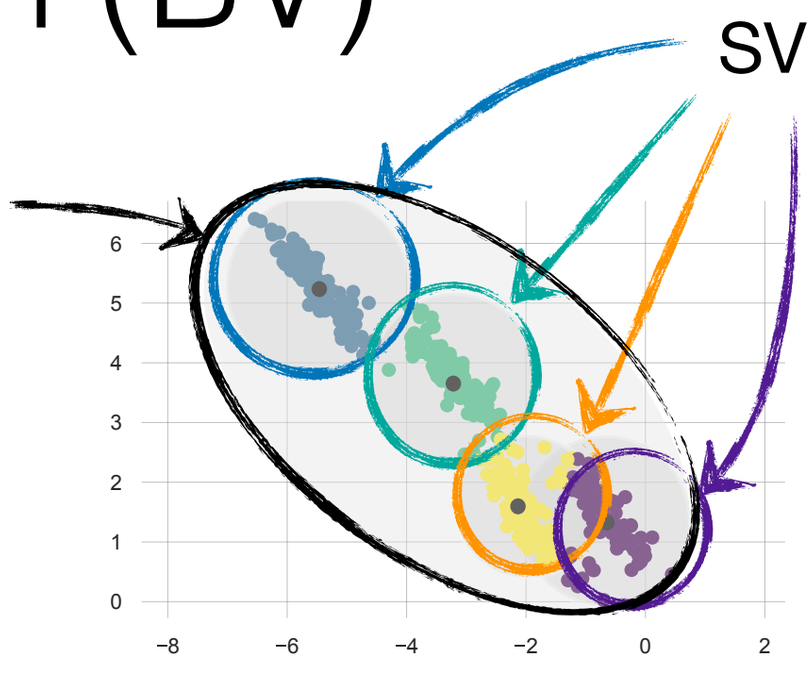
Biological variation (BV)

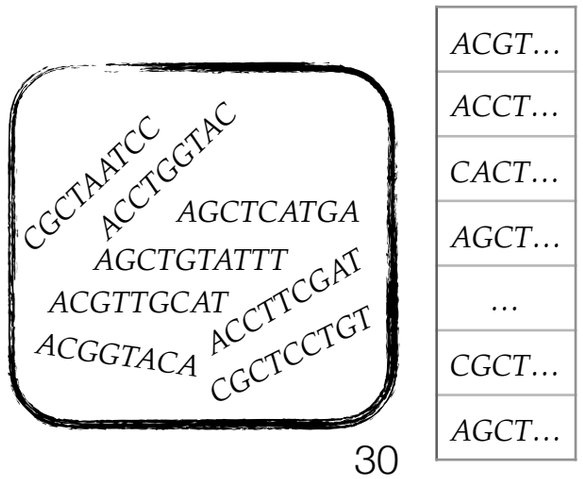
- Considering variations between samples

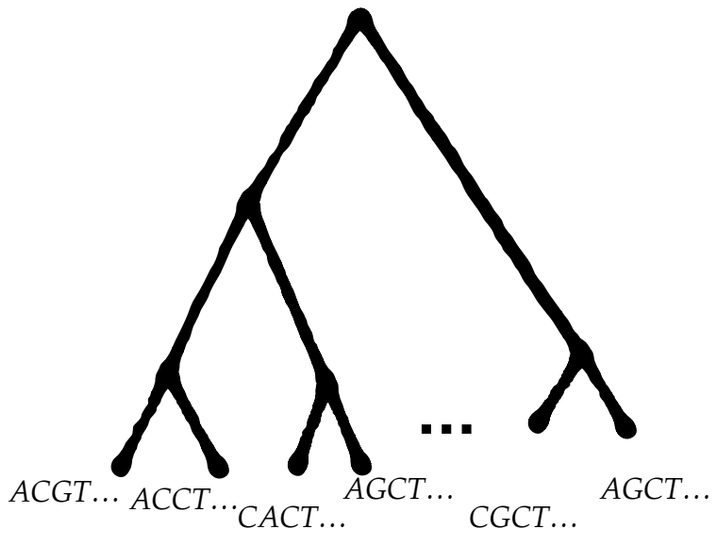


Biological variation (BV)

- Considering variations between samples

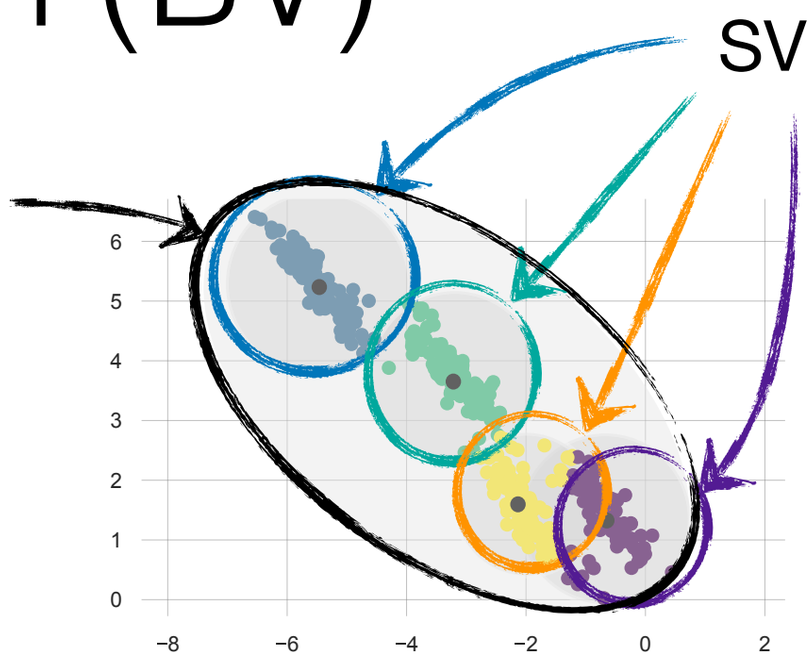


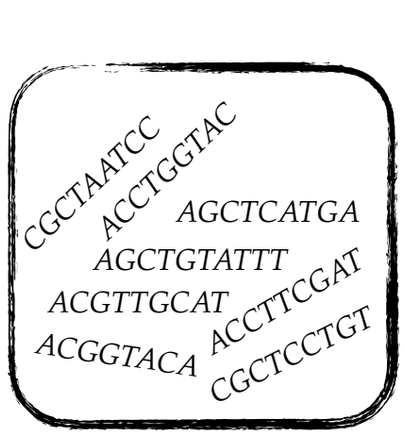




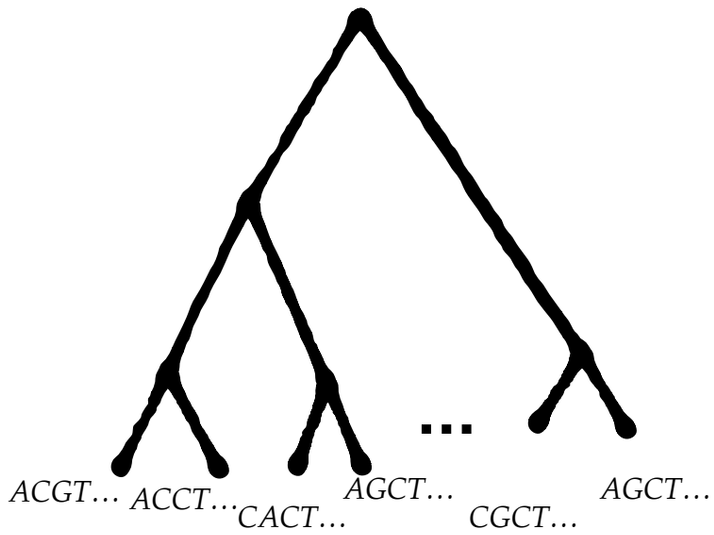
Biological variation (BV)

- Considering variations between samples





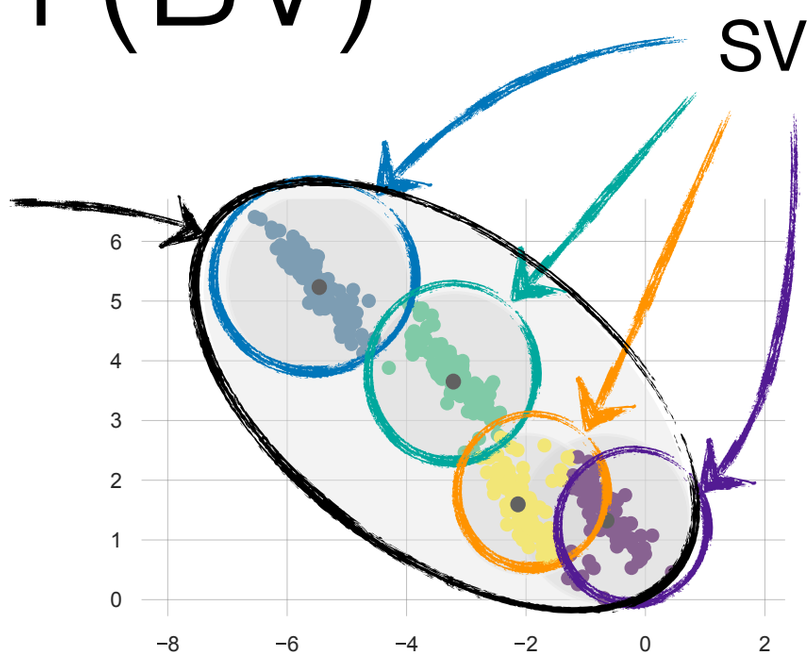
ACGT…	3
ACCT…	100
CACT…	20
AGCT…	10
…	…
CGCT…	30
AGCT…	20

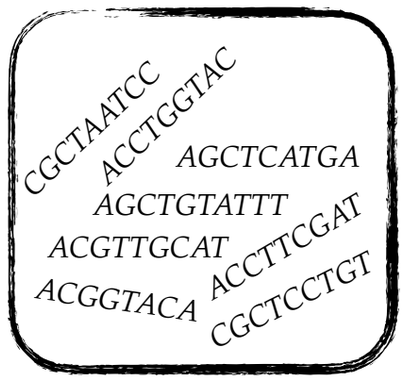


3	100	20	10	30	20

Biological variation (BV)

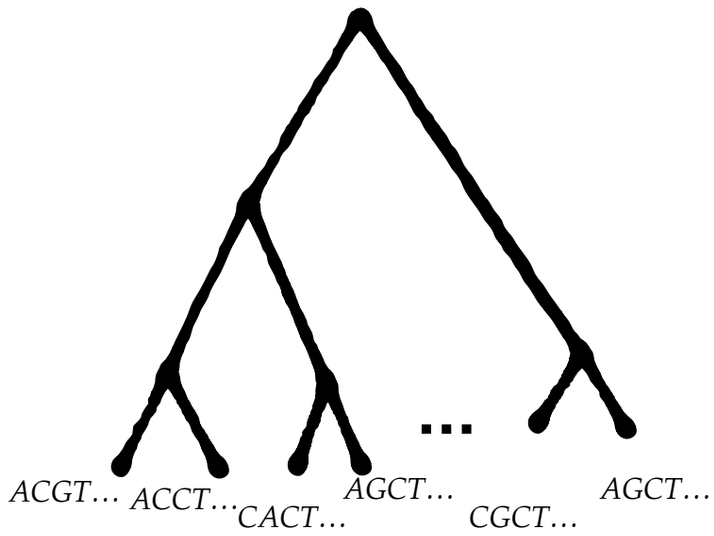
- Considering variations between samples





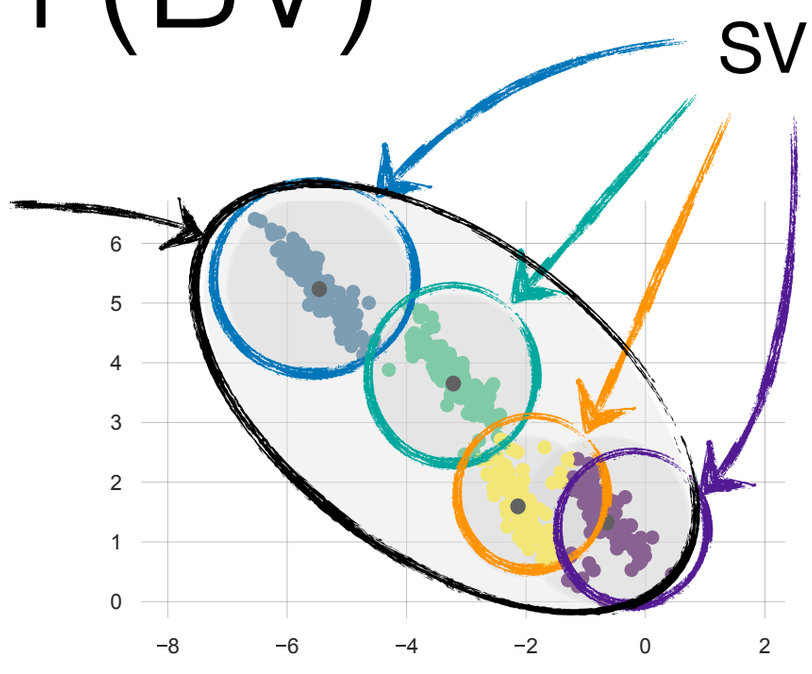
ACGT…
ACCT…
CACT…
AGCT…
…
CGCT…
AGCT…

3	…	5
100	…	100
20	…	25
10	…	1
…	…	…
30	…	10
20	…	20

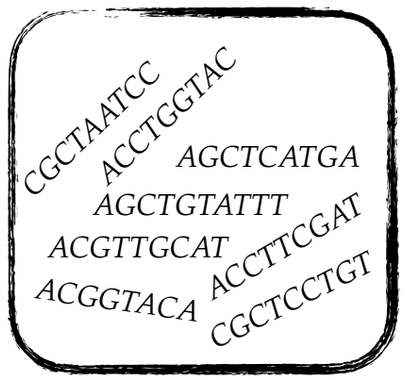


3	100	20	10	30	20
…	…	…	…	…	…
5	100	25	1	10	20

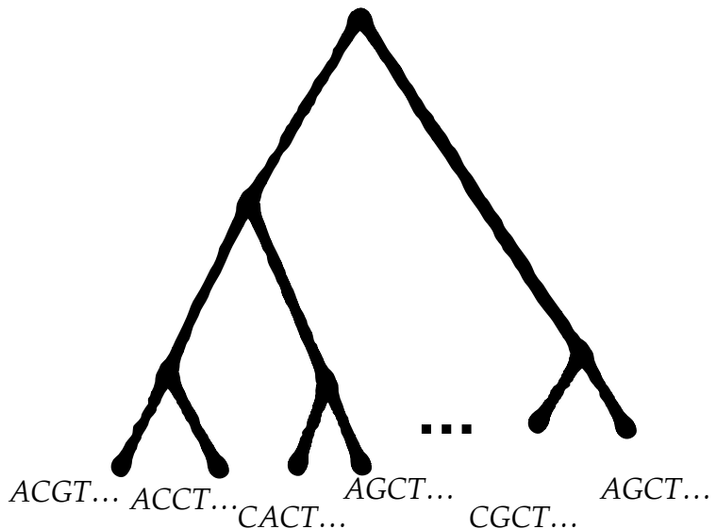
Biological variation (BV)



- Considering variations between samples

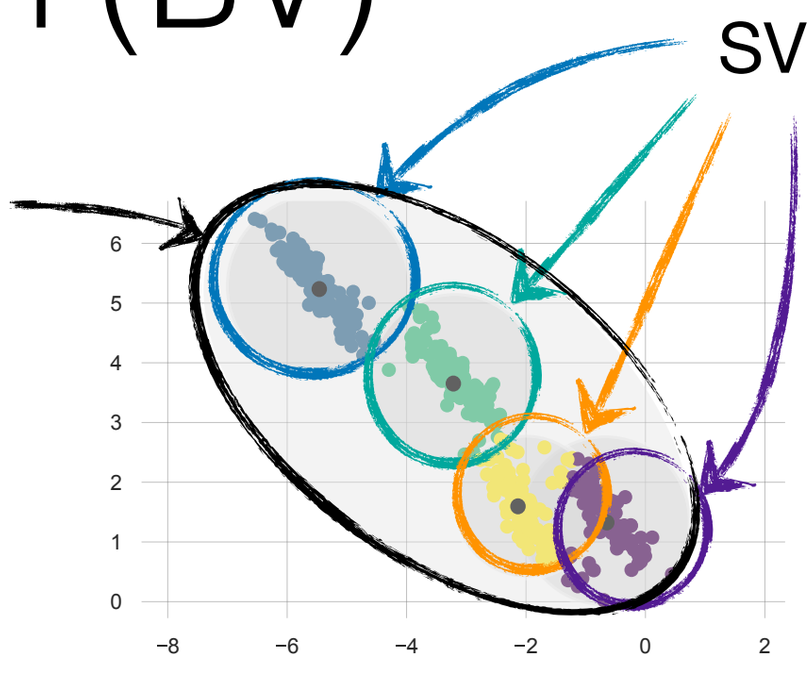


ACGT…	4	3	…	5
ACCT…	98	100	…	100
CACT…	23	20	…	25
AGCT…	23	10	…	1
…		…	…	…
CGCT…	20	30	…	10
AGCT…	19	20	…	20

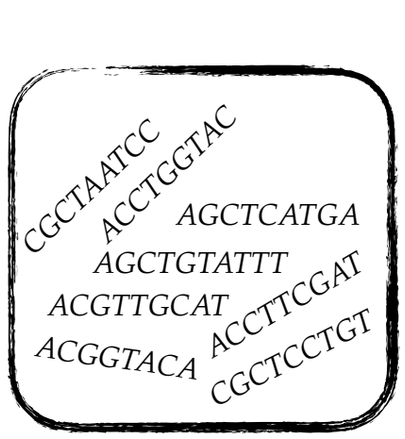


4	98	23	23	20	19
3	100	20	10	30	20
…	…	…	…	…	…
5	100	25	1	10	20

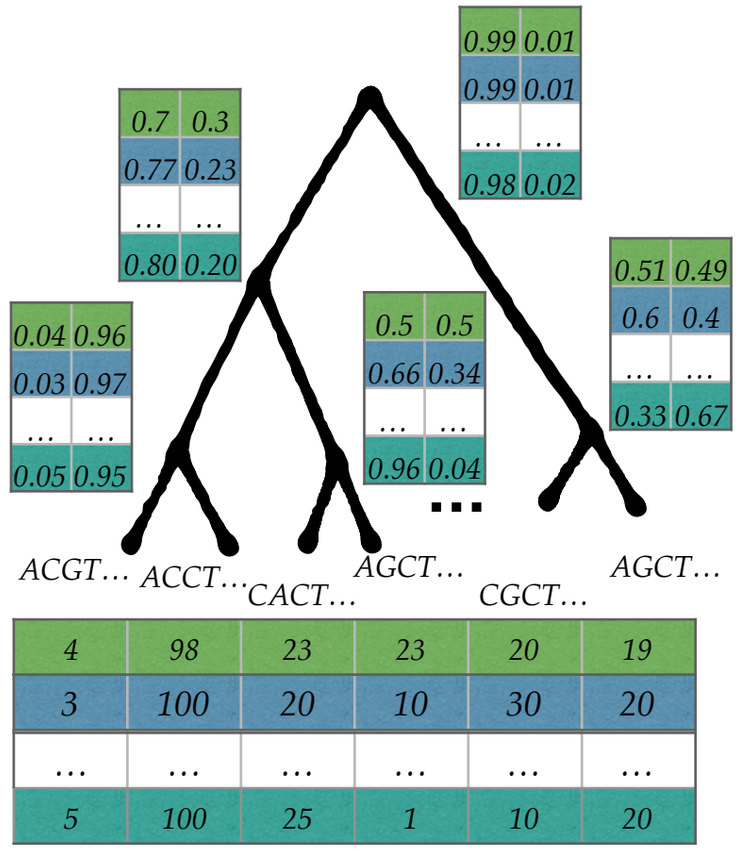
Biological variation (BV)



- Considering variations between samples



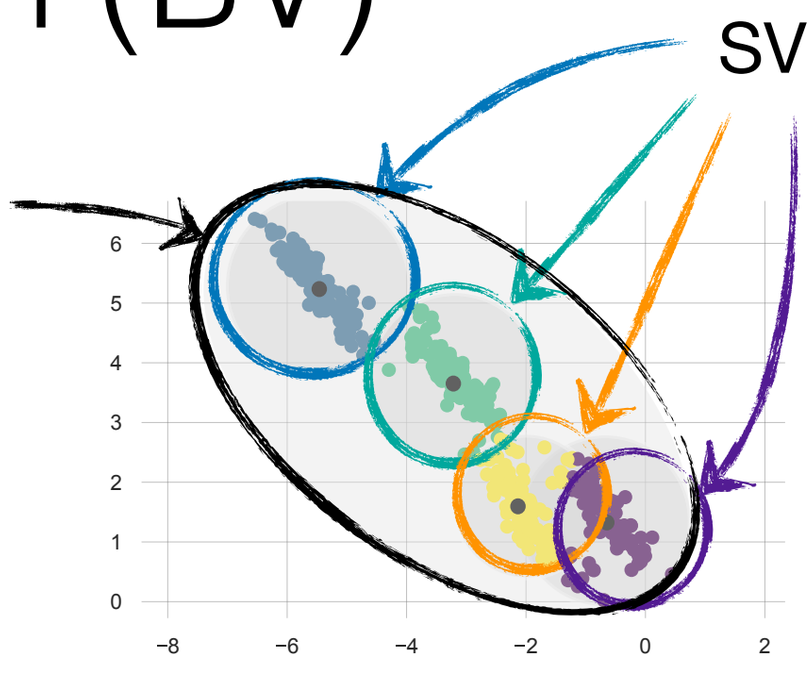
ACGT…	4	3	…	5
ACCT…	98	100	…	100
CACT…	23	20	…	25
AGCT…	23	10	…	1
…		…	…	…
CGCT…	20	30	…	10
AGCT…	19	20	…	20

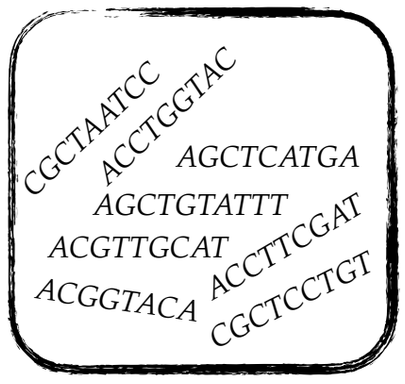


4	98	23	23	20	19
3	100	20	10	30	20
…	…	…	…	…	…
5	100	25	1	10	20

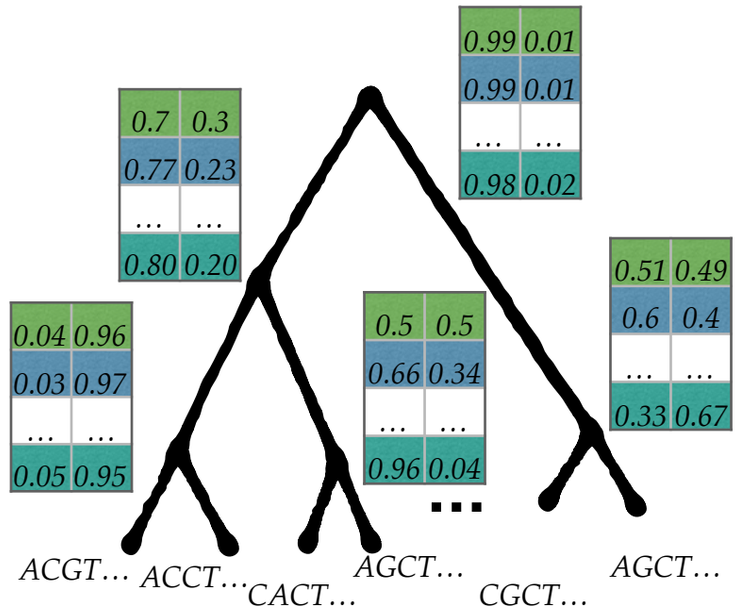
Biological variation (BV)

- Considering variations between samples
- BV: Beta distributions



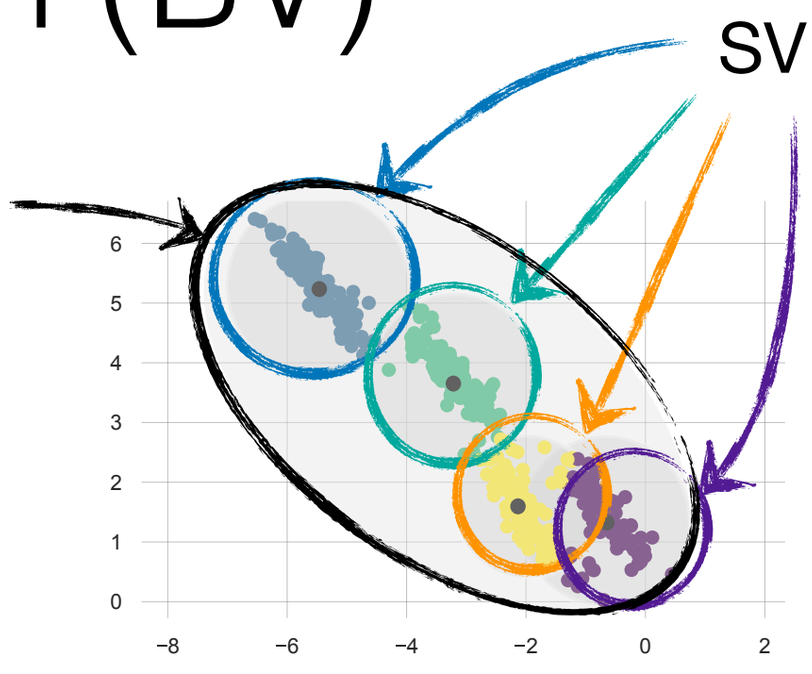


ACGT…	4	3	…	5
ACCT…	98	100	…	100
CACT…	23	20	…	25
AGCT…	23	10	…	1
…		…	…	…
CGCT…	20	30	…	10
AGCT…	19	20	…	20

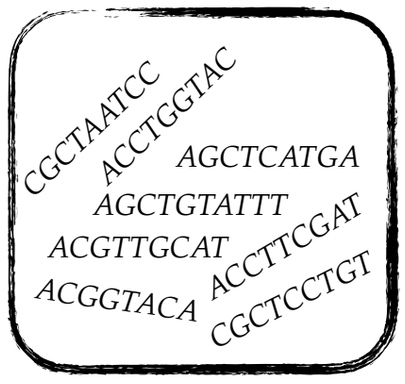


4	98	23	23	20	19
3	100	20	10	30	20
…	…	…	…	…	…
5	100	25	1	10	20

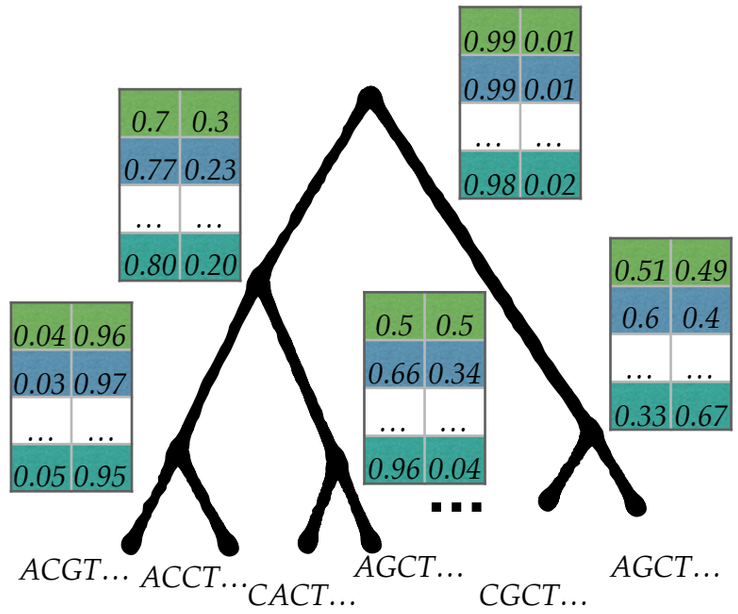
Biological variation (BV)



- Considering variations between samples
- BV: Beta distributions
- SV: Binomial distributions



ACGT…	4	3	…	5
ACCT…	98	100	…	100
CACT…	23	20	…	25
AGCT…	23	10	…	1
…		…	…	…
CGCT…	20	30	…	10
AGCT…	19	20	…	20



4	98	23	23	20	19
3	100	20	10	30	20
…	…	…	…	…	…
5	100	25	1	10	20